Université de Montréal

Faculté des Arts et Sciences

Département de Mathématiques et de Statistique

MCMC adaptatifs à essais multiples

par **Simon Fontaine**

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M. Sc.)
en **Statistique**

Dépôt initial
21 mai 2019
Révisions
13 août 2019
Dépôt final
4 septembre 2019

Université de Montréal

Ce mémoire, intitulé

MCMC adaptatifs à essais multiples,

par **Simon Fontaine,**

a été évalué par un jury constitué des personnes suivantes :

Directrice de recherche	**Mylène Bédard** Professeure agrégée en Statistique
Membre du jury	**Florian Maire** Professeur adjoint en Statistique
Membre du jury	**François Perron** Professeur titulaire en Statistique

Acceptation
14 août 2019

Résumé

Ce mémoire a pour but d'intégrer une composante adaptative au sein des algorithmes Metropolis à essais multiples (MTM) qui sont un cas particulier des méthodes de Monte Carlo par chaîne de Markov (MCMC). Les méthodes MCMC ainsi que leurs extensions adaptatives et à essais multiples sont explorées en profondeur (tant au niveau des variations possibles que de leurs propriétés théoriques) afin de bien ancrer l'étude de l'algorithme Metropolis à essais multiples adaptatif (aMTM) proposé. De plus, certains résultats dans la littérature sur les méthodes à essais multiples sont généralisés permettant alors l'obtention de résultats plus généraux à propos de l'algorithme aMTM. L'ergodicité de l'algorithme est ensuite démontrée en utilisant des résultats bien connus tirés de Roberts et Rosenthal (2007), d'Andrieu et Moulines (2006) et de Craiu et collab. (2015) et sa performance empirique est étudiée à travers une série d'expériences de simulation. L'algorithme aMTM arrive notamment à surpasser substantiellement des échantillonneurs plus simples (sans adaptation ou à un seul essai) pour des distributions cibles multimodales ou à géométrie complexe. Enfin, différentes variantes de l'algorithme sont proposées et comparées afin d'identifier des réglages particulièrement plus efficaces. Une implémentation de l'algorithme aMTM est fournie dans un progiciel `R` appelé `aMTM` disponible au `https://github.com/fontaine618/aMTM`.

Mots clés. Monte Carlo par chaînes de Markov, Monte Carlo par chaînes de Markov adaptatif, Metropolis à essais multiples, ergodicité.

Abstract

This memoir aims at introducing adaptation within the Multiple-Try Metropolis (MTM) algorithms which are a special case of the Markov chain Monte Carlo (MCMC) methods. The MCMC methods, along with their adaptive and multiple-try extensions, are thoroughly explored (both in their possible variations and in their theoretical properties) in order to firmly anchor the study of the proposed adaptive Multiple-Try Metropolis (aMTM) algorithm. Moreover, some existing results on the properties of MTM algorithms are generalized to enable more general results about the aMTM algorithm. The ergodicity of the algorithm is then established using well known results of Roberts and Rosenthal (2007), Andrieu and Moulines (2006) and Craiu et al. (2015) and its empirical performance is studied through a series of simulation experiments. The aMTM algorithm achieves notably better performance than simpler samplers (non-adaptive or single-try) when applied to distributions that are multimodal or that exhibit complex geometry. Finally, many variations of the algorithm are proposed and compared to identify settings that are particularly more efficient. An implementation of the algorithm is provided in a `R` *package called* `aMTM` *available at* `https://github.com/fontaine618/aMTM`.

Keywords. *Markov chain Monte Carlo, Adaptive Markov chain Monte Carlo, Multiple-Try Metropolis, Ergodicity.*

Table des matières

Liste des figures

Liste des tableaux

Liste des algorithmes

Sigle	Définition	
ACF	Fonction d'autocorrélation (*Autocorrelation Function*)	Section 2.4.1
ACMTM	Metropolis adaptatif à essais multiples par composante (*Adaptive Component-wise MTM*)	Algorithme 5.3
ACT	Temps d'autocorrélation (*Autocorrelation Time*)	Section 2.4.3.1
ADS	Échantillonnage directionnel adaptatif (*Adaptive Direction Sampling*)	Section 3.1.3.2
AIMH	Metropolis-Hastings indépendant adaptatif (*Adaptive Independent MH*)	Algorithme 3.5
AIMTM	(*Annealed Interacting Multiple Try Metropolis*)	Section 5.2
AirMCMC	MCMC adapté de plus en plus rarement (*Adapted Increasingly Rarely MCMC*)	Algorithme 3.6
AISMTM	(*Adaptive Independent Sticky MTM*)	Algorithme 5.1
AM	Metropolis adaptatif (*Adaptive Metropolis*)	Algorithme 3.1
aMTM	Metropolis adaptatif à essais multiples (*Adaptive Multiple Try Metropolis*)	Algorithme 5.4
ASWAM	Metropolis adaptatif avec échelle adaptative (*Adaptive Scaling within AM*)	Algorithme 3.10
BAM	Metropolis à adaptation bornée (*Bounded Adaptation Metropolis*)	Algorithme 3.13
CMTM	Metropolis à essais multiples par composante (*Component-wise MTM*)	Section 4.1.3.3
CUSUM	Somme cumulative (*Cumulative Sum*)	Section 2.4.1
DR	Rejet retardé (*Delayed rejection*)	Algorithme 2.8
DRAM	Metropolis adaptatif à rejet retardé (*Delayed Rejection AM*)	Section 3.1.3.4
EA	Extrêmement antithétique (*Extremely Antithetic*)	Section 4.1.1.4
EE	Équi-énergie (*Equi-Energy*)	Section 3.1.3.3
EM	Espérance-maximisation (*Expectation-Maximization*)	Section 3.1.4.4
ESEJD	Saut euclidien moyen (*Expected Squared Euclidian Jumping Distance*)	Section 2.4.3.3
ESJD	Saut quadratique moyen (*Expected Squared Jumping Distance*)	Section 2.4.3.3
ESS	Taille échantillonnale effective (*Effective Sample Size*)	Section 2.4.3.2
FFT	Transformée de Fourier rapide (*Fast Fourier Transform*)	Section 2.4.2.2
HR	(*Hit-and-Run*)	Exemple 4.1
IMH	Metropolis-Hastings indépendant (*Independent Metropolis-Hastings*)	Section 2.3.1.1
IMTM	(*Interacting Multiple Try Metropolis*)	Section 4.1.3.1
INCA	Adaptation entre chaînes (*Inter-chain Adaptation*)	Section 3.1.2
IR-MCMC	MCMC à rééchantillonnage par importance (*Importance Resampling MCMC*)	Section 3.1.3.3
KL	Kullback-Leibler	Section 3.1.4.4
MAE	Erreur absolue moyenne (*Mean Absolue Error*)	Section 6.2.2
MALA	Algorithme Metropolis Langevin(*Metropolis-Adjusted Langevin Algorithm*)	Section 2.3.1.4
MH	Metropolis-Hastings	Section 2.3.1
MCMC	Monte-Carlo par chaînes de Markov (*Markov Chain Monte Carlo*)	Section 2.3
MCMCMC	MCMC couplé Metropolis (*Metropolis-coupled MCMC*)	Section 3.1.3.3
MCTM	Metropolis à essais multiples corrélés (*Multiple Correlated Try Metropolis*)	Section 4.2.3
MPSRF	(*Multivariate Potential Scale Reduction Factor*)	Section 6.2.1
MSE	Erreur quadratique moyenne (*Mean Squared Error*)	Section 6.2.2
MTM	Metropolis à essais multiples (*Multiple Try Metropolis*)	Section 4.1
MwG	Metropolis dans Gibbs (*Metropolis-within-Gibbs*)	Algorithme 2.7
NKC	(*Normal Kernel Coupler*)	Algorithme 3.8
NOLBM	Moyenne par lots sans chevauchement (*Nonoverlapping Batch Means*)	Section 2.4.2.3
QMCR	Quasi-Monte Carlo randomisé (*Randomized Quasi-Monte Carlo*)	Section 4.1.1.5
RAM	Metropolis adaptatif robuste (*Robust Adaptive Metropolis*)	Algorithme 3.11
RMSE	La racine carrée de l'erreur quadratique moyenne (*Root Mean Squared Error*)	Section 4.3.2
RWM	Metropolis marche aléatoire (*Random Walk Metropolis*)	Section 2.3.1.2
TINCA	INCA tempéré (*Tempered INCA*)	Section 3.1.3.3
TV	Variation totale (*Total Variation*)	Définition 2.16

Tableau 0.1 *Sigles et abréviations utilisés dans le texte. Les sigles provenant de l'anglais sont accompagnés de leur définition originale entre parenthèse.*

Symbole	Définition
$\pi\left(f\right)$	L'espérance de la fonction f sous la distribution π
$a_n \overset{n\to\infty}{\longrightarrow} a$	La convergence de la suite $\{a_n\}_{n\geqslant 1}$ vers a
$\delta_x(\cdot)$	La fonction de masse delta de Dirac en x
$\overset{\mathcal{D}}{\longrightarrow}$	La convergence en distribution
i.i.d., $\overset{\text{i.i.d.}}{\sim}$	Indépendants et identiquement distribués
$\mathbb{E}\{\cdot\}$	L'opérateur d'espérance
$\mathbb{E}\{\cdot\|\cdot\}$	L'opérateur d'espérance conditionnelle
$x_{j:k}$	La sous-suite ou le sous-vecteur $\{x_i\}_{i=j}^{k}$
$\mathbb{P}(\cdot)$	L'opérateur de probabilité
$\mathcal{B}(\mathcal{X})$	La σ-algèbre de Borel sur $\mathcal{X}$
$\mathbb{P}(\cdot\|\cdot)$	L'opérateur de probabilité conditionnelle
λ_{Leb}	La mesure de Lebesgue
$\overline{\mathbb{R}}$	La droite réelle achevée
$\mathbb{1}_A$	La fonction indicatrice de l'ensemble A
$\overset{\text{p.s.}}{\longrightarrow}$	La convergence presque sûre
Φ	La fonction de répartition de la loi normale centrée réduite
$\mathcal{N}(\mu,\sigma^2)$	La loi normale d'espérance μ et de variance σ^2
$\left\|\cdot\right\|_2$	La norme euclidienne
$\exp(\cdot)$	La fonction exponentielle de base naturelle
∇	L'opérateur de gradient
$\mathcal{N}_d(\boldsymbol{\mu},\Sigma)$	La loi multinormale en d dimensions d'espérance $\boldsymbol{\mu}$ et de covariance Σ
$x_{-j}, x^{(-j)}$	Le sous-vecteur de x excluant la j-ième composante
$\text{Cov}(\cdot,\cdot)$	L'opérateur de covariance
$\text{Corr}(\cdot,\cdot)$	L'opérateur de corrélation
$x^\top, A^\top$	La transposée du vecteur x ou de la matrice A
$\text{Var}(\cdot)$	L'opérateur de variance
$\lfloor\cdot\rfloor$	La fonction plancher
$\mathbf{0}_d, \mathbf{0}$	La vecteur de 0 de dimension d ou de dimension donnée par le contexte
I_d	La matrice identité en d dimensions
logit	La fonction logistique
det	L'opérateur déterminant
$\overset{\mathcal{L}_2}{\longrightarrow}$	La convergence en espérance d'ordre 2
$\mathcal{O}$	La comparaison asymptotique de domination
$\mathcal{L}_V$	L'espace des fonctions à V-norme finie
$\left\|\cdot\right\|_F$	La norme de Frobenius
$\mathbf{1}_d, \mathbf{1}$	Le vecteur de 1 de dimension d ou de dimension donnée par le contexte
$\otimes$	Le produit de Kronecker

Tableau 0.2 *Notation mathématique utilisée qui n'est pas explicitement définie dans le texte. La notation plus communément utilisée en mathématique est supposée connue lorsqu'il n'y a pas d'ambiguïté et n'apparait donc pas dans cette description.*

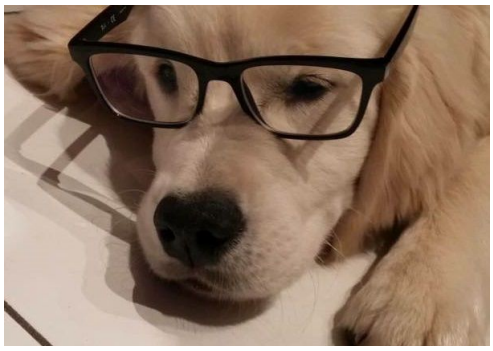
Remerciements

Je tiens tout d'abord à remercier ma formidable superviseure de recherche, Mylène. Elle a été pour moi une guide extraordinaire dans toute cette aventure. D'une part, ses judicieux conseils, ses corrections attentives et sa vaste expertise m'ont beaucoup appris non seulement sur le domaine des MCMC, mais aussi sur le processus de recherche et d'écriture en général. D'autre part, elle m'a été d'une grande utilité dans mon entrée dans le domaine académique alors que j'ai pu profiter de toute son expérience pour bien orienter mon parcours à la suite de cette maîtrise. Merci infiniment !

Je dois également une immense gratitude à l'égard de mes parents, Manon et Normand. Leur support absolu tout au long de mes études a été une source constante de motivation. Je leur serai pour toujours reconnaissant de leur amour et de leur appui qui ont fortement contribué à mon épanouissement.

Il est également important pour moi de remercier mon frère Mathieu et tous mes amis, trop nombreux pour être énumérés, qui m'ont toujours soutenu et encouragé dans mes ambitieux projets.

Je veux finalement laisser à ma compagne de tous les jours Laïka quelques mots qu'elle ne pourra malheureusement jamais lire. Ta présence a été et sera toujours d'un très grand réconfort ; ton amour inconditionnel est la pierre d'assise de mon bonheur. Merci de me réchauffer les pieds dans mon froid sous-sol et merci de me rappeler gentiment, de ton museau, qu'il est temps de prendre une pause pour aller marcher.



Chapitre 1

Introduction

Bien que le concept même fût introduit quelques années plus tôt, les méthodes de Monte Carlo trouvent leurs origines au milieu du 20e siècle dans les travaux de Metropolis et Ulam (1949). Initialement proposée pour résoudre numériquement des problèmes d'équations différentielles en physique, cette méthode s'est rapidement prêtée à de nombreuses autres applications puisqu'il s'agit fondamentalement d'une méthode générale d'intégration numérique. Metropolis et collab. (1953) ont ensuite proposé une version séquentielle des méthodes Monte Carlo, éventuellement généralisée par Hastings (1970), qui a, par le fait même, lancé le domaine des méthodes de Monte Carlo par chaînes de Markov (MCMC). Puis, les années 1990 ont donné lieu à de nombreuses avancées dans la théorie et l'utilisation des méthodes MCMC ainsi qu'à une augmentation rapide de la puissance de calcul disponible : ces deux conditions ont propulsé les méthodes MCMC parmi les algorithmes indispensables à divers milieux scientifiques.

Les méthodes MCMC comportent cependant une importante composante de mise au point devant être prise en charge par l'utilisateur. De par leur nature numérique, ces algorithmes ont donc une performance intimement liée aux différents choix que l'utilisateur doit faire. Une avancée importante à ce sujet est l'introduction des méthodes MCMC adaptatives. Ces méthodes, principalement inspirées de Haario et collab. (2001) (et basées sur divers travaux d'échelle optimale tels que ceux de Roberts et collab. (1997)), automatisent une certaine partie de la mise au point requise pour optimiser la performance des algorithmes MCMC. Depuis, ces idées ont suscité beaucoup d'intérêt dans la communauté MCMC et les méthodes adaptatives sont désormais grandement répandues.

Une autre classe de méthodes MCMC, les algorithmes Metropolis à essais multiples (MTM), a été proposée par Liu et collab. (2000) dans le but d'augmenter l'efficacité de l'algorithme Metropolis en générant plusieurs candidats à l'intérieur d'une itération donnée. Par construction, ces algorithmes améliorent la version à candidat unique (l'algorithme Metropolis) ; ils sont cependant associés à un coût computationnel plus élevé et nécessitent davantage de mise au point.

Le but principal de ce mémoire est d'incorporer les idées des méthodes adaptatives au sein de l'algorithme MTM afin d'automatiser la mise au point de cette classe d'algorithmes. L'algorithme Metropolis à essais multiples adaptatif (aMTM) proposé facilite l'application de l'algorithme MTM en minimisant l'intervention de l'utilisateur et améliore sa performance en trouvant automatiquement un ensemble de paramètres à efficacité supérieure, voire optimale. Ainsi, l'augmentation de la performance peut être suffisante pour valoir l'augmentation du temps de calcul requis par l'algorithme MTM.

Le texte sera divisé comme suit. Au chapitre 2, on introduit formellement les méthodes MCMC : le

principe Monte Carlo est d'abord expliqué puis la théorie des chaînes de Markov est étudiée. Ensuite, l'algorithme Metropolis-Hastings, duquel découlent la majorité des algorithmes MCMC, est défini et ses propriétés sont abordées : l'ergodicité, les théorèmes limites et l'échelle optimale. Puis, certains outils de diagnostique et certaines mesures de performance sont introduits.

Le chapitre 3 se veut une introduction au vaste monde des algorithmes MCMC adaptatifs. De nombreuses techniques ont été proposées au cours des dernières décennies afin d'améliorer les algorithmes MCMC et plusieurs d'entre elles se retrouvent dans ce chapitre. Les propriétés théoriques des algorithmes adaptatifs requièrent une étude particulière comparativement aux algorithmes MCMC non-adaptatifs de sorte qu'une seconde partie théorique s'y retrouve.

La famille des algorithmes Metropolis à essais multiples est considérée au chapitre 4. D'abord, les différentes variantes possibles sont définies puis les propriétés théoriques sont étudiées. Ces algorithmes sont cependant plus coûteux computationnellement et on discute du compromis entre temps de calcul et performance pour ce type d'algorithme.

Au chapitre 5, on propose l'algorithme Metropolis à essais multiples adaptatif (aMTM). D'abord, on survole les premières tentatives similaires d'introduction d'adaptation dans les algorithmes MTM. L'algorithme est ensuite défini et les différentes variantes possibles, tant dans les essais multiples que dans l'adaptation, sont explorées. Ses propriétés théoriques sont enfin abordées : en particulier, on dégage des ensembles de conditions suffisantes à l'ergodicité de l'algorithme et toutes les variantes proposées sont couvertes par l'un ou l'autre des résultats produits. La méthode de preuve est inspirée des travaux de Roberts et Rosenthal (2007), de Andrieu et Moulines (2006) et de Craiu et collab. (2015).

Une série d'expériences numériques est effectuée au chapitre 6 afin d'analyser les propriétés empiriques de l'algorithme aMTM dans différentes situations et comparativement à des algorithmes plus simples. Simultanément, les différentes variantes proposées sont étudiées afin de bien comprendre le comportement de l'algorithme par rapport au choix de distribution cible.

Principales contributions. La contribution principale de cette recherche est l'introduction de l'algorithme aMTM qui se veut une version adaptative de l'algorithme MTM. Bien qu'on retrouve quelques exemples de ce type d'extensions dans la littérature (voir section 5.2), l'algorithme aMTM est la première tentative de la sorte qui ne soit pas limitée à des mises à jour par composante ou limitée à une adaptation constituée seulement de chaînes parallèles interactives. En particulier, il s'agit de la première tentative d'adaptation des densités instrumentales en pleine dimension. Afin de justifier la validité de l'algorithme, l'ergodicité est démontrée d'une manière relativement générale de sorte que d'éventuels algorithmes MTM adaptatifs peuvent tomber sous la couverture des résultats produits. Les expériences numériques accompagnant la proposition ajoutent à l'assurance d'une performance accrue ainsi qu'à la compréhension de certaines lacunes d'autres algorithmes. Les résultats théoriques et empiriques obtenus améliorent aussi la compréhension du comportement des algorithmes MTM non-adaptatifs : certains résultats théoriques sont généralisés et la compréhension de la performance empirique des variantes de cet algorithme est approfondie.

Note au lecteur. Ce manuscrit est remarquablement long et particulièrement lourd en contenu dans le but de produire un texte qui soit relativement autonome. Certaines parties ne sont évidemment pas nécessaires à la compréhension des contributions apportées par cette recherche. D'abord, un lecteur déjà familier avec les méthodes MCMC peut se passer du chapitre 2 qui n'est qu'un rappel des bases de

la discipline. Le chapitre 3 comporte un survol presqu'exhaustif des différentes méthodes d'adaptation des méthodes MCMC : cette revue fut nécessaire à la recherche pour bien comprendre comment effectuer l'adaptation, mais n'est pas requise à l'exposition. La section 3.2 est cependant pertinente à la compréhension des démonstrations du chapitre 5 puisque plusieurs éléments en sont directement tirés. Le chapitre 4 sur les algorithmes MTM est quant à lui important dans la mesure où les différentes variantes sont expliquées et où certains résultats théoriques seront requis pour démontrer l'ergodicité de l'algorithme aMTM. Les chapitres 5 et 6 sont évidemment les plus importants puisqu'ils contiennent la grande majorité des contributions contenues dans ce mémoire.

Chapitre 2

Méthodes MCMC

Les méthodes de **Monte Carlo par chaînes de Markov** (MCMC) ont reçu une grande quantité d'attention au cours des dernières décennies en raison de l'avancement de leur compréhension ainsi que de l'augmentation de la puissance de calcul disponible. Il s'agit d'algorithmes permettant principalement l'approximation d'espérances (intégrales de probabilité) au moyen de simulations numériques. Comme leur nom l'indique, ces méthodes sont basées à la fois sur le principe Monte Carlo—approcher une intégrale par la moyenne d'un échantillon— et à la fois sur le concept de chaîne de Markov—une suite temporelle de variables aléatoires liées entre elles— : la chaîne de Markov produit l'échantillon Monte Carlo servant à approcher l'espérance recherchée.

Ce chapitre se veut une introduction aux méthodes MCMC et, pour ce faire, une approche progressive est empruntée. D'abord, le principe Monte Carlo est expliqué à la section 2.1. Ensuite, la théorie des chaînes de Markov est abordée en détail à la section 2.2. Les propriétés des chaînes de Markov se transposent directement au contexte MCMC de sorte que la théorie des MCMC est fortement basée sur celle des chaînes de Markov. Les méthodes MCMC sont alors introduites à la section 2.3 et un des algorithmes principaux, l'algorithme Metropolis-Hastings, est considéré en détail. La réalisation d'une méthode MCMC est la réalisation d'une chaîne de Markov dont les propriétés doivent être analysées afin de bien évaluer l'estimation Monte Carlo : ces considérations sont abordées à la section 2.4. Enfin, le problème de l'échelle optimale dans les méthodes MCMC est considéré à la section 2.5 ; ce problème est au cœur de l'efficacité de ces algorithmes.

2.1 Approche Monte Carlo

En statistique, plusieurs situations mènent au calcul d'espérances de la forme

$$\pi(f) = \int_{\mathcal{X}} f(x)\pi(\,\mathrm{d}x), \tag{2.1}$$

où π est une mesure de probabilité sur l'espace d'états $\mathcal{X}$ et $f : \mathcal{X} \to \mathbb{R}$ est une fonction π-mesurable. Il arrive souvent que le calcul analytique de (2.1) ne soit pas possible et des méthodes alternatives doivent être considérées afin de calculer $\pi(f)$, ou du moins d'en obtenir une approximation. Étant donné que les espérances ne sont que des intégrales ayant une forme particulière, il est donc possible

d'arriver à des approximations de $\pi(f)$ en utilisant une méthode d'intégration numérique. Ce type d'algorithme divise typiquement l'espace $\mathcal{X}$ en rectangles où la fonction f est approchée par un certain choix de fonctions plus simples (e.g. constant, linéaire, etc.) Cependant, cette partition en rectangle devient problématique dès que la dimension de $\mathcal{X}$ est moyennement élevée : le nombre de rectangle requis pour obtenir une certaine précision dans l'approximation est exponentiel par rapport à la dimension. De plus, lorsque le domaine d'intégration $\mathcal{X}$ n'est pas borné, ce type d'approche peut également rencontrer des problèmes. Des méthodes d'intégration numérique spécifiquement adaptées aux espérances, les méthodes de **Monte Carlo**, souffrent généralement moins de ces problématiques.

2.1.1 Le principe Monte Carlo

Le principe fondamental derrière les méthodes de Monte Carlo repose sur la loi des grands nombres. En effet, une loi des grands nombres est tout résultat du genre

$$\frac{1}{N}\sum_{n=1}^{N} f(x_n) \xrightarrow{c} \pi(f), \qquad n \to \infty,$$

c'est-à-dire que la moyenne de la fonction f prise sur un échantillon $\{x_n\}_{n=1}^N$ converge, sous certaines conditions, vers l'espérance de cette fonction prise sous la distribution π pour un mode de convergence $c \in \{\mathbb{P}, \text{p.s.}\}$. Cette moyenne échantillonnale peut, quant à elle, être vue comme l'espérance suivante :

$$\frac{1}{N}\sum_{n=1}^{N} f(x_n) = \int_{\mathcal{X}} f(x)\hat{\pi}_N(\,\mathrm{d}x) = \hat{\pi}_N(f), \tag{2.2}$$

où $\hat{\pi}_N$ est la fonction empirique de masse de l'**échantillon Monte Carlo** $\{x_n\}_{n=1}^N$ donnée par

$$\hat{\pi}_N(x) = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}(x),$$

et où $\delta_y(\cdot)$ dénote la fonction de masse delta en y. C'est donc dire que la mesure empirique $\hat{\pi}_N$ est utilisée à la place de π dans l'espérance. Il devient évident que les propriétés de l'échantillon $\{x_n\}_{n=1}^N$ requises pour que l'**estimateur Monte Carlo** $\hat{\pi}_N(f)$ satisfasse une loi des grands nombres seront telles que la mesure empirique $\hat{\pi}_N$ fournisse une bonne approximation de la distribution π. Les différentes méthodes de Monte Carlo diffèrent donc généralement seulement de par la manière dont l'échantillon est produit.

Une seconde propriété souvent recherchée pour un estimateur est un théorème limite central. Ce type de résultat montre que la distribution asymptotique de l'estimateur est une loi gaussienne :

$$N^{-1/2}\hat{\pi}_N(f) \xrightarrow{\mathcal{D}} \mathcal{N}(\pi(f),\sigma_f^2), \qquad n \to \infty,$$

où $\xrightarrow{\mathcal{D}}$ dénote la convergence en distribution et où σ_f^2 est la **variance asymptotique** de l'estimateur. Lorsqu'un théorème limite central est satisfait, il est possible de joindre à l'estimé ponctuel $\hat{\pi}_N(f)$ une **erreur standard Monte Carlo** donnée par $\hat{\sigma}_f/\sqrt{N}$ où $\hat{\sigma}_f^2$ est un estimé de σ_f^2 tel que la variance échantillonnale,

$$\hat{\sigma}_f^2 = \frac{1}{N}\sum_{n=1}^{N} \left(f(x_n) - \hat{\pi}_N(f)\right)^2$$

. Il est donc possible de fournir une appréciation de la qualité de l'estimation faite par l'estimateur Monte Carlo lorsqu'un tel résultat est vérifié.

Notons que la plupart des termes utilisés dans les méthodes Monte Carlo sont également utilisés plus généralement en statistique (estimateur, erreur standard, échantillon, etc.) Afin de distinguer les deux concepts, on ajoute souvent la mention « Monte Carlo » après ces termes. Cette distinction est particulièrement pertinente lorsque l'espérance $\pi(f)$ dépend elle-même d'un réel échantillon provenant d'une expérience ; l'échantillon Monte Carlo est une collection de points dans $\mathcal{X}$ et non l'ensemble des unités de l'expérience. Similairement, l'erreur standard Monte Carlo ne correspond pas à l'erreur standard de la moyenne de l'échantillon.

2.1.2 Algorithmes de Monte Carlo

Les versions les plus simples de la loi des grands nombres requièrent que chacun des éléments de l'échantillon soient des réalisations de variables aléatoires indépendantes et identiquement distribuées selon la distribution cible π, ce que l'on dénote par $X_n \overset{\text{i.i.d.}}{\sim} \pi$. La distribution empirique sera donc bien représentative de la distribution cible puisque l'échantillon est généré à même π. Ce type d'échantillonnage, appelé **échantillonnage i.i.d.** et décrit à l'algorithme 2.1, constitue la méthode de Monte Carlo standard. Deux problèmes empêchent cependant souvent l'utilisation de l'échantillonnage i.i.d.

Algorithme 2.1 Algorithme Monte Carlo standard i.i.d.

Données	Distribution cible π et taille de l'échantillon Monte Carlo N.
Procédure	Pour $n = 1, \ldots, N$, échantillonner $X_n \overset{\text{i.i.d.}}{\sim} \pi$.
Sortie	L'échantillon $x_{1:N}$ et l'estimé Monte Carlo (2.2).

D'abord, il arrive que l'échantillonnage directement à partir de π soit impossible. En fait, l'échantillonnage doit être fait par ordinateur où seul des nombres aléatoires uniformes sont généralement disponibles. La méthode de la transformation d'intégrale de probabilité permet de faire le passage entre une distribution uniforme et une distribution plus complexe, mais cette transformation n'est pas toujours connue ou possible. La méthode par **échantillonnage pondéré** permet d'éviter ce problème. On considère une seconde distribution q qui est telle que son support contienne le support de π et par rapport à laquelle il est possible d'obtenir un échantillon i.i.d. Puis, en observant l'identité suivante,

$$\pi(f) = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x = \int_{\mathcal{X}} f(x)\frac{\pi(x)}{q(x)}q(x)\,\mathrm{d}x = q\left(f\frac{\pi}{q}\right),$$

on constate qu'un estimateur Monte Carlo de $\pi(f)$ peut être obtenu par la moyenne échantillonnale de la fonction $f\pi/q$ sur un échantillon de distribution $q(\cdot)$:

$$\hat{\pi}_N(f) = \hat{q}_N\left(f\frac{\pi}{q}\right) = \frac{1}{N}\sum_{n=1}^{N} f(x_n)\frac{\pi(x_n)}{q(x_n)} = \int_{\mathcal{X}} f(x)\frac{\pi(x)}{q(x)}\hat{q}_N(\,\mathrm{d}x),$$

où $\hat{q}_N(\cdot) = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}(\cdot)$ est la distribution empirique de l'échantillon généré à partir de q. On peut

également réinterpréter cette expression comme une distribution empirique de π pondérée :

$$\hat{\pi}_N(x) = \frac{1}{N}\sum_{n=1}^{N} w(x_n)\delta_{x_n}(x), \qquad w(x) = \frac{\pi(x)}{q(x)}. \tag{2.3}$$

Le rapport $w = \pi/q$ est appelé le **rapport de vraisemblance** ou le **poids d'importance**. En plus de ne pas avoir à échantillonner directement à partir de π, cette méthode possède l'avantage d'une réduction possible de la variance asymptotique de l'estimateur Monte Carlo par rapport un échantillonnage i.i.d. Bien que cette remarque soit théoriquement vraie, elle ne l'est pas en pratique puisqu'une réduction de la variance nécessite la connaissance de $\pi(|f|)$ qui est généralement inconnu vu qu'on cherche déjà $\pi(f)$.

L'échantillonnage préférentiel exige cependant le calcul explicite de $\pi(x)$ dans les poids d'importance, ce qui requiert l'expression exacte de π. Une seconde problématique commune aux échantillonnages i.i.d. et préférentiel est que la distribution π peut ne pas être complètement connue. En effet, certaines situations sont telles qu'une espérance est recherchée pour une certaine distribution qui n'est connue seulement qu'à une constante de proportionnalité près. Par exemple, le contexte des statistiques bayésiennes produit souvent des estimateurs donnés par une espérance à postériori de la forme

$$\hat{\theta}(x) = \mathbb{E}\{\theta|X = x\} = \int \theta\pi(\theta|x)\,\mathrm{d}\theta,$$

où l'on cherche à estimer un paramètre θ à l'aide de x et où π est la distribution à postériori de θ sachant x :

$$\pi(\theta|x) = \frac{p(x|\theta)\pi_0(\theta)}{\int p(x|\theta)\pi_0(\theta)\,\mathrm{d}\theta}, \tag{2.4}$$

où $\pi_0(\cdot)$ est la distribution à priori sur θ et p est la distribution de X sachant θ. L'intégrale au dénominateur de (2.4) étant elle-même souvent difficile à évaluer, l'expression de $\pi(\theta|x)$ n'est donc pas complètement connue.

Plus généralement, on obtient donc la situation suivante : $\pi(x) \propto \tilde{\pi}(x)$. Dans ce cas, il est possible d'adapter l'échantillonnage préférentiel afin de produire un estimateur Monte Carlo valide. Les poids d'importance sont alors connus seulement à une constante de proportionnalité près :

$$\tilde{w}(x) = \frac{\tilde{\pi}(x)}{q(x)} \propto \frac{\pi(x)}{q(x)} = w(x).$$

Dans ce cas, la mesure empirique $\frac{1}{N}\sum_{n=1}^{N}\tilde{w}(x_n)\delta_{x_n}(x)$ n'est plus une mesure de probabilité puisqu'elle n'intègre pas à 1 sur $\mathcal{X}$ par rapport à π. Il faut donc normaliser cette mesure afin d'obtenir les bons poids d'importance :

$$\hat{\pi}_N(x) = \frac{1}{N}\sum_{n=1}^{N} w(x_n)\delta_{x_n}(x), \qquad w(x_n) = \frac{\tilde{w}(x_n)}{\sum_{n=1}^{N}\tilde{w}(x_n)}.$$

Ceci correspond à la réécriture suivante de $\pi(f)$ en notant que $\tilde{\pi}(x) = \tilde{w}(x)q(x)$:

$$\begin{aligned}\pi(f) &= \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} f(x)\frac{\pi(x)}{q(x)}q(x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} f(x)\tilde{w}(x)\frac{\pi(x)}{\tilde{\pi}(x)}q(x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} f(x)\frac{\tilde{w}(x)q(x)}{\int_{\mathcal{X}}\tilde{\pi}(x)\,\mathrm{d}x}\,\mathrm{d}x \\ &= \int_{\mathcal{X}} f(x)\frac{\tilde{w}(x)q(x)}{\int_{\mathcal{X}}\tilde{w}(x)q(x)\,\mathrm{d}x}\,\mathrm{d}x.\end{aligned}$$

Ainsi, le calcul de π n'est plus nécessaire lorsque les poids d'importance sont normalisés ; l'algorithme 2.2 détaille la procédure. À noter qu'avoir $\tilde{w} = w$ (c.-à-d., $\tilde{\pi} = \pi$) se réduit à l'échantillonnage préférentiel régulier puisqu'on trouve alors

$$\int_{\mathcal{X}} \tilde{w}(x)q(x)\,\mathrm{d}x = \int_{\mathcal{X}} w(x)q(x)\,\mathrm{d}x = \int_{\mathcal{X}} \pi(x)\,\mathrm{d}x = 1.$$

Algorithme 2.2 Algorithme Monte Carlo à échantillonnage pondéré

Données Distribution cible $\pi \propto \tilde{\pi}$, taille de l'échantillon Monte Carlo N et densité instrumentale q à support incluant le support de π.

Procédure

1. Pour $n = 1, \ldots, N$,
   (a) *Échantillonnage.* Générer $X_n \sim q$ indépendamment de $X_{1:n-1}$;
   (b) *Poids.* Calculer le poids d'importance $\tilde{w}(x_n) = \tilde{\pi}(x_n)/q(x_n)$;
2. Normaliser les poids d'importance,

$$w(x_n) = \tilde{w}(x_n) \Big/ \sum_{n=1}^{N} \tilde{w}(x_n), \qquad n = 1, \ldots, N.$$

Sortie L'échantillon $x_{1:N}$ et l'estimé Monte Carlo (2.3).

Un autre type d'échantillonnage permet d'éviter le problème de la densité cible connue seulement à une constante de proportionnalité près. On considère donc à nouveau $\pi \propto \tilde{\pi}$ et une densité instrumentale q. L'**échantillonnage par rejet** suppose qu'il existe une constante $M < \infty$ telle que $\tilde{\pi} \leqslant Mq$, c'est-à-dire qu'un multiple de q enveloppe complètement π. Un candidat $X \sim q$ est inclus dans l'échantillon Monte Carlo avec probabilité $\tilde{\pi}(x)/Mq(x)$. La procédure (algorithme 2.3) est répétée jusqu'à ce que l'échantillon contienne le nombre requis de points. Chacun des points acceptés dans l'échantillon sera distribué selon π ; en effet, pour un ensemble mesurable $B \in \mathcal{B}(\mathcal{X})$, on a

$$\mathbb{P}(X \in B|\text{accept.}) = \frac{\mathbb{P}(X \in B, \text{accept.})}{\mathbb{P}(\text{accept.})} = \frac{\int_B \frac{\tilde{\pi}(x)}{Mq(x)}q(x)\,\mathrm{d}x}{\int_{\mathcal{X}} \frac{\tilde{\pi}(x)}{Mq(x)}q(x)\,\mathrm{d}x} = \frac{\int_B \tilde{\pi}(x)\,\mathrm{d}x}{\int_{\mathcal{X}} \tilde{\pi}(x)\,\mathrm{d}x} = \int_B \pi(x)\,\mathrm{d}x = \pi(B).$$

Notons que l'efficacité de l'échantillonnage par rejet dépend fortement du choix de q. En effet, la

Algorithme 2.3 Algorithme Monte Carlo à échantillonnage par rejet

Données	Distribution cible $\pi \propto \tilde{\pi}$, taille de l'échantillon Monte Carlo N et densité instrumentale q telle que $\tilde{\pi} \leqslant Mq$ pour un $M < \infty$.
Procédure	1. Initialiser $n = 1$; 2. Tant que $n \leqslant N$, (a) *Échantillonnage.* Générer $X \sim q$ et $U \sim$ uniforme(0,1) ; (b) *Acceptation.* Si $u < \tilde{\pi}(x)/Mq(x)$, accepter $x_n = x$ et poser $n = n + 1$.
Sortie	L'échantillon $x_{1:N}$ et l'estimé Monte Carlo (2.2).

probabilité d'accepter un candidat est donnée par

$$\mathbb{P}(\text{accept.}) = \int_{\mathcal{X}} \frac{\tilde{\pi}(x)}{Mq(x)} q(x)\,\mathrm{d}x = \frac{1}{M} \int_{\mathcal{X}} \tilde{\pi}(x)\,\mathrm{d}x =: \frac{1}{\widetilde{M}}.$$

C'est donc dire qu'il faudra en moyenne $\widetilde{M}$ candidats afin de produire une nouveau point de l'échantillon. Lorsque $\widetilde{M}$ est grand, plusieurs itérations de l'algorithme sont alors perdues (notons qu'il est possible de recycler les rejets par Rao-Blackwellisation, Casella et Robert, 1998). L'algorithme Metropolis-Hastings, présenté à la sous-section 2.3.1, utilise également le concept d'acceptation/rejet de candidats afin de produire un algorithme de Monte Carlo. Les restrictions sur la densité instrumentale y sont par contre moins fortes de sorte que le problème précédemment exposé est évité. Cependant, cet algorithme requiert l'utilisation de chaînes de Markov ; la prochaine section se voit une introduction à ces concepts préalables à l'étude des algorithmes tels que celui de Metropolis-Hastings.

2.2 Rappels sur les chaînes de Markov

Les méthodes MCMC utilisent les propriétés des chaînes de Markov afin de produire un échantillon qui possède certaines garanties quant à l'estimation de $\pi(f)$ par l'estimateur Monte Carlo. Il est donc de mise de parcourir certaines de ces propriétés avant de se lancer dans l'étude des méthodes MCMC. Cette section constitue donc une revue de la théorie requise à l'analyse de ces méthodes. L'exposition est inspirée de diverses sources telles que Meyn et Tweedie (2009), Robert et Casella (2004), Roberts et Rosenthal (2004) ainsi que de Tierney (1994).

Les notions de probabilité et de théorie de la mesure seront présumées connues du lecteur. On réfère à des ouvrages généraux tels que Williams (1991), Billingsley (2012), Halmos (2013) ou Chung (2001) pour des rappels à ce sujet.

2.2.1 Définitions

Une chaîne de Markov est une suite de variables aléatoire telle que la distribution d'une de ces variables est entièrement déterminée par la réalisation de la variable aléatoire précédente. Il y a donc un ordre naturel au sein de la suite qui s'apparente au temps : la distribution actuelle ne dépend du passé seulement qu'à travers la dernière valeur de la chaîne.

Définition 2.1 (Chaîne de Markov) *Un processus stochastique à temps discret $\{X_n\}_{n\in\mathbb{N}}$ à valeurs dans l'espace d'états $\mathcal{X} \subseteq \mathbb{R}^d$ équipé d'une σ-algèbre $\mathcal{B}(\mathcal{X})$ engendrée par $\{X_n\}_{n\in\mathbb{N}}$ satisfait la **propriété markovienne** si la distribution conditionnelle de X_{n+1} sachant l'état actuel $X_n = x_n$ est indépendante du passé de la chaîne $\{x_i\}_{i<n}$, c.-à-d.,*

$$\mathbb{P}\left(X_{n+1} \in B \,|\, X_i = x_i, i \leqslant n\right) = \mathbb{P}\left(X_{n+1} \in B \,|\, X_n = x_n\right), \qquad \forall B \in \mathcal{B}(\mathcal{X}), x_n \in \mathcal{X}, n \in \mathbb{N}. \quad (2.5)$$

*On dit alors que le processus $\{X_n\}_{n\in\mathbb{N}}$ est une **chaîne de Markov** à valeurs dans l'espace d'états $\mathcal{X}$.*

L'expression du côté droit de (2.5) est une instance du **noyau de transition** de la chaîne au temps n et est souvent écrite d'une façon plus concise lorsqu'il n'y a pas d'ambiguïté : par exemple, on peut noter

$$P_n(B|x) = \mathbb{P}\left(X_{n+1} \in B \,|\, X_n = x\right), \qquad \forall B \in \mathcal{B}(\mathcal{X}), x \in \mathcal{X}. \quad (2.6)$$

Lorsque la distribution de transition est constante dans le temps, c.-à-d., $P_n(B|x) = P(B|x)$ pour tout $n \in \mathbb{N}$, alors la chaîne est dite **homogène** (dans le temps.) Sinon, la chaîne est dite **inhomogène** (dans le temps.) L'ensemble de toutes ces distributions conditionnelles forme une famille appelée le noyau de transition :

Définition 2.2 (Noyau de transition) *La famille*

$$P = \{P(B|x) \,|\, B \in \mathcal{B}(\mathcal{X}), x \in \mathcal{X}\}$$

*est appelée le **noyau de transition** (de Markov) de la chaîne si elle satisfait aux conditions suivantes :*

(i) Pour tout $B \in \mathcal{B}(\mathcal{X})$, $P(B|\cdot)$ est une fonction mesurable positive sur $\mathcal{X}$;

(ii) Pour tout $x \in \mathcal{X}$, $P(\cdot|x)$ est une mesure de probabilité sur $\mathcal{B}(\mathcal{X})$.

Remarque 2.1 *La notation utilisée pour le noyau de transition varie selon les sources. Par exemple, on retrouve souvent les écritures $P(x,B)$, $P(x;B)$ ou même $P(x \to B)$ dans la littérature pour*

exprimer la même quantité. La notation $P(B|x)$ sera privilégiée ici afin de mettre l'emphase sur le fait qu'il s'agit d'une distribution conditionnelle. De plus, le choix de la lettre P se justifie par la concaténation dans (2.6), mais on retrouve parfois T pour « transition » ou bien K de l'anglais « kernel » signifiant « noyau ».

Dans bien des cas, l'espace d'états $\mathcal{X}$ sera contenu dans $\mathbb{R}^d$ pour un certain $d \in \mathbb{N}$ et toutes les distributions conditionnelles de la famille supposeront des densités. Dans ce cas, on parlera de densité de transition.

Définition 2.3 (Densité de transition) *Soit $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov homogène à espace d'états $\mathcal{X} \subset \mathbb{R}^d$. Si le noyau de transition P admet une densité (la dérivée de Radon-Nykodym) par rapport à la mesure de Lebesgue λ_{Leb} sur $\mathcal{X}$, alors il est possible de considérer la **densité de transition** (de Markov) $p(\cdot|x) = \mathrm{d}P(\cdot|x)/\,\mathrm{d}\lambda_{\text{Leb}}$ satisfaisant*

$$P(B|x) = \int_B p(y|x)\,\mathrm{d}y, \qquad B \in \mathcal{B}(\mathcal{X}).$$

Les distributions conditionnelles du noyau de transition expriment la distribution du prochain état de la chaîne sachant l'état actuel. Lorsque la chaîne est homogène, la distribution conditionnelle de l'état de la chaîne dans m pas s'exprime d'une façon récursive appelée le noyau de transition itéré.

Définition 2.4 (Noyau de transition itéré) *Soit $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov homogène à espace d'états $\mathcal{X}$. Le **noyau de transition itéré de m pas**, noté P^m, correspond à la distribution conditionnelle de X_{n+m} sachant $X_n = x_n$ et peut être défini de la manière récursive suivante. Pour $B \in \mathcal{B}(\mathcal{X})$ et $x \in \mathcal{X}$*

$$P^m(B|x) = \int_{\mathcal{X}} P^{m-1}(B|y)P(\,\mathrm{d}y|x), \qquad m > 1, \tag{2.7}$$

$$P^1(B|x) = P(B|x). \tag{2.8}$$

Une propriété importante dans l'étude des chaînes de Markov est la ϕ-irréductibilité ; cette propriété s'interprète de la façon suivante. Pour une certaine mesure de probabilité ϕ sur $\mathcal{X}$, alors tout ensemble important (de probabilité ϕ non-nulle) peut être atteint, avec probabilité positive, en un nombre fini de pas à partir de n'importe quel point $x \in \mathcal{X}$. Ainsi, la chaîne de Markov visitera avec probabilité positive toute région de probabilité non-nulle ce qui garantit que l'espace visité par la chaîne ne se réduise pas avec n.

Définition 2.5 (ϕ-irréducibilité) *Soient $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov homogène à espace d'états $\mathcal{X}$, P son noyau de transition et ϕ une mesure sur $\mathcal{X}$. Alors, la chaîne est dite **ϕ-irréductible** si, pour tout ensemble mesurable $B \in \mathcal{B}(\mathcal{X})$ avec $\phi(B) > 0$ et pour tout $x \in \mathcal{X}$, il existe un entier $m \in \mathbb{N}$, dépendant possiblement de x et de B, satisfaisant $P^m(B|x) > 0$.*

Proposition 2.2 (Meyn et Tweedie, 2009, proposition 4.2.1) *Un noyau de transition P est ϕ-irréductible si et seulement si, pour tout $x \in \mathcal{X}$ et tout $B \in \mathcal{B}(\mathcal{X})$ tel que $\phi(B) > 0$, il existe $m \in \mathbb{N}$ tel que $P^m(B|x) > 0$.*

Remarque 2.3 *La proposition 2.2 est différente de la définition de la ϕ-irréductibilité en ce que l'entier m peut dépendre à la fois de x et de B. Cette propriété permettra de simplifier certaines preuves éventuellement.*

Une seconde propriété nécessaire à l'analyse des chaînes de Markov est la récurrence. Intuitivement, une chaîne de Markov récurrente est telle que tout ensemble important (de probabilité non-nulle) sera visité infiniment souvent peu importe l'état initial de la chaîne. Il s'agit donc d'une propriété plus forte que la ϕ-irréductibilité.

Définition 2.6 (Temps d'occupation, premier retour et première visite) *Soient $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov à espace d'états $\mathcal{X}$ et $B \in \mathcal{B}(\mathcal{X})$ un ensemble mesurable. Le **temps d'occupation***

de B est défini comme le nombre de visites de l'ensemble B par la chaîne, c.-à-d.,

$$\eta_B := \sum_{n \geqslant 1} \mathbb{1}(X_n \in B). \tag{2.9}$$

*Le **premier retour** à B est défini selon*

$$\tau_B := \begin{cases} \min\{n \geqslant 1 \,|\, X_n \in B\}, & \bigcup_{n \geqslant 1} \{X_n \in B\} \neq 0 \\ \infty, & \textit{sinon.} \end{cases} \tag{2.10}$$

Définition 2.7 (Récurrence) *Soit $\{X_n\}_{n \in \mathbb{N}}$ une chaîne de Markov à espace d'états $\mathcal{X}$. La chaîne est dite **récurrente** si les conditions suivantes sont satisfaites :*

(i) Il existe une mesure ϕ telle que la chaîne est ϕ-irréductible ;

(ii) Pour tout $B \in \mathcal{B}(\mathcal{X})$ tel que $\phi(B) > 0$ et tout état initial $x_0 \in \mathcal{X}$, $\mathbb{E}\{\eta_B \,|\, X_0 = x_0\} = \infty$.

La propriété d'Harris-récurrence renforce celle de récurrence au sens où l'on exige que la chaîne de Markov passe infiniment souvent par l'ensemble B avec probabilité un plutôt qu'en moyenne. Ainsi, pour une chaîne Harris-récurrente, seulement un ensemble de réalisations ayant une probabilité nulle évitera complètement un ensemble de probabilité non-nulle. Il en résulte que le processus visitera avec probabilité un tout l'espace important (par rapport à ϕ.)

Définition 2.8 (Harris-récurrence) *Soit $\{X_n\}_{n \in \mathbb{N}}$ une chaîne de Markov à espace d'états $\mathcal{X}$. Un ensemble mesurable $B \in \mathcal{B}(\mathcal{X})$ est dit **Harris-récurrent** si $\mathbb{P}(\eta_B = \infty \,|\, X_0 = x_0) = 1$ pour toute valeur initiale x_0. La chaîne est dite **Harris-récurrente** si les conditions suivantes sont satisfaites :*

(i) Il existe une mesure ϕ telle que la chaîne est ϕ-irréductible ;

(ii) Pour tout $B \in \mathcal{B}(\mathcal{X})$ tel que $\phi(B) > 0$, B est Harris-récurrent.

Le but d'utiliser des chaînes de Markov est de produire un échantillon Monte Carlo dont la distribution des éléments est (ou du moins s'approche de) celle d'une certaine distribution cible. Une propriété intéressante des chaînes de Markov homogènes à cet effet est l'existence d'une distribution stationnaire. Ce type de distribution est tel que, lorsque le processus se trouve dans cette distribution, alors le processus restera dans cette distribution malgré la transition.

Définition 2.9 (Mesure invariante et chaîne positive) *Soit un noyau de transition P sur un espace d'états $\mathcal{X}$. Une mesure σ-finie Π sur $\mathcal{B}(\mathcal{X})$ est dite **invariante** pour P si le noyau P préserve la mesure Π, c.-à-d.,*

$$\Pi(B) = \int_{\mathcal{X}} P(B|x)\Pi(\mathrm{d}x), \qquad \forall B \in \mathcal{B}(\mathcal{X}).$$

*Lorsqu'il existe une mesure de probabilité qui est invariante pour une chaîne de Markov homogène ϕ-irréductible, alors la chaîne est dite **positive**.*

Proposition 2.4 (Meyn et Tweedie, 2009, proposition 10.1.1) *Une chaîne de Markov homogène positive $\{X_n\}_{n \in \mathbb{N}}$ est récurrente. De plus, si la chaîne est Harris-récurrente et positive, alors la chaîne est dite **Harris-positive**.*

Le théorème suivant établit la (presque) suffisance de la récurrence pour assurer la positivité de la chaîne. En pratique, les chaînes de Markov sont construites pour admettre une certaine mesure invariante et la récurrence sera ensuite à prouver, mais ce résultat est intéressant dans la mesure où on obtient l'unicité de la mesure invariante.

Définition 2.10 (Petit ensemble) *Soit $\{X_n\}_{n \in \mathbb{N}}$ une chaîne de Markov homogène à noyau de transition P. Un ensemble mesurable $C \in \mathcal{B}(\mathcal{X})$ est dit **petit** s'il existe $m > 0$ et une mesure non-triviale ν_m sur $\mathcal{B}(\mathcal{X})$ telle que, pour tout $x \in C$ et $B \in \mathcal{B}(\mathcal{X})$, la **condition de minorisation** suivante est satisfaite*

$$P^m(B|x) \geqslant \nu_m(B).$$

*Spécifiquement, on dit que C est ν_m**-petit**.*

Théorème 2.5 (Meyn et Tweedie, 2009, théorème 10.0.1) *Une chaîne de Markov homogène récurrente $\{X_n\}_{n\in\mathbb{N}}$ admet une mesure invariante Π qui est unique à une constante de proportionnalité près et qui satisfait la représentation suivante*

$$\Pi(B) = \int_A \mathbb{E}\left\{\sum_{n=1}^{\tau_A} \mathbb{1}(X_n \in B) \,\middle|\, X_0 = x_0\right\} \Pi(\mathrm{d}x_0), \qquad \forall A,B \in \mathcal{B}(\mathcal{X}).$$

De plus, la mesure Π est finie s'il existe un petit ensemble C tel que

$$\sup_{x\in C} \mathbb{E}\left\{\tau_C \,|\, X_0 = x_0\right\} < \infty,$$

et (Meyn et Tweedie, 2009, théorème 10.4.9) la mesure Π est équivalente à toute mesure ϕ telle que la ϕ-irréductibilité implique la récurrence de la chaîne.

Lorsque cette unicité est établie, la constante de proportionnalité peut être choisie de sorte à définir une mesure de probabilité ; on parle alors de distribution stationnaire de la chaîne.

Définition 2.11 (Distribution stationnaire) *Soit $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov homogène positive (donc récurrente) à mesure invariante Π finie. La mesure Π, après mise à l'échelle, correspond à une mesure de probabilité sur $\mathcal{B}(\mathcal{X})$ qui constitue la **distribution stationnaire** de la chaîne. Lorsque $X_n \sim \pi$ pour un certain $n \in \mathbb{N}$, alors on trouve $X_m \sim \pi$ pour tout $m \geqslant n$ et la chaîne est alors dite **stationnaire**.*

Afin de vérifier la positivité de la chaîne, une condition suffisante est la réversibilité. Une chaîne réversible sera stationnaire et progressera avec la même distribution vers le futur que vers le passé. De plus, cette propriété peut être elle-même démontrée à l'aide de la condition d'équilibre.

Définition 2.12 (Réversibilité) *Une chaîne de Markov stationnaire $\{X_n\}_{n\in\mathbb{N}}$ est dite **réversible** si la distribution conditionnelle de X_{n+1} sachant $X_{n+2} = x$ est la même que la distribution conditionnelle de de X_{n+1} sachant $X_n = x$.*

Définition 2.13 (Condition d'équilibre) *Un noyau de transition P sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ satisfait la **condition d'équilibre** s'il existe une mesure Π satisfaisant*

$$\int_A P(B|x)\Pi(\mathrm{d}x) = \int_B P(A|x)\Pi(\mathrm{d}x), \qquad \forall A,B \in \mathcal{B}(\mathcal{X}). \tag{2.11}$$

Proposition 2.6 *Soit P un noyau de transition sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ admettant une densité $p(\cdot|x)$ pour tout $x \in \mathcal{X}$ et soit Π une mesure sur $\mathcal{B}(\mathcal{X})$ admettant une densité π. Si la condition*

$$p(y|x)\pi(x) = p(x|y)\pi(y), \qquad \forall x,y \in \mathcal{X} \tag{2.12}$$

est satisfaite, alors P satisfait la condition d'équilibre pour Π.

Démonstration. On vérifie directement l'égalité (2.11). Soit $A,B \in \mathcal{B}(\mathcal{X})$, alors en appliquant la condition d'équilibre des densités puis en changeant l'ordre d'intégration (théorème de Fubini), on obtient

$$\begin{aligned}
\int_A P(B|x)\Pi(\mathrm{d}x) &= \int_A \left[\int_B p(y|x)\,\mathrm{d}y\right] \pi(x)\,\mathrm{d}x \\
&= \int_A \int_B p(y|x)\pi(x)\,\mathrm{d}y\mathrm{d}x \\
&= \int_A \int_B p(x|y)\pi(y)\,\mathrm{d}y\mathrm{d}x \\
&= \int_B \left[\int_A p(x|y)\,\mathrm{d}x\right] \pi(y)\,\mathrm{d}y \\
&= \int_B P(A|y)\Pi(\mathrm{d}y),
\end{aligned}$$

ce qui prouve la condition d'équilibre en changeant les étiquettes des variables d'intégration. □

Théorème 2.7 (Robert et Casella, 2004, théorème 6.2.2) *Soit $\{X_n\}_{n\in\mathbb{N}}$, une chaîne de Markov homogène à noyau de transition P satisfaisant la condition d'équilibre pour une distribution Π. Alors, la chaîne est réversible et admet donc Π comme distribution invariante.*

Finalement, une dernière propriété utile à l'étude des chaînes de Markov est la périodicité. Une chaîne de Markov sera périodique s'il existe une suite finie d'ensembles de probabilité positive dans laquelle la chaîne peut rester prise. Alternativement, l'apériodicité de la chaîne est une propriété recherchée qui exige l'absence de tels cycles périodiques.

Définition 2.14 (Périodicité) *Soit $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov homogène, ϕ-irréductible et à noyau de transition P. La chaîne est dite **périodique** s'il existe un entier $m \geqslant 2$ et une suite d'ensembles mesurables $(S_0,S_1,\ldots,S_m) \subseteq \mathcal{B}(\mathcal{X})$ tous disjoints excepté $S_0 = S_m$ satisfaisant, pour tout $i \in \{0,1,\ldots,m-1\}$, $\phi(S_i) > 0$ et*

$$P(S_{i+1}|x) = 1, \qquad \forall x \in S_i.$$

*Un tel ensemble cyclique est appelé un **m-cycle**. La **période** de la chaîne $\{X_n\}_{n\in\mathbb{N}}$ est le plus grand m tel qu'un m-cycle existe.*

*Si la chaîne n'est pas périodique, c.-à-d., si la période de la chaîne est $m = 1$, alors on dit qu'elle est **apériodique**.*

Une condition suffisante à l'apériodicité d'une chaîne est l'apériodicité forte où l'on requiert l'existence d'un ensemble de probabilité positive tel que la chaîne a une probabilité positive de rester dans l'ensemble lorsqu'elle s'y trouve déjà.

Définition 2.15 (Apériodicité forte) *Soit $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov homogène. La chaîne est dite **fortement apériodique** s'il existe un ν_1-petit ensemble $B \in \mathcal{B}(\mathcal{X})$ avec $\nu_1(B) > 0$.*

2.2.2 Convergence

La notion de distribution stationnaire suppose que le processus a déjà atteint la distribution stationnaire. Cependant, s'il est possible d'obtenir un processus déjà dans cette situtation, il ne sera souvent pas nécessaire de produire une chaîne de Markov pour générer l'échantillon Monte Carlo. En général, il ne sera pas possible d'atteindre directement la stationnarité et l'on cherchera donc plutôt à s'en approcher. On peut s'attendre intuitivement à ce qu'une transition invariante modifie une distribution non-stationnaire vers une distribution un peu plus près de la distribution stationnaire. À la longue, on peut donc espérer que la chaîne converge vers cette distribution stationnaire cible. Cette convergence peut être définie de diverses manières et cette sous-section en explore quelques-unes.

2.2.2.1 Ergodicité

Lorsque la chaîne de Markov n'est pas stationnaire mais que le noyau de transition admet une distribution invariante, on peut s'attendre intuitivement à ce que la distribution marginale de la chaîne converge vers la distribution stationnaire. C'est ce que la propriété d'**ergodicité** d'une chaîne de Markov signifie.

Définition 2.16 (Ergodicité) *Une chaîne de Markov $\{X_n\}_{n\in\mathbb{N}}$ est dite **ergodique** à Π si la transition itérée à n pas converge en variation totale vers Π, c.-à-d.,*

$$\lim_{n\to\infty} \|P^n(\cdot|x) - \Pi(\cdot)\|_{\mathrm{TV}} = 0,$$

où $\|\mu\|_{\mathrm{TV}}$ *dénote la norme de variation totale de la mesure signée* μ *et est donnée par*

$$\|\mu\|_{\mathrm{TV}} = \sup_{B \in \mathcal{B}(\mathcal{X})} |\mu|(B), \tag{2.13}$$

où $|\mu|$ *est la mesure de variation totale induite par* μ*.*

Remarque 2.8 *La norme de variation totale possède une définition alternative qui est équivalente à (2.13). En effet, on retrouve parfois l'expression suivante :*

$$\|\mu\|_{\mathrm{TV}} = \left[\sup_{B \in \mathcal{B}(\mathcal{X})} \mu(B)\right] - \left[\inf_{B \in \mathcal{B}(\mathcal{X})} \mu(B)\right]. \tag{2.14}$$

Remarque 2.9 *Selon le contexte, l'ergodicité est définie d'une manière différente par rapport à* $x \in \mathcal{X}$*. On trouve parfois l'ergodicité pour un certain choix de* x*, parfois pour* Π*-presque tous les* x*, parfois pour tout* $x \in \mathcal{X}$ *et parfois pour toute distribution initiale* μ *sur* $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$*. Les résultats d'ergodicité préciseront la condition sur la valeur initiale* x*. Notons que la condition la plus forte est de supposer l'ergodicité pour toute distribution initiale, puisque celles-ci contiennent les distributions dégénérées* δ_x *pour tout* $x \in \mathcal{X}$ *ainsi que pour la distribution stationnaire* Π*.*

Un ensemble de trois conditions, ou des variantes de celles-ci, est suffisant pour assurer l'ergodicité d'une chaîne de Markov : la positivité, la récurrence et l'apériodicité. L'apériodicité assure que la chaîne ne reste pas prise dans une certain cycle d'ensembles. La récurrence assure quant à elle que tous les ensembles à probabilité positive soient visités par la chaîne. Ensemble, ces deux conditions permettent à la chaîne d'atteindre la mesure invariante. Enfin, une chaîne positive admet une distribution invariante ; ainsi, lorsque la chaîne atteint la bonne distribution, elle tend à conserver cette distribution ce qui correspond alors à la convergence de la distribution. Ceci justifie intuitivement la suffisance de ces trois conditions pour l'ergodicité de la chaîne.

Théorème 2.10 (Théorème d'ergodicité I, Tierney, 1994, théorème 1) *Soit* $\{X_n\}_{n \in \mathbb{N}}$*, une chaîne de Markov* Π*-irréductible et admettant* Π *comme distribution invariante. Alors, la chaîne de Markov est positive récurrente et, si la chaîne est apériodique, alors elle est ergodique à* Π *pour* Π*-presque toute valeur initiale* x*.*

Théorème 2.11 (Théorème d'ergodicité II, Meyn et Tweedie, 2009, théorème 13.3.3 et Tierney, 1994, théorème 1) *Soit* $\{X_n\}_{n \in \mathbb{N}}$*, une chaîne de Markov apériodique, Harris-récurrente et admettant* Π *comme distribution invariante. Alors, la chaîne de Markov satisfait l'ergodicité par rapport à toute distribution initiale* μ *sur* $\mathcal{B}(\mathcal{X})$*, c.-à-d.,*

$$\lim_{n \to \infty} \left\| \int_{\mathcal{X}} P^n(\cdot|x_0)\mu(\,\mathrm{d}x_0) - \Pi(\cdot) \right\|_{\mathrm{TV}} = 0.$$

En particulier, la chaîne est ergodique à Π *pour toute valeur initiale* $x \in \mathcal{X}$*.*

Proposition 2.12 (Tierney, 1994, corollaire 1) *Soit* P *un noyau de transition Markov qui soit* Π*-irréductible et qui admet* Π *comme distribution invariante (donc positive.) Si* $P(\cdot|x)$ *admet une densité par rapport à* Π *pour tout* $x \in \mathcal{X}$*, alors* P *est Harris-recurrent.*

En résumé, afin de démêler toutes ces définitions et résultats, une chaîne de Markov sera ergodique à condition que

(a) la chaîne admette Π comme distribution invariante, ce qui peut être assuré par la condition d'équilibre ;

(b) la chaîne soit Π-irréductible ;

(c) la chaîne soit apériodique.

Sous ces conditions, l'ergodicité sera pour Π-presque toute valeur initiale. Pour que l'ergodicité soit pour toute valeur initiale, on requiert la condition plus forte d'Harris-récurrence ; notons cependant que des résultats tels que la proposition 2.12 peuvent être utilisés afin de montrer l'Harris-récurrence à partir de la Π-irréductibilité.

La vérification de ces trois conditions dépend de P et de sa distribution stationnaire Π. Lorsqu'il sera question des algorithmes MCMC, la distribution Π sera donnée et l'algorithme Monte Carlo sera construit de sorte à produire une transition P qui respecte ces trois conditions. Bien que la condition d'équilibre ne soit pas nécessaire, il s'agit d'une manière particulièrement simple d'assurer que la chaîne admette la bonne distribution stationnaire. Ainsi, la plupart des algorithmes MCMC reposent sur la condition d'équilibre ; les deux autres conditions – la Π-irréductibilité et l'apériodicité – se vérifient relativement facilement dans la majorité des cas. Certaines applications de ces résultats seront considérées à la section 2.3 afin de démontrer la validité de certains algorithmes MCMC couramment utilisés.

2.2.2.2 V-ergodicité

L'ergodicité d'une chaîne de Markov est définie en fonction de la norme induite par la variation totale. Il est possible de généraliser cette définition à des normes plus générales et dont la variation totale est un cas particulier.

Définition 2.17 (V-norme d'une mesure signée) *Soit $V : \mathcal{X} \to [1,\infty)$ une fonction. Alors, la **V-norme d'une mesure signée** $\mu : \mathcal{B}(\mathcal{X}) \to \overline{\mathbb{R}}$ sur une σ-algèbre $\mathcal{B}(\mathcal{X})$ de l'espace $\mathcal{X}$ est définie selon*

$$\|\mu\|_V = \sup_{g:|g|\leqslant V} |\mu(g)|,$$

Définition 2.18 (V-ergodicité) *Soit $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov Harris-positive admettant Π comme distribution invariante. S'il existe une fonction $V \geqslant 1$ avec $\Pi(V) < \infty$ telle que*

$$\lim_{n\to\infty} \|P^n(\cdot|x) - \Pi(\cdot)\|_V = 0, \qquad \forall x \in \mathcal{X},$$

*alors la chaîne est dite **V-ergodique** à Π.*

Remarque 2.13 *Lorsque $V \equiv 1$, la V-ergodicité correspond à l'ergodicité en terme de la norme de variation totale. En effet, il est possible de montrer l'équivalence suivante :*

$$\|\mu\|_{\mathrm{TV}} = \frac{1}{2} \sup_{h:|h|\leqslant 1} |\mu(h)| = \frac{1}{2}\|\mu\|_1.$$

Les théorèmes de V-ergodicité demandent un ensemble de définitions supplémentaires afin de détailler les conditions nécessaires. Cependant, la supposition additionnelle principale par rapport aux théorèmes d'ergodicité (e.g. théorèmes 2.10 et 2.11) sera d'exiger que $\Pi(V) < \infty$ (voir Meyn et Tweedie, 2009, chapitre 14, pour plus de détails.)

Remarque 2.14 *La V-ergodicité implique également (Robert et Casella, 2004, théorème 4.6.7) que l'espérance d'une certaine fonction f sous la distribution marginale converge vers l'espérance sous la distribution stationnaire de cette même fonction à condition que $|f| \leqslant |V|$. En particulier, l'ergodicité implique la convergence de l'espérance pour toute fonction bornée.*

2.2.2.3 Ergodicité V-géométrique

L'ergodicité et la V-ergodicité supposent uniquement la convergence de la distribution itérée vers la distribution stationnaire : aucune supposition n'est faite sur le rythme de convergence. Quant à elle, l'ergodicité V-géométrique suppose un rythme de convergence géométrique, c'est-à-dire exponentiel décroissant en n.

Définition 2.19 (Ergodicité V-géométrique) *Soit $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov Harris-positive avec Π comme distribution invariante. S'il existe une fonction $V \geqslant 1$ avec $\Pi(V) < \infty$ et*

une constante $r_V > 1$ telles que, pour tout $x \in \mathcal{X}$,

$$\sum_{n \geqslant 1} r_V^n \|P^n(\cdot|x) - \Pi\|_V < \infty, \tag{2.15}$$

*alors la chaîne est dite V-**géométriquement ergodique** à Π. Lorsque $V \equiv 1$ alors la chaîne est dite **géométriquement ergodique**. Lorsque la condition (2.15) n'est vérifiée que pour Π-presque tout x, alors la chaîne est dite Π-**p.s.** V-**géométriquement ergodique**.*

Remarque 2.15 *L'ergodicité géométrique est parfois définie d'une manière différente, mais équivalente. Tierney (1994) la définit par l'existence d'une fonction $M : \mathcal{X} \to [0,\infty]$ avec $\Pi(M) < \infty$ et d'un $r < 1$ tels que, pour tout $x \in \mathcal{X}$,*

$$\|P^n(\cdot|x) - \Pi\|_{\mathrm{TV}} \leqslant M(x) r^n,$$

alors que Roberts et Rosenthal (2004) la définissent plutôt par l'existence d'une fonction $M : \mathcal{X} \to [0,\infty)$ et d'un $r < 1$ tels que

$$\|P^n(\cdot|x) - \Pi\|_{\mathrm{TV}} \leqslant M(x) r^n, \qquad \text{presque partout } \Pi.$$

Le lien entre ces définitions et la définition 2.19 est relativement évident puisque l'ergodicité V-géométrique implique une décroissance géométrique :

$$\|P^n(\cdot|x) - \Pi\|_V \leqslant M(x) r_V^{-n},$$

où

$$M(x) = \sum_{n \geqslant 1} r_V^n \|P^n(\cdot|x) - \Pi\|_V.$$

Afin de vérifier l'ergodicité V-géométrique d'une chaîne de Markov, certaines conditions de dérive géométrique, parfois appelées condition de dérive de Foster-Lyapunov, peuvent être utilisées. On considère ici un cas particulier de la condition générale introduite par Meyn et Tweedie (2009, section 14.2.1, condition V3).

On introduit d'abord une notation pratique. Pour une fonction $V : \mathcal{X} \to \mathbb{R}$ et une transition P sur $\mathcal{X}$, l'espérance conditionnelle de $V(X_{n+1})$ sachant $X_n = x$ est notée $PV(x)$, ce qui définit la fonction $PV : \mathcal{X} \to \mathbb{R}$ et qui s'exprime explicitement par

$$PV(x) = \int_{\mathcal{X}} V(y) P(\mathrm{d}y|x).$$

Définition 2.20 (Dérive géométrique vers un ensemble C, Jarner et Hansen, 2000) *Un noyau de transition Markov P satisfait une **condition de dérive géométrique vers** C s'il existe des constantes $0 < \lambda < 1$ et $b < \infty$, possiblement dépendantes sur C, ainsi qu'une fonction $V : \mathcal{X} \to [1,\infty]$ finie pour au moins un x telles que*

$$PV(x) \leqslant \lambda V(x) + b \mathbb{1}_C(x), \qquad \forall\, x \in \mathcal{X}\,. \tag{2.16}$$

Remarque 2.16 *Dans une notation plus compacte, on écrit souvent (2.16) dans la forme $PV \leqslant \lambda V + b\,\mathbb{1}_C$. De plus, une condition suffisante à (2.16) couramment utilisée et plus intuitive à vérifier est la suivante :*

$$PV(x) \leqslant \begin{cases} \lambda V(x), & x \in \mathcal{X} \setminus C, \\ b, & x \in C. \end{cases}$$

Théorème 2.17 (Meyn et Tweedie, 2009, théorème 15.0.1) *Soit P un noyau de transition Markov ϕ-irréductible et apériodique. S'il existe une petit ensemble $C \in \mathcal{B}(\mathcal{X})$ tel que P satisfait à la condition de dérive géométrique vers C pour une certaine fonction $V \geqslant 1$ telle que $\pi(V) < \infty$, alors la chaîne est π-p.s. V-géométriquement ergodique à π. En particulier, la convergence géométrique du noyau itéré vers π est pour tout x tel que $V(x) < \infty$.*

2.2.2.4 Ergodicité V-uniforme

L'ergodicité V-géométrique demande la convergence de la distribution itérée vers la distribution cible en norme V pour tout état initial de la chaîne. Une telle convergence est uniforme au sens où le rythme de convergence doit être uniforme par rapport à l'état initial alors que ceci n'est pas nécessaire à la V-ergodicité. L'*ergodicité V-uniforme* est plus forte que l'ergodicité V-géométrique puisque la constante de proportionnalité associée au rythme de convergence géométrique, $M(x)$, peut varier selon $x \in \mathcal{X}$ et est possiblement non-bornée.

Définition 2.21 (V-norme d'un noyau de transition signé) *Soit $V : \mathcal{X} \to [1,\infty)$ une fonction. Alors, la V-**norme d'un noyau de transition signé** $P : \mathcal{B}(\mathcal{X}) \times \mathcal{X} \to \overline{\mathbb{R}}$ sur une σ-algèbre $\mathcal{B}(\mathcal{X})$ et l'espace $\mathcal{X}$ est définie selon*

$$\left|\!\left|\!\left| P \right|\!\right|\!\right|_V = \sup_{x \in \mathcal{X}} \frac{\|P(\cdot|x)\|_V}{V(x)}.$$

Définition 2.22 (Ergodicité V-uniforme) *Soit $\{X_n\}_{n\in\mathbb{N}}$ une chaîne de Markov ergodique à Π. S'il existe une fonction $V \geqslant 1$ telle que*

$$\lim_{n\to\infty} \left|\!\left|\!\left| P^n - \Pi \right|\!\right|\!\right|_V = 0, \tag{2.17}$$

*avec la convention $\Pi(B) \equiv \Pi(B|x)$, alors la chaîne est dite V-**uniformément ergodique** à Π. Lorsque $V \equiv 1$ alors la chaîne est dit **uniformément ergodique** et cette définition correspond à l'existence d'un $M < \infty$ et d'un $r < 1$ tels que*

$$\|P^n(\cdot|x) - \Pi\|_{\mathrm{TV}} \leqslant M r^n, \qquad \forall\, x \in \mathcal{X}\,.$$

Pour bien comprendre le concept d'ergodicité V-uniforme par rapport à celui de V-ergodicité, on peut dresser un parallèle avec la convergence d'une suite de fonction. En effet, l'ergodicité V-uniforme est à la V-ergodicité ce que la convergence uniforme est à la convergence point à point. L'ergodicité V-uniforme est donc un concept global—uniforme—par rapport à $x \in \mathcal{X}$. Ainsi, la vitesse de convergence ne dépends donc naturellement pas de x pour ce mode de convergence alors qu'elle en dépends pour la V-ergodicité et cette distinction est la même que pour la convergence de suite de fonctions.

2.2.3 Théorèmes limites

L'ergodicité assure que la chaîne de Markov converge en distribution vers une certaine distribution Π. Dans le contexte des méthodes de Monte Carlo, ce genre de propriété n'est pas exactement ce qui est recherché, bien qu'il s'agisse d'une garantie intéressante. Il est plus pertinent de se renseigner sur les propriétés de l'estimateur Monte Carlo lui-même plutôt que sur celles de l'échantillon. Deux types de résultats permettent de connaître le comportement asymptotique de cet estimateur : une loi des grands nombres et un théorème limite central.

2.2.3.1 Loi des grands nombres

Les résultats du type « loi des grands nombres fonctionnelle » établissent que la moyenne d'une fonction de la chaîne converge vers l'espérance de cette fonction sous une certaine distribution. Il y a cependant un lien étroit avec le concept d'ergodicité puisque l'Harris-positivité est suffisante pour assurer la loi forte des grands nombres pour toute fonction Π-intégrable. De plus, l'ergodicité est parfois suffisante pour assurer l'Harris-positivité.

Théorème 2.18 (Loi forte des grands nombres I, Meyn et Tweedie, 2009, théorème 17.1.7) *Soit $\{X_n\}_{n\in\mathbb{N}}$, une chaîne de Markov à espace d'états $\mathcal{X}$ qui soit Harris-positive avec Π comme distribution stationnaire. Pour toute fonction $f : \mathcal{X} \to \mathbb{R}$ telle que $\Pi(|f|) < \infty$ on a la* ***loi forte des grands nombres****, c.-à-d.,*

$$\frac{1}{N}\sum_{n=1}^{N} f(X_n) \xrightarrow{\text{p.s.}} \Pi(f), \qquad n \to \infty,$$

où la convergence presque sûre est par rapport à la mesure de probabilité sur le processus au complet.

Théorème 2.19 (Loi forte des grands nombres II, Meyn et Tweedie, 2009, théorème 17.3.2) *Soit $\{X_n\}_{n\in\mathbb{N}}$, une chaîne de Markov à espace d'états $\mathcal{X}$ qui soit Harris-récurrente avec Π comme mesure invariante σ-finie. Pour toute fonction f,$g : \mathcal{X} \to \mathbb{R}$ telle que $\Pi(|f|)$,$\Pi(|g|) < \infty$ on a la* ***loi forte des grands nombres*** *pour le ratio, c.-à-d.,*

$$\frac{\sum_{n=1}^{N} f(X_n)}{\sum_{n=1}^{N} g(X_n)} \xrightarrow{\text{p.s.}} \frac{\Pi(f)}{\Pi(g)}, \qquad n \to \infty.$$

En particulier, ce résultat tient lorsque Π est une mesure de probabilité.

Remarque 2.20 *Par abus de langage et afin d'alléger le texte, on utilisera l'expression « loi des grands nombres » pour désigner plusieurs types de résultats. En effet, une distinction est souvent faite entre une loi des grands nombres pour un échantillon indépendant—comme il fut question à la section 2.1.1 sur les méthodes MC standards i.i.d.—et une* ***loi des grands nombres pour chaînes de Markov****. Le second type est en fait un cas particulier des théorèmes ergodiques. Un* ***théorème ergodique****, dans le contexte de la théorie de la mesure, est un résultat du type*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} f\left(T^t x\right) = \frac{1}{\mu(\mathcal{X})}\int f \,\mathrm{d}\mu, \tag{2.18}$$

où $T : \mathcal{X} \to \mathcal{X}$ est un système dynamique mesuré (c.-à-d. une transformation μ-invariante) sur un espace mesuré $(\mathcal{X},\Sigma,\mu)$, où f est μ-intégrable ($f \in \mathcal{L}^1(\mu)$) et où $x \in \mathcal{X}$ est l'état initial du système. L'interprétation d'un théorème ergodique est la suivante : la moyenne temporelle, donnée par le côté gauche de l'égalité dans (2.18), est égale à la moyenne dans l'espace, donnée par le côté droit de l'égalité dans (2.18).

Dans le cas qui nous concerne, T est la transition de Markov qui construit l'échantillon $\{x_t\}_{t\geqslant 0}$ avec $x_k = T^t x$, μ est la distribution cible qui intègre donc à 1 ($\mu(\mathcal{X}) = 1$), $\mu(f) = \int f \,\mathrm{d}\mu$ est l'espérance recherchée et $x = x_0$ est l'état initial de la chaîne de Markov.

On voit dès lors le lien direct entre ergodicité d'une chaîne de Markov et la loi des grands nombres. En effet, l'ergodicité de la chaîne correspond à l'ergodicité de la transformation T qui assure du coup l'égalité (2.18) et donc la loi des grands nombres.

2.2.3.2 Théorème limite central

La loi des grands nombres assure que la moyenne échantillonnale converge vers l'espérance. Cependant, ce type de résultat ne renseigne pas sur la distribution asymptotique de la moyenne. Cette distribution est particulièrement intéressante afin de mesurer la précision de l'estimation de l'espérance produite par la moyenne. La classe des théorèmes limites centraux permettent d'établir la normalité de la distribution asymptotique, ce qui permet ensuite de calculer des erreurs standards Monte Carlo ou des régions de confiance pour l'estimation. On considère ici des conditions sous lesquelles une chaîne de Markov satisfait à un théorème central limite. Afin de simplifier la notation, on écrira $\bar{f} = f - \Pi(f)$ de sorte que $\Pi(\bar{f}) = 0$.

Définition 2.23 (Théorème limite central) *Pour une chaîne de Markov $\{X_n\}_{n\in\mathbb{N}}$, on dit qu'un* ***théorème central limite*** *est satisfait pour une fonction f s'il existe une constante $0 < \sigma_f^2 < \infty$, la* ***variance asymptotique****, telle que pour toute valeur initiale $x_0 \in \mathcal{X}$, on a*

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{N}\sigma_f}\sum_{n=1}^{N} \bar{f}(X_n) \leqslant z \,\middle|\, X_0 = x_0\right) = \Phi(z),$$

où Φ dénote la fonction de répartition d'une variable aléatoire normale centrée réduite. En particulier,

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} f(X_n) \xrightarrow{\mathcal{D}} \mathcal{N}(\Pi(f), \sigma_f^2), \qquad N \to \infty.$$

En considérant la distribution dégénérée comme une distribution normale à variance nulle, cette définition se généralise au cas $\sigma_f^2 = 0$.

Remarque 2.21 *Bien que la notation semble indiquer le contraire, σ_f^2 ne dépend pas que de f : la variance asymptotique dépend évidemment de Π, mais surtout du noyau de transition P. Un domaine important des MCMC est justement l'étude du lien entre σ_f et le choix de P, particulièrement la minimisation de la variance asymptotique parmi une certaine classe de noyaux (section 2.5).*

Théorème 2.22 (Théorème limite central I, Meyn et Tweedie, 2009, théorèmes 17.5.3 et 17.5.4) *Soit $\{X_n\}_{n\in\mathbb{N}}$, une chaîne de Markov à espace d'états $\mathcal{X}$ qui soit Harris-positive avec Π comme distribution stationnaire de sorte que la chaîne est V-géométriquement ergodique pour une certaine fonction $V \geqslant 1$ telle que $\pi(V^2) < \infty$. Alors, pour toute fonction $f : \mathcal{X} \to \mathbb{R}$ telle que $f^2 \leqslant V$, la chaîne de Markov satisfait la loi forte des grands nombres pour une certaine variance asymptotique $0 \leqslant \sigma_f^2 < \infty$ qui est donnée par*

$$\sigma_f^2 := \|f\|_\Pi^2 + 2 \sum_{n\geqslant 1} \langle \bar{f}, P^n(\bar{f}) \rangle_\Pi. \tag{2.19}$$

où $\langle f,g \rangle_\Pi = \Pi(fg)$ définit le produit scalaire sur l'espace $\mathcal{L}^2(\Pi)$ des fonctions qui sont Π-intégrables lorsqu'élevées au carré avec la norme $\|f\|_\Pi^2 = \langle f,f \rangle_\Pi$.

Théorème 2.23 (Théorème limite central II, Roberts et Rosenthal, 2004, théorème 27) *Soit $\{X_n\}_{n\in\mathbb{N}}$, une chaîne de Markov ϕ-irréductible, apériodique et réversible. Alors, le théorème limite central est vérifié pour toute fonction f telle que la variance asymptotique (2.19) est finie.*

2.3 Monte Carlo par chaînes de Markov

Dans l'exposition des propriétés des chaînes de Markov (section 2.2), on a vu que certaines conditions sur la transition de Markov P de la chaîne peuvent assurer une loi forte des grands nombres et même un théorème limite central. Ainsi, on peut envisager l'utilisation d'une chaîne de Markov comme échantillon Monte Carlo afin d'estimer une espérance $\pi(f)$.

Les méthodes de Monte Carlo considérées à la section 2.1 possédaient toutes un élément commun : les valeurs de l'échantillon Monte Carlo utilisé pour l'estimation sont produites d'une manière indépendante. Quant à elles, les méthodes **Monte Carlo par chaîne de Markov** (MCMC) génèrent l'échantillon Monte Carlo séquentiellement à l'aide d'une chaîne de Markov qui contient alors de la dépendance séquentielle. L'algorithme 2.4 détaille la procédure générale.

Algorithme 2.4 Algorithme MCMC général

Données Distribution cible π, transition de Markov P et taille de l'échantillon Monte Carlo N.

Procédure
1. *Initialisation.* Valeur initiale de la chaîne x_0.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonnage.* Générer le nouvel état de la chaîne :
   $$X_{n+1}|X_n = x_n \sim P(\cdot|x_n).$$

Sortie L'échantillon $x_{0:N}$.

Cette section considère certains exemples d'algorithmes MCMC ainsi que leurs propriétés théoriques relativement à la convergence vers la distribution cible ainsi qu'aux théorèmes limites.

2.3.1 L'algorithme Metropolis-Hastings

Un des algorithmes MCMC les plus utilisés et dont les propriétés sont le mieux connues est l'algorithme **Metropolis-Hastings** (MH). Il s'agit d'un algorithme basé sur le principe d'acceptation/rejet de candidats, à l'image de l'échantillonnage par rejet (algorithme 2.3), mais en utilisant une chaîne de Markov. Comme pour toute chaîne de Markov, un nouvel état de la chaîne est généré à partir de l'état actuel. Pour ce faire, un candidat généré conditionnellement à l'état actuel est proposé comme nouvel état de la chaîne ; ensuite, le nouvel état de la chaîne est choisi selon une certaine probabilité entre cette proposition et l'état actuel. La procédure exacte est décrite à l'algorithme 2.5. Pour un certain choix de probabilité d'acceptation et pour certaines conditions sur la distribution instrumentale générant les candidats, la chaîne de Markov produite sera ergodique à la distribution cible.

Définition 2.24 (Noyau Metropolis-Hastings) *Soit* Π*, une distribution cible qui admet une densité* π *par rapport à une mesure* σ*-finie* μ *et soit* Q *un noyau de transition de Markov admettant une densité* q *par rapport à* μ*, appelée la* ***densité instrumentale****, c.-à-d.,*

$$Q(\mathrm{d}y|x) = q(y|x)\mu(\mathrm{d}y).$$

On définit la ***probabilité d'acceptation Metropolis-Hastings*** *selon l'expression*

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}. \tag{2.20}$$

Un ***noyau Metropolis-Hastings*** *prend alors la forme suivante*

$$P(B|x) = \int_B \alpha(y|x)Q(\,\mathrm{d}y|x) + r(x)\,\mathbb{1}(x \in B),$$

et admet la (pseudo-) densité suivante

$$p(y|x) = \alpha(y|x)q(y|x) + r(x)\delta_x(y),$$

où $\delta_x(\cdot)$ est la fonction de masse delta de Dirac en x et où $r(x)$ est la probabilité de la chaîne de demeurer en x, donnée par

$$r(x) = 1 - \int_{\mathcal{X}} \alpha(y|x)Q(\,\mathrm{d}y|x).$$

Algorithme 2.5 Algorithme Metropolis-Hastings (MH)

Données Densité cible π, densité instrumentale q et taille de l'échantillon Monte Carlo N.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Proposition.* Générer la proposition
   $$Y|X_n = x_n \sim q(\cdot|x_n);$$
   (b) *Acceptation.* Avec probabilité
   $$\alpha(y|x_n) = \min\left\{1, \frac{\pi(y)q(x_n|y)}{\pi(x_n)q(y|x_n)}\right\},$$
   accepter la proposition $(x_{n+1} = y)$; sinon rejeter la proposition $(x_{n+1} = x_n.)$

Sortie L'échantillon $x_{0:N}$.

Dans l'étude des propriétés des chaînes de Markov, une des conditions qui est souvent utilisée afin de vérifier l'ergodicité est la condition d'équilibre (définition 2.13.) En choisissant la probabilité d'acceptation Metropolis-Hastings (2.20), une condition suffisante à la condition d'équilibre est que la densité instrumentale soit positive sur le support de π.

Proposition 2.24 *Soit P un noyau Metropolis-Hastings pour une densité π à support $\mathcal{X} = \{x : \pi(x) > 0\}$. Alors, le noyau P satisfait la condition d'équilibre 2.13 dès que la densité instrumentale q est positive pour toute paire de points du support de π, c.-à-d.,*

$$q(y|x) > 0, \qquad \forall\, x,y \in \mathcal{X}\,.$$

Démonstration. La preuve utilise la proposition 2.6 où la condition d'équilibre de la densité est suffisante. Soient $x,y \in \mathcal{X}$, alors $\pi(x),\pi(y) > 0$ et, par hypothèse, $q(y|x), q(x|y) > 0$. La division par ces quantités est donc permise. Ainsi, pour $x \neq y$, on a $\delta_x(y) = 0$ et on trouve directement

$$\begin{aligned}
p(y|x)\pi(x) &= \left[\alpha(y|x)q(y|x) + r(x)\delta_x(y)\right]\pi(x) \\
&= \alpha(y|x)q(y|x)\pi(x) \\
&= \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\} q(y|x)\pi(x) \\
&= \min\left\{\pi(y)q(x|y), \pi(x)q(y|x)\right\},
\end{aligned}$$

qui est exactement symétrique en (x,y), ce qui montre la condition d'équilibre. Puis, pour $x = y$, la condition d'équilibre est triviale,

$$p(y|x)\pi(x) = p(x|x)\pi(x) = p(x|y)\pi(y).$$

□

Afin de vérifier l'ergodicité d'une chaîne de Markov, la condition d'équilibre n'est pas suffisante à elle seule : le théorème 2.7 montre seulement que la chaîne de Markov admet π comme distribution stationnaire. En effet, la π-irréductibilité et l'apériodicité de la chaîne doivent également être démontrées (théorème 2.10). En supposant que la densité instrumentale soit bornée par le bas pour des pas bornés, il est possible de vérifier ces propriétés pour un algorithme Metropolis-Hastings.

Proposition 2.25 (Robert et Casella, 2004, lemme 6.2.7) *Soit P un noyau Metropolis-Hastings pour une densité π à support $\mathcal{X}$ qui soit connecté. Supposons que π soit bornée par le haut et par le bas sur tout sous-ensemble compact de $\mathcal{X}$ et qu'il existe $\delta, \varepsilon > 0$ tels que*

$$q(y|x) > \varepsilon, \qquad \forall \, \|x - y\|_2 < \delta,$$

alors le noyau P est π-irréductible et apériodique. De plus, tout ensemble compact non-nul est un petit ensemble.

La proposition suivante énonce un ensemble de conditions suffisantes à l'apériodicité d'une chaîne de Markov provenant d'un algorithme Metropolis-Hastings. On requiert la π-irréductibilité et une probabilité de rejet soit non-nulle sur le support de π.

Proposition 2.26 *Soit P un noyau Metropolis-Hastings pour une distribution Π à support $\mathcal{X}$ satisfaisant la condition de Π-irréductilibité et tel que $r(x) > 0$ pour tout $x \in \mathcal{X}$. Alors, le noyau P est apériodique.*

Démonstration. Supposons le contraire : il existe un m-cycle pour $m \geqslant 2$. On a donc $(S_0, S_1, \ldots, S_m)$ tous disjoints excepté $S_0 = S_m$ tel que, pour tout i, $\Pi(S_i) > 0$ et $P(S_{i+1}|x) = 1$, $x \in S_i$. Par hypothèse, on a

$$\mathbb{P}(X_1 = X_0) = r(X_0) > 0.$$

Cependant, puisque les ensembles sont disjoints, le passage de X_0 à S_1 doit s'effectuer par une acceptation M.-H. vu que $X_0 \notin S_1$. On trouve alors

$$P(S_1|x) = \mathbb{P}(X_1 \in S_1 | X_0 = x) \leqslant 1 - r(x) < 1,$$

ce qui contredit $P(S_1|x) = 1$. □

Enfin, la proposition suivante établit un lien direct entre la récurrence et l'Harris-récurrence des noyaux Metropolis-Hastings. Ainsi, par le théorème 2.11, on trouve que ces noyaux seront ergodiques pour toute distribution initiale dès que la condition d'équilibre, la π-irréductibilité et l'apériodicité sont vérifiées.

Proposition 2.27 (Tierney, 1994, corollaire 2) *Soit P un noyau de transition Metropolis-Hastings pour une certaine distribution stationnaire cible π. Si P est π-irréductible, alors P est Harris-récurrent.*

2.3.1.1 L'algorithme Metropolis-Hastings indépendant

Lorsque la densité instrumentale est indépendante de l'état actuel de la chaîne, l'algorithme M.-H. est alors dit **indépendant** (IMH : *Independent Metropolis-Hastings*), parfois appelé l'algorithme **Hastings**. Dans ce cas, l'algorithme ressemble à la méthode Monte Carlo par acceptation/rejet où des propositions i.i.d. sont successivement ajoutées avec une certaine probabilité à l'échantillon. La

différence est dans le calcul de la probabilité d'acceptation et dans le fait qu'un rejet correspond plutôt à répéter l'état actuel dans l'échantillon.

Définition 2.25 (Noyau Metropolis-Hastings indépendant) *Soit P un noyau Metropolis-Hastings de densité instrumentale q. Si la densité instrumentale est indépendante de l'état actuel de la chaîne, c.-à-d., $q(y|x) \equiv q(y)$, alors l'algorithme est dit* ***indépendant****. La probabilité d'acceptation M.-H. prend alors la forme suivante :*

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)q(x)}{\pi(x)q(y)}\right\}. \tag{2.21}$$

Tout comme l'échantillonneur par rejet (algorithme 2.3), le rapport des densités est au cœur de la validité de l'algorithme. En effet, lorsque la densité instrumentale peut envelopper la densité cible, alors l'algorithme sera uniformément ergodique.

Théorème 2.28 (Robert et Casella, 2004, théorème 6.3.1 et lemme 6.3.2) *Soit P un noyau M.-H. indépendant de densité instrumentale q indépendante et de densité cible π ayant $\mathcal{X}$ comme support. S'il existe $M < \infty$ tel que*

$$\pi(x) \leqslant Mq(x),$$

alors l'algorithme est uniformément ergodique avec

$$\|P^n(\cdot|x) - \pi\|_{\mathrm{TV}} \leqslant 2\left(1 - \frac{1}{M}\right)^n.$$

De plus, la probabilité d'acceptation M.-H. (2.21) sera supérieure ou égale à $\frac{1}{M}$ dès que la chaîne est stationnaire.

2.3.1.2 L'algorithme Metropolis-Hastings marche aléatoire

Un second cas particulier de l'algorithme Metropolis-Hastings est le cas **marche aléatoire**. Dans ce cas, la proposition est générée en perturbant l'état actuel x, c'est-à-dire en y ajoutant un pas ε venant d'une densité q indépendante de l'état actuel. Il est alors possible d'écrire $y = x + \varepsilon$, où $\varepsilon \sim q$, et la distribution de y sachant x prend alors la forme suivante :

$$q(y|x) = q(\varepsilon) = q(y - x).$$

Le nouvel état, choisi entre y et x selon la probabilité d'acceptation M.-H., suit donc la densité suivante :

$$p(y|x) = q(y - x)\alpha(y|x) + r(x)\delta_x(y),$$

qui définit une marche aléatoire homogène.

Définition 2.26 (Noyau Metropolis-Hastings marche aléatoire) *Soit P un noyau Metropolis-Hastings de densité instrumentale q. Si la densité instrumentale est une marche aléatoire, c.-à-d.,*

$$q(y|x) = q(y - x),$$

alors l'algorithme est dit du type ***marche aléatoire****.*

La proposition 2.25 peut s'appliquer d'une manière triviale aux algorithme M.-H. marche aléatoire. Si la densité instrumentale de la marche aléatoire est bornée par le bas dans une boule centrée à l'origine, c.-à-d., si

$$\|z\|_2 < \delta \quad \Rightarrow \quad q(z) > \varepsilon$$

alors le noyau M.-H. résultant est π-irréductible et apériodique. En général ce type de condition est facilement vérifiée en choisissant une densité instrumentale relativement régulière. Par la condition

d'équilibre de l'algorithme, le noyau admet également π comme distribution stationnaire et le théorème 2.10 implique l'ergodicité de la chaîne pour π-presque toute valeur initiale.

2.3.1.3 L'algorithme Metropolis

Un cas particulier de l'algorithme M.-H. marche aléatoire est l'algorithme **Metropolis** (RWM : *Random Walk Metropolis*) (Metropolis et collab., 1953) où la densité instrumentale est supposée symétrique, c.-à-d., $q(y-x) = q(x-y)$. L'avantage de cette condition est la simplification de la probabilité d'acceptation au rapport des densités :

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\} = \min\left\{1, \frac{\pi(y)q(x-y)}{\pi(x)q(y-x)}\right\} = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}.$$

Contrairement à l'algorithme M.-H. indépendant, l'algorithme M.-H. marche aléatoire ne satisfait généralement pas à l'ergodicité uniforme. En effet, Mengersen et Tweedie (1996, théorème 3.1) montrent qu'un noyau Metropolis sur $\mathbb{R}^d$ n'est jamais uniformément ergodique, et ce, pour n'importe quelle densité cible π. Par contre, la propriété d'ergodicité géométrique peut être vérifiée sous certaines conditions. Par exemple, ceci peut être fait en dimension $d = 1$ en supposant la **log-concavité des ailes** de la densité cible.

Définition 2.27 (Log-concavité des ailes) *Soit π une densité à support $\mathcal{X} \subseteq \mathbb{R}^d$. S'il existe $0 < \alpha < \infty$ et $0 < M < \infty$ tels que*

$$\log \pi(x) - \log \pi(y) \geqslant \alpha \|y - x\|_2, \qquad \forall\, \|x\|_2, \|y\|_2 \geqslant M,$$

alors on dit que π a des ailes log-concaves.

Théorème 2.29 (Mengersen et Tweedie, 1996, théorème 3.1) *Soit π une densité symétrique à support $\mathcal{X} \subseteq \mathbb{R}$ et aux ailes log-concaves de constante α et soit un noyau Metropolis de densité instrumentale q symétrique et positive. Alors, la chaîne produite par le noyau Metropolis est V-géométriquement ergodique pour la fonction $V(x) = \exp(s|x|)$, où $0 < s < \alpha$.*

Si π n'est pas symétrique, alors la même conclusion tient en supposant de plus que la densité instrumentale satisfait

$$q(z) \leqslant b \cdot \exp\left(-\alpha|x|\right),$$

pour un certain $0 < b < \infty$.

La log-concavité des ailes de la densité cible assure des ailes à décroissance au moins exponentielle. Pour le cas général (en dimension d arbitraire), en plus d'une condition sur le rythme de décroissance des ailes, une ergodicité géométrique pourra être assurée en supposant la condition supplémentaire de **contours réguliers** qui exige, intuitivement, que la densité cible soit décroissante vers les grands $\|x\|_2$.

Théorème 2.30 (Jarner et Hansen, 2000, lemme 3.5) *Soit P un noyau Metropolis pour une densité cible π positive et continue et soit q une densité instrumentale telle que*

$$q(y|x) = q(\|y - x\|_2), \tag{2.22}$$

$$|z| \leqslant \delta \quad \Rightarrow \quad q(z) \geqslant \varepsilon. \tag{2.23}$$

Supposons qu'il existe un petit ensemble C pour lequel P satisfait la dérive géométrique pour une certaine fonction $V \geqslant 1$ continue et que les conditions suivantes sont satisfaites ,

$$\limsup_{\|x\|_2 \to \infty} \frac{PV(x)}{V(x)} < 1, \qquad \sup_{x \in \mathcal{X}} \frac{PV(x)}{V(x)} < \infty.$$

Alors, la chaîne est V-géométriquement ergodique à π.

Définition 2.28 (Ailes super-exponentielles) *On dit qu'une densité π à support dans $\mathbb{R}^d$ admet des* ***ailes super-exponentielles*** *si elle est positive et si elle admet des premières dérivées continues telles que*

$$\lim_{\|x\|_2 \to \infty} \frac{x}{\|x\|_2} \cdot \nabla \log \pi(x) = -\infty.$$

Théorème 2.31 (Jarner et Hansen, 2000, théorème 4.1) *Soit P un noyau Metropolis pour une densité cible π aux ailes super-exponentielles et soit q une densité instrumentale satisfaisant (2.22) et (2.23). Alors, la chaîne est V-géométriquement ergodique à π si et seulement si*

$$\liminf_{\|x\|_2 \to \infty} Q(A(x)|x) > 0,$$

où $A(x) = \{y \,|\, \pi(y) \geqslant \pi(x)\}$ est la région d'acceptation automatique. En particulier, une condition de dérive géométrique est satisfaite avec $V \propto \pi^{-1/2}$.

Définition 2.29 (Contours réguliers) *On dit qu'une densité π à support dans $\mathbb{R}^d$ admet des* ***contours réguliers*** *si elle est positive et si elle admet des premières dérivées continues telles que*

$$\limsup_{\|x\|_2 \to \infty} \frac{x}{\|x\|_2} \cdot \frac{\nabla \pi(x)}{\|\nabla \pi(x)\|_2} < 0.$$

Théorème 2.32 (Jarner et Hansen, 2000, théorème 4.3) *Soit P un noyau Metropolis pour une densité cible π aux ailes super-exponentielles et aux contours réguliers et soit q une densité instrumentale satisfaisant (2.22) et (2.23). Alors, la chaîne est V-géométriquement ergodique à π.*

2.3.1.4 L'algorithme MALA

L'**échantillonneur MALA** (pour *Metropolis-Adjusted Langevin Algorithm*, Roberts et Tweedie, 1996a) est une version de l'algorithme Metropolis-Hastings de type marche aléatoire où l'incrément est biaisé dans la direction du gradient de π. Spécifiquement, la proposition est donnée par

$$Y|X = x \sim \mathcal{N}_d\left(x + \frac{\sigma^2}{2}\nabla \log \pi(x), \sigma^2 I_d\right).$$

Une justification théorique de ce choix d'incrément est possible, mais elle requiert certains concepts théoriques hors de l'étendue de cette exposition. Intuitivement, modifier la marche aléatoire de sorte à favoriser des pas vers une densité cible plus élevée fait en sorte que les candidats proposés risquent d'être de meilleure qualité. Ainsi, l'exploration de l'espace $\mathcal{X}$ sera généralement plus efficace.

Il sera question à la section 2.5 de l'efficacité des algorithmes MCMC et on y verra que l'algorithme MALA est théoriquement plus efficace qu'un algorithme Metropolis.

2.3.2 Extensions et variantes

L'algorithme Metropolis-Hastings produit une chaîne de Markov aux propriétés souhaitées en imposant des conditions relativement faibles sur la densité instrumentale. Cette densité instrumentale est la seule partie de l'algorithme qui peut être modifiée ; cette section considère donc d'autres algorithmes utilisant une chaîne de Markov afin de produire un échantillon Monte Carlo. Ces méthodes seront souvent similaires à l'algorithme Metropolis-Hastings, mais leurs distinctions respectives peuvent les rendre pertinentes dans certaines situations.

2.3.2.1 Compositions de noyaux

Supponsons que la transition de Markov P utilisée pour produire la chaîne se décompose en une suite de transitions $P_1, \ldots, P_m$ elles-mêmes de Markov. Si x_n est l'état actuel de la chaîne, le prochain état est généré séquentiellement :

$$\begin{aligned} X_{n+1}^{(1)}|X_n = x_n &\sim P_1(\cdot|x_n) \\ X_{n+1}^{(2)}|X_{n+1}^{(1)} = x_{n+1}^{(1)} &\sim P_2(\cdot|x_{n+1}^{(1)}) \\ &\vdots \\ X_{n+1}^{(m)}|X_{n+1}^{(m-1)} = x_{n+1}^{(m-1)} &\sim P_m(\cdot|x_{n+1}^{(m-1)}). \end{aligned}$$

pour finalement poser $x_{n+1} = x_{n+1}^{(m)}$. Par exemple, si $\mathcal{X}$ est de dimension d, alors chacune des transitions P_j, $j = 1, \ldots, p$, peut correspondre à la mise à jour de la j-ième composante de x_n.

D'abord, si chacune des transitions P_j admet π comme distribution invariante, alors il est clair que P admettra également π comme distribution invariante : soit $B \in \mathcal{B}(\mathcal{X})$,

$$\pi P(B) = \pi(P_m \cdots P_1)(B) = \pi P_m(P_{m-1} \cdots P_1)(B) = \pi(P_{m-1} \cdots P_1)(B) = \ldots = \pi P_1(B) = \pi(B).$$

Cependant, si chacune des transition P_j est réversible, il ne sera pas nécessairement le cas que la composition P le sera également. C'est une des raisons pour lesquelles les noyaux MCMC non-réversibles doivent également être considérés et que les résultats de la section 2.2 sont majoritairement énoncés en terme d'invariance et non de réversibilité. Par contre, il est possible d'assurer la réversibilité d'une composition en procédant à une composition palindrome $P = P_1 \cdots P_m P_m \cdots P_1$.

Tierney (1994, proposition 4) montre que les compositions P_1P_2 et P_2P_1 de deux transitions P_1,P_2 admettant π comme distribution invariante et telles que P_1 ou P_2 satisfait la condition de minorisation (définition 2.10) sont uniformément ergodiques (définition 2.22).

L'**échantillonneur de Gibbs** (algorithme 2.6) effectue une mise à jour cyclique des composantes de l'état actuel en échantillonnant la distribution conditionnelle au sous-espace engendré en fixant toutes les autres composantes. Explicitement, soit

$$\pi(x^{(j)}|x^{(-j)}) = \frac{\pi(x)}{\int_{\mathbb{R}} \pi(x)\,\mathrm{d}x^{(j)}},$$

la densité conditionnelle de la j-ième composante sachant les autres composantes $x^{(-j)}$, où $x = (x^{(j)},x^{(-j)})$. Alors, si l'état actuel est donné par x_n, le nouvel état est obtenu par la séquence suivante de mise-à-jour :

$$X_{n+1}^{(1)} \sim \pi(\cdot|x_n^{(-1)}), \quad X_{n+1}^{(2)} \sim \pi(\cdot|x_{n+1}^{(1)},x_n^{(3:d)}), \quad \ldots, \quad X_{n+1}^{(d)} \sim \pi(\cdot|x_{n+1}^{(-d)}).$$

L'échantillonneur de Gibbs peut être réinterprété comme un algorithme Metropolis-Hastings même si, en apparence, il n'y a aucune étape d'acceptation/rejet. En effet, si l'on considère la densité conditionnelle comme une densité instrumentale, alors la probabilité d'acceptation lors de la mise-à-jour d'une composante est d'exactement 1. Pour voir ceci, on considère la mise-à-jour de la j-ième composante de x et une proposition $Y^{(j)} \sim \pi(\cdot|x^{(-j)})$ avec $y^{(-j)} = x^{(-j)}$. Alors, en posant

Algorithme 2.6 Échantillonneur de Gibbs

Données Distribution cible π à support $\mathcal{X} \subseteq \mathbb{R}^d$ et taille de l'échantillon Monte Carlo N.

Procédure
1. *Initialisation.* Valeur initiale de la chaîne $x_0 \in \mathcal{X}$;
2. *Échantillonnage.* Pour $n = 0, \ldots, N-1$,
   (a) Pour $j = 1, \ldots, d$, mettre à jour la j-ième composante :

$$X_{n+1}^{(j)} \sim \pi(\cdot | x_{n+1}^{(1:j-1)}, x_n^{(j+1:d)}).$$

Sortie L'échantillon $x_{1:N}$ et l'estimé Monte Carlo (2.2).

$y = (y^{(j)}, y^{(-j)})$, le ratio dans la probabilité d'acceptation est donné par

$$\frac{\pi(y)\pi(x^{(j)}|x^{(-j)})}{\pi(x)\pi(y^{(j)}|x^{(-j)})} = \frac{\pi(y)}{\pi(y^{(j)}|y^{(-j)})}\frac{\pi(x^{(j)}|x^{(-j)})}{\pi(x)} = \int_{\mathbb{R}} \pi(y)\,\mathrm{d}y^{(j)} \times \frac{1}{\int_{\mathbb{R}} \pi(x)\,\mathrm{d}x^{(j)}} = 1.$$

L'algorithme de Gibbs requiert qu'il soit possible d'échantillonner des distributions conditionnelles $\pi(\cdot|x^{(-j)})$. Bien que certaines situations le permettent, ce ne sera pas généralement le cas. La réinterprétation décrite au paragraphe précédent suggère que la structure de cet algorithme est similaire à celle d'un algorithme Metropolis-Hastings où la densité instrumentale est choisie comme la densité conditionnelle. Si ces dernières ne sont pas disponibles, il est facile de remplacer ces densités conditionnelles par une autre densité instrumentale et de procéder réellement à une étape d'acceptation rejet. C'est le principe derrière l'algorithme **Metropolis-within-Gibbs** 2.7. La densité instrumentale est alors une densité à une dimension, souvent une marche aléatoire afin de profiter de l'état actuel de la chaîne. De plus, la densité instrumentale univariée peut différer entre les composantes, ce qui peut être pertinent lorsque la variance des composantes de π diffèrent largement entre elles. Ces densités peuvent dépendre d'une seule ou de plusieurs composantes de l'état actuel. Cette technique peut également être étendue à la mise à jour par bloc des composantes plutôt qu'une seule composante à la fois.

Algorithme 2.7 Algorithme *Metropolis-within-Gibbs* (MwG)

Données Distribution cible π à support $\mathcal{X} \subseteq \mathbb{R}^d$, taille de l'échantillon Monte Carlo N et densités instrumentales $q_j(\cdot|\cdot) : \mathbb{R} \times \mathbb{R} \to [0,\infty)$.

Procédure
1. *Initialisation.* Valeur initiale de la chaîne $x_0 \in \mathcal{X}$;
2. *Échantillonnage.* Pour $n = 0, \ldots, N-1$,
   (a) Pour $j = 1, \ldots, d$, mettre à jour la j-ième composante :
      i. *Proposition.* Le nouvel état est proposé :

$$Y_j \sim q_j(\cdot|x^{(j)});$$

      ii. *Acceptation.* Avec probabilité

$$\min\left\{1, \frac{\pi(x_{n+1}^{(1:j-1)}, y^{(j)}, x_n^{(j+1,d)})q_j(y^{(j)}|x_n^{(j)})}{\pi(x_{n+1}^{(1:j-1)}, x_n^{(j)}, x_n^{(j+1,d)})q_j(x_n^{(j)}|y^{(j)})}\right\},$$

      accepter la proposition ($x_{n+1}^{(j)} = y^{(j)}$) ; sinon rejeter la proposition ($x_{n+1}^{(j)} = x_n^{(j)}$.)

Sortie L'échantillon $x_{1:N}$ et l'estimé Monte Carlo (2.2).

2.3.2.2 Mélange de noyaux MCMC

Les algorithmes Gibbs et Metropolis-within-Gibbs parcourent tous deux les différentes composantes de l'espace d'une manière cyclique. Alternativement à cette approche parfois appelée **balayage déterministe**, il est aussi possible de procéder à un **balayage aléatoire** des composantes. Ainsi, plutôt que de parcourir les composantes dans l'ordre $j = 1, \ldots, d$, un indice j est sélectionné au hasard (uniformément) et seulement cette composante est mise à jour. Si l'on dénote la mise-à-jour de la j-ième composante par la transition P_j alors la transition globale est donnée par le mélange

$$P = \frac{1}{d}(P_1 + \cdots + P_d).$$

Si chacune des transitions P_j admet π comme distribution invariante, alors le mélange admettra aussi π comme distribution invariante : pour $B \in \mathcal{B}(\mathcal{X})$, on a

$$\pi P(B) = \pi\left(\frac{1}{d}(P_1 + \cdots P_d)\right)(B) = \frac{1}{d}\left(\pi P_1(B) + \cdots \pi P_d(B)\right) = \frac{1}{d}\left(\pi(B) + \cdots + \pi(B)\right) = \pi(B).$$

Roberts et Rosenthal (1997, section 3) considèrent ce type de mélange dans le cas de mises à jour Gibbs (via les distributions conditionnelles) ainsi que MwG (via des densités instrumentales) et prouvent certains résultats quant à l'ergodicité V-géométrique de ces algorithmes.

En fait, cette construction s'applique à des cas plus généraux que les mises à jour par composantes : chacune des transitions P_j, $j = 1, \ldots, m$, peuvent être de pleine dimensions. Par exemple, des transitions Metropolis-Hastings avec différentes densités instrumentales peuvent être utilisées aléatoirement dans un tel mélange. Ceci peut être pertinent lorsque l'espace $\mathcal{X}$ contient des régions à covariances différentes ; une certaine densité peut être bien ajustée à π dans une région, mais pas dans une autre. Si une des transitions est uniformément ergodique, alors le mélange le sera également (Tierney, 1994, proposition 3).

2.3.2.3 Algorithmes à plusieurs candidats

L'algorithme Metropolis-Hastings propose un seul candidat à chaque itération. Une extension possible de cet algorithme est de considérer plutôt un ensemble de candidats puis de sélectionner un des candidats les plus intéressants.

L'algorithme MH à **rejet retardé** (DR : *Delayed Rejection*, Mira, 2001a) est un exemple d'algorithme qui considère plusieurs candidats à chaque itération. Un nombre maximal de candidats K est choisi ; à chaque itération, un candidat est proposé soit jusqu'à ce qu'il soit accepté, soit jusqu'à ce que le nombre maximal soit atteint. Dans le second cas, la chaîne reste sur place tel un rejet MH. La manière séquentielle de produire les candidats fait en sorte que la probabilité d'acceptation M.-H. des candidats supplémentaires doit être adaptée afin de prendre en compte la conditionnalité sur les candidats précédemment rejetés. Par exemple, pour $K = 2$, on trouve que la probabilité d'acceptation du premier candidat $y^{(1)}$ (notée $\alpha^{(1)}(y^{(1)}|x)$) est la même que celle d'un algorithme MH, puis que celle du second candidat $y^{(2)}$ est donnée par

$$\alpha^{(2)}(y^{(2)}|y^{(1)},x) = \min\left\{1, \frac{\pi(y^{(2)})q_1(y^{(1)}|y^{(2)})\left(1-\alpha^{(1)}(y^{(1)}|y^{(2)})\right)q_2(x|y^{(2)},y^{(1)})}{\pi(x)q_1(y^{(1)}|x)\left(1-\alpha^{(1)}(y^{(1)}|x)\right)q_2(y^{(2)}|y^{(1)},x)}\right\}, \tag{2.24}$$

où q_1 et q_2 sont les densités instrumentales respectivement du premier et du second candidat. L'algorithme 2.8

contient la procédure pour $K = 2$ candidats. Le choix des densités instrumentales peut jouer un rôle important dans l'efficacité d'un tel algorithme : par exemple, Bédard et collab. (2014, corollaire 5) montrent qu'un second candidat opposé au premier dans une marche aléatoire est optimal parmi une certaine classe de densités instrumentales.

Algorithme 2.8 Algorithme Metropolis-Hastings à rejet retardé (DR) à deux étapes

Données Densité cible π, densités instrumentales q_1 et q_2 et taille de l'échantillon Monte Carlo N.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Proposition 1.* Générer le premier candidat
   $$Y^{(1)}|X_n = x_n \sim q_1(\cdot|x_n);$$
   (b) *Acceptation 1.* Avec probabilité
   $$\alpha^{(1)}(y^{(1)}|x_n) = \min\left\{1, \frac{\pi(y^{(1)})q_1(x_n|y^{(1)})}{\pi(x_n)q_1(y^{(1)}|x_n)}\right\},$$
   accepter la proposition $(x_{n+1} = y^{(1)})$ et passer à l'itération $n+1$; sinon rejeter la proposition et :
   i. *Proposition 2.* Générer le second candidat
   $$Y^{(2)}|X_n = x_n \sim q_2(\cdot|y^{(1)}, x_n);$$
   ii. *Acceptation 1.* Avec probabilité $\alpha^{(2)}(y^{(2)}|y^{(1)}, x_n)$ donnée par (2.24), accepter la proposition $(x_{n+1} = y^{(2)})$; sinon rejeter la proposition $(x_{n+1} = x_n)$.

Sortie L'échantillon $x_{0:N}$.

Plutôt que de considérer un ensemble de propositions générées séquentiellement, les algorithmes **à essais multiples** procèdent plutôt à la génération simultanée de K candidats pour ensuite sélectionner un de ces candidats, sur lequel le test d'acceptation/rejet Metropolis-Hastings est effectué. Plusieurs aspects de ces algorithmes peuvent être modifiés : la manière dont les candidats sont générés simultanément peut inclure de la corrélation entre les candidats, la manière de sélectionner la proposition parmi les candidats peut favoriser certains comportements de l'algorithme, etc. Le chapitre 4 est dédié complètement à l'étude de ces algorithmes.

2.4 Diagnostiques et performance

Jusqu'ici, l'étude des méthodes MCMC fut concentrée sur trois propriétés fondamentales : l'ergodicité, la loi des grands nombres et le théorème limite central. Dans le contexte de l'estimation de $\pi(f)$ par sa version empirique $\hat{\pi}_N(f)$, la loi des grands nombres assure que l'estimateur est asymptotiquement sans biais alors que le théorème limite central nous renseigne sur la distribution asymptotique de l'estimateur pour un certain choix de noyau de transition P. Cependant, ces résultats ne constituent pas à eux seuls la boîte à outils nécessaire à l'application des méthodes MCMC.

Les estimateurs MCMC sont généralement biaisés pour toute taille d'échantillon MCMC finie même s'ils sont néanmoins des estimateurs convergents par une loi des grands nombres. Notons cependant que de récents développements (Jacob et collab., 2017) permettent de construire des chaînes MCMC par couplage permettant une estimation Monte Carlo qui soit sans biais pour une taille d'échantillon MC finie.

D'une part, un théorème limite central supposera toujours que la chaîne est stationnaire, c.-à-d. que $X_0 \sim \pi$. Cette supposition n'est généralement pas satisfaite : s'il est possible d'échantillonner directement $X_0 \sim \pi$, alors il est possible d'obtenir un échantillon i.i.d. de π et les méthodes MCMC deviennent inutiles. Par contre, l'ergodicité de la chaîne assure que la distribution marginale de la chaîne converge vers la distribution cible π. Ainsi, pour appliquer le théorème limite central, il faudra être en mesure de vérifier que la chaîne a (approximativement) atteint la stationnarité ; la sous-section 2.4.1 sera consacrée à quelques-uns de ces diagnostiques.

D'autre part, pour un noyau fixe P, le théorème limite central nous indique que la distribution asymptotique de l'estimateur Monte Carlo est normale, avec une certaine variance asymptotique. Bien qu'il soit parfois possible de la calculer directement—par exemple, lorsque l'espace d'état est discret—,le calcul de cette variance asymptotique est généralement impossible dans la mesure où l'on cherche déjà à estimer $\pi(f)$, qui est nécessaire au calcul. Il est donc souvent nécessaire d'estimer cette variance asymptotique si l'on veut ensuite produire des régions de confiance associées à l'estimation de $\pi(f)$; la sous-section 2.4.2 explore certaines méthodes d'estimation.

En l'absence d'un théorème limite central, il est malgré tout parfois possible d'évaluer l'incertitude autour d'un estimé MCMC. Rosenthal (2017) propose un intervalle de confiance conservatif, basé sur un estimeur convergent de la variance asymptotique, qui est valide sous des conditions plus faibles que celles vérifiant un TLC. Łatuszyński et Niemiro (2011) proposent quant à eux une méthode permettant la construction d'intervalles de confiance en supposant une condition de dérive géométrique (définition 2.20). Plutôt que d'estimer directement la variance asymptotique, il est également possible de l'encadrer par des bornes relativement précises à l'aide de représentations variationnelles par les formes de Dirichlet (voir e.g. Sherlock (2018) pour une introduction).

Enfin, le noyau de transition P doit être choisi avant l'échantillonnage de la chaîne. Ce choix détermine entièrement les propriétés de la chaîne et, en particulier, la variance asymptotique. Un choix différent de noyau produira donc une variance asymptotique différente. Dans le but d'obtenir une estimation plus efficace, par exemple avec une région de confiance plus serrée, il sera approprié d'avoir des moyens de comparer les noyaux entre eux. Différentes mesures de performance sont proposées dans la littérature à ce sujet et la sous-section 2.4.3 aborde les plus courantes. Ces idées formeront les préalables à la section suivante, où l'on cherchera à déterminer le choix optimal de noyau de transition.

2.4.1 Diagnostiques de convergence à la stationnarité

On considère quelques méthodes permettant l'évaluation de la convergence vers la stationnarité d'une chaîne produite par un algorithme MCMC. La plupart de ces diagnostiques permettent seulement d'identifier une chaîne n'ayant pas atteint la stationnarité ; il est généralement difficile d'établir avec confiance qu'une chaîne a bel et bien convergé. De plus, ces méthodes vérifient parfois deux aspects distincts de la stationnarité : que la chaîne ait convergé et que la distribution limite soit celle visée. Par exemple, lorsque la densité cible est multimodale, la chaîne peut rester prise dans un mode : la chaîne convergera alors vers une portion de la distribution cible, sur un sous-ensemble du support seulement. Ainsi, la vérification de la convergence à la stationnarité exige de multiples confirmations selon différents critères puisqu'aucun diagnostique seul ne suffit à vérifier la convergence.

Les approches se distinguent également selon le nombre de chaînes nécessaires à leur application. Certains diagnostiques analysent directement la chaîne produite par l'algorithme, alors que d'autres nécessitent un ensemble de chaînes indépendantes qui sont ensuite comparées, combinées et analysées. Évidemment, l'utilisation de multiples chaînes augmente considérable le coût computationnel de la vérification, mais permet de relever certains problèmes de convergence que les méthodes sur chaîne unique ne peuvent déceler.

Seules certaines méthodes seront présentées ici ; pour plus de détails et pour des critères additionnels, on réfère le lecteur à diverses revues de ces diagnostiques telles que celles par Cowles et Carlin (1996), Brooks et Roberts (1998), Mengersen et collab. (1999) ou Sinharay (2003).

2.4.1.1 La période de chauffe

Une notion fréquemment rencontrée dans l'analyse de la sortie d'un algorithme MCMC, tant pour évaluer la convergence que pour l'estimation elle-même, est la période de chauffe (l'expression « *burn-in* » est communément utilisée en anglais et parfois en français). L'ergodicité des chaînes de Markov assure que la distribution marginale du processus convergera vers la distribution limite cible pour (presque) toute valeur initiale de la chaîne. Ainsi, cette propriété assure que la chaîne « oubliera » ses conditions initiales à la longue. Cependant, bien que le comportement à long terme soit celui voulu, une partie considérable de la chaîne sera tout de même affectée par les conditions initiales avant d'atteindre la distribution limite. Il est alors difficile de supposer la stationnarité de cette portion initiale et, puisque la plupart des résultats et des diagnostiques se basent sur la stationnarité, il serait imprudent d'inclure ces premières itérations dans la sortie de l'algorithme.

Par exemple, si l'état initial de la chaîne est généré à partir d'une certaine distribution qui soit particulièrement différente de la distribution cible, alors la chaîne de Markov débutera fort probablement dans une région de faible probabilité. Un algorithme de type marche aléatoire pourrait requérir de nombreuses itérations avant d'arriver à un région de haute probabilité. Les premières itérations ne refléteront donc pas bien la distribution cible et le remède le plus intuitif est de tout simplement écarter une portion initiale de la chaîne, appelée période de chauffe. L'intérêt principal de jeter un certain nombre d'itérations initiales est donc d'effacer le passé, pendant lequel la chaîne n'avait pas encore « oublié » les conditions initiales.

Par contre, si la valeur initiale est choisie de sorte à être probable par rapport à la distribution cible, alors une telle période de chauffe n'est souvent pas nécessaire puisque la dépendance par rapport

à l'initialisation s'estompera presqu'instantanément. Comme le mentionne Geyer (2011, section 1.11.4), « *Any point you don't mind having in a sample is a good starting point.* » En effet, la période de chauffe est tout simplement une méthode pour trouver un bon point de départ à la chaîne de sorte que le processus soit déjà relativement près de la stationnarité.

Les méthodes qui seront présentées ensuite ne seront pas exprimées explicitement en fonction d'une période de chauffe pour alléger la notation. Afin d'évaluer la convergence à la stationnarité d'une chaîne, on considérera la réalisation de la chaîne au complet : on suppose donc qu'une période de chauffe fut effectué si nécessaire.

2.4.1.2 Méthodes exploratoires

Les méthodes exploratoires permettent de relever des problèmes majeurs de convergence. Sans être constituées de tests statistiques, ces méthodes vont tout de même soulever une exploration lente du support, une exploration incomplète du support ou bien une dépendance trop forte sur le choix des valeurs initiales. Lorsque ces techniques ne soulèvent pas d'accrocs importants, l'utilisateur peut avoir davantage confiance que la distribution limite sera la distribution cible.

Méthodes graphiques. Tout d'abord, l'analyse graphique de la série temporelle induite par la chaîne peut révéler un mélange trop lent de la chaîne. Par exemple, un algorithme Metropolis peut avoir des pas trop petits par rapport à la densité cible de sorte que seule une partie du support est explorée. Alternativement, des pas trop grands feront en sorte que la proposition sera souvent rejetée et la chaîne demeurera alors sur place pour de longues périodes. Les histogrammes et les moyennes cumulatives peuvent également être utilisés à ces fins.

En tant que série temporelle corrélée, la chaîne produite par l'algorithme peut être analysée par sa **fonction d'autocorrélation** (ACF : *Autocorrelation Function*) $k \mapsto \rho_k(f)$: pour $X_0 \sim \pi$, on définit l'autocovariance et l'autocorrélation d'ordre $k \in \mathbb{N}$ selon

$$\gamma_k(f) := \mathrm{Cov}(f(X_0), f(X_k)), \qquad \rho_k(f) := \mathrm{Corr}(f(X_0), f(X_k)) = \frac{\gamma_k(f)}{\gamma_0(f)},$$

Afin d'estimer ces quantités, on considère la **fonction d'autocorrélation empirique** définie par $k \mapsto \hat{\rho}_k(f)$ où

$$\hat{\gamma}_k(f) := \frac{1}{N} \sum_{n=1}^{N-k} \left(f(x_n) - \bar{f}_N\right)\left(f(x_{n+k}) - \bar{f}_N\right), \qquad \bar{f}_N := \frac{1}{N} \sum_{n=1}^{N} f(x_n) \qquad \hat{\rho}_k(f) := \frac{\hat{\gamma}_k(f)}{\hat{\gamma}_0(f)}.$$

Lorsque l'autocorrélation au sein de la chaîne ne décroît pas assez rapidement avec le délai, ceci peut indiquer que le mélange est lent. En effet, si la chaîne reste dans une même région pour de longues périodes, l'autocorrélation sera d'autant plus élevée. L'ACF convergera généralement vers 0 sur une période plus ou moins longue : le rythme de convergence peut être étudié (via le graphe de $\hat{\rho}_k$ pour $k = 1, \ldots, K$) afin d'évaluer l'exploration de la chaîne.

De plus, lorsque plusieurs chaînes sont générées indépendamment à partir de valeurs initiales dispersées sur le support de la distribution cible, il est possible d'observer des comportements problématiques. Par exemple, certains modes peuvent n'être accessibles qu'à partir de certaines valeurs initiales. Ainsi, la comparaison des nuages de points ou de moyennes cumulées peuvent mettre en évidence des distributions limites distinctes.

Évidemment, ce type de vérifications est qualitatif et n'est donc pas suffisant à l'évaluation de la convergence ; des méthodes quantitatives sont alors requises.

Sommes cumulées. Pour une certaine fonction cible f (dont l'espérance $\pi(f)$ est recherchée), une évaluation graphique introduite par Yu et Mykland (1998), appelée CUSUM (pour *Cumulative sum*) peut également être utilisée. On considère la suite des différences entre les évaluations de la fonction et l'estimé Monte Carlo final,

$$D_N^n = \sum_{i=1}^{n} [f(x_i) - \hat{\pi}_N(f)] = n\left(\hat{\pi}_n(f) - \hat{\pi}_N(f)\right), \qquad \hat{\pi}_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x_i).$$

Puisque $\hat{\pi}_N(f)$ s'approche de l'espérance $\pi(f)$ par la loi des grands nombres, la suite de D_N^n devrait être centrée en 0 sur toute la durée $n = 1, \ldots, N$. Lorsque cette suite s'éloigne de 0 pour de longues périodes, on observe alors un mélange lent de la chaîne. Il est également possible de quantifier ce critère : Brooks et Roberts (1998) proposent un test statistique pour évaluer si la suite des différences cumulatives varie suffisamment à chaque itération.

Distance entre les distributions. En théorie, la convergence de la chaîne vers sa distribution stationnaire est définie par rapport à la variation totale (ou plus généralement par une V-norme.) Ainsi, la convergence pourrait être évaluée en calculant $\|P^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}}$; cependant, le calcul de cette quantité est généralement irréaliste et il est donc nécessaire d'obtenir un estimé de cette norme. Lorsque les distributions marginales et cible sont des densités, il est également possible de considérer les normes L_1 ou L_2 afin de mesurer l'adéquation entre les distributions.

Plusieurs types d'estimation ont été proposés dans la littérature. Liu et collab. (1993) considèrent un estimé de la norme L_2 entre les densité marginale et cible pour un algorithme Gibbs alors que Roberts (1996) estime la norme L_2 pour la classe d'algorithmes à transition réversible, tous deux utilisant des chaînes parallèles indépendantes. Yu (1995) construisent un estimé de la norme L_1 pour des algorithmes généraux qui n'utilise qu'une seule chaîne.

Notons finalement que Chauveau et Vandekerkhove (2014) proposent une méthode permettant l'estimation de la divergence KL entre la transition itérée et la distribution cible. L'estimateur proposé est construit à partir de la méthode des plus proches voisins sur un échantillon i.i.d. de chaînes parallèles.

Ces méthodes ne permettent généralement pas de tester statistiquement la convergence vers π, mais il est possible de choisir un seuil sous lequel la distance entre les densités est jugée suffisamment petite. Il restera que le choix du seuil est souvent arbitraire et que l'atteinte du seuil est fortement dépendante du problème.

2.4.1.3 Évaluation de la convergence

Rapport de variance. Pour une fonction univariée cible f, Gelman et Rubin (1992) proposent une manière d'évaluer la convergence de M chaînes via la construction d'un estimateur de la variance de la fonction f, c.-à-d., $\sigma^2 = \pi(f - \pi(f))^2$. L'estimateur est constitué d'une moyenne pondérée entre les variances de chaque chaîne et la variance entre les chaînes. Après convergence, on s'attend à ce que l'estimateur de la variance s'approche de σ^2 ; le ratio des variances permet alors d'estimer l'efficacité

de l'algorithme. On considère ici la généralisation de Brooks et Gelman (1998) dans le cas où f est multivariée. La matrice de covariance est estimée par

$$\widehat{\boldsymbol{V}} = \frac{N-1}{N}\boldsymbol{W} + \left(1 - \frac{1}{M}\right)\frac{\boldsymbol{B}}{N},$$

où $\boldsymbol{W}$ est la moyenne des variances empiriques de chaque chaîne,

$$\boldsymbol{W} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N-1}\sum_{n=1}^{N}\left(f(x_{n,m}) - \bar{f}_m\right)\left(f(x_{n,m}) - \bar{f}_m\right)^\top,$$

$\bar{f}_m = \frac{1}{N}\sum_{n=1}^{N} f(x_{n,m})$ est la moyenne de la chaîne m, $\bar{f} = \frac{1}{M}\sum_{m=1}^{M}\bar{f}_m$ est la moyenne globale et $\boldsymbol{B}/N$ est l'estimé de la covariance entre les chaînes,

$$\frac{\boldsymbol{B}}{N} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{f}_m - \bar{f}\right)\left(\bar{f}_m - \bar{f}\right)^\top.$$

Ensuite, soit λ_1 la plus grande valeur propre de la matrice $\boldsymbol{W}^{-1}\boldsymbol{B}/N$ symétrique et définie positive. Alors, la statistique

$$\widehat{R}^p = \frac{N-1}{N} + \left(\frac{M+1}{M}\right)\lambda_1,$$

appelée le MPSRF (pour *Multivariate Potential Scale Reduction Factor*), constitue une borne sur les p statistiques univariées de Gelman et Rubin (1992). Un ensemble de chaînes ayant chacune convergé vers une même distribution limite aura une valeur de $\widehat{R}^p$ près de 1, ce qui permet d'évaluer la convergence de l'algorithme.

Notons que Brooks et Gelman (1998) proposent également une méthode similaire pour des moments autres que l'ordre deux ainsi qu'une autre méthode basée sur des intervalles de couverture plutôt qu'un moment.

Méthodes spectrales. Pour une fonction cible f univariée, la suite $\{f(X_n)\}_{n\in\mathbb{N}}$ constitue une série temporelle qui est stationnaire lorsque la chaîne de Markov $\{X_n\}_{n\in\mathbb{N}}$ est positive récurrente. Ainsi, l'analyse spectrale de cette suite peut être utilisée afin d'évaluer la convergence à la stationnarité. On considère ici la méthode proposée par Geweke (1992), mais notons que plusieurs autres auteurs ont construit des diagnostics semblables. Voir, par exemple, Garren et Smith (2000) ou Heidelberger et Welch (1983).

La méthode de Geweke repose sur l'intuition suivante : si la chaîne a bel et bien convergé, on ne devrait pas trouver de différence entre deux portions distinctes de la chaîne. Concrètement, on considère les n_A premiers états ainsi que les n_B derniers états. La moyenne dans chacune des deux parties est calculée,

$$\bar{f}_A = \frac{1}{n_A}\sum_{n=1}^{n_A} f(x_n), \qquad \bar{f}_B = \frac{1}{n_B}\sum_{n=N-n_B+1}^{N} f(x_n),$$

et un estimé convergent de la variance est produit dans chacune des chaînes, dénotés respectivement $\hat{S}_f^A$ et $\hat{S}_f^B$. Ces estimés sont également convergent pour la variance de f sous π. En supposant que le

processus soit stationnaire, que n_A/N et n_B/N soient fixes avec N et que $(n_A + n_B)/N < 1$, alors

$$\frac{\bar{f}_A - \bar{f}_B}{\left(\frac{1}{n_A}\hat{S}_f^A + \frac{1}{n_B}\hat{S}_f^B\right)^{1/2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \qquad N \to \infty.$$

En choisissant un certain niveau α, il est alors possible de définir un test statistique qui permet de rejeter l'égalité des moyennes entre les deux portions de la chaine, ce qui serait un fort indicateur que l'hypothèse de stationnarité n'est pas vérifiée.

Comparaison de distributions. En simulant simultanément plusieurs chaînes de façon indépendante, on obtient plusieurs instances de distributions marginales. Lorsque ces distributions marginales sont semblables entre les chaînes, on comprend que toute ces chaînes ont atteint la même distribution limite. Dans le cas contraire, c.-à-d., lorsque deux chaînes parallèles ont des distributions particulièrement différentes, la convergence d'une ou de toutes les chaînes est mise en doute. Plusieurs méthodes d'évaluation de la convergence se basent sur ce principe.

Brooks et collab. (1997) proposent de comparer les distributions marginales entre elles via leur distance. Il s'agit d'une estimation de la variation totale qui fournit du même coup une borne supérieure à la distance L_1. Soit M chaînes indépendantes $x_{1:N,1}, \ldots, x_{1:N,M}$ produites en parallèle ; chaque chaîne est partitionnée en blocs de longueur $N_0 < N$. Alors, pour le l-ième bloc de la m-ième chaîne, on définit

$$K_{ml}(x) = \frac{1}{N_0} \sum_{n=(l-1)N_0+1}^{lN_0} P(x|x_{n,m}),$$

qui constitue un estimé de la densité de X par le l-ième bloc de la m-ième chaîne. Puis, pour chaque l-ième bloc et chaque paire de chaînes $m \neq m'$, on considère l'estimé suivant de la distance L_1 entre les densités $K_{ml}(\cdot)$ et $K_{m'l}(\cdot)$.

$$\hat{r}_{m,m'}(l) = 1 - \min\left\{1, \frac{K_{ml}(x)}{K_{m'l}(x)}\right\},$$

pour un x échantillonné de la densité $K_{m'l}(\cdot)$. Enfin, pour chaque bloc l, on estime la distance moyenne entre les chaînes par

$$B_l = \frac{1}{M(M-1)} \sum_{m \neq m'} \hat{r}_{m,m'}(l).$$

Au fil où les chaînes progressent, la valeur de B_l sera plus basse lorsque les différentes chaînes ont des distributions semblables, ce qui indique que les chaînes ont atteint leur distribution limite.

2.4.2 Estimation de la variance asymptotique

Lorsqu'un algorithme MCMC vérifie un théorème limite central, on connait alors le comportement asymptotique de l'estimateur Monte Carlo $\hat{\pi}_N(f) = \frac{1}{N}\sum_{n=1}^{N} f(x_n)$. Cependant, la variance asymptotique, bien que finie, n'est pas connue ; il faudra donc l'estimer afin de pouvoir utiliser les propriétés de la loi normale pour calculer l'erreur standard Monte Carlo ou construire une région de confiance. Dans

cette section, on explorera quelques estimateurs ainsi que leurs propriétés asymptotiques. Notamment, il sera important que ces estimateurs soient convergents vers la vraie variance asymptotique.

2.4.2.1 Estimation naïve par l'ACF empirique

La variance asymptotique de l'estimateur Monte Carlo de l'espérance $\pi(f)$ prend la forme suivante : pour $X_0 \sim \pi$,

$$\sigma^2(f) = \mathrm{Var}(f(X_0)) + 2\sum_{k\geqslant 1} \mathrm{Cov}(f(X_0), f(X_k)) \tag{2.25}$$

$$= \gamma_0(f) + 2\sum_{k\geqslant 1} \gamma_k(f). \tag{2.26}$$

En inspectant cette dernière expression, il serait tentant d'estimer la variance asymptotique en remplaçant la variance et les autocorrélations par leurs versions empiriques, c.-à-d.,

$$\hat{\sigma}^2_{N,\text{naïve}}(f) = \hat{\gamma}_0(f) + 2\sum_{k\geqslant 1} \hat{\gamma}_k(f).$$

Bien que ce genre de substitution puisse parfois donner de bons estimateurs en statistique, ce n'est malheureusement pas le cas pour les méthodes MCMC. En effet, cet estimateur n'est pas convergent (Geyer, 1992, section 3.1) étant donné que la variance des autocorrélations de grands délais ($k \to \infty$) ne décroit pas assez rapidement pour balancer la sommation infinie de l'estimateur.

2.4.2.2 Estimation spectrale

Bien que l'estimateur naïf ne constitue pas un bon choix d'estimateur, on peut tout de même s'en inspirer pour produire un estimateur similaire qui soit convergent. En s'attaquant directement au problème causant la non-convergence, il est possible de construire un estimateur qui soit convergent. Un **estimateur spectral** ou **par fenêtre** est de la forme

$$\hat{\sigma}^2_{N,\text{spec}}(f) = \sum_{k=-\infty}^{\infty} w_N(k)\hat{\gamma}_k(f),$$

où $w_N : \mathbb{Z} \to [0,1]$ est la fonction de poids parfois appelée la **fenêtre de délai**. Sous certaines conditions sur les autocovariances et sur w_N, il est possible de montrer la convergence forte de cet estimateur. Par exemple, Flegal et Jones (2010, théorème 1) exigent essentiellement une ergodicité géométrique, un quatrième moment de f fini, ainsi que certaines conditions techniques sur les poids. Un exemple de fonction de poids pouvant être utilisée est la fenêtre de Blackman-Tukey :

$$w_N(k) = \left(1 - 2a + 2a\cos\frac{\pi|k|}{b_n}\right)\mathbb{1}(|k| < b_n), \qquad b_n = \lfloor n^v \rfloor,$$

pour un certain choix de $a > 0$ et de $0 < v < 1$.

Dans tous les cas, les autocovariances doivent être calculées jusqu'à un certain délai b_N. L'approche par convolution (c.-à-d., par le calcul direct) est généralement lente alors que l'approche par la transformée de Fourier rapide (FFT : *Fast Fourier Transform*) est généralement plus rapide.

2.4.2.3 Estimation par moyennes de lots

L'idée Monte Carlo est d'utiliser un échantillon dont la distribution provient de la distribution cible pour approximer cette distribution par une version empirique. La même approche peut être utilisée pour estimer la variance. La chaîne $\{x_n\}_{n=1}^N$ est partitionnée en a_N lots de longueur b_N (tels que $N = a_N b_N$) et un estimé de $\pi(f)$ est produit dans chacun des lots. La collection de ces estimés agit en quelque sorte comme un échantillon de l'estimateur $\hat{\pi}_N(f)$; la distribution empirique de cet échantillon pourrait permettre d'estimer la distribution de l'estimateur. Puisque la distribution est connue pour être normale et que cet estimateur est asymptotiquement sans biais, seule la variance reste à déterminer. On estime enfin la variance asymptotique par la variance empirique de cet échantillon en argumentant que l'échantillon est pratiquement indépendant lorsque les lots sont suffisamment grands. Soit les estimés de chaque lot,

$$\hat{\pi}_{N,a}(f) = \frac{1}{b_N} \sum_{n=(a-1)b_N+1}^{ab_N} f(x_n), \qquad a = 1, \ldots, a_N.$$

Puis, la variance asymptotique est estimée par

$$\hat{\sigma}^2_{N,\mathrm{NOLBM}}(f) = \frac{b_N}{a_N - 1} \sum_{a=1}^{a_N} (\hat{\pi}_{N,a}(f) - \hat{\pi}_N(f))^2.$$

Cet estimateur, appelé **moyenne par lots sans chevauchement** ou NOLBM (pour *Nonoverlapping Batch Means*), n'est généralement pas convergent pour la variance asymptotique lorsque la taille de l'échantillon N est variable (par exemple, lorsque l'algorithme est arrêté par un certain critère), mais des conditions sur a_N et sur b_N peuvent assurer la convergence forte de l'estimateur. En effet, Jones et collab. (2006, section 3.2, voir aussi Flegal et Jones, 2010, section 3) montrent que si le nombre de lots et la longueur des lots augmentent avec N, alors l'estimateur est fortement convergent. Le rythme d'augmentation dépend de la fonction f et de l'algorithme, mais il est possible de choisir $b_N = \lfloor N^\theta \rfloor$ et $a_N = \lfloor N^{1-\theta} \rfloor$ pour un certain choix de $0 < \theta < 1$. À défaut de pouvoir calculer θ directement à partir des conditions, les auteurs suggèrent d'utiliser $\theta = 1/2$.

Maintenant, si la taille de l'échantillon est fixée, le nombre de lots peut être lui aussi fixé. Pour assurer l'indépendance des lots, leur taille doit être la plus grande possible pour que l'autocorrélation disparaisse. Ainsi, l'estimé de la variance asymptotique peut être utilisé pour construire des intervalles de confiance t qui ne requièrent pas un échantillon particulièrement grand. Plusieurs auteurs (Schmeiser, 1982 ou Geyer, 1992) recommandent d'utiliser de 10 à 30 lots : moins de 10 lots donne des intervalles instables, alors que plus de 30 lots ne procure pas d'intervalles de confiance particulièrement plus petits (vu que les quantiles t tendent rapidement vers les quantiles z.)

Un avantage de l'estimateur par lots est la réduction de la taille de la sortie de l'algorithme. En effet seule la moyenne de la fonction cible dans chaque lot est requise et non la valeur de la fonction évaluée à chaque point de l'échantillon. Un second avantage de cet estimateur est qu'il peut être utilisé en combinaison avec une méthode par séquence initiale (à la section suivante), de sorte que la taille des lots n'affecte plus l'estimation (Geyer, 2011, section 1.10.3).

Meketon et Schmeiser (1984) montrent que l'utilisation de lots avec chevauchement peut améliorer l'efficacité de l'estimateur en réduisant la variance asymptotique (de l'estimateur de la variance) d'un facteur pouvant aller jusqu'à 1/3. Geyer (2011, section 1.10.3) argue cependant contre l'utilisation de ce type d'estimateur puisque les estimateurs avec chevauchement ne bénéficient pas des deux avantages

exposés précédemment. De plus, ces estimateurs exigent plus de calculs : les observations sont utilisées plusieurs fois dans le calcul plutôt qu'une seule fois par l'estimateur sans chevauchement.

2.4.2.4 Estimation par séquence initiale

On a vu que l'estimation de la variance asymptotique ne pouvait se faire directement par l'estimation des autocovariances. Cependant, le résultat suivant permet de produire trois estimateurs possédant de bonnes propriétés en utilisant l'approche *plug-in* directe.

Théorème 2.33 (Geyer, 1992, théorème 3.1) *Soit une chaîne de Markov stationnaire, irréductible et réversible et soit $\gamma_k(f)$ les autocovariances de la chaîne. Alors, la fonction $m \mapsto \Gamma_m(f) = \gamma_{2m}(f) + \gamma_{2m+1}(f)$ est strictement positive, strictement décroissante et convexe.*

On estime d'abord $\Gamma_m(f)$ par $\hat{\Gamma}_{m,N}(f) = \hat{\gamma}_{2m,N}(f) + \hat{\gamma}_{2m+1,N}(f)$. Par la positivité de Γ_m, on s'attend à ce que les estimés $\hat{\Gamma}_{m,N}(f)$ soient positifs également. On note cependant que ces quantités tendent vers 0 et que le bruit introduit par l'estimation peut faire en sorte que $\hat{\Gamma}_{m,N}(f)$ soit négatif à un certain point ; après ce point on peut s'attendre à ce que tous les autres $\hat{\Gamma}_{m,N}(f)$ soient près de zéro, ce qui suggère de tronquer la somme dès que l'on trouve un $\hat{\Gamma}_{m,N}(f)$ négatif. L'estimateur par **séquence initiale positive** (IPS) est donc donné par

$$\hat{\sigma}^2_{N,\mathrm{IPS}}(f) = -\hat{\gamma}_0(f) + 2\sum_{m=0}^{M} \hat{\Gamma}_{m,N}(f),$$

où M est le plus grand entier tel que $\hat{\Gamma}_{m,N}(f) > 0$ pour tout $m = 0, \ldots, M$. Ceci correspond à tronquer la somme de l'estimateur naïf après $2M+1$ autocovariances :

$$\hat{\sigma}^2_{N,\mathrm{IPS}}(f) = \hat{\gamma}_0(f) + 2\sum_{k=1}^{2M+1} \hat{\gamma}_{m,N}(f).$$

La suite des estimés ne possède pas les garanties de positivité, de monotonicité ou de convexité comme c'est le cas pour la suite des Γ_m. On peut donc modifier l'estimateur IPS, qui n'assure que la positivité, pour assurer d'abord la monotonicité puis la convexité. L'estimateur par **séquence initiale monotone** (IMS) modifie les estimés de sorte à forcer la monotonicité,

$$\tilde{\Gamma}_{m,N}(f) = \min\left\{\tilde{\Gamma}_{m-1,N}(f), \hat{\Gamma}_{m,N}(f)\right\}, \qquad m = 0, \ldots, M,$$

pour produire l'estimé suivant,

$$\hat{\sigma}^2_{N,\mathrm{IMS}}(f) = -\hat{\gamma}_0(f) + 2\sum_{m=0}^{M} \tilde{\Gamma}_{m,N}(f).$$

Enfin, l'estimateur par **séquence initiale convexe** assure la convexité de la séquence initiale en modifiant les estimés utilisés par l'estimateur IMS. Ces trois estimateurs jouissent d'une même propriété (Geyer, 1992, théorème 3.2, voir aussi Kosorok, 2000, théorème 2) : ce sont des sur-estimateurs convergents pour la variance asymptotique. Dans le contexte de construction de régions de confiance, la sur-estimation n'est pas autant à craindre que la sous-estimation puisque le résultat ne sera que des régions possiblement trop vastes. Ainsi, ces trois estimateurs offrent donc des garanties intéressantes. Thompson (2010) observe que les trois estimateurs performent d'une manière similaire en pratique.

2.4.2.5 Fonctions multivariées

Lorsque f est une fonction multivariée (de dimension $p > 1$), l'estimateur Monte Carlo de $\pi(f)$ est tout simplement le vecteur des estimateurs Monte Carlo de chacune des composante. Les conditions pour un théorème limite central sont sensiblement les mêmes et, dans ce cas, la variance asymptotique est une matrice de dimension $p \times p$ donnée par

$$\Sigma(f) = \mathrm{Var}(f(X_0)) + 2\sum_{k\geqslant 1} \mathrm{Cov}\left(f(X_0), f(X_k)\right).$$

On peut généraliser la définition de fonction d'autocovariance dans le cas multivarié,

$$\Gamma_k(f) = \mathrm{Cov}\left(f(X_0), f(X_k)\right),$$

de sorte que l'on trouve la même expression pour la variance asymptotique :

$$\Sigma(f) = \Gamma_0(f) + 2\sum_{k\geqslant 1} \Gamma_k(f).$$

Afin d'estimer $\Sigma(f)$, la plupart des méthodes présentées peuvent être trivialement adaptées et les garanties théoriques sont souvent préservées. Vats et collab. (2019, théorème 2) démontrent la convergence forte de l'estimateur NOLBM ; Vats et collab. (2018, théorème 2) démontrent la convergence forte des estimateurs spectraux. Dans les deux cas, la chaîne de Markov est supposée polynomialement ergodique. De plus, Kosorok (2000, théorème 2, voir aussi Dai et Jones, 2017, théorème 2) établit que des versions multivariées des estimateurs par séquence initiale sont des sur-estimateurs convergents, comme le sont leurs équivalents univariés.

2.4.3 Comparaison et mesures de performance

Afin de comparer des chaînes MCMC entres elles, on doit avoir recours à la notion de l'efficacité d'une chaîne. Les algorithmes couramment utilisés satisfont généralement une loi des grands nombres ainsi qu'un théorème central limite de sorte que les propriétés asymptotiques de l'estimateur Monte Carlo sont connues. Ainsi, dans le contexte de l'estimation de $\pi(f)$, l'estimateur sera asymptotiquement sans biais et la variance asymptotique sera finie et estimable. Puisque le biais sera nul et qu'il ne reste que la variance pour complètement définir la distribution asymptotique de l'estimateur, une mesure naturelle de l'efficacité d'un algorithme MCMC est donc la variance asymptotique de la chaîne.

Pour deux noyaux de transition P_1 et P_2, on dira que P_1 a une **variance (asymptotique) inférieure** à celle de P_2 si $\sigma^2(f)$ est inférieur lorsque la chaîne utilise P_1 comme noyau de transition. Ceci permet alors de définir l'**efficacité relative** entre P_1 et P_2 :

$$\mathrm{eff}_{P_1,P_2}(f) = \frac{\sigma^2(f,P_2)}{\sigma^2(f,P_1)},$$

où $\sigma^2(f,P)$ correspond à la variance asymptotique avec le noyau P. Lorsque P_1 a une variance asymptotique inférieure, alors le ratio sera supérieur à 1 et l'efficacité relative sera élevée. De plus, on

considère également l'**efficacité d'un noyau au sein d'une famille** de noyaux $\mathcal{P}$:

$$\text{eff}_{P\in\mathcal{P}}(f) = \frac{\inf_{Q\in\mathcal{P}} \sigma^2(f,Q)}{\sigma^2(f,P)}.$$

Lorsque $P = \arg\min_{Q\in\mathcal{P}} \sigma^2(f,Q)$, alors on obtient une efficacité de 1 ; sinon, l'efficacité sera inférieure à 1.

Ces définitions admettent que la fonction f soit choisie d'avance et ne permettent que de comparer des chaînes pour une même fonction cible f. Plus généralement, on dira que P_1 est **uniformément au moins aussi efficace** que P_2 si $\sigma^2(f,P_1) \leqslant \sigma^2(f,P_2)$ pour tout $f \in L^2(f)$. Mira (2001b) et Tierney (1998) montrent que, sous certaines conditions (assez fortes), un noyau peut être uniformément plus efficace selon ce critère. La condition principale sera qu'un noyau domine l'autre dans l'**ordre de Peskun** (Peskun, 1973).

2.4.3.1 Temps d'autocorrélation

On note d'abord la formulation alternative suivante de la variance asymptotique en notant que la variance sous stationnarité ne dépend pas du noyau de transition P :

$$\begin{aligned}\sigma^2(f,P) &= \text{Var}\left(f(X_0)\right) + 2\sum_{k\geqslant 1} \text{Cov}\left(f(X_0),f(X_k)\right) \\ &= \text{Var}\left(f(X_0)\right)\left(1 + 2\sum_{k\geqslant 1} \text{Corr}\left(f(X_0),f(X_k)\right)\right) =: \sigma^2(f)\tau(f,P),\end{aligned}$$

où $\sigma^2(f)$ correspond à la variance de f à la stationnarité et $\tau(f,P)$ est appelé le **temps d'autocorrélation (intégré)** ((I)ACT : *(Integrated) Autocorrelation Time*) de la chaîne. Cette factorisation est particulièrement utile pour le calcul de l'efficacité étant donné que le facteur $\sigma^2(f)$ sera partagé par tous les noyaux et le ratio des variances se réduit alors au ratio des temps d'autocorrélation :

$$\text{eff}_{P_1,P_2}(f) = \frac{\sigma^2(f)\tau(f,P_2)}{\sigma^2(f)\tau(f,P_1)} = \frac{\tau(f,P_2)}{\tau(f,P_1)}.$$

Ainsi, pour comparer des chaînes entre elles, il est suffisant de calculer leur temps d'autocorrélation respectif. Cependant, l'estimation du temps d'autocorrélation est affectée par les mêmes problèmes que celle de la variance asymptotique : des méthodes analogues à celle de la section 2.4.2 doivent donc être utilisées.

2.4.3.2 Taille échantillonnale effective

Les méthodes MCMC sont généralement employées lorsqu'il n'est pas possible d'obtenir échantillon i.i.d. Les propriétés de ce type d'échantillon sont bien connues en statistique et une manière de mesurer la performance de chaînes MCMC serait par comparaison avec un échantillon i.i.d. En particulier, si deux échantillons suivent un théorème limite central, alors la variance asymptotique sera la seule propriété qui différera entre les deux estimateurs Monte Carlo respectifs. Un échantillon i.i.d. selon π

de taille N aura une variance asymptotique donnée par

$$N \operatorname{Var}_\pi \left(\frac{1}{N} \sum_{n=1}^{N} f(X_n) \right) = \operatorname{Var}_\pi \left(f(X_0) \right) = \sigma^2(f).$$

Ainsi, l'efficacité d'un algorithme MCMC par rapport à un échantillonneur i.i.d. sera donné par

$$\operatorname{eff}_{P,\text{iid}}(f) = \frac{\sigma^2(f)}{\sigma^2(f)\tau(f,P)} = \frac{1}{\tau(f,P)}.$$

Cette quantité peut prendre des valeurs dans $[0,\infty)$, mais il est assez rare que la variance asymptotique MCMC soit inférieure à celle i.i.d de sorte que cette efficacité est généralement inférieure à 1.

Le temps d'autocorrélation est donc une mesure de performance en soi qui compare la variance asymptotique d'une chaîne à celle d'une chaîne i.i.d. L'interprétation de cette quantité correspond au ratio des variances, mais une transformation permet une seconde interprétation plus intuitive. La **taille échantillonnale effective** (ESS : *Effective Sample Size*) d'une chaîne MCMC correspond à la taille échantillonnale i.i.d. requise pour produire un échantillon ayant la même variance asymptotique que celle de la chaîne MCMC. On doit donc résoudre l'égalité de variance des deux estimateurs Monte Carlo,

$$\frac{1}{N_{\text{i.i.d.}}} \sigma^2(f) = \frac{1}{N} \sigma^2(f)\tau(f,P)$$

d'où l'on définit

$$\operatorname{ESS}_P(f) := \frac{N_{\text{i.i.d.}}}{\tau(f,P)}. \tag{2.27}$$

En terme de comparaison et de mesure de performance, cette définition sera équivalente à l'utilisation de l'ACT puisque l'ordre sera préservé, mais elle permet cependant une interprétation plus pratique. Enfin, l'ESS est estimé en substituant un estimateur fortement convergent dans (2.27) :

$$\widehat{\operatorname{ESS}}_P(f) := \frac{N}{\hat{\tau}(f,P)}.$$

2.4.3.3 Saut quadratique moyen

Une autre mesure de performance est définie d'une manière intuitive, mais peut être réinterprétée comme une simplification de l'ACT. Une chaîne MCMC qui effectue de grands pas à chaque itération devrait naturellement être efficace. En effet, de grands sauts réduira l'autorrélation au sein de la chaîne et réduira du coup le temps d'autocorrélation. Ceci suggère donc de considérer le **saut euclidien moyen** (ESEJD : *Expected Squared Euclidian Jumping Distance*) défini comme étant l'espérance de la distance euclidienne au carré moyenne, c.-à-d.,

$$\operatorname{ESEJD}_P = \mathbb{E}\left\{ \|X_1 - X_0\|_2^2 \right\},$$

où l'espérance est prise par rapport à $X_0 \sim \pi$ et pour une transition P. Lorsque la densité cible présente de fortes corrélations ou des variances d'échelles différentes, il peut être préférable d'utiliser la distance de Mahalanobis afin de corriger les sauts. Ceci définit le **saut quadratique moyen**

(ESJD : *Expected Squared Jumping Distance*), qui satisfait

$$\mathrm{ESJD}_P = \mathbb{E}\left\{(X_1 - X_0)^\top \Sigma_\pi^{-1}(X_1 - X_0)\right\},$$

où Σ_π est la covariance de π.

Ces définitions ne prennent pas en compte la fonction cible f mais, tel que mentionné à la sous-section 2.4.1, certains résultats permettent d'établir une efficacité uniforme par rapport à $f \in L^2(\pi)$ de sorte que l'indépendance sur f n'est pas problématique dans bien des cas. Puis, en dimension $d = 1$, on retrouve le lien avec l'ACT. D'une part, on note l'égalité suivante :

$$\mathbb{E}\left\{\|X_0 - X_1\|_2^2\right\} = \mathrm{Var}\,(X_0 - X_1) = 2\,\mathrm{Var}(X_0) - 2\,\mathrm{Cov}(X_0,X_1) = 2\,\mathrm{Var}(X_0)\,(1 - \mathrm{Corr}(X_0,X_1))\,.$$

(Notons que l'ESEJD et l'ESJD sont équivalent, à une constante multiplicative près en dimension 1.) Ainsi, le saut euclidien moyen sera maximisé lorsque l'autocorrélation d'ordre 1 sera minimisée. D'autre part, la troncature de l'ACT au premier délai est donnée par

$$\tau_P^1(f) = 1 + 2\,\mathrm{Corr}(X_0,X_1),$$

et sera minimale elle aussi lorsque l'autocorrélation d'ordre 1 sera minimale. On peut voir la troncation au premier délai comme une estimation spectrale où les poids sont donnés par $w_N(k) = \mathbb{1}(|k| \leqslant 1)$, de sorte que les deux critères, l'ESJD et l'ACT, sont équivalents pour $d = 1$ et ce choix d'estimateur. Un des avantages du saut quadratique (ou euclidien) comme critère est qu'il est extrêmement simple à calculer comparativement à l'ACT, particulièrement lorsque f est multivariée. La correspondance précédente montre que la simplification produira des comparaisons similaires.

2.4.3.4 Fonctions multivariées

Lorsque f est une fonction multivariée, il est possible de calculer chacune de ces mesures puis de les agréger (e.g. en choisissant le minimum des composantes) afin de comparer des méthodes MCMC. Il existe cependant des extensions multivariées à ces mesures qui prennent en compte la codépendance.

D'abord, la variance asymptotique sera alors une matrice symétrique définie positive de dimension $p \times p$. Parmi cette classe, il n'existe pas d'ordre unique comme dans les réels. Cette famille de matrices fait partie du cône des matrices définies semi-positives dans lequel un ordre partiel est donné par l'ordre de Loewner : pour A,B deux matrices définies semi-positives, on dit que $A \succcurlyeq B$ si $A - B$ est définie semi-positive. D'autres ordres sont également possibles, tels que les ordres induits par le déterminant ou la trace de la matrice.

Le temps d'autocorrélation trouve lui aussi une extension multivariée. Soit $\Sigma(f)$ la covariance de $f(X_0)$ à stationnarité et soit $S(f)$ une racine carrée de $\Sigma(f)$, c.-à-d., $S(f)S(f)^\top = \Sigma(f)$. Alors, on

peut écrire la covariance asymptotique d'une chaîne à noyau P selon :

$$\begin{aligned}\Sigma_P(f) &= N \operatorname{Cov}\left(\frac{1}{N}\sum_{n=1}^{N} f(X_n)\right) \\ &= \operatorname{Var}(f(X_0)) + 2\sum_{k\geqslant 1}\left(1-\frac{k}{N}\right)\operatorname{Cov}(f(X_0),f(X_k)) \\ &\overset{N\to\infty}{\longrightarrow} \operatorname{Var}(f(X_0)) + 2\sum_{k\geqslant 1}\operatorname{Cov}(f(X_0),f(X_k)) \\ &= S(f)\left(I_p + 2\sum_{k\geqslant 1}\operatorname{Corr}(f(X_0),f(X_k))\right)S(f)^\top =: S(f)\tau_P(f)S(f)^\top\end{aligned}$$

où $\tau_P(f)$ est l'**ACT multivarié**. Maintenant, tout comme pour la variance asymptotique, $\tau_P(f)$ est une matrice $d \times d$ et la comparaison doit être effectuée selon un ordre prédéterminé.

Pour ce qui est de l'ESS, il est possible d'estimer un ESS par composante et de choisir le plus petit comme mesure agrégée, mais Vats et collab. (2019) propose d'utiliser une généralisation de l'ESS utilisant l'ordre induit par le déterminant :

$$\operatorname{ESS}_P(f) := N\left(\frac{\det\Sigma(f)}{\det\Sigma_P(f)}\right)^{1/p} = \frac{N}{\det^{1/p}\tau_P(f)}.$$

Notons que cette définition correspond à définition univariée de l'ESS lorsque $p = 1$. Cette quantité s'estime en remplaçant $\Sigma(f)$ et $\Sigma_P(f)$ par leur estimateur fortement convergent respectif.

Finalement, puisque les critères du ESEJD et du ESJD ne reposent pas sur le choix de f, alors ils s'appliquent trivialement au cas multivarié tout comme au cas univarié. On trouve cependant un lien avec une version particulière du critère de l'ACT pour la fonction identité $f(x) = x$. On remarque que le saut quadratique est en fait une forme quadratique, ce qui permet d'évaluer directement l'espérance : soit $\Sigma_\pi = SS^\top$, alors

$$\begin{aligned}\mathbb{E}\left\{(X_0-X_1)^\top\Sigma_\pi^{-1}(X_0-X_1)\right\} &= \operatorname{tr}\left(\Sigma_\pi^{-1}\operatorname{Var}(X_0-X_1)\right) + \mathbb{E}\left\{(X_0-X_1)\right\}^\top\Sigma_\pi^{-1}\mathbb{E}\left\{(X_0-X_1)\right\} \\ &= \operatorname{tr}\left(\Sigma_\pi^{-1}\left(2\Sigma_\pi - 2\operatorname{Cov}(X_0,X_1)\right)\right) + 0 \\ &= 2\operatorname{tr}\left(\Sigma_\pi^{-1}S\left(I_d - \operatorname{Corr}(X_0,X_1)\right)S^\top\right) \\ &= 2\operatorname{tr}\left(I_d - \operatorname{Corr}(X_0,X_1)\right) \\ &= 2\left(d - \sum_{j=1}^{d}\operatorname{Corr}(X_{0,j},X_{1,j})\right).\end{aligned}$$

Cette quantité est maximale lorsque $\operatorname{Corr}(X_{0,j},X_{1,j}) = -1$ pour tout $j = 1,\dots,d$. Si l'on considère plutôt l'ACT estimé spectralement par les poids $w_N(k) = \mathbb{1}(|k| \leqslant 1)$, alors on trouve

$$\tau_P^1 = I_d + 2\operatorname{Corr}(X_0,X_1).$$

En choisissant la trace comme ordre sur les matrices, on trouve alors

$$\operatorname{tr}(\tau_P^1) = d + 2\sum_{j=1}^{d}\operatorname{Corr}(X_{0,j},X_{1,j}),$$

qui est lui aussi minimisé lorsque $\text{Corr}(X_{0,j},X_{1,j}) = -1$.

2.5 Échelle optimale des algorithmes MCMC

Les différents résultats théoriques sur les algorithmes Metropolis-Hastings (section 2.3.1) posent certaines conditions sur la densité instrumentale Q afin d'assurer la validité de l'algorithme. La positivité locale de la densité, entre autres, permet d'assurer l'ergodicité de l'algorithme (proposition 2.25). Cependant, une large classe de densités satisfait cette condition et on a donc peu d'information quant à laquelle choisir. Ensuite, pour vérifier un théorème limite central, l'ergodicité V-géométrique est une condition suffisante souvent utilisée ; celle-ci peut être vérifiée à l'aide de conditions de dérive géométrique sur la densité instrumentale. Ceci réduit à nouveau la classe de densités instrumentales pertinentes, mais beaucoup de liberté persiste.

Un théorème limite central englobe la totalité des propriétés asymptotiques de l'estimateur Monte Carlo en déterminant sa distribution limite. Ainsi, les paramètres de cette distribution asymptotique, qui dépendent de la densité instrumentale, peuvent aider dans le choix de Q. En effet, la variance asymptotique est le seul paramètre de la distribution asymptotique de l'estimateur qui variera d'une densité instrumentale à l'autre. Dans un contexte d'estimation, une variance minimale est souhaitable de sorte que le choix de Q puisse être guidé par une volonté de réduction de la variance de l'estimateur.

Par contre, minimiser la variance au sein d'une si grande classe de densités (définie seulement par la dérive géométrique, par exemple) est généralement impossible. Alors, la recherche dans ce domaine se limite généralement à une famille paramétrique gaussienne isotropique, c.-à-d., $\mathcal{N}_d(\mathbf{0}, \sigma^2 I_d)$, $\sigma > 0$. Dans ce cas, la minimisation s'effectue sur le paramètre d'échelle unidimensionnel : ce problème est souvent appelé **échelle optimale** pour cette raison, bien que plusieurs résultats s'expriment plus simplement en terme de probabilité d'acceptation optimale. En effet, le calcul de l'échelle optimale peut être difficile : il est plus simple de mettre au point manuellement l'échelle de sorte à obtenir un taux d'acceptation empirique près de la probabilité optimale. Cette section portera donc sur divers résultats d'échelle optimale pour quelques algorithmes MCMC et par rapport à la forme de la densité cible π.

Notons que ce domaine de recherche est beaucoup plus vaste que la courte exposition qui suit. Seuls les résultats sur la valeur même de l'échelle optimale ainsi que la probabilité d'acceptation seront présentés. Les références aux notions de diffusion limite, de vitesse et d'efficacité, quoiqu'essentielles aux résultats, ne seront pas abordées afin de ne pas alourdir le texte en introduisant davantage de théorie qui ne sera pas utilisée ultérieurement.

2.5.1 L'algorithme de Metropolis

La majeure partie des résultats d'échelle optimale apparaissent dans le contexte de l'algorithme de Metropolis 2.3.1, où la densité instrumentale est une loi normale de covariance proportionnelle à l'identité, c.-à-d. $\mathcal{N}_d(\mathbf{0}, \sigma^2 I_d)$, $\sigma > 0$.

2.5.1.1 Densité cible à composantes i.i.d.

Roberts et collab. (1997) ont obtenu le premier résultat théorique d'échelle optimale pour une densité cible à composantes i.i.d., c.-à-d.,

$$\pi(x) = \prod_{j=1}^{d} g(x_j), \qquad x \in \mathcal{X} \subseteq \mathbb{R}^d, \tag{2.28}$$

où $g(\cdot)$ est une densité sur $\mathbb{R}$. Dans ce cas, l'efficacité est définie comme l'inverse du temps d'autocorrélation de la première composante de la chaîne.

Théorème 2.34 (Roberts et collab., 1997, théorème 1.1 et corollaire 1.2) *Soit un algorithme de Metropolis de type marche aléatoire à densité instrumentale $\mathcal{N}_d(\mathbf{0}, \sigma_d^2 I_d)$, $\sigma_d > 0$, soit π, une densité cible à composantes i.i.d.* (2.28) *pour une densité marginale g et soit*

$$\mathcal{I}_g = \mathbb{E}_g\left\{\left(\frac{\mathrm{d}}{\mathrm{d}x}\log g(X)\right)^2\right\} = \mathbb{E}_g\left\{\left(\frac{g'(X)}{g(X)}\right)^2\right\}.$$

Sous certaines conditions de régularité sur g, alors, pour $d \to \infty$, l'échelle optimale, c.-à-d. celle qui maximise l'efficacité (section 2.4.3), est donnée par $\sigma_d^2 = \ell^2/d$ où $\ell = (2{,}38)^2/\mathcal{I}_g$ et la probabilité d'acceptation optimale associée est de 0,234.

Corollaire 2.35 (Roberts et Rosenthal, 2001, section 2.2) *À la limite, toutes les mesures d'efficacité basées sur l'ACT sont équivalentes jusqu'à une constante multiplicative donnée par le choix de la fonction f pour laquelle l'ACT est calculé.*

Quelques remarques s'imposent sur ce théorème. D'abord, la supposition de composantes i.i.d. est particulièrement forte étant donné qu'une seule composante serait suffisante pour étudier π : ainsi, il serait suffisant d'étudier la fonction g univariée, ce qui est généralement un problème beaucoup plus simple. En effet, des méthodes d'intégration numériques pourront être utilisée pour évaluer $\pi(f)$. Divers travaux de recherche tentent de relâcher cette supposition : quelques cas seront abordés par la suite.

Ensuite, ce résultat est asymptotique par rapport à la dimension d. En pratique, la dimension du problème sera fixée et finie de sorte qu'il n'est pas clair, à premier abord, si ce résultat présente un intérêt. Gelman et collab. (1996, section 3.2) ont effectué une étude numérique de l'échelle optimale en petite dimension ($d \leqslant 10$) et ont obtenu les résultats présentés au tableau 2.1 pour une densité cible gaussienne en minimisant l'ACT d'une des composantes. En comparant avec le résultat théorique, on voit que l'échelle optimale en dimension finie est similaire à la valeur optimale asymptotiquement. La probabilité d'acceptation optimale est par contre relativement différente pour les petites dimensions ($d \leqslant 5$) et en particulier en dimension 1 où la probabilité d'acception optimale est de 0,441. Cette remarque est importante dans l'application de ce résultat à des algorithmes où le nouvel état de la chaîne consiste en une mise-à-jour d'une seule composante (e.g. l'algorithme de *Metropolis-wihtin-Gibbs* 2.7.)

De plus, il est important de noter que l'ajustement de l'échelle à sa valeur optimale n'a pas besoin d'être exact. En effet, on observe que l'efficacité est relativement constante autour de l'échelle optimale ; similairement, l'efficacité est stable près de la probabilité d'acceptation optimale (voir figure 2.1). Par exemple, l'efficacité est supérieure à 95% de sa valeur optimale pour une échelle dans l'intervalle $[1{,}97; 2{,}82]$ ou bien une probabilité d'acceptation dans l'intervalle $[0{,}158; 0{,}324]$.

Enfin, la conclusion additionnelle que toutes les mesures d'efficacité sont équivalentes est intéressante au sens ou elle retire la dépendance sur le choix de la fonction cible f pour laquelle $\pi(f)$ est recherchée.

Dimension (d)	Échelle opt. (σ_d)	$2{,}38/\sqrt{d}$	Pr. d'acceptation
1	2,40	2,38	0,441
2	1,70	1,68	0,352
3	1,39	1,37	0,316
4	1,25	1,19	0,279
5	1,10	1,06	0,275
6	1,00	0,97	0,266
7	0,93	0,90	0,261
8	0,87	0,84	0,255
9	0,80	0,79	0,261
10	0,74	0,75	0,267
$\to \infty$	$2{,}38/\sqrt{d}$	–	0,234

Tableau 2.1 *Échelle optimale de l'algorithme Metropolis à densité instrumentale $\mathcal{N}_d(\mathbf{0}, \sigma_d^2 I_d)$ pour une densité cible à composante i.i.d. avec $\mathcal{I}_g = 1$ en fonction de la dimension (Gelman et collab., 1996; Roberts et collab., 1997).*

Ce résultat d'échelle optimale est donc uniforme par rapport à f. Bien que les conditions soient assez restrictives et que le résultat soit asymptotique, on obtient tout de même une confirmation que les différentes mesures d'efficacité seront généralement en accord.

2.5.1.2 Densité cible plus générale

Depuis les travaux de Roberts et collab. (1997) sur une densité cible à composantes i.i.d., plusieurs tentatives de généralisation à des classes de densités cibles moins restreintes ont eu lieu.

D'abord, Roberts et Rosenthal (2001) relâchent légèrement la supposition de distributions identiques en introduisant un facteur d'échelle diffèrent pour chaque composante :

$$\pi(x) = \prod_{j=1}^{d} C_j g(C_j x_j), \tag{2.29}$$

où $g(\cdot)$ est à nouveau une densité sur $\mathbb{R}$ et $C_j > 0$ sont les facteurs d'échelle possiblement aléatoires. Dans ce cas, on retrouve des conclusions similaires à celles du théorème 2.34.

Théorème 2.36 (Roberts et Rosenthal, 2001, théorème 5 et 6) *Soit un algorithme Metropolis à densité instrumentale $\mathcal{N}_d(\mathbf{0}, \sigma_d^2 I_d)$, $\sigma_d > 0$, et soit π, une densité cible de la forme (2.29) pour une densité marginale g et des facteurs d'échelle i.i.d. satisfaisant $\mathbb{E}\{C_j\} = 1$ et $\mathrm{Var}(C_j) = b < \infty$.*

Sous certaines conditions de régularité sur g, alors, pour $d \to \infty$, la probabilité d'acceptation optimale est de 0,234 pour l'estimation de toute fonction de la première composante. L'échelle optimale est alors la même qu'au théorème 2.28, mais multipliée par C_1^2/b.

De plus, ces conclusions demeurent en utilisant plutôt une densité instrumentale non-homogène donnée par la covariance $\sigma_d^2 \,\mathrm{diag}(C_1, \ldots, C_d)$. Dans ce cas, l'estimation sera plus efficace par un facteur $\mathbb{E}\{C_i^2\} / \mathbb{E}\{C_i\}^2$.

La nature aléatoire des facteurs d'échelle—qui peut paraître étonnante à première vue puisque la densité cible est connue en pratique et d'autant plus que les résultats sont asymptotiques ($d \to \infty$) impliquant que l'analyse requiert un nombre potentiellement infini de paramètres aléatoires—permet d'étendre la portée pratique du résultat. En effet, comme Roberts et Rosenthal (2001) le mentionnent, même si la densité cible n'est qu'*approximativement* de la forme (2.29) avec des C_i variant significativement, alors le théorème 2.36 suggère qu'il peut être profitable d'obtenir des estimés préliminaires des C_i pour définir la densité instrumentale.

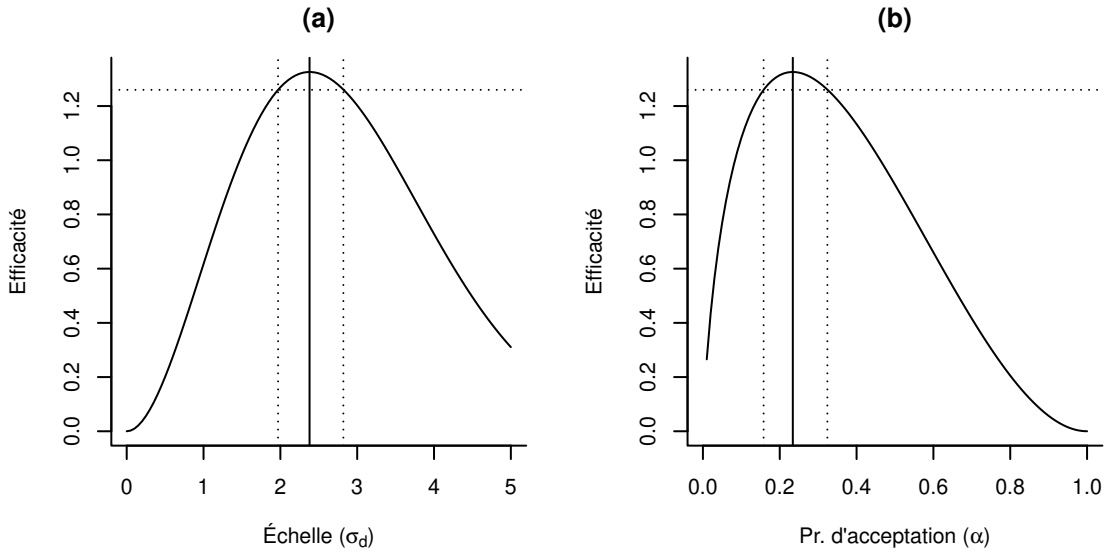


Figure 2.1 *Efficacité de l'algorithme Metropolis à densité instrumentale $\mathcal{N}_d(\mathbf{0}, \sigma_d^2 I_d)$ pour une densité cible à composante i.i.d. avec $\mathcal{I}_g = 1$ en fonction (a) de l'échelle et (b) de la probabilité d'acceptation. La ligne horizontale hachurée représente 95% de l'efficacité maximale les lignes verticales hachurées représentent les bornes de la région 95% et la ligne verticale pleine indique la valeur optimale.*

Les conditions i.i.d. sur les facteurs d'échelle font en sorte qu'ils seront relativement tous du même ordre peu importe j et peu importe la dimension d. Bédard et Rosenthal (2008, voir aussi Bédard, 2007, Bédard, 2008 et Bédard, 2008) considèrent un cas plus général où les facteurs d'échelle $\theta_j(d)$ peuvent dépendre à la fois de j et de d :

$$\pi(x) = \prod_{j=1}^{d} \frac{1}{\sqrt{\theta_j^2(d)}} g\left(\frac{x_j}{\sqrt{\theta_j^2(d)}}\right).$$

Des conditions sur les propriétés de $\theta_j(d)$ influenceront directement les résultats d'échelle optimale. Sans entrer dans trop de détails, la probabilité d'acception optimale de l'algorithme sera de 0,234 si et seulement si

$$\lim_{d\to\infty} \frac{\sum_{j=1}^{n} \theta_j^{-2}(d)}{\sum_{j=1}^{d} \theta_j^{-2}(d)} = 0,$$

pour un certain $n < d$ fixé. L'interprétation de cette condition est que les facteurs d'échelle des premières composantes ne doivent pas être dominées par le reste des facteurs (c.-à-d., d'ordre trop petit.) Lorsque cette limite est entre 0 et 1, alors la probabilité d'acceptation optimale sera plus petite que 0,234.

Breyer et Roberts (2000) relâchent quant à eux la supposition d'indépendance en introduisant une corrélation partielle entre les composantes. Des résultats d'échelle optimale asymptotique peuvent alors être obtenus à condition que l'étendue de la corrélation soit finie, c.-à-d., que la corrélation soit nulle au-delà d'une certaine différence d'indice de composante. Dans le contexte des champs aléatoires, Beskos et collab. (2009, voir aussi Mattingly et collab., 2012) arrivent à un résultat d'échelle optimale pour des densités cibles à dimensions infinies comprenant de la corrélation.

Sherlock et Roberts (2009) considèrent la classe des densités elliptiquement symétriques unimodales.

Une telle densité π est le résultat d'une transformation linéaire orthogonale $T : \mathbb{R}^d \to \mathbb{R}^d$ d'une densité sphériquement symétrique $p : \mathbb{R}^d \to [0,\infty)$, c'est à dire

$$\pi(x) \propto p(Tx), \quad p(y) = \frac{1}{\sigma^d} g\left(\frac{\|y-\mu\|_2^2}{\sigma^2}\right),$$

pour une certaine densité univariée $g : \mathbb{R} \to [0,\infty)$ et des paramètres de centralité $\mu \in \mathbb{R}^d$ et de dispersion $\sigma > 0$. Pour de telles densités, l'algorithme Metropolis admet une échelle optimale asymptotique (et une probabilité d'acceptation optimale) parmi les densités instrumentales symétriques de la forme

$$q(y|x) = \frac{1}{\lambda^d} r\left(\frac{y-x}{\lambda}\right),$$

pour une certaine densité symétrique $r : \mathbb{R}^d \to [0,\infty)$ et un paramètre d'échelle $\lambda > 0$, en admettant que la transformation ne soit pas trop excentrique en terme d'ellipse (cette condition est similaire à celle sur les facteurs d'échelle dans Bédard et Rosenthal, 2008.) De plus, un résultat d'échelle optimale est également obtenu en dimension finie pour des densités sphériquement symétriques ; la mesure d'efficacité utilisée dans ce cas est l'ESJD.

2.5.2 L'algorithme Langevin et autres

L'algorithme Langevin (MALA, section 2.3.2) jouit également de résultats théoriques d'échelle optimale. Roberts et Rosenthal (1998) démontrent un résultat similaire au théorème 2.34 pour une densité cible à composantes i.i.d.

Théorème 2.37 (Roberts et Rosenthal, 1998, théorèmes 1 et 2) *Soit un algorithme Langevin à propositions*

$$Y|X = x \sim \mathcal{N}_d\left(x + \frac{\sigma_d^2}{2}\nabla \log \pi(x), \sigma_d^2 I_d\right),$$

où $\sigma_d > 0$ et π est la densité cible à composantes i.i.d. (2.28) pour une densité marginale g. Sous certaines conditions de régularité sur g, alors, pour $d \to \infty$, l'échelle optimale est donnée par $\sigma_d^2 = \ell^2/d^{1/3}$ pour un certain ℓ et la probabilité d'acceptation optimale est de 0,574.

On note deux différences majeures entre le résultat pour l'algorithme Metropolis et celui pour l'algorithme Langevin. D'abord, la probabilité d'acceptation optimale est plus élevée – 0,574 plutôt que 0,234 – pour l'algorithme Langevin. Ceci s'explique par le fait que l'algorithme Langevin est construit de sorte à produire des candidats de meilleure qualité puisque la marche aléatoire est biaisée dans la direction du gradient de π. Ensuite, l'exposant affectant la dimension d dans l'échelle σ_d^2 passe de 1 à 1/3 : le temps de convergence (en nombre d'itérations) de l'algorithme est de l'ordre de $\mathcal{O}(d^{1/3})$ plutôt que de l'ordre de $\mathcal{O}(d)$. Ainsi, le calcul de $\nabla \log \pi$ lors des propositions augmente théoriquement l'efficacité de l'algorithme et il reste donc à voir si, selon le choix de π, l'augmentation du temps de calcul n'est pas significativement importante en comparaison.

La généralisation de Roberts et Rosenthal (2001) aux densités cible à composantes non-homogène (théorème 2.36) est également valide pour l'algorithme Langevin. En effet, ils trouvent un résultat similaire (théorème 7) où $\sigma_d^2 = \ell^2/d^{1/3}$ et où la probabilité d'acceptation optimale est toujours de 0,574. Breyer et collab. (2004) arrivent à ces mêmes conclusions pour une densité cible provenant d'un

contexte de régression non-linéaire. Plus récemment, Pillai et collab. (2012) trouvent qu'une probabilité d'acception de 0,574 est optimale également pour une classe de densités cibles à dimensions infinies (champs aléatoires.)

Pour ce qui est d'autres algorithmes, on relève quelques résultats d'échelle optimale présentés dans la littérature. Bédard et collab. (2012) proposent une étude de certaines variantes de l'algorithme Metropolis à essais multiples (section 4.1) dont il sera question à la section 4.2.3. Bédard et collab. (2014) considèrent quant à eux l'algorithme Metropolis-Hastings à rejet retardé (section 2.3.2). Pour une densité cible à composantes i.i.d., ils obtiennent une probabilité d'acceptation optimale de 0,234 lorsque deux étapes indépendantes ou antithétiques sont utilisées. Enfin, notons que Beskos et collab. (2013) étudient le problème d'échelle optimale pour un algorithme Monte Carlo hybride ; dans ce cas, le probabilité d'acceptation optimale est de 0,651 et l'exposant de la dimension dans l'échelle optimale est de 1/4 de sorte que cet algorithme est encore plus efficace que l'algorithme MALA.

Chapitre 3

MCMC adaptatifs

Soit $\mathcal{X}$ un espace d'états, π une distribution cible à support dans $\mathcal{X}$ et $f : \mathcal{X} \to \mathbb{R}$ une fonction dont l'espérance sous π est recherchée. On cherche donc à calculer

$$\pi(f) = \int_{\mathcal{X}} f(x)\pi(\,\mathrm{d}x).$$

Les méthodes Monte Carlo par chaînes de Markov (MCMC) abordées au Chapitre 2 permettent d'estimer ce genre d'intégrale en produisant un échantillon dont la distribution asymptotique est celle de π ; puis $\pi(f)$ est estimé par l'estimateur Monte Carlo

$$\hat{\pi}_N(f) = \frac{1}{N}\sum_{n=1}^{N} f(x_n).$$

À la Section 2.5 , il fut question de l'optimisation d'algorithmes MCMC par rapport à l'efficacité de l'estimation de $\pi(f)$ par $\hat{\pi}_N(f)$ pour un certain choix de distribution de transition. Par exemple, le théorème 2.34 (dû à Roberts et collab., 1997) indique qu'une densité de proposition gaussienne $\mathcal{N}_d(\mathbf{0}, \sigma^2 I_d)$ dans un algorithme Metropolis est optimale pour une densité cible $\pi(x) = \prod_{i=1}^{d} f(x_i)$ lorsque l'on choisit la covariance $\sigma^2 = (2{,}38)^2/d$ (asymptotiquement par rapport à la dimension d et sous certaines conditions de régularités pour f.) Le résultat est ensuite étendu aux distributions cibles non-homogènes $\pi(x) = \prod_{i=1}^{d} C_i f(C_i x_i)$ (théorème 2.36) sous certaines conditions additionnelles sur les échelles C_i. Ainsi, si la distribution cible est gaussienne, c.-à-d., $\pi = \mathcal{N}_d(\mu_\pi, \Sigma_\pi)$, il est possible de montrer, après changement de base, qu'une densité de proposition $\mathcal{N}_d(\mathbf{0}, s_d\Sigma)$ sera optimale (asymptotiquement) lorsque $\Sigma = \Sigma_\pi$ et $s_d = (2{,}38)^2/d$.

En pratique, la densité cible π ne sera pas gaussienne et, même si on le suppose, la matrice de covariance Σ_π ne sera pas connue. Il faut donc procéder par « essais-erreurs » afin de chercher une matrice suffisamment proche de Σ_π. En petite dimension, cette méthode peut être envisageable, mais elle devient rapidement impraticable lorsque d augmente : $d(d+1)/2$ paramètres devront être ajustés manuellement. Alternativement, la covariance de la distribution cible peut être grossièrement estimée par une première chaîne de Markov et la chaîne servant à produire l'estimé de $\pi(f)$ sera produite à l'aide de cet estimé. Cette seconde méthode peut cependant s'avérer problématique dans la mesure où la covariance utilisée dans la chaîne estimant la covariance doit elle-même être relativement bien ajustée afin de produire un bon estimé ; on peut donc se retrouver dans un impasse lorsqu'une première covariance raisonnable est difficile à trouver.

Afin d'intégrer l'estimation de Σ_π à même l'algorithme, Haario et collab. (2001) proposent d'utiliser les échantillons précédents d'un algorithme Metropolis pour en apprendre davantage sur Σ_π tout en produisant la chaîne estimant $\pi(f)$. Concrètement, ils suggèrent de faire évoluer la densité de proposition gaussienne au fur et à mesure que l'on progresse dans les itérations, en remplaçant la matrice de covariance par la covariance empirique de l'échantillon obtenu jusqu'à présent. L'algorithme Metropolis Adaptatif (AM) ainsi construit se trouve à l'Algorithme 3.1. Ainsi, on espère que la succession d'échantillons représente bien la densité cible et, donc, que la covariance empirique $\Sigma_n = \mathrm{Cov}(x_{0:n})$ s'approche de Σ_π. Enfin, si les nouveaux échantillons sont obtenus à partir d'une marche aléatoire gaussienne de covariance $s_d\Sigma_n$, alors on peut s'attendre à ce que l'algorithme soit optimal dans l'esprit des résultats de la section 2.5.

Algorithme 3.1 Metropolis Adaptatif (AM, Haario et collab., 2001)

Données Densité cible π à support dans $\mathbb{R}^d$, paramètre d'échelle $s_d = (2.38)^2/d$, paramètre assurant la non-singularité $\varepsilon > 0$ et durée de non-adaptation initiale n_0.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0 et covariance initiale Σ_0.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonage Metropolis.*
      i. Proposition $Y \sim \mathcal{N}_d(x_n, s_d\Sigma_n)$;
      ii. Calcul de la probabilité d'acceptation
      $$\alpha(x_n, y) = \min\left\{1, \frac{\pi(y)}{\pi(x_n)}\right\};$$
      iii. Avec probabilité $\alpha(x_n, y)$, acceptation de la proposition $(x_{n+1} = y)$; sinon rejet $(x_{n+1} = x_n.)$
   (b) *Adaptation.* Mise à jour de la covariance
   $$\Sigma_{n+1} = \begin{cases} \Sigma_0, & n+1 < n_0; \\ s_d\left[\mathrm{Cov}\left(x_{0:n+1}\right) + \varepsilon I_d\right], & n+1 \geqslant n_0. \end{cases}$$

Sortie L'échantillon $x_{1:N}$.

On remarque cependant un problème potentiel à cette construction : chacune des densités de proposition dépend alors de tout le passé de la chaîne qui n'est dès lors plus markovienne. Toute la théorie sur les méthodes MCMC semblerait donc s'écrouler, mais ce n'est pas le cas : il est possible de produire un cadre théorique justifiant la validité d'un tel algorithme, ce qui constituera l'objet de ce chapitre. En parallèle avec les algorithmes proposés dans Gilks et collab. (1994) et dans Gilks et collab. (1998), l'article phare de Haario et collab. (2001) a lancé le développement de ce qui est désormais connu comme les *algorithmes MCMC adaptatifs* dont la popularité ne cesse de croître depuis.

Bien qu'il soit possible de démontrer la validité de tels algorithmes (ergodicité, loi des grands nombres, théorème limite central), aucun résultat théorique ne permet de montrer la supériorité des algorithmes adaptatifs vis-à-vis les algorithmes MCMC non-adaptatifs. En fait, l'utilisation des algorithmes MCMC adaptatifs repose sur l'heuristique suivante : sachant un résultat d'optimalité—par exemple, d'échelle optimale—, on produit l'échantillon Monte Carlo tout en modifiant automatiquement le noyau de transition de façon à ce qu'il s'approche de l'optimalité. Il s'agit donc plutôt d'un argument de facilité d'utilisation qu'un argument de performance. On comprend donc que les algorithmes

adaptatifs présenteront de meilleurs résultats empiriques dès que comparés a leurs versions non-adaptatives avec mise au point sous-optimale.

Ce chapitre sera divisé comme suit. À la Section 3.1, il sera question de la définition générale des algorithmes adaptatifs ainsi que d'une revue des différentes manières d'effectuer l'adaptation. Plusieurs exemples d'algorithmes dans la littérature y seront donc présentés. La Section 3.2 abordera la question de l'ergodicité des algorithmes adaptatifs et la Section 3.3 comportera différentes propriétés de l'estimation effectuée par les algorithmes adaptatifs.

3.1 Définition

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité sur lequel toutes les variables aléatoires considérées sont définies. On considère une densité cible π à support $\mathcal{X} \subseteq \mathbb{R}^d$ et une famille de densités de transition $\{P_\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+\}_{\theta \in \Theta}$ où $P_\theta(\cdot|x)$ est une densité sur $\mathcal{X}$ pour tout $x \in \mathcal{X}$ et Θ est un espace indexant les densités de transition. On privilégiera la notation $P_\theta(y|x)$ afin de mettre l'emphase sur la transition de x à y.

L'espace Θ correspondra souvent à l'espace des paramètres possibles pour un choix de distribution. Par exemple, si l'on choisit une densité de proposition gaussienne dans une marche aléatoire (Metropolis-Hastings), alors la densité de proposition sera celle de $\mathcal{N}_d(x,\Sigma)$ et l'on peut choisir Θ comme l'espace des matrices $d \times d$ symétriques définies semi-positives.

Dans les algorithmes MCMC homogènes, comme ceux présentés au chapitre 2, le paramètre θ est laissé fixe pour toute la durée de l'échantillonnage et le processus $\{X_n\}_{n \geqslant 0}$ vérifie la propriété markovienne, c.-à-d., pour toute fonction $f : \mathcal{X} \to \mathbb{R}$ bornée et mesurable

$$\mathbb{E}\{f(X_{n+1})|X_{0:n} = x_{0:n}\} = \mathbb{E}\{f(X_{n+1})|X_n = x_n\} = \int_{\mathcal{X}} f(y) P_\theta(\,\mathrm{d}y|x_n) =: P_\theta f(x_n).$$

Dans les algorithmes MCMC adaptatifs, θ est plutôt une variable aléatoire et la densité de transition P_θ dépend de la réalisation de θ. On considère donc le processus conjoint $\{(X_n,\theta_n)\}_{n \geqslant 0}$ adapté à une certaine filtration $\mathcal{F} = \{\mathcal{F}_n\}_{n \geqslant 0}$ satisfaisant

$$\mathbb{E}\{f(X_{n+1})|\,\mathcal{F}_n\} = \int_{\mathcal{X}} f(y) P_{\theta_n}(\,\mathrm{d}y|x_n) =: P_{\theta_n} f(x_n), \tag{3.1}$$

pour toute fonction $f : \mathcal{X} \to \mathbb{R}$ bornée et mesurable. La propriété (3.1) ne correspond pas exactement à la propriété markovienne du processus conjoint, mais cette propriété sera suffisante pour étudier rigoureusement les MCMC adaptatifs. Dans bien des applications, $\{(X_n,\theta_n)\}_{n \geqslant 0}$ sera tout de même markovien, mais ce ne sera généralement jamais le cas du processus $\{X_n\}_{n \geqslant 0}$ seul. Donc, on dira que l'adaptation est **markovienne** si θ_{n+1} est $\sigma(X_n,\theta_n)$-mesurable. Dans ce cas, le processus conjoint $\{(X_n,\theta_n)\}_{n \geqslant 0}$ est une chaîne de Markov non-homogène dans le temps puisque X_n sera $\sigma(X_n,\theta_n)$-mesurable.

On dira que l'adaptation est **indépendante** si, pour tout n, θ_n est indépendante de X_n. Évidemment, on peut voir les algorithmes MCMC réguliers comme des algorithmes à adaptation indépendante avec $\theta_n \equiv \theta$ pour tout n. L'exemple 3.1 montre que l'on peut réinterpréter l'échantillonneur de Gibbs, ainsi que sa version à balayage aléatoire, comme des algorithmes à adaptation indépendante.

Définition 3.1 (Adaptation indépendante, Holden et collab., 2009, section 2) *Un algorithme MH adaptatif est dit **indépendant** si la densité instrumentale $Q_n(\cdot|x)$ est indépendante de l'état actuel de la chaîne $x \in \mathcal{X}$ pour tout $n \geqslant 1$.*

Exemple 3.1 Réinterprétation de l'échantilloneur de Gibbs comme algorithme à adaptation indépendante

Soit $\Theta = \{1, \ldots, d\}$ et les transitions associées

$$P_j(y|x) = \pi(y_j|x_{-j})\, \mathbb{1}\{y_{-j} = x_{-j}\}, \qquad j = 1, \ldots, d,$$

où $\pi(y_j|x_{-j})$ correspond à la distribution conditionnelle complète de X_j sachant X_{-j} lorsque $X \sim \pi$, c.-à-d., la mise à jour de la j-ième composante par la méthode de Gibbs.

Lorsque le balayage est cyclique, alors $\mathbb{P}(\theta_n = j) = 1$ si et seulement si $j - 1 = n \mod d$. Lorsque le balayage est aléatoire, alors $\mathbb{P}(\theta_n = j) = 1/d$. Ces deux probabilités sont évidemment indépendantes de la valeur observée de X_n pour tout n. Ainsi, l'échantillonneur de Gibbs à balayage cyclique ou aléatoire est bien un algorithme à adaptation indépendante.

On distingue deux types d'algorithmes MCMC adaptatifs (Atchadé et collab., 2011) : les algorithmes à adaptation interne et ceux à adaptation externe.

3.1.1 Adaptation interne

Les algorithmes MCMC à **adaptation interne** sont tels que la valeur de θ_n dépend de tout le passé de la chaîne conjointe $(X_{0:n}, \theta_{0:n-1})$ (on inclut X_n puisque l'adaptation est généralement faite après la génération de X_n et celle-ci est utilisée dans le calcul de θ_n.) L'algorithme 3.2 détaille la procédure générale pour ce type d'adaptation. La mise à jour de $\theta_{n+1} = \theta(x_{0:n+1}, \theta_{0:n})$ peut être faite selon n'importe quelle construction $\theta(\cdot)$, mais on cherchera évidemment à améliorer l'efficacité de l'estimation de $\pi(f)$.

Algorithme 3.2 MCMC à adaptation interne

Données Densité cible π, famille de densités de transition $\{P_\theta\}_{\theta \in \Theta}$, $\theta(\cdot)$ une fonction de mise à jour de θ.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0 et paramètre initial θ_0.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonage.* Génération de $X_{n+1} \sim P_{\theta_n}(\cdot|x_n)$;
   (b) *Adaptation.* Calcul du nouveau paramètre

$$\theta_{n+1} = \theta\left(x_{0:n+1}, \theta_{0:n}\right).$$

Sortie L'échantillon $X_{0:N}$.

La plupart des algorithmes à adaptation interne entrent dans le contexte des **approximations stochastiques**, habituellement sous la forme de récursions de Robbins-Monro. Ces concepts seront brièvements introduits ici, mais une exposition plus complète peut être trouvée dans Benveniste et collab. (1987).

L'idée est de changer notre point de vue en considérant le processus $\{\theta_n\}_{n \geqslant 0}$ comme le processus d'intérêt et l'échantillon $\{X_n\}_{n \geqslant 0}$ comme le processus auxiliaire. En effet, on verra $\{\theta_n\}_{n \geqslant 0}$ comme une approximation stochastique à la solution d'une équation $h(\theta) = 0$ (Andrieu et Robert, 2001) où

$$\begin{aligned} &h : \Theta \to \mathbb{R}, \\ &\theta \mapsto h(\theta) = \int_{\mathcal{X} \times \mathcal{X}} H(\theta, (x,y)) P_\theta(\mathrm{d}y|x) \pi(\mathrm{d}x), \end{aligned} \tag{3.2}$$

pour une certaine fonction $H : \Theta \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ sensée fournir une mesure de l'efficacité du choix de P_θ comme densité de transition de x vers y. Ainsi, la fonction $h(\theta)$ correspond donc à une mesure de l'efficacité de P_θ lorsqu'on intègre par rapport à l'ensemble des paires de points (x,y).

Tout comme l'on est forcé d'utiliser une méthode MCMC pour estimer l'intégrale $\pi(f)$, on ne pourra espérer ni calculer directement $h(\theta)$ ni résoudre analytiquement $h(\theta) = 0$. En fait, $h(\theta)$ correspond au choix à $\pi(f)$ où $f(x) = \int_{\mathcal{X}} H(\theta, (x,y)) P_\theta(\,\mathrm{d}y|x)$. On choisira alors une définition de θ_n qui constituera une approximation stochastique d'une solution θ^* satisfaisant l'équation $h(\theta^*) = 0$.

On reconnait (3.2) comme une cas particulier d'un champ moyen de type Robbins-Monro (Benveniste et collab., 1987). En effet, on identifie

$$w = (x,y) \in \mathcal{W} = \mathcal{X} \times \mathcal{X}\,, \qquad \mu_\theta(\,\mathrm{d}w) = P_\theta(\,\mathrm{d}y|x)\pi(\,\mathrm{d}x),$$

pour réécrire

$$h(\theta) = \int_{\mathcal{W}} H(\theta,w)\mu_\theta(\,\mathrm{d}w).$$

Pour résoudre $h(\theta) = 0$, on procède donc à l'approximation stochastique donnée par les récursions de Robbins-Monro suivantes :

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, w_{n+1}),$$

où γ_{n+1} est une suite décroissante telle que $\sum_{n\geqslant 1} \gamma_n = \infty$ et $W_{n+1} \sim \mu_\theta$. Dans notre cas, on obtient donc

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, (x_{n+1}, y_{n+1})),$$

où $Y_{n+1} \sim P_{\theta_n}(\cdot|x_{n+1})$, c'est-à-dire y_{n+1} est généré selon la densité de transition de l'algorithme MCMC adaptatif. Notons que P_θ correspond à un noyau de transition général ; par exemple, il peut s'agir d'un noyau Metropolis-Hastings où un candidat est proposé puis accepté selon une certaine probabilité d'acceptation. Enfin, la chaîne principale est poursuivie en posant $X_{n+1} = Y_n$. L'algorithme 3.3 résume la procédure générale des MCMC à adaptation interne.

Algorithme 3.3 MCMC à adaptation interne par Robbins-Monro

Données Densité cible π, famille de densités de transition $\{P_\theta\}_{\theta\in\Theta}$, une suite $\{\gamma_n\}_{n\geqslant 1}$ décroissante telle que $\sum_{n\geqslant 1} \gamma_n = \infty$ et $H(\theta,(x,y))$ une fonction telle que décrite en (3.2).

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0 et paramètre initial θ_0.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonage.* Génération de $X_{n+1} \sim P_{\theta_n}(\cdot|x_n)$;
   (b) *Adaptation.* Calcul du nouveau paramètre

$$\theta_{n+1} = \theta_n + \gamma_n H(\theta_n, (x_n, x_{n+1})).$$

Sortie L'échantillon $X_{0:N}$ et l'approximation stochastique θ_N comme solution de l'équation $h(\theta) = 0$.

Les deux processus $\{X_n\}_{n\geqslant 0}$ et $\{\theta_n\}_{n\geqslant 0}$ progresseront donc symbiotiquement en résolvant deux problèmes simultanément : estimer $\pi(f)$ efficacement et résoudre $h(\theta) = 0$ stochastiquement. La chaîne $\{X_n\}_{n\geqslant 0}$ permet d'améliorer l'approximation de θ^* ; cette meilleure approximation produit alors une estimation plus efficace de $\pi(f)$.

Il est possible de voir l'algorithme AM (3.1) comme un algorithme MCMC à adaptation interne utilisant les récursions de Robbins-Monro.

Exemple 3.2 Interprétation Robbins-Monro de l'algorithme AM

Soit $q(\cdot|\mu,\Sigma)$ la densité d'une loi normale multivariée de moyenne μ et de covariance Σ. La densité de transition Metropolis de type marche aléatoire pour des propositions normales de covariance Σ est donnée par

$$P_\Sigma(y|x) = \alpha(x,y)q(y|x,\Sigma) + \delta_x(y)\int_{\mathcal{X}} (1-\alpha(x,y))\, q(y|x,\Sigma)\,\mathrm{d}y,$$

où $\delta_x(\cdot)$ est la fonction de masse delta de Dirac en x. On considère $\mu_n = \overline{x}_n$, qui suit une récursion simple :

$$\overline{x}_{n+1} = \frac{1}{n+2}\sum_{i=0}^{n+1} x_i = \frac{n+1}{n+2}\overline{x}_n + \frac{1}{n+2}x_{n+1} = \overline{x}_n + \frac{1}{n+2}\left(x_{n+1} - \overline{x}_n\right)$$

Ceci permet d'écrire la récursion pour $\Sigma_n = \mathrm{Cov}(x_{0:n})$:

$$\begin{aligned}
\Sigma_{n+1} &= \frac{1}{n+2}\sum_{i=0}^{n+1}(x_i-\mu_{n+1})(x_i-\mu_{n+1})^\top \\
&= \frac{1}{n+2}\left[\sum_{i=0}^{n+1} x_i x_i^\top - (n+2)\mu_{n+1}\mu_{n+1}^\top\right] \\
&= \frac{1}{n+2}\left[x_{n+1}x_{n+1}^\top + \sum_{i=0}^{n} x_i x_i^\top - (n+1)\mu_n\mu_n^\top + (n+1)\mu_n\mu_n^\top - (n+2)\mu_{n+1}\mu_{n+1}^\top\right] \\
&= \frac{1}{n+2}\left[x_{n+1}x_{n+1}^\top + (n+1)\Sigma_n + (n+1)\mu_n\mu_n^\top - (n+2)\mu_{n+1}\mu_{n+1}^\top\right] \\
&= \Sigma_n + \frac{1}{n+2}\left[x_{n+1}x_{n+1}^\top - \left(\Sigma_n + \mu_n\mu_n^\top\right)\right] + \mu_n\mu_n^\top - \mu_{n+1}\mu_{n+1}^\top,
\end{aligned}$$

ce qui nous permet d'écrire finalement,

$$\Sigma_{n+1} + \mu_{n+1}\mu_{n+1}^\top = \Sigma_n + \mu_n\mu_n^\top + \frac{1}{n+2}\left[x_{n+1}x_{n+1}^\top - \left(\Sigma_n + \mu_n\mu_n^\top\right)\right].$$

On identifie donc les paramètres $\theta = (\theta_1,\theta_2) = (\mu,\Sigma+\mu\mu^\top)$, $\gamma_{n+1} = \frac{1}{n+2}$ et la fonction

$$H(\theta,y) = \begin{bmatrix} y - \theta_1 \\ yy^\top - \theta_2 \end{bmatrix}.$$

En posant l'espérance de $H(\theta,Y)$ nulle, on comprend que la solution à l'équation $h(\theta) = 0$ sera exactement l'espérance et le second moment non-centré de π. Ainsi, l'approximation stochastique estimera $\mathbb{E}\{Y\} = \mu_\pi$ et $\mathbb{E}\{YY^\top\} = \Sigma_\pi + \mu_\pi\mu_\pi^\top$ respectivement par $\hat{\theta}_1$ et $\hat{\theta}_2$ avec lesquels on estimera

$$\Sigma_\pi = \mathbb{E}\left\{(Y-\mu_\pi)(Y-\mu_\pi)^\top\right\} = \mathbb{E}\left\{YY^\top\right\} - \mu_\pi\mu_\pi^\top$$

par $\hat{\theta}_2 - \hat{\theta}_1\hat{\theta}_1^\top$. Cet appariement des moments est l'un des types de critère d'optimisation utilisé dans les algorithmes adaptatifs. La sous-section 3.1.4 généralise ce critère et en aborde d'autres.

3.1.2 Adaptation externe

Comparativement aux algorithmes à adaptation interne, qui utilisent le passé de la chaîne $x_{0:n+1}$ pour mettre à jour la densité de transition, les algorithmes MCMC à **adaptation externe** utilisent en plus un (des) processus auxiliaire(s) $\{Y_n\}_{n\geqslant 0}$ généré(s) en parallèle à la chaîne principale $\{X_n\}_{n\geqslant 0}$.

La mise à jour θ_{n+1} est choisie comme fonction des réalisations $y_{0:n+1}$ du processus auxiliaire (et possiblement de la chaîne principale $x_{0:n+1}$). La transition de x_n à x_{n+1} ne dépend que de θ_n, comme pour tout algorithme de la forme de l'algorithme 3.3. Le processus auxiliaire est « parallèle » à la chaîne principale au sens où Y_{n+1} est généré aléatoirement à partir de θ_n et/ou des réalisation $x_{0:n}$. L'algorithme 3.4 détaille une procédure générale.

Algorithme 3.4 MCMC à adaptation externe

Données Densité cible π, famille de densités de transition $\{P_\theta\}_{\theta\in\Theta}$, $\theta(x_{0:n+1}, y_{0:n+1}, \theta_{0:n})$ une fonction de mise à jour de θ.

Procédure

1. *Initialisation.* Valeur initiale des chaînes x_0 et y_0, paramètre initial θ_0.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonage.* Génération de $X_{n+1} \sim P_{\theta_n}(\cdot|x_n)$ et de Y_{n+1} (à définir) ;
   (b) *Adaptation.* Calcul du nouveau paramètre
   $$\theta_{n+1} = \theta\left(x_{0:n+1}, y_{0:n+1}, \theta_{0:n}\right).$$

Sortie L'échantillon $X_{0:N}$.

La définition technique est la suivante. Soit $\mathcal{F} = \{\mathcal{F}_n\}_{n\geqslant 0}$ la filtration naturelle de $\{Y_n\}_{n\geqslant 0}$, c.-à-d., $\mathcal{F}_n = \sigma(Y_{0:n})$. Alors, un algorithme à **adaptation externe** est tel que le processus $\{\theta_n\}_{n\geqslant 0}$ soit adapté à la filtration $\mathcal{F}$, de sorte que θ_n soit une fonction des réalisations $y_{0:n}$. De plus, on requiert également la condition (presque markovienne) des algorithmes adaptatifs généraux (3.1). Dans le cas de l'adaptation externe, cette condition se traduit par

$$\mathbb{E}\left\{f(X_{n+1})|X_{0:n}, Y_{0:n}, \theta_{0:n}\right\} = \int_{\mathcal{X}} f(y) P_{\theta_n}(\,\mathrm{d}y|x_n) =: P_{\theta_n} f(x_n), \tag{3.3}$$

pour toute fonction $f : \mathcal{X} \to \mathbb{R}$ bornée.

Ce type d'algorithme offre beaucoup de liberté : d'abord dans le choix de l'espace des densités de transition indexé par Θ et dans le choix de la mise à jour de θ, comme pour l'adaptation interne, mais ensuite dans le choix du processus auxiliaire $\{Y_n\}_{n\geqslant 0}$.

Un cas particulier mais général de l'algorithme 3.4 est l'algorithme INCA (pour *Inter-chain Adaptation*) de Craiu et collab. (2009) où K chaînes sont produites de la façon suivante. Pour chaque $k = 1, \ldots, K$, on génère $X_{n+1}^{[k]} \sim P_{\theta_n}(\cdot|x_n^{[k]})$ où θ_n est le même entre les chaînes. Ensuite, le paramètre θ_n, commun à toutes les chaînes, est mis à jour en utilisant tous les nouveaux échantillons $x_{n+1}^{[1:K]}$ selon une manière à déterminer $\theta_{n+1} = \theta(x_{n+1}^{[1:K]}, \theta_n)$. Si P_θ correspond à une marche aléatoire M.-H., alors les chaînes parallèles peuvent explorer simultanément plusieurs région du support de π.

Comme second exemple, Holden et collab. (2009) proposent d'adapter la proposition d'un algorithme Metropolis-Hastings indépendant (définition 2.25) de la façon suivante. Une chaîne auxiliaire $\tilde{y}_n$ est maintenue en lui ajoutant les point rejetés dans l'étape d'acception/rejet Metropolis-Hastings. Cette chaîne est ensuite utilisée pour mettre à jour la densité de proposition. La procédure exacte se trouve à l'algorithme 3.5. En pratique, le choix de la proposition et celui de la mise à jour en fonction de la chaîne auxiliaire doivent être spécifiés, voir Holden et collab. (2009) pour quelques exemples.

Algorithme 3.5 Metropolis-Hastings indépendant adaptatif (AIMH) à chaîne auxiliaire (Holden et collab., 2009)

Données Densité cible π à support dans $\mathbb{R}^d$, $q_n(\cdot|x,\tilde{y})$ une densité de proposition pour tout n.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0 et de la chaîne auxiliaire $\tilde{y}_0 = \emptyset$.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonage Metropolis-Hastings.*
      i. Proposition $Y \sim q_n(\cdot|x_n, \tilde{y}_n)$;
      ii. Calcul de la probabilité d'acceptation
      $$\alpha(x_n, y) = \min\left\{1, \frac{\pi(y) q_n(y|x_n, \tilde{y}_n)}{\pi(x_n) q_n(x_n|y, \tilde{y}_n)}\right\};$$
      iii. Avec probabilité $\alpha(x_n, y)$, acceptation de la proposition $(x_{n+1} = y)$; sinon rejet $(x_{n+1} = x_n.)$
   (b) *Adaptation.* Mises à jour de la chaîne auxiliaire :
   — Si q_n est indépendante de x_n (c.-à-d., $q_n(\cdot|x,\tilde{y}) = q_n(\cdot|\tilde{y})$),
      — Si la proposition fut acceptée, poser $\tilde{y}_{n+1} = \tilde{y}_n \cup x_n$;
      — Sinon, poser $\tilde{y}_{n+1} = \tilde{y}_n \cup y$.
   — Sinon, poser $\tilde{y}_{n+1} = \tilde{y}_n$.

Sortie L'échantillon $x_{1:N}$.

3.1.3 Stratégies d'adaptation

On parcourt ici diverses stratégies souvent utilisées pour construire des algorithmes adaptatifs.

3.1.3.1 Temps d'adaptation

Jusqu'ici, les algorithmes adaptatifs considérés affichaient tous un point en commun : l'adaptation est effectuée à chaque nouvel état de la chaîne. D'autres méthodes sont envisageables et parfois préférables.

D'abord, il est possible d'arrêter l'adaptation après une certaine période de B itérations. Cette stratégie, appelée **adaptation finie** a initialement été énoncée par Pasarica et Gelman (2010). En poursuivant ensuite la chaîne avec la densité de transition obtenue à la dernière étape d'adaptation, on obtient un algorithme MCMC régulier et la théorie standard s'applique. Le travail effectué dans les B premières itérations est donc seulement pour trouver une densité de transition relativement optimale, évitant ainsi la mise au point manuelle nécessaire aux méthodes MCMC non-adaptatives. Cette technique est particulièrement intéressante pour des algorithmes adaptatifs dans lesquels l'adaptation tend à converger avec N. Dans ce cas, il n'y aura pas une grande différence entre la densité de transition fixée au temps B et celles obtenues plus tard. L'algorithme sera alors presqu'optimal et le travail computationnel supplémentaire relié à l'adaptation est évité pour le reste de la simulation. À la sous-section 3.2.3, il sera question de la convergence de l'adaptation dans le contexte des algorithmes par approximation stochastique.

D'autre part, le rythme d'adaptation peut être modifié. Plutôt que d'adapter à chaque itération,

il est possible d'adapter seulement à certaines itérations fixées à l'avance ou selon un rythme aléatoire. Chimisov et collab. (2018) produisent un cadre théorique général pour ce type d'algorithme, qu'ils nomment AirMCMC (pour *Adapted Increasingly Rarely MCMC*). On considère une suite croissante de temps entre deux adaptations $\{n_j\}_{j\geqslant 0}$ telle que $n_k \to \infty$; l'adaptation est donc effectuée à chaque $N_j = \sum_{i=0}^{j} n_j$, $j \geqslant 0$ (en posant $N_0 = n_0 = 0$). Dans le contexte des récursions de Robbins-Monro, la mise à jour de θ pourrait prendre la forme (Atchadé et collab., 2011)

$$\theta_{j+1} = \theta_j + \frac{1}{n_{j+1}} \sum_{n=N_j+1}^{N_{j+1}} H(\theta_j, (x_n, x_{n+1})) \qquad j \geqslant 0.$$

Algorithme 3.6 MCMC adapté de plus en plus rarement *Adapted Increasingly Rarely MCMC* (AirMCMC, Chimisov et collab., 2018)

Données Densité cible π, famille de densités de transition $\{P_\theta\}_{\theta\in\Theta}$, $\theta(\cdot)$ une fonction de mise à jour de θ et $\{N_j\}_{j\geqslant 0}$ la suite de temps d'adaptation.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0 et paramètre initial θ_0.
2. Pour $j = 0, \ldots,$
   (a) *Échantillonage.* Pour $n = N_j + 1, \ldots, N_{j+1}$,
      i. Génération de $X_n \sim P_{\theta_j}(\cdot|x_{n-1})$;
      ii. Si $n = N$, arrêter l'algorithme.
   (b) *Adaptation.* Calcul du nouveau paramètre
   $$\theta_{j+1} = \theta\left(x_{0:n}, \theta_{0:j}\right).$$

Sortie L'échantillon $x_{0:N}$.

Enfin, une troisième méthode d'adaptation considérant des périodes variables d'adaptation est due à Gilks et collab. (1998). Plutôt que d'adapter de moins en moins souvent, l'adaptation est effectuée seulement lorsque la chaîne se trouve dans des états choisis de sorte à séparer la chaînes en sections indépendantes. Cette indépendance aidera grandement au traitement théorique de l'algorithme puisqu'on y retrouve en quelque sorte la propriété markovienne perdue en intégrant l'adaptation. Concrètement, on doit supposer l'existence d'un ensemble $A \in \mathcal{B}(\mathcal{X})$ tel que $\pi(A) > 0$ et qui satisfait la condition suivante : X_{n+1},$X_{n+2}, \ldots$ est conditionnellement indépendant de $X_{0:n}$ sachant $X_n \in A$. Dans ce cas, l'ensemble S est appelé un **atome propre** de la chaîne de Markov et, chaque n tel que $X_n \in A$ est appelé un **temps de régénération** de la chaîne. Les temps de régénération séparent donc la chaîne en sections, appelées **tours**, qui sont alors indépendantes. À l'intérieur d'un tour, la transition est gardée fixe ; lorsque la chaîne passe par A, alors la transition est adaptée. Avec cette construction en tête, il est possible de définir des algorithmes tels qu'un atome propre existe et que l'adaptation ait lieu relativement souvent : voir Gilks et collab. (1998), Brockwell et Kadane (2005) et Sahu et Zhigljavsky (2003) pour divers exemples.

3.1.3.2 Estimation non-paramétrique

L'un des premier algorithme MCMC adaptatifs, proposé par Gilks et collab. (1994), est un exemple d'adaptation interne non-paramétrique. L'algorithme ADS (pour *Adaptive Direction Sampling*) et sa version simplifiée, l'algorithme *Snooker*, consistent en un échantillonneur de Gibbs (algorithme 2.6)

dans lequel la direction d'échantillonnage est déterminée par un passé fini de la chaîne. Dans le cas de l'algorithme *Snooker*, afin d'échantillonner x_{n+1}, deux points sont choisis parmi le passé de la chaîne $x_{0:n}$. Puis, la droite passant entre ces deux points détermine le domaine d'échantillonnage de la prochaine valeur, qui est généré selon la densité conditionnelle de la droite. L'Algorithme 3.7 détaille la procédure.

Algorithme 3.7 *Snooker* (Gilks et collab., 1994)

Données Densité cible π à support dans $\mathbb{R}^d$, m le nombre de points à conserver.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0 et l'ensemble courant $\theta_0 = \{x_0^{(1)}, \dots, x_0^{(m)}\} \in \mathbb{R}^{d \times m}$.
2. Pour $n = 0, \dots, N-1$,
   (a) *Échantillonage.*
      i. Sélectionner $x_n^{(c)}$ et $x_n^{(a)}$ uniformément parmi $\theta_n = \{x_n^{(1)}, \dots, x_n^{(m)}\}$ et calculer la direction $e_{n+1} = x_n^{(a)} - x_n^{(c)}$;
      ii. Générer $r_{n+1} \sim f(\cdot)$, où*
      $$f(r) \propto \pi\left(x_n^{(c)} + r e_{n+1}\right) |1 - r|^{d-1}$$
      est la densité conditionnelle complète sur la droite $\{x_n^{(c)} + r e_{n+1} | r \in \mathbb{R}\}$;
      iii. Poser $x_{n+1} = x_{n+1}^{(c)} = x_n^{(c)} + r_{n+1} e_{n+1}$ et $x_{n+1}^{(i)} = x_n^{(i)}$, $i \neq c$.

Sortie L'échantillon $x_{1:N}$.

* Le terme $|1-r|^{d-1}$ correspond au Jacobien de la transformation induite dans le calcul de $x_{n+1}^{(c)} = x_n^{(c)} + r e_{n+1} = (1-r) x_n^{(c)} + r x_n^{(a)}$, ajusté de sorte à obtenir la bonne densité invariante de la chaîne conjointe $\{x_n^{(1:m)}\}_{n \geqslant 0}$ (Roberts et Gilks, 1994).

L'adaptation externe est particulièrement propice à l'utilisation de densités de transition non-paramétriques où la chaîne auxiliaire permet l'estimation non-paramétrique de la densité. Par exemple, Chauveau et Vandekerkhove (2002) proposent d'utiliser l'histogramme de plusieurs chaînes générées en parallèle pour produire la densité de transition. Similairement, Warnes (2000), par l'algorithme NKC (pour *Normal Kernel Coupler*), propose d'utiliser une densité estimée par noyaux gaussiens à partir de l'état précédent de chaînes parallèles ; l'algorithme 3.8 détaille la procédure. Des exemples de constructions similaires peuvent être trouvées dans Ter Braak (2006) et dans Tran et collab. (2016).

3.1.3.3 Densité invariante non-cible et chaînes parallèles

Dans les algorithmes à adaptation externe, les chaînes auxiliaires ne doivent pas nécessairement admettre π comme distribution invariante. En fait, il peut être judicieux de définir des chaînes auxiliaires de sorte que leurs distributions invariantes respectives soient différentes de π.

Lorsque le support de π présente une géométrie complexe (multiples modes, formes particulières, etc.), les algorithmes MCMC réguliers ou même adaptatifs peuvent éprouver de la difficulté à visiter tout le support de π, et ce, rapidement et efficacement. Le **tempérage en parallèle** (Geyer, 1991; Marinari et Parisi, 1992; Geyer et Thompson, 1995) tente de résoudre ce problème en considérant des chaînes parallèles à la chaîne principale qui échantillonnent des distribution semblables à π, mais plus simples. Lorsque la distribution invariante est une version « adoucie » de π, ces chaînes parcourront

Algorithme 3.8 *Normal Kernel Coupler* (NKC, Warnes, 2000)

Données Densité cible π à support dans $\mathbb{R}^d$, K le nombre de chaînes, V la matrice de covariance du noyau gaussien et h^2 l'échelle du noyau.

Procédure

1. *Initialisation.* Valeurs initiales de la chaîne $x_0^{[k]}$, $k = 1, \ldots, K$.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonage.*
      i. Choisir $k \in \{1, \ldots, K\}$ l'indice de la chaîne à mettre à jour (par cycle ou aléatoirement) ;
      ii. Échantillonner la proposition en sélectionnant une composante du mélange $U \sim \text{uniforme}\{1, \ldots, K\}$, puis en générant $Y|U = u \sim \mathcal{N}(x_n^{[u]}, h^2 V)$;
      iii. Calculer la probabilité d'acceptation M.-H.

$$\alpha\left(x_n^{[i]}, y\right) = \min\left\{1, \frac{\pi(y) q(x_n^{[i]}|y, x_n^{[-i]})}{\pi(x_n^{[i]}) q(y|x_n^{[i]}, x_n^{[-i]})}\right\};$$

où

$$q(\cdot|x^{[i]}, x^{[-i]}) = \sum_{k \neq i} \frac{1}{K} \varphi(\cdot|x^{[k]}, h^2 V) + \frac{1}{K} \varphi(\cdot|x^{[i]}, h^2 V),$$

est l'estimé non-paramétrique de π par noyaux gaussiens à partir de l'échantillon $\{x^{[i]}, x^{[-i]}\}$ et où $\varphi(\cdot|\mu, \Sigma)$ correspond à la densité d'une loi normale de moyenne μ et de covariance Σ. Avec probabilité $\alpha\left(x_n^{[i]}, y\right)$, acceptation de la proposition $(x_{n+1} = y)$; sinon rejet $(x_{n+1} = x_n.)$

   (b) *Adaptation.* Copie des chaînes non-utilisées $x_{n+1}^{[k]} = x_{n+1}^{[k]}$, $k \neq i$, et mise à jour de la chaîne k, $x_{n+1}^{[i]} = x_{n+1}$.

Sortie L'échantillon $x_{1:N}$.

plus aisément le support de leur propre distribution, ce qui renseigne du même coup sur la distribution cible. Alors, par adaptation externe, la nouvelle information peut être transmise à la chaîne principale pour améliorer la simulation.

L'algorithme *Metropolis-coupled MCMC* (MCMCMC, Geyer (1991)) est un exemple de tempérage en parallèle. On considère K chaînes parallèles, chacune admettant π_k comme distribution invariante, où

$$\pi_k(x) = \pi^{1/k}(x), \qquad k = 1, \ldots, K. \tag{3.4}$$

Lorsque $k = 1$, on retrouve $\pi_1 = \pi^1 = \pi$, ce qui constitue alors la chaîne principale de laquelle l'échantillon final sera obtenu. Lorsque $k > 1$ alors la densité π_k est une version adoucie de π et l'échantillonnage de tout l'espace est facilité dans cette chaîne. La méthode de simulation dans chaque chaîne est arbitraire ; on peut considérer une marche aléatoire M.-H. par exemple. Afin de transmettre l'information entre les chaînes, l'étape d'adaptation consiste en une permutation des états entre deux des chaînes selon une probabilité M.-H. Intuitivement, alors que la chaîne principale peut rester prise dans un des modes de π, un saut entre les modes est potentiellement possible pour une des chaînes auxiliaires. L'échange permettra alors à la chaîne principale de visiter tout l'espace. L'algorithme 3.9 explique en détail la procédure.

Évidemment, cette description est générale. Le nombre de chaînes ainsi que les transitions dans

Algorithme 3.9 *Metropolis-coupled MCMC* (MCMCMC, Geyer, 1991)

Données Densité cible π à support dans $\mathbb{R}^d$, nombre de chaînes K, densités tempérées $\pi^{[k]} = \pi^{1/k}$ et de transition $P^{[k]}$, $k = 1, \ldots, K$.

Procédure

1. *Initialisation.* Valeur initiale des chaînes $x_0^{[1:K]}$.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonnage.* Pour $k = 1, \ldots, K$, générer $X_{n+1}^{[k]} \sim P^{[k]}$ (cette étape contient possiblement une étape d'acceptation rejet lorsque $P^{[k]}$ est un noyau M.-H.)
   (b) *Adaptation.* L'échange $x_{n+1}^{[i]} \leftrightarrow x_{n+1}^{[j]}$, $i \neq j$ est proposé et accepté avec probabilité

$$\min\left\{1, \frac{\pi^{[i]}(x_{n+1}^{[j]})\pi^{[j]}(x_{n+1}^{[i]})}{\pi^{[i]}(x_{n+1}^{[i]})\pi^{[j]}(x_{n+1}^{[j]})}\right\}$$

Sortie L'échantillon $x_{1:N}^{[1]}$.

chaque chaîne sont à déterminer (et possiblement à adapter) et la définition des distributions invariantes (3.4) peut être changée. Des constructions plus précises incluent les MCMC à rééchantillonnage par importance (IR-MCMC, Atchadé, 2009; Andrieu et collab., 2011) et l'échantilloneur Equi-Énergie (EE, Kou et collab., 2006), desquels se dégage une généralisation dans Atchadé (2010), et l'algorithme PTEEM (pour *Parallel Tempering with Equi-Energy Moves*) de Baragatti et collab. (2013) qui est à mi-chemin entre le tempérage en parallèle régulier et l'échantilloneur EE. Miasojedow et collab. (2013) proposent d'adapter le choix des distributions invariantes ainsi que celui des densités de transitions (une marche aléatoire M.-H.) dans chaque chaîne par une covariance commune adaptée comme dans l'algorithme AM 3.1. Roberts et Rosenthal (2014) considèrent le choix optimal théorique des densités tempérées, alors que Woodard et collab. (2009) dérivent des conditions pour vérifier l'optimalité par rapport à des densités cibles multi-modales. Casarin et collab. (2013) proposent une variante de leur algorithme à essais multiples (AIMTM, pour *Annealed Interacting Multiple Try Metropolis*) où des chaînes parallèles à distributions invariantes tempérées sont utilisées.

La stratégie de **tempérage** peut également être effectuée d'une autre façon. Plutôt que de considérer des chaînes parallèles, la distribution invariante de la chaîne est modifiée dans le temps, d'une manière déterminée et jusqu'à atteindre la densité cible. Au début de l'algorithme, on considère une version simplifiée de π, e.g. $\pi^{1/K}$ pour un grand K. Ensuite la distribution invariante est rapprochée de la distribution cible en approchant K de 1. À travers cela, le support complet de π sera parcouru plus rapidement, ce qui permet une adaptation de la densité de transition plus efficace. Finalement, la chaîne principale est simulée avec π comme distribution invariante, mais les valeurs initiales ont été optimisées rapidement à l'aide du tempérage (l'adaptation peut se poursuivre, mais l'optimialité sera déjà près d'être atteinte.) Par exemple, Craiu et collab. (2009) augmentent leur algorithme INCA avec ce type de tempérage pour obtenir l'algorithme TINCA (pour *Tempered INCA*.)

3.1.3.4 Choix de la densité de transition

Les méthodes adaptatives requièrent une certaine famille de densités de transition $\{P_\theta\}_{\theta \in \Theta}$. On explore ici quelques choix répandus venant des MCMC non-adaptatifs auxquels l'adaptation peut être ajoutée.

L'algorithme AM (Haario et collab., 2001), basé sur une marche aléatoire a démarré le domaine des MCMC adaptatifs. Depuis, de nombreux autres algorithmes M.-H. de type marche aléatoire ont été developpés ; voir e.g. Andrieu et Thoms (2008); Atchadé et Rosenthal (2005); Bai et collab. (2011a); Chen et collab. (2016); Craiu et collab. (2009); Garthwaite et collab. (2016); Shaby et Wells (2010); Vihola (2011a, 2012). Lorsque le gradient de π est disponible et que son calcul est computationnellement rapide, l'utilisation d'apadatation au sein d'un algorithme MALA peut s'avérer pertinent (Atchadé, 2006; Marshall et Roberts, 2012; Shaby et Wells, 2010). Voir aussi Sejdinovic et collab. (2014) pour une marche aléatoire dans un espace transformé par noyaux non-paramétriques et Tak et collab. (2018) pour une marche aléatoire construite pour des distributions multimodales à l'aide d'une chaîne auxiliaire.

Pour ce qui est des algorithmes Metropolis-Hastings indépendants, l'adaptation est souvent utilisée afin de modifier la densité de proposition pour qu'elle ressemble le plus possible à la densité cible. Plusieurs constructions utilisent une mixture pour approximer la densité cible (Andrieu et Moulines, 2006; Giordani et Kohn, 2010; Luengo et Martino, 2013; Brockwell et Kadane, 2005; Pompe et collab., 2018; Keith et collab., 2008). Voir aussi Martino et collab. (2018) pour un algorithme M.-H. indépendant non-paramétrique et Holden et collab. (2009) pour un algorithme indépendant à chaînes parallèles.

En dimensions plus élevées, le coût en calcul de l'adaptation d'une matrice de covariance (e.g. algorithme AM 3.1) peut être trop élevé et des transitions (e.g. Metropolis-dans-Gibbs) peuvent être préférables (Bai, 2009a; Roberts et Rosenthal, 2009; Rosenthal, 2011; Ł atuszyński et collab., 2013; Gåsemyr, 2003; Haario et collab., 2005; Levine et Casella, 2006). De telles propositions en une dimension se prêtent bien à l'adaptation non-paramétrique (Gilks et collab., 1995; Cai et collab., 2008; Zhang et collab., 2016; Martino et collab., 2015) alors que normalement ce type d'estimation souffre de la malédiction de la dimensionnalité. Cependant, ce type de mise-à-jour peuvent ne pas bien fonctionner lorsque la densité cible est fortement corrélé.

La plupart des extensions aux méthodes MCMC peuvent également incorporer une certaine dose d'adaptation dans leur procédure pour améliorer l'efficacité et réduire la quantité d'ajustement manuel. Haario et collab. (2006), par l'algorithme DRAM (pour *Delayed Rejection Adaptive Metropolis*) proposent un algorithme Metropolis (marche aléatoire) à rejet retardé (algorithme 2.8) où les propositions sont adaptés différemment à chaque étape. À la première étape, l'adaptation est faite comme dans l'algorithme AM 3.1 ; aux étapes suivantes, l'adaptation consiste en une mise à l'échelle de la covariance. Si l'itération se rend à la seconde étape, alors la probabilité d'acceptation devait être faible (en moyenne) à la première étape ce qui invite à réduire l'échelle de la covariance, et ainsi de suite. Les algorithmes à essais multiples sont également propices à l'adaptation des propositions de chacun des essais. Le chapitre 4 portera sur les méthodes à essais multiples et le Chapitre 5 sur l'intégration de l'adaptation dans les méthodes à essais multiples.

3.1.4 Critères d'optimisation

À l'exemple 3.2, on a vu un premier type de critère utilisé en adaptation interne. L'adaptation des paramètres de la densité de proposition est construite de sorte que ses premiers et second moments soient appariés à ceux de la densité cible. On généralise cette technique, puis l'on détaille d'autres critères utilisés pour construire des algorithmes adaptatifs.

3.1.4.1 Optimisation par appariement des moments

La méthode par **appariement des moments** (ou **méthode des moments**) considère une fonction H de la forme

$$H(\theta, x) = \phi(x) - \theta,$$

pour une certaine fonction ϕ de même dimension que θ. Alors, on trouve

$$h(\theta) = \int_{\mathcal{X}} (\phi(x) - \theta)\, \pi(\,\mathrm{d}x) = \int_{\mathcal{X}} \phi(x)\pi(\,\mathrm{d}x) - \theta,$$

c'est-à-dire qu'une solution à $h(\theta^*) = 0$ sera telle que $\theta^* = \mathbb{E}\{\phi(X)\}$: l'approximation stochastique estimera donc le moment généralisé $\mathbb{E}\{\phi(X)\}$. À l'Exemple 3.2, on avait simplement $\phi(x) = (x, xx^\top)$.

3.1.4.2 Optimisation par probabilité d'acceptation forcée

Dans le cas de transitions de type Metropolis-Hastings, une seconde méthode, dite par **probabilité d'acceptation forcée**, peut être utilisée. En effet, plusieurs études théoriques ou empiriques (section 2.5) nous renseignent sur la probabilité d'acceptation Metropolis-Hastings optimale. Pour une certaine densité de proposition $q_\theta(y|x)$ dépendant d'un paramètre θ, la probabilité d'acceptation moyenne est donnée par

$$\bar{\alpha}_\theta = \mathbb{E}\{\alpha(X,Y)\} = \int_{\mathcal{X}\times\mathcal{X}} \alpha_\theta(x,y) q_\theta(\,\mathrm{d}y|x)\pi(\,\mathrm{d}x),$$

où

$$\alpha_\theta(x,y) = \min\left\{1, \frac{q_\theta(y|x)\pi(x)}{q_\theta(x|y)\pi(y)}\right\}$$

est la probabilité d'acceptation Metropolis-Hastings. Sachant que $\bar{\alpha}_*$ est optimal dans une certaine situation, on aimerait trouver un θ^* tel que $\bar{\alpha}_\theta = \bar{\alpha}_*$. Il s'agit en fait d'une solution au champ moyen

$$h(\theta) = \int_{\mathcal{X}\times\mathcal{X}} (\alpha_\theta(x,y) - \bar{\alpha}_*)\, q_\theta(\,\mathrm{d}y|x)\pi(\,\mathrm{d}x)$$

sur les variables aléatoire (X,Y) (notons ici que Y est la proposition M.-H. plutôt que le prochain état de la chaîne), le paramètre θ et la fonction $H(\theta,(x,y)) = \alpha_\theta(x,y) - \bar{\alpha}_*$. L'approximation stochastique de Robbins-Monro (algorithme 3.3) suggère donc de générer le processus $\{(X_n,Y_n)\}_{n\geqslant 1}$ par la méthode de Metropolis-Hastings puis de mettre à jour le paramètre selon la récursion

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\,(\alpha_\theta(x,y) - \bar{\alpha}_*)\,.$$

D'une manière plus générale, on peut remplacer $\alpha_\theta(x,y) - \bar{\alpha}_*$ par n'importe quelle fonction de $\alpha_\theta(x,y)$ et de $\bar{\alpha}_*$ qui s'annule lorsque ces deux valeurs sont égales (certaines conditions seront à considérer pour assurer la validité de l'adaptation, cf. sous-section 3.2.3.) Par exemple, Gilks et collab. (1998) proposent de mettre à jour le paramètre d'échelle σ de la proposition $\mathcal{N}(\mathbf{0}, \sigma^2 I_d)$ d'un algorithme Metropolis-Hatings de type marche aléatoire par la récursion

$$\log\sigma_{i+1} = \log\sigma_i + \frac{1}{n}\,(\operatorname{logit}\alpha_\theta(x,y) - \operatorname{logit}\bar{\alpha}_*)\,.$$

Les indices distincts i et n découlent de l'adaptation faite à des temps de régénération (section 3.1.3.1). Shaby et Wells (2010) estiment plutôt $\alpha_\theta(x,y)$ par la proportion d'acceptation au cours des dernières itérations en adaptant à certains intervalles.

3.1.4.3 Optimisation directe de la variance de l'estimation

Une troisième méthode pour définir le critère d'optimalité d'un algorithme d'adaptation serait d'optimiser directement la variance asymptotique de l'estimateur de $\pi(f)$. En effet, en estimant cette espérance par son estimateur Monte Carlo $\hat{\pi}_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$, où les échantillons sont générés avec la densité de transition P_θ fixée, et en supposant qu'un théorème de limite centrale tel que

$$\sqrt{N}\left(\hat{\pi}_N(f) - \pi(f)\right) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}_d\left(0, \Sigma_\theta(f)\right), \qquad N \to \infty,$$

est valide, alors l'efficacité de l'estimation est directement liée à $\Sigma_\theta(f)$. Afin d'obtenir une estimation plus efficace de $\pi(f)$, on cherchera donc à minimiser la variance asymptotique $\Sigma_\theta(f)$ par rapport au choix de $\theta \in \Theta$. Dans le cas où f est multidimensionnelle, $\Sigma_\theta(f)$ est une matrice et on cherchera plutôt à minimiser un certain critère tel que la norme de Frobenius ou bien la trace de la matrice

Dans la dérivation du temps d'autocorrélation intégré (section 2.4.3.4), on a montré que la variance asymptotique est donnée par

$$\Sigma_\theta(f) = \mathrm{Var}\left(f(X_0)\right) + 2\sum_{i=1}^{N} \mathrm{Cov}_\theta\left(f(X_0), f(X_i)\right). \tag{3.5}$$

Puisque l'on doit avoir recours à une méthode MCMC pour estimer $\pi(f)$, qui est nécessaire au calcul des variances et covariances impliquées, on se doute que le calcul de la covariance asymptotique ne sera pas possible en général ; l'optimisation analytique de $\Sigma_f(\theta)$ n'est généralement pas envisageable. Comme pour la méthode par probabilité d'acceptation forcée, on pourrait effectuer une approximation stochastique afin de trouver θ^* qui minimise la norme de Frobenius de la variance asymptotique par exemple. Cependant, puisque $\Sigma_f(\theta)$ dépend fortement de f, on ne pourrait se fier à des résultats généraux pour trouver une valeur de $||\Sigma_\theta(f)||_F$ optimale pour f, contrairement à la méthode par probabilité d'acceptation forcée.

Afin d'adapter directement la covariance sous ce critère, Andrieu et Robert (2001) suggèrent de considérer le gradient de la norme de Frobenius, $\nabla_\theta ||\Sigma_f(\theta)||_F^2$, qui devrait s'annuler au minumum atteint en θ^*. Ceci définit un champ moyen dont une solution peut être obtenu itérativement via une mise à jour du paramètre

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\, \nabla_\theta ||\Sigma_f(\theta)||_F^2\big|_{\theta=\theta_n}.$$

Le gradient n'étant généralement pas simple à obtenir, les auteurs suggèrent de l'estimer à l'aide de chaînes générées en parallèles, constituant un algorithme à adaptation externe. Cette solution est computationnellement coûteuse et dépend à nouveau du choix de f.

Néanmoins, il est possible d'utiliser des versions simplifiées du critère de la variance asymptotique menant à une adaptation possiblement sous-optimale, mais plus simple à réaliser. Par exemple, le saut quadratique moyen se calcule beaucoup plus rapidement que la variance asymptotique (ou que l'ACT). Cette mesure est similaire à l'ACT d'ordre 1 et est uniforme par rapport au choix de f (section 2.4.3.3.) Pasarica et Gelman (2010) proposent d'utiliser ce critère dans un algorithme M.-H. de type marche aléatoire où le paramètre d'échelle λ des propositions gaussiennes de covariance $\lambda\Sigma$ est mis à jour en optimisant le ESJD. La covariance est quant à elle mise à jour d'une manière similaire à celle de l'algorithme AM 3.1.

3.1.4.4 Optimisation de la divergence

Dans le cas de densités de transition indépendantes, par exemple via un algorithme M.-H. indépendant, l'efficacité de l'estimation sera directement liée à l'adéquation entre la densité de transition et la densité cible. Il serait donc pertinent d'utiliser une méthode adaptative qui fera en sorte que la densité de transition $P_\theta(\cdot)$ épouse le mieux possible la densité cible π, résultant en une meilleure efficacité d'estimation.

Un critère général pour comparer deux densités est la **divergence de Kullback-Leibler** définie selon

$$\mathrm{KL}(\pi,P_\theta) = \mathbb{E}\left\{\log\frac{\pi(X)}{P_\theta(X)}\right\}.$$

Minimiser la divergence de Kullback-Leibler peut être faite en cherchant une solution à $\nabla_\theta \mathrm{KL}(\pi,P_\theta) = 0$. Dans le cas où l'on peut intervertir espérance et dérivée, on trouve

$$0 = \nabla_\theta \mathbb{E}\left\{\log\frac{\pi(X)}{P_\theta(X)}\right\} = \mathbb{E}\left\{\nabla_\theta \log\frac{\pi(X)}{P_\theta(X)}\right\} =: \mathbb{E}\left\{H(\theta,X)\right\}$$

On reconnait alors le champ moyen $h(\theta) = \mathbb{E}\left\{H(\theta,X)\right\}$ duquel une solution optimale peut être trouvée à l'aide d'une approximation stochastique de Robbins-Monro.

Par exemple, Andrieu et Atchadé (2007) considèrent un mélange (généralisé) de densités d'une famille exponentielle pour approximer la densité cible. La mise à jour des paramètres des densités est faite par un algorithme EM en ligne (Cappé et Moulines, 2009) qui peut être réinterprété comme une approximation stochastique de Robbins-Monro. Dans Andrieu et Thoms (2008)(voir aussi Bai et collab., 2011a), un mélange fini de densités gaussiennes ou Student est construit dans ce contexte d'algorithme EM en ligne ; l'exemple 3.3 détaille la récursion. Alternativement, Giordani et Kohn (2010) procèdent à l'optimisation en utilisant plutôt des moyennes harmoniques pour estimer le mélange.

Exemple 3.3 Algorithme Metropolis-Hastings indépendant – Mélange de densités gaussiennes adapté par un algorithme EM en ligne (Andrieu et Thoms, 2008)

Soit d la dimension de $\mathcal{X}$, K le nombre de composantes du mélange et

$$\varphi(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}(\det\Sigma)^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right\}, \tag{3.6}$$

la densité gaussienne de moyenne μ et de covariance Σ. On réécrit (3.6) pour obtenir une paramétrisation de la famille exponentielle :

$$\log\varphi(x|\mu,\Sigma) = -\frac{1}{2}\left[d\log(2\pi) + \log(\det\Sigma) + x^\top\Sigma^{-1}x - 2\mu^\top\Sigma^{-1}x + \mu^\top\Sigma^{-1}\mu\right].$$

Puis, on écrit $x^\top\Sigma^{-1}x - 2\mu^\top\Sigma^{-1}x$ sous la forme d'un produit scalaire entre un vecteur de fonctions des paramètres (μ,Σ) et un vecteur de statistiques $S(x)$. Si $\Sigma^{-1} = [s_{ij}]_{i,j=1}^d$, on a

$$\begin{aligned} x^\top\Sigma^{-1}x &= \sum_{i,j=1}^d s_{ij}x_ix_j \\ &= (s_{11},\ldots,s_{1d},\ldots,s_{d1},\ldots,s_{dd})\,(x_1x_1,\ldots,x_1x_d,\ldots,x_dx_1,\ldots,x_dx_d)^\top \\ &=: \mathrm{vec}(\Sigma^{-1})^\top\mathrm{vec}(xx^\top) \end{aligned}$$

où vec$(\cdot)$ représente la vectorisation (par ligne) d'une matrice. Donc,

$$\log \varphi(x|\mu,\Sigma) = \log h(x) + \eta(\mu,\Sigma)^\top S(x) + \psi(\mu,\Sigma)$$

où

$$\begin{aligned} h(x) &= (2\pi)^{-d/2}, \\ \psi(\theta) &= -\frac{1}{2}\left(\log(\det\Sigma) + \mu^\top\Sigma^{-1}\mu\right), \\ \eta(\mu,\Sigma)^\top S(x) &= \begin{pmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\mathrm{vec}(\Sigma^{-1}) \end{pmatrix}^\top \begin{pmatrix} x \\ \mathrm{vec}(xx^\top) \end{pmatrix}. \end{aligned}$$

On considère alors la proposition M.-H. indépendante donnée par un mélange de ces densités,

$$q_\theta(x) = \sum_{k=1}^K w^{[k]}\varphi(x|\mu^{[k]},\Sigma^{[k]}), \qquad \theta = \{w^{[k]},\mu^{[k]},\Sigma^{[k]}\}_{k=1}^K. \tag{3.7}$$

Dans le contexte de l'algorithme EM, on considère une variable aléatoire Z correspondant à la composante du mélange de laquelle provient l'observation x, telle que $\mathbb{P}_\theta(Z=k) = w^{[k]}$. On obtient donc la vraisemblance complète

$$q_\theta(x,z) = \prod_{k=1}^K \left[w^{[k]}\varphi(x|\mu^{[k]},\Sigma^{[k]})\right]^{\delta_k(z)}, \tag{3.8}$$

où $\delta_k(\cdot)$ représente la fonction de masse delta de Dirac en k. Soit $\theta = \{(w^{[k]},\mu^{[k]},\Sigma^{[k]})\}_{k=1}^K$ le vecteur de paramètres. Contrairement à la densité du mélange (3.7), la vraisemblance complète (3.8) forme une famille exponentielle :

$$\begin{aligned} \log q_\theta(x,z) &= \sum_{k=1}^K \delta_k(z)\log(w^{[k]}) + \delta_k(z)\log\varphi(x|\mu^{[k]},\Sigma^{[k]}) \\ &= \sum_{k=1}^K \delta_k(z)\log(w^{[k]}) + \delta_k(z)\left[\log h(x) + \eta(\mu^{[k]},\Sigma^{[k]})^\top S(x) + \psi(\mu^{[k]},\Sigma^{[k]})\right] \\ &= \sum_{k=1}^K \delta_k(z)\log h(x) + \sum_{k=1}^K \begin{pmatrix} \log(w^{[k]}) + \psi(\mu^{[k]},\Sigma^{[k]}) \\ \eta(\mu^{[k]},\Sigma^{[k]}) \end{pmatrix}^\top \begin{pmatrix} \delta_k(z) \\ \delta_k(z)S(x) \end{pmatrix} \\ &=: h(x,z) + \eta(\theta)^\top S(x,z). \end{aligned}$$

où $S(x,z) = (S^{[1]}(x,z)^\top,\ldots,S^{[K]}(x,z)^\top)^\top$ avec $S^{[k]}(x,z) = (\delta_k(z),\delta_k(z)S(x)^\top)^\top$ et $\eta(\theta) = (\eta^{[1]}(\theta)^\top,\ldots,\eta^{[K]}(\theta)^\top)^\top$ avec $\eta^{[k]}(\theta) = (\log(w^{[k]}) + \psi(\mu^{[k]},\Sigma^{[k]}),\eta(\mu^{[k]},\Sigma^{[k]}))^\top$. Il est donc possible de considérer l'algorithme EM en ligne de Cappé et Moulines (2009). On définit ensuite les probabilités postérieures des composantes selon

$$q_\theta(k|x) = \mathbb{P}_\theta\left(Z=k|X=x\right) = \frac{w^{[k]}\varphi(x|\mu^{[k]},\Sigma^{[k]})}{q_\theta(x)} = \frac{w^{[k]}\varphi(x|\mu^{[k]},\Sigma^{[k]})}{\sum_{j=1}^K w^{[j]}\varphi(x|\mu^{[j]},\Sigma^{[j]})}.$$

On calcule ensuite

$$\bar{s}_\theta^{[k]}(x) = \mathbb{E}_\theta\left\{S^{[k]}(X,Z)|X=x\right\} = \begin{pmatrix} q_\theta(k|x) \\ q_\theta(k|x)S(x) \end{pmatrix}, \qquad k=1,\ldots,K.$$

Ainsi, l'étape E de l'algorithme EM est faite par l'approximation stochastique de Robbins-Monro suivante :

$$\hat{s}_{n+1} = \hat{s}_n + \gamma_{n+1}\left(\bar{s}_{\hat{\theta}_n}(x_{n+1}) - \hat{s}_n\right),$$

où x_{n+1} est généré par l'algorithme Metropolis-Hastings indépendant avec la proposition $q_{\theta_n}(\cdot)$. Ceci correspond à

$$\hat{s}_{n+1}^{[k]} = \hat{s}_n^{[k]} + \gamma_{n+1}\left(q_\theta(k|x_{n+1})\begin{pmatrix} 1 \\ x_{n+1} \\ x_{n+1}x_{n+1}^\top \end{pmatrix} - \hat{s}_n^{[k]}\right), \qquad k=1,\ldots,K.$$

Ensuite, l'étape M de l'algorithme consiste à trouver $\hat{\theta}_{n+1} = \bar{\theta}(\hat{s}_{n+1})$ où

$$\bar{\theta}(s) = \arg\max_{\theta\in\Theta} \ell_\theta(s), \qquad \ell_\theta(s) = \eta(\theta)^\top s,$$

qui est séparable selon k :

$$\bar{\theta}^{[k]}(s) = \arg\max_{\theta\in\Theta} \ell_\theta^{[k]}(s), \qquad \ell_\theta^{[k]}(s) = \eta^{[k]}(\theta)^\top s^{[k]}. \tag{3.9}$$

Puis, on a

$$\begin{aligned}
\mathbb{E}_{\theta^{[k]}}\left\{S^{[k]}(X,Z)\right\} &= \mathbb{E}_{\theta^{[k]}}\left\{\mathbb{E}_{\theta^{[k]}}\left\{S^{[k]}(X,Z)\middle| Z\right\}\right\} \\
&= \mathbb{E}_{\theta^{[k]}}\left\{\begin{pmatrix} \delta_k(Z) \\ \delta_k(Z)\mu^{[k]} \\ \delta_k(Z)(\Sigma^{[k]} + \mu^{[k]}\mu^{[k]\top}) \end{pmatrix}\right\} \\
&= \begin{pmatrix} w^{[k]} \\ w^{[k]}\mu^{[k]} \\ w^{[k]}(\Sigma^{[k]} + \mu^{[k]}\mu^{[k]\top}) \end{pmatrix}
\end{aligned}$$

La solution à (3.9) sera donnée par en posant $\mathbb{E}_{\theta^{[k]}}\left\{S^{[k]}(X,Z)\right\} = \bar{s}^{[k]}(x)$. Après manipulations algébriques, on trouve la solution suivante

$$\begin{pmatrix} w_{n+1}^{[k]} \\ \mu_{n+1}^{[k]} \\ \Sigma_{n+1}^{[k]} \end{pmatrix} = \begin{pmatrix} w_n^{[k]} \\ \mu_n^{[k]} \\ \Sigma_n^{[k]} \end{pmatrix} + \gamma_{n+1} \begin{pmatrix} q_\theta(k|x_{n+1}) - w_n^{[k]} \\ q_\theta(k|x_{n+1})\left(x_{n+1} - \mu_n^{[k]}\right) \\ q_\theta(k|x_{n+1})\left((x_{n+1} - \mu_n^{[k]})(x_{n+1} - \mu_n^{[k]})^\top - \Sigma_n^{[k]}\right) \end{pmatrix}, \tag{3.10}$$

ce qui constitue les récursions pour la mise à jour des paramètres du mélange.

Cappé et Moulines (2009) montrent que cet algorithme converge, sous certaines conditions, vers un point stationnaire de la divergence de Kullback-Leibler entre la densité cible π et la densité du mélange q_θ. On voit donc bien qu'un algorithme Metropolis-Hastings indépendant utilisant une proposition constituée d'une mixture gaussienne adaptée selon les récursions (3.10) cherche à minimiser la divergence entre la proposition et la cible.

Une extension intéressante de l'utilisation de mélanges est l'**adaptation régionale**. En plus d'ajuster un mélange global sur tout l'espace $\mathcal{X}$, Bai et collab. (2011a), via l'algorithme RAPTOR, proposent de partitionner adaptivement $\mathcal{X}$ en plusieurs régions définies par la composantes dominante du mélange, puis d'utiliser la covariance de la composante dominant la région pour produire une marche aléatoire. Craiu et collab. (2009) et Roberts et Rosenthal (2009), par leurs algorithmes RAPT et RAMA, offrent des constructions similaires où les régions ne sont cependant pas adaptées.

Ce critère justifie également l'utilisation de propositions non-paramétriques. Lorsque $P_{x_{0:n}}$ est une estimation non-paramétrique de π basée sur le passé de la chaîne $x_{0:n}$ qui joue alors le rôle du paramêtre θ, l'algorithme est en fait un algorithme à adaptation interne qui optimise l'adéquation entre la densité cible et la densité de transition. La variation totale (2.13) permet également de mesurer la « distance » entre deux densités ; cette mesure (ou bien la divergence) est souvent utilisée pour les algorithmes à propositions non-paramétriques. On trouve plusieurs exemples de ce type de construction dans la littérature, cf. algorithme NPAIC de Gåsemyr (2003), algorithme ARMS de Gilks et collab. (1995) et ses déclinaisons ATRIMS et ATRAMS de Cai et collab. (2008), *Hit and Run* ARMS de Zhang et collab. (2016), AISM de Martino et collab. (2018) et IA^{2}RMS de Martino et collab. (2015).

3.1.4.5 Optimisation par critères composés

Les stratégies exposées précédemment peuvent également être combinées, formant un **critère composé**. Particulièrement, plusieurs algorithmes sont construit en adaptant les paramètres de la densité de transition différemment selon le rôle du paramètre.

L'algorithme AM 3.1 pose un paramètre d'échelle $s_d = (2.38)^2/d$ et adapte la covariance Σ par les récursions dérivées à l'exemple 3.2, correspondant au critère par appariement des moments. Cette valeur du paramètre d'échelle est optimale dans certaines situations, mais ces situations sont assez restrictives et ce choix de paramètre d'échelle n'est pas nécessairement optimal pour tout choix de densité cible π. On pourrait alors plutôt considérer s_d comme un paramètre de la densité de transition et l'adapter selon le critère de probabilité d'acceptation forcée, qui est moins sensible au choix de π. On obtient ainsi l'algorithme ASWAM (pour *Adaptive Scaling Within Adaptive Metropolis*) de Andrieu et Thoms (2008) (voir aussi Atchadé et Fort (2010)) décrit à l'algorithme 3.10.

Algorithme 3.10 Metropolis adaptatif à échelle adaptée (ASWAM, Andrieu et Thoms, 2008)

Données Densité cible π à support dans $\mathbb{R}^d$, probabilité d'acceptation cible $\bar{\alpha}_* \in (0,1)$ et $\{\gamma_n\}_{n\geqslant 1}$ décroissante telle que $\sum_{n\geqslant 1}\gamma_n = \infty$.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0, du paramètre d'échelle $\lambda_0 > 0$ et covariance initiale Σ_0.
2. Pour $n = 0,\ldots,N-1$,
   (a) *Échantillonage Metropolis.*
      i. Proposition $Y \sim \mathcal{N}_d(x_n, \lambda_n\Sigma_n)$;
      ii. Calcul de la probabilité d'acceptation
      $$\alpha(x_n, y) = \min\left\{1, \frac{\pi(y)}{\pi(x_n)}\right\};$$
      iii. Avec probabilité $\alpha(x_n, y)$, acceptation de la proposition $(x_{n+1} = y)$; sinon rejet $(x_{n+1} = x_n.)$
   (b) *Adaptation.* Mises à jour
   $$\begin{aligned}\mu_{n+1} &= \mu_n + \gamma_{n+1}(x_{n+1} - \mu_n)\\ \Sigma_{n+1} &= \Sigma_n + \gamma_{n+1}\left[(x_{n+1} - \mu_n)(x_{n+1} - \mu_n)^\top - \Sigma_n\right]\\ \log\lambda_{n+1} &= \log\lambda_n + \gamma_{n+1}(\alpha(x_n, y) - \bar{\alpha}_*)\end{aligned}$$

Sortie L'échantillon $x_{1:N}$.

D'une manière similaire, il est possible de combiner un critère d'appariement des moments et un critère de probabilité d'acceptation forcée dans un algorithme M.-H. de type marche aléatoire Gausienne sans utiliser un paramètre d'échelle. En effet, Vihola (2012), par l'algorithme RAM (pour *Robust Adaptive Metropolis*), suggère une récursion ajustant un critère combiné duquel on peut retrouver les deux critères d'une manière marginale. L'algorithme 3.11 décrit la procédure et l'exemple 3.4 explique le lien avec les critères en écrivant la mise à jour sous forme de récursion de Robbins-Monro.

Algorithme 3.11 Metropolis adaptatif robuste (RAM, Vihola, 2012)

Données Densité cible π à support dans $\mathbb{R}^d$, probabilité d'acceptation cible $\bar{\alpha}_* \in (0,1)$ et $\{\gamma_n\}_{n\geqslant 1}$ décroissante telle que $\sum_{n\geqslant 1}\gamma_n = \infty$.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0 et de la covariance $\Sigma_0 = S_0 S_0^\top$, où S_0 est diagonale inférieure.
2. Pour $n = 0,\ldots,N-1$,
   (a) *Échantillonage Metropolis.*
      i. Proposition $y = x_n + S_n u_{n+1}$, où $U_{n+1} \sim \mathcal{N}_d(\mathbf{0}, I_d)$;
      ii. Calcul de la probabilité d'acceptation
      $$\alpha(x_n, y) = \min\left\{1, \frac{\pi(y)}{\pi(x_n)}\right\};$$
      iii. Avec probabilité $\alpha(x_n, y)$, acceptation de la proposition $(x_{n+1} = y)$; sinon rejet $(x_{n+1} = x_n.)$
   (b) *Adaptation.* Trouver S_{n+1} diagonale inférieure telle que
   $$S_{n+1}S_{n+1}^\top = S_n\left(I_d + \gamma_{n+1}(\alpha(x_n, y) - \bar{\alpha}_*)\frac{u_{n+1}u_{n+1}^\top}{||u_{n+1}||_2^2}\right)S_n^\top \tag{3.11}$$

Sortie L'échantillon $x_{1:N}$.

Exemple 3.4 Étude de l'algorithme Metropolis adaptatif robuste (RAM, Vihola, 2012)

On réécrit (3.11) sous forme de récursion de Robbins-Monro :

$$\begin{aligned} S_{n+1}S_{n+1}^\top &= S_n S_n^\top + \gamma_{n+1} S_n(\alpha(x_n, x_n + S_n u_{n+1}) - \bar{\alpha}_*)\frac{u_{n+1}u_{n+1}^\top}{||u_{n+1}||_2^2} S_n^\top \\ &= S_n S_n^\top + \gamma_{n+1} H(S_{n+1}, x_n, u_{n+1}) \end{aligned}$$

où

$$H(S,x,u) = S(\alpha(x, x+Su) - \bar{\alpha}_*)\frac{uu^\top}{||u||_2^2}S^\top.$$

Ainsi, l'algorithme RAM (3.11) cherche une solution, par approximation stochastique, à $h(S) = 0$ où

$$h(S) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} H(S,x,u)\varphi(\mathrm{d}u|\,\mathbf{0}, I_d)\pi(\mathrm{d}x).$$

Vihola (2012) montre que, pour une densité cible elliptiquement symétrique, c.-à-d., $\pi(x) = \det(\Sigma_\pi)^{-1} p(||\Sigma^{-1}x||)$ pour une certaine fonction $p : \mathbb{R}^+ \to \mathbb{R}^+$, une solution S_* à $h(S) = 0$ sera telle que $S_* S_*^\top \propto \Sigma_\pi$ et que la probabilité d'acceptation moyenne est bien $\bar{\alpha}_*$. Ainsi, cet algorithme se comporte exactement comme si l'on appariait les moments et que l'on forçait la probabilité d'acceptation par un paramètre d'échelle correspondant au multiple entre $S_* S_*^\top$ et Σ_π. Cette propriété n'est cependant pas vraie pour π générale, mais cette remarque donne une intuition sur le comportement général de l'algorithme.

3.2 Ergodicité dans les algorithmes adaptatifs

On rappelle d'abord la définition de l'ergodicité d'une chaîne de Markov homogène. Une transition P est dite **ergodique pour** π pour la valeur initiale $x \in \mathcal{X}$ si P admet π comme distribution stationnaire et

$$\lim_{n\to\infty} \|P^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}} = 0, \tag{3.12}$$

où $P^n(\cdot|x)$ est la distribution marginale après n transitions P à partir de x et $\|\cdot\|_{\mathrm{TV}}$ dénote la variation totale (2.13). Ceci correspond à la convergence de la distribution marginale vers la distribution cible lorsque le nombre de pas tend vers l'infini. D'une manière plus générale, pour une fonction $V : \mathcal{X} \to [1,\infty)$, on rappelle la définition de la V-ergodicité d'une chaîne (section 2.2.2) par

$$\lim_{n\to\infty} |||P^n(\cdot|\cdot) - \pi(\cdot)|||_V = 0,$$

où $|||\cdot|||_V$ est la V-norme d'une mesure signée $\mu : \mathcal{B}(\mathcal{X}) \to \overline{\mathbb{R}} = [-\infty,\infty]$ donnée par

$$|||\mu|||_V = \sup_{\{g:\|g\|_V \leqslant 1\}} |\mu(g)| \tag{3.13}$$

et $\|g\|_V$ est la V-norme d'une fonction $g : \mathcal{X} \to \mathbb{R}^q$ donnée par

$$\|g\|_V = \sup_{x\in\mathcal{X}} \frac{\|g(x)\|_2}{V(x)}. \tag{3.14}$$

On note que l'ergodicité définie par (3.12) est équivalente à la 1-ergodicité, c.-à-d., la V-ergodicité avec $V \equiv 1$. L'ergodicité d'une chaîne de Markov, comme il en est question à la Section 2.2.2, est assurée dès lors que la transition P est apériodique, positive récurrente et qu'elle admet π comme distribution invariante.

Dans le cas d'algorithmes MCMC adaptatifs, la transition P change entre les itérations ; on devra donc généraliser l'ergodicité pour prendre en compte cette non-homogénéité. L'ergodicité de l'algorithme adaptatif dépendra donc à la fois des propriétés de la famille $\{P_\theta\}_{\theta\in\Theta}$ ainsi que de la manière dont l'adaptation est effectuée.

On pourrait être tenté de croire qu'un algorithme adaptatif soit un algorithme MCMC valide dès que chacune des densité de transition P_θ est ergodique pour π. Cependant, cette condition n'est pas suffisante pour assurer l'ergodicité de la chaîne produite. En effet, plusieurs contre-exemples montrent qu'un algorithme adaptatif sur une famille de transition, où chaque transition admet π comme distribution stationnaire, peut ne pas satisfaire la condition d'ergodicité (voir e.g. Roberts et Rosenthal (2007), exemple 1). Cette section abordera différents résultats montrant l'ergodicité d'algorithmes adaptatifs selon certains ensembles de conditions plus ou moins fortes.

Un résultat obtenu par Roberts et Rosenthal (2007, théorème 2) propose des conditions suffisantes, mais générales, pour assurer l'ergodicité d'une méthode MCMC adaptative. Il sera d'abord question de ce résultat fondamental ; on considérera ensuite la nécessité ainsi que la vérification des conditions. Puis, les méthodes d'adaptation par approximation stochastique (sous-section 3.1.1) seront soumises à ce résultat pour établir des conditions plus précises dans ce cas particulier.

Dans cette section, on supposera que la famille de transitions $\{P_\theta\}_{\theta\in\Theta}$ est telle que chaque P_θ admette π comme distribution invariante. Cette condition n'est cependant parfois pas voulue : par exemple, lorsque le tempérage (sous-section 3.1.3) est intégré à l'algorithme. Dans ce cas, des conditions supplémentaires sont requises, voir Atchadé et collab. (2011), Fort et collab. (2011), Andrieu et collab. (2011) et Atchadé et collab. (2011) pour une étude de l'ergodicité d'algorithmes MCMC adaptatifs où P_θ admet π_θ comme distribution invariante et où π_θ peut donc varier selon $\theta \in \Theta$.

3.2.0.1 Premières définitions

On débute donc par certaines définitions. Pour $n \geqslant 0$ soit X_n une variable aléatoire à valeurs dans $\mathcal{X} \subseteq \mathbb{R}^d$ et ϑ_n une variable aléatoire à valeur dans Θ. Pour éviter de surcharger la notation, on utilisera $\mathbb{P}$ et $\mathbb{E}$ pour noter respectivement la probabilité et l'espérance par rapport à la loi de toutes les variables aléatoires comprises dans l'argument. On considère la filtration naturelle induite par le processus conjoint $\{(X_n,\vartheta_n)\}_{n\geqslant 0}$, c.-à-d.,

$$\mathcal{G}_n = \sigma(X_{0:n}, \vartheta_{0:n}), \qquad n \geqslant 0.$$

Ainsi, la distribution de la transition dans le processus $\{X_n\}_{n\geqslant 0}$ satisfait

$$\mathbb{P}\left(X_{n+1} \in B | X_n = x, \vartheta_n = \theta, \mathcal{G}_n\right) = P_\theta(B|x), \qquad B \in \mathcal{B}(\mathbb{R}^d),$$

alors que la distribution de ϑ_{n+1} conditionnellement à $\mathcal{G}_n$ dépend de la méthode d'adaptation utilisée. Bien souvent, ϑ_{n+1} sera directement fonction de $X_{n+1} = x_{n+1}$ et de $\vartheta_n = \theta_n$, ce qui produira une distribution conditionnelle dégénérée. La distribution marginale de X_n conditionnelle aux valeurs initiales $X_0 = x_* \in \mathcal{X}$ et $\vartheta_0 = \theta_* \in \Theta$ est définie selon

$$A^{(n)}((x_*,\theta_*),B) = \mathbb{P}(X_n \in B | X_0 = x_*, \vartheta_0 = \theta_*), \qquad B \in \mathcal{B}(\mathbb{R}^d).$$

La variation totale entre cette distribution et celle de la densité cible est définie par

$$T(x,\theta,n) = \|A^{(n)}((x,\theta),\cdot) - \pi(\cdot)\|_{\mathrm{TV}}.$$

On dira donc qu'un algorithme adaptatif défini par $\{(X_n,\vartheta_n)\}_{n\geqslant 0}$ et les distribution précédentes est **ergodique** si

$$\lim_{n\to\infty} T(x,\theta,n) = 0, \qquad \forall x \in \mathcal{X}, \theta \in \Theta. \tag{3.15}$$

Comme pour les chaînes de Markov régulières, il est possible de généraliser cette définition à la **V-ergodicité** où l'on requiert

$$\lim_{n\to\infty} \left\| A^{(n)}((\cdot,\theta),\cdot) - \pi(\cdot) \right\|_V = 0, \qquad \forall \theta \in \Theta.$$

À nouveau, le cas $V \equiv 1$ correspond à l'ergodicité régulière définie par (3.15).

3.2.0.2 Résultat principal

Deux conditions sont centrales afin d'assurer l'ergodicité d'un algorithme adaptatif. La première condition, l'**adaptation diminuante**, requiert que le changement entre deux transitions consécutives

tende vers 0 lorsque $n \to \infty$. On écrit

$$\operatorname{dist}(\theta, \theta') = \sup_{x \in \mathcal{X}} \|P_\theta(\cdot|x) - P_{\theta'}(\cdot|x)\|_{\mathrm{TV}} = \|\!|P_\theta - P_{\theta'}|\!\|_1,$$

qui mesure la variation entre deux distributions conditionnelles d'une manière uniforme par rapport à l'état de la chaîne. Ainsi, $\operatorname{dist}(\vartheta_n,\vartheta_{n+1})$ mesure la variation, induite par l'adaptation, entre deux transitions consécutives. Alors, on définit

Condition 3.1 (Adaptation diminuante) $\lim_{n\to\infty} \operatorname{dist}(\vartheta_n,\vartheta_{n+1}) = 0$ *en probabilité, c.-à-d.,*

$$\forall \varepsilon > 0, \delta > 0, \exists N \in \mathbb{N} \text{ tel que } \forall n \geqslant N, \mathbb{P}(\operatorname{dist}(\vartheta_n,\vartheta_{n+1}) > \varepsilon) < \delta.$$

Cette condition impose une adaptation entre deux transitions successives P_{ϑ_n} et $P_{\vartheta_{n+1}}$ tendant vers 0 avec probabilité arbitrairement élevée par rapport au processus $\{\vartheta_n\}_{n\geqslant 0}$. Ceci n'impose cependant pas que ϑ_n converge, seulement que les distribution induites ne diffèrent pas significativement entre deux itérations successives ; à long terme, la somme des changements peut devenir significative. Il sera question à la sous-section 3.2.3 du cas où le processus $\{\vartheta_n\}_{n\geqslant 0}$ converge vers une valeur fixe lorsque l'algorithme utilise une approximation stochastique.

La seconde condition, la **convergence bornée**, requiert que le temps de convergence de chaque P_θ vers π (donnée par l'ergodicité de chaque transition) soit uniformément borné en probabilité. Explicitement, on écrit

$$M_\varepsilon(x,\theta) = \inf_j \left\{ j \geqslant 1 : \|P_\theta^j(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}} \leq \varepsilon \right\},$$

où $P_\theta^j(\cdot|x)$ est la distribution marginale de X_{n+j} à partir de $X_n = x$ par les transitions successives P_θ, c.-à-d., pour $A \in \mathcal{F}$,

$$P_\theta^j(A|x) = \mathbb{P}(X_{n+j} \in A | X_n = x) = \begin{cases} \int P_\theta^{j-1}(A|y) P_\theta(\,\mathrm{d}y|x) & j \geqslant 0; \\ \delta_x(A), & j = 0. \end{cases}$$

La quantité $M_\varepsilon(x,\theta)$ représente ainsi le temps de convergence au niveau ε (en variation totale) entre la distribution marginale et la distribution cible à partir d'un point x et en utilisant la transition P_θ. Alors, on définit

Condition 3.2 (Convergence bornée) *Pour tout $\varepsilon > 0$, la suite $\{M_\varepsilon(X_n,\vartheta_n)\}_{n\geqslant 0}$ est bornée en probabilité conditionnellement à $X_0 = x_*$ et $\vartheta_0 = \theta_*$, c.-à-d.,*

$$\forall \delta > 0, \exists M_* \in \mathbb{N} \text{ tel que } \forall n \in \mathbb{N}, \mathbb{P}\left(M_\varepsilon(X_n,\vartheta_n) \leq M_* | X_0 = x_*, \vartheta_0 = \theta_*\right) \geqslant 1 - \delta.$$

Cette condition impose donc que, pour un certain choix de valeurs initiales $X_0 = x_*$ et $\vartheta_0 = \theta_*$, le temps de convergence soit borné avec probabilité arbitrairement élevée par rapport au processus $\{(X_n,\vartheta_n)\}_{n\geqslant 0}$.

Intuitivement, on comprend que si la convergence bornée (condition 3.2) et l'adaptation diminuante (condition 3.1) sont satisfaites simultanément, alors d'une part on arrivera à une distribution marginale arbitrairement près de la distribution cible et d'autre part cette distribution marginale tendra à ne plus changer entre deux itérations. Ainsi, il n'est pas surprenant d'obtenir la convergence de la distribution marginale vers la distribution cible décrite dans le résultat suivant.

Théorème 3.1 (Roberts et Rosenthal, 2007, Théorème 2) *Soit $\{P_\theta\}_{\theta\in\Theta}$ une famille de transitions sur $\mathcal{X}$ telle que P_θ admette π comme distribution stationnaire pour tout $\theta \in \Theta$ et soit les*

valeurs initiales $X_0 = x_*$ et $\vartheta_0 = \theta_*$. Sous les conditions 3.2 et 3.1, $\lim_{n\to\infty} T(x_*,\theta_*,n) = 0$, c.-à-d., l'algorithme adaptatif est ergodique à π pour ce choix de valeurs initiales.

Le théorème 3.1 n'implique pas que l'algorithme adaptatif est ergodique puisque la conclusion $\lim_{n\to\infty} T(x,\theta,n) = 0$ n'est valide que pour ce choix de valeurs initiales. Ceci découle directement du fait que la condition 3.2 n'est exprimée qu'en fonction d'un choix de valeurs initiales. Pour obtenir l'ergodicité telle que définie dans (3.15), la condition 3.2 doit être vérifiée pour tout choix de valeurs initiales $(x_*,\theta_*) \in \mathcal{X} \times \Theta$. En général, la plupart des résultats qui seront abordés auront l'uniformité par rapport aux valeurs initiales (souvent implicitement) et l'on dira que l'algorithme est ergodique à ce moment.

3.2.0.3 Nécessité des conditions

Le théorème 3.1 propose des conditions suffisantes pour assurer l'ergodicité d'un algorithme adaptatif pour un choix de valeurs initiales. On considère, dans cette section, la question de savoir si ces conditions sont également nécessaires et, dans la négative, nous énoncerons des conditions sous lesquelles nécessaires à l'ergodicité.

Roberts et Rosenthal (2007, exemple 2) montrent également qu'un algorithme peut ne pas être ergodique sans adaptation diminuante (condition 3.1) mais où chaque transition est stationnaire pour la distribution cible. En assurant ensuite l'adaptation diminuante, l'algorithme devient ergodique. D'autres exemples semblables sont construits par Atchadé et Rosenthal (2005, section 2), par Yang (2008b, section 3.2), par Ł atuszyński et collab. (2013, exemple 3.1, proposition 3.2), par Atchadé et collab. (2011, section 1.3), par Rosenthal (2011, section 3.1), par Andrieu et Thoms (2008, section 2) et par Fort et collab. (2011, exemple 1.). Bien que la nécessité de l'adaptation diminuante ne soit pas prouvée, on constate en pratique qu'on ne peut s'en passer.

Yang (2009, théorème 5.2) propose un exemple montrant que la convergence bornée (condition 3.2) n'est pas nécessaire à l'ergodicité d'un algorithme adaptatif. En effet, l'algorithme considéré satisfait la condition d'adaptation diminuante et est ergodique, mais est tel que $M_\varepsilon(X_n,\vartheta_n)$ n'est pas borné en probabilité.

Bai et collab. (2011b, exemple 3.1, proposition 3.2) proposent quant à eux un exemple où l'algorithme satisfait l'adaptation diminuante, mais n'est pas ergodique. Ainsi, la convergence bornée n'est pas respectée et l'on conclut que l'adaptation diminuante n'est pas suffisante à elle seule.

3.2.1 Sur l'adaptation diminuante

Suite à la discussion de la section précédente, on comprend que l'adaptation diminuante (condition 3.1) est une condition (pratiquement) nécessaire à l'ergodicité d'un algorithme adaptatif. Intuitivement, cette condition requiert que l'adaptation entre deux densités de transition consécutives P_{ϑ_n} et $P_{\vartheta_{n+1}}$ tende vers 0.

Au moment de développer un algorithme adaptatif, il est relativement facile d'assurer la satisfactioncette condition directement par construction. Plusieurs méthodes d'adaptation de la section 3.1 satisfont cette condition.

La sous-section 3.2.3 contient une discussion de l'ergodicité des algorithmes adaptatifs par approximations stochastiques. Dans ce cas, l'index ϑ est en fait le paramètre de la distribution de transition et peut varier dans un espace métrique. Ce contexte particulier permet de dégager des conditions précises sur la manière dont l'adaptation est effectuée pour assurer l'adaptation diminuante.

On considère ici deux cas particuliers d'algorithmes adaptatifs où l'adaptation diminuante peut être facilement vérifiée.

3.2.1.1 Adaptation finie

Tel que mentionné à la sous-section 3.1.3, il est possible de considérer un algorithme adaptatif à adaptation finie. D'abord, le temps d'adaptation peut être fixé d'avance à τ itérations. Dans ce cas, dès que chaque P_θ est ergodique, alors l'algorithme sera ergodique.

Proposition 3.2 (Roberts et Rosenthal, 2007, Proposition 2) *Soit un algorithme MCMC adaptatif à adaptation finie (fixée) tel que chaque P_θ est ergodique à π. Alors, l'algorithme est également ergodique.*

D'une manière plus générale, le temps d'adaptation τ peut lui-aussi être une variable aléatoire : par exemple, l'adaptation peut être arrêtée lorsqu'un certain critère de convergence est rempli. Dans ce cas, l'adaptation diminuante (condition 3.1) sera satisfaite dès que τ est borné en probabilité.

Proposition 3.3 *Soit un algorithme MCMC adaptatif à adaptation finie (aléatoire) tel que chaque P_θ est ergodique à π et que le temps d'adaptation τ est borné en probabilité. Alors, l'algorithme est également ergodique.*

3.2.1.2 Adaptation de plus en plus rare

Deux autres manières d'assurer l'adaptation diminuante (condition 3.1) consistent à contrôler seulement la fréquence de l'adaptation au cours des itérations. À la sous-section 3.1.3, il fut question d'adapter de moins en moins souvent ou selon une probabilité décroissante vers 0. Ces deux stratégies peuvent assurer cette condition.

Chimisov et collab. (2018) développent une théorie complète des algorithmes adaptatifs lorsque l'adaptation est effectuée à des intervalles croissants. En particulier, ils montrent qu'un temps entre les adaptations randomisés peut assurer l'adaptation diminuante. Soit $\{n_k^*\}_{k\geqslant 0}$ une suite croissante et fixée d'intervalles ; on considère ensuite $n_k = n_k^* + N$, où $N \sim \text{uniforme}\{0, \ldots, \lfloor (n_k^*)^\delta \rfloor\}$ pour un certain $\delta \in (0,1)$ et où $\lfloor x \rfloor$ dénote le plus grand entier inférieur à x. Pour ce choix d'intervalles dans une version de leur algorithme AirMCMC 3.6, l'adaptation diminuante est satisfaite, et ce, peu importe comment l'adaptation est effectuée. Sans la randomisation introduite par N, les temps d'adaptation sont fixes et la méthode d'adaptation à ces moments doit être choisie pour satisfaire l'adaptation diminuante.

Il est également possible d'adapter à chaque itération selon une certaine probabilité p_n. Dans ce cas, $\text{dist}(\vartheta_n, \vartheta_{n+1}) = 0$ avec probabilité $1 - p_n$ et, donc, pour $\varepsilon > 0$,

$$\mathbb{P}\left(\text{dist}(\vartheta_n, \vartheta_{n+1}) > \varepsilon\right) \leqslant p_n.$$

En choisissant $p_n \to 0$, alors on trouvera $\lim_{n\to\infty} \text{dist}(\vartheta_n, \vartheta_{n+1}) = 0$ en probabilité, ce qui satisfait directement l'adaptation diminuante, indépendamment de la manière dont l'adaptation est effectuée. En fait, on peut voir qu'il s'agit d'une généralisation de la première stratégie. La probabilité d'effectuer

l'adaptation entre n et $n + n_k^*$ est nulle puisqu'aucune adaptation n'est effectuée avant n_k^* itérations. Entre $n + n_k^*$ et le moment où l'adaptation est effectuée, cette probabilité est de $\lfloor (n_k^*)^\delta \rfloor^{-1}$ qui tend vers 0 avec k puisque les n_k^* sont croissants.

3.2.2 Sur la convergence bornée

La convergence bornée (condition 3.2) n'est pas nécessaire à l'ergodicité des algorithmes adaptatifs. Cependant, les exemples illustrant ceci sont des cas pathologiques construits à cet effet ; en pratique, il s'agit d'une condition très générale qui est utilisée massivement dans la littérature pour justifier l'ergodicité des algorithmes adaptatifs. Cette sous-section constitue un résumé d'une discussion, présentée à l'annexe 3.4.1, sur différents moyens de vérifier la convergence bornée.

Contrairement à l'adaptation diminuante (condition 3.1), cette condition n'est ni simple à implémenter dans la construction de l'algorithme, ni simple à vérifier directement : il s'agit d'une condition dépendant des densités de transition, de la distribution cible ainsi que de la méthode d'adaptation. Ainsi, des conditions sur chacun de ces objets ainsi que sur leurs interactions sont nécessaires.

La convergence bornée correspond à un temps de convergence $M_\varepsilon(x,\theta)$ borné en probabilité par rapport à la loi conjointe de $\{(X_n,\vartheta_n)\}_{n\geqslant 1}$. Une manière d'assurer cette propriété est de contrôler le rythme de convergence (vers la distribution cible) d'une manière uniforme par rapport à $\theta \in \Theta$ et à $x \in \mathcal{X}$. C'est l'idée derrière l'ergodicité uniforme simultanée (condition 3.5.) Cette condition est cependant elle-même difficile à vérifier étant donné que les espaces $\mathcal{X}$ et Θ peuvent contenir des régions ou le rythme de convergence ralentit considérablement.

Lorsqu'une ergodicité uniforme ne peut être démontrée ou est tout simplement fausse, d'autres manières de contrôler le rythme de convergence sont également possibles. En effet, il est possible de vérifier la convergence bornée en supposant principalement que toutes les transitions P_θ ont un rythme de convergence commun qui soit géométrique (condition 3.8.) L'uniformité par rapport à $x \in \mathcal{X}$ est donc affaiblie en imposant un certain rythme de convergence par rapport à $\theta \in \Theta$. De plus, le rythme géométrique peut être affaibli à un rythme polynomial (condition 3.9.)

Ce type de conditions est décrit d'une manière générale ; lorsqu'une certaine classe d'algorithmes MCMC est considérée, il est possible de dégager des conditions beaucoup plus sensibles qui sont suffisantes. Par exemple, un algorithme Metropolis satisfera la convergence bornée en supposant principalement que la densité cible satisfait certaines conditions de régularité et que la densité instrumentale soit symétrique, positive près de l'origine et aux ailes lourdes comparativement à la densité cible.

3.2.3 Adaptation par approximations stochastiques

Lorsque l'ensemble Θ indexant la famille paramétrique correspond à l'espace des paramètres lui-même, la mise à jour de θ se prête particulièrement bien aux approximations stochastiques (cf. sous-section 3.1.1.) D'une manière générale, on considère $\Theta \subseteq \mathbb{R}^{n_\theta}$ et une mise à jour de la forme

$$X_{n+1}|\theta_n,x_n \sim P_{\theta_n}(\cdot|x_n); \tag{3.16}$$

$$\vartheta_{n+1}|\theta_n,x_{n+1} = \theta_n + \gamma_{n+1}H(\theta_n, x_{n+1}), \qquad n \geqslant 0, \tag{3.17}$$

où $\{\gamma_n\}_{n\geqslant 1}$ est une suite non-négative de pas d'adaptation telle que $\gamma_n \to 0$ et $H : \Theta \times \mathcal{X} \to \mathbb{R}^{n_\theta}$ est une fonction telle que

$$h(\theta) := \int_{\mathcal{X}} H(\theta, x)\pi(\,\mathrm{d}x)$$

corresponde à une mesure de l'efficacité de θ dans l'estimation de $\pi(f)$. On cherche ainsi à optimiser $h(\theta)$ par rapport à θ par approximation stochastique. Par exemple, l'algorithme AM 3.1 utilise le nouvel état de la chaîne afin de mettre à jour la matrice de covariance de la marche aléatoire M.-H.

La valeur passée de la chaîne x_n, en plus de la nouvelle valeur x_{n+1}, est souvent utilisée dans le calcul de H, ce qui permet de considérer une fonction H qui rend bien compte du rôle de θ comme paramètre de la transition de x_n à x_{n+1}. On obtient alors

$$\vartheta_{n+1}|\theta_n,x_n,x_{n+1} = \theta_n + \gamma_{n+1}H(\theta_n,(x_n, x_{n+1})), \qquad n \geqslant 0,$$

où

$$h(\theta) := \int_{\mathcal{X}\times\mathcal{X}} H(\theta, (x,y))P_\theta(\,\mathrm{d}y|x)\pi(\,\mathrm{d}x).$$

Similairement, dans le cas d'algorithmes Metropolis-Hastings à proposition Q_θ, la proposition $Y_{n+1} \sim Q_{\theta_n}(\cdot|x_n)$ est parfois utilisée dans le calcul de la mise à jour :

$$\vartheta_{n+1}|\theta_n,x_n,x_{n+1},y_{n+1} = \theta_n + \gamma_{n+1}H(\theta_n, (x_n,x_{n+1}, y_{n+1})), \qquad n \geqslant 0.$$

Par exemple, l'algorithme ASWAM 3.10 utilise x_n et y_{n+1} pour la mise à jour du paramètre d'échelle via la probabilité d'acceptation et x_{n+1} pour la mise à jour de la matrice de covariance. La notation utilisée dans cette sous-section suggérera principalement les mises-à-jour du type $H(\theta,x)$, mais les extensions précédentes peuvent être réécrite d'une manière triviale pour se conformer à cette notation (e.g. en écrivant $w_{n+1} = (x_n,x_{n+1}, y_{n+1})$ et $H(\theta_n,w_{n+1})$).

La relation entre les algorithmes MCMC adaptatifs et les approximations stochastiques a initialement été rapportée par Andrieu et Robert (2001). Cependant, l'étude comprise dans cet ouvrage considère principalement les propriétés du processus $\{\vartheta_n\}_{n\geqslant 0}$ et non celles du processus conjoint $\{(X_n,\vartheta_n)\}_{n\geqslant 0}$. Afin d'assurer la convergence (presque sûrement) de ϑ—ce qui ne sera pas requis dans notre cas—, ils requièrent que $H(\vartheta, X)$ soit (asymptotiquement) sans-biais pour $h(\vartheta)$ et que la suite des pas d'adaptation soit décroissante et satisfasse

$$\sum_{n\geqslant 0} \gamma_n = +\infty, \qquad \sum_{n\geqslant 1} \gamma_n^{2\eta} < +\infty,$$

pour un $\eta < 1$ (Benveniste et collab., 1987, théorème 17, utilisent plutôt $1 + \lambda$ pour un $\lambda \in (0,1)$ au lieu de 2η.) Un exemple de suite qui satisfait ces condition est $\gamma_n = n^{-\gamma}$ pour un certain $\gamma \in (1/2,1]$. La première condition permet une adaptation qui est infinie, permettant l'accès à l'ensemble de Θ et donc à une valeur optimale de θ ; la seconde condition assure quant à elle la convergence du processus. Il sera question de la convergence (ou de la non-convergence) du paramètre θ plus loin dans cette sous-section.

L'étude des propriétés de l'algorithme adaptatif, donc principalement du processus $\{X_n\}_{n\geqslant 0}$, requiert évidemment plus d'attention. La forme particulière des mises-à-jour par approximation stochastique permet une étude de l'ergodicité de l'algorithme plus approfondie ; il sera possible de dégager des conditions plus précises sur les différentes composantes de l'algorithme. En utilisant le théorème 3.1

pour montrer l'ergodicité d'un algorithme adaptatif, on rappelle que l'adaptation diminuante (condition 3.1) et la convergence bornée (condition 3.2) sont des conditions suffisantes.

L'adaptation diminuante peut être directement implémentée dans l'algorithme en imposant principalement que la suite des pas d'adaptation $\{\gamma_n\}_{n\geqslant 0}$ converge vers 0. En effet, lorsque $\gamma_n \to 0$, on aura également $\gamma_{n+1}H(\vartheta_n, X_{n+1}) \to 0$ à condition que $H(\vartheta_n, X_{n+1})$ soit borné en probabilité. Ainsi, entre deux itérations, le paramètre θ variera d'une quantité tendant vers 0. Puis, si la transition induite P_θ varie d'une manière régulière en fonction de θ (e.g. Lipschitzienne), alors la variation entre P_{θ_n} et P_{θ_n+1} tendra elle aussi vers 0. On trouve donc que $P_{\theta_n} - P_{\theta_{n+1}} \to 0$ lorsque $\theta_n - \theta_{n+1} \to 0$, ce qui peut être assuré par $\gamma_n \to 0$. On remarque que cette condition n'implique pas la convergence de $\{\theta_n\}_{n\geqslant 0}$ vers une certaine valeur θ_* ; il faudrait plutôt exiger que la somme de l'adaptation converge :

$$\sum_{n\geqslant 0}(\theta_n - \theta_{n+1}) < \infty \quad \Rightarrow \quad \theta_n \to \theta_* \in \Theta.$$

La condition de convergence bornée, comme la sous-section 3.2.2 le montre, est plus complexe à établir. La plupart des résultats exigent une sorte d'uniformité, par rapport à $\theta \in \Theta$ et à $x \in \mathcal{X}$, dans l'ergodicité de chaque transition $P_\theta(\cdot|x)$. Cette sous-section traitera de quelques stratégies pour s'attaquer à se problème.

3.2.3.1 Espace d'états et espace des paramètres compacts

La condition d'ergodicité uniforme simultanée (condition 3.5), qui est suffisante pour la convergence bornée, exige l'ergodicité de chaque P_θ d'une manière uniforme sur $\mathcal{X}\times\Theta$. D'une manière générale, la quantité en question,

$$\|P_\theta^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}} \tag{3.18}$$

sera telle qu'un rythme de convergence uniforme n'existe pas sur $\mathcal{X}\times\Theta$. En effet, lorsque θ tend vers certaines valeurs problématiques de la fermeture de Θ, le rythme de convergence peut tendre à ralentir jusqu'à perdre l'ergodicité.

Cependant, une fonction continue sur un domaine compact atteint son maximum. Donc, si (3.18) varie d'une manière continue en fonction de (x,θ) et que $\mathcal{X}\times\Theta$ est un espace produit compact, alors (3.18) atteindra son maximum sur $\mathcal{X}\times\Theta$ et on pourra ainsi borner uniformément l'ergodicité des transitions. Spécifiquement, il s'agit du corollaire 3.13 ; le corollaire 3.14 donne des conditions suffisantes en supposant entre autres que les transitions P_θ admettent des densités bornées continues par rapport à (x,θ) tandis que le corollaire 3.15 considère le cas de transitions Metropolis-Hastings. Dans le cas particulier des approximations stochastiques, θ représente le paramètre de la distribution de transition P_θ ; il adviendra souvent que P_θ variera d'une façon continue par rapport à θ et à x, ce qui permettra l'application de ces résultats.

En pratique, une certaine dualité existe par rapport à ces résultats. D'une part, le support des distributions cibles est souvent non-borné dans $\mathbb{R}^d$ rendant l'application de ce résultat techniquement impossible, mais, d'autre part, le support *effectif* de π est bien souvent quant à lui borné. Similairement, le paramètre θ admet souvent des valeurs non-bornées, mais seules certaines régions de Θ produisent des transitions P_θ pertinentes à l'algorithme adaptatif.

Par exemple, lorsque qu'une densité cible admet des ailes légères, la probabilité d'une région éloignée (e.g. de l'origine) diminue rapidement vers 0 de sorte que la chaîne de Markov ne visitera

probablement jamais une telle région. On observe donc un support virtuellement compact. Dans un algorithme adaptatif, on peut donc imposer à la chaîne de ne jamais sortir d'un sous-ensemble compact $\widetilde{\mathcal{X}} \subset \mathcal{X}$ sans que cela ne cause problème ; la chaîne n'aurait qu'une infime probabilité de sortir de cet ensemble. Théoriquement, ceci revient à produire une chaîne de Markov pour simuler de la distribution $\pi|_{\widetilde{\mathcal{X}}}$ plutôt que de π ; on doit donc espérer que la spécification de $\widetilde{\mathcal{X}}$ sera telle que l'approximation de $\pi(f) = \int_{\mathcal{X}} f(x)\pi(\,\mathrm{d}x)$ par $\pi|_{\widetilde{\mathcal{X}}}(f) = \int_{\widetilde{\mathcal{X}}} f(x)\pi(\,\mathrm{d}x)$ sera précise.

Notons qu'une chaîne contraintes dan $\widetilde{\mathcal{X}}$ ne peut plus être π-irréductible puisqu'il est impossible d'atteindre certaines régions à probabilité positive (quoique virtuellement nulle). En fait, lorsqu'on restreint l'espace d'état à un sous-espace compact $\widetilde{\mathcal{X}} \subset \mathcal{X}$, on doit tout simplement « oublier » π et $\mathcal{X}$ pour considérer $\pi|_{\widetilde{\mathcal{X}}}$ et $\widetilde{\mathcal{X}}$ comme la vraie densité cible et le vrai espace d'états. Le résultat de la simulations MCMC est donc une estimation MC de $\pi|_{\widetilde{\mathcal{X}}}(f)$ plutôt que de $\pi(f)$ et l'interprétation des résultats doit alors tenir compte de ce changement.

Une discussion similaire peut être faite sur Θ : on peut imposer θ à demeurer dans un sous-ensemble compact de Θ. On suppose alors que ce sous-ensemble contient avec très forte probabilité toutes les valeurs de θ qui pourraient être visitées par l'algorithme. À nouveau, ceci requiert une connaissance préalable des valeurs probables de θ. Contrairement à limiter x dans un sous-ensemble de $\mathcal{X}$, limiter θ à un sous-ensemble $\widetilde{\Theta} \subset \Theta$ ne posera pas de problèmes théoriques puisque ceci revient tout simplement à définir l'algorithme adaptatif sur $\widetilde{\Theta}$ plutôt que sur Θ. Cependant, l'efficacité de l'estimation de $\pi(f)$ pourrait être affectée lorsque $\widetilde{\Theta}$ est mal spécifié.

Explicitement, une fois que $\widetilde{\Theta}$ est spécifié, l'algorithme doit prévoir le cas où θ_{n+1} tomberait à l'extérieur de $\widetilde{\Theta}$. Trois options sont possibles.

Premièrement, on peut redéfinir la mise à jour selon

$$\theta^*_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1}) \tag{3.19}$$

$$\theta_{n+1} = \begin{cases} \theta^*_{n+1}, & \theta^*_{n+1} \in \widetilde{\Theta}; \\ \theta_n, & \text{sinon}, \end{cases} \tag{3.20}$$

c'est-à-dire que la mise-à-jour proposée θ^*_{n+1} est acceptée seulement lorsqu'elle est dans $\widetilde{\Theta}$, sinon le paramètre n'est pas mis à jour. Il est clair que si l'adaptation diminuante est satisfaite sur Θ, alors elle tiendra également sur $\widetilde{\Theta}$.

Deuxièmement, il est possible d'augmenter $\widetilde{\Theta}$ d'un point cimetière θ_c qui réinitialise le paramètre θ lorsque la mise-à-jour sort de $\widetilde{\Theta}$:

$$\theta_{n+1} = \begin{cases} \theta^*_{n+1}, & \theta^*_{n+1} \in \widetilde{\Theta}; \\ \theta_c, & \text{sinon}. \end{cases}$$

Dans ce cas, l'adaptation diminuante requiert que ce genre de réinitialisation survient avec probabilité tendant vers 0, puisque la différence entre P_{θ_n} et P_{θ_c} pourrait être considérable. À une itération de l'algorithme, le processus conjoint $\{(X_n,\theta_n)\}_{n\geqslant 0}$ suit donc la transition suivante

$$\begin{aligned} Q_\gamma(A \times B|x,\theta) = &\int_A P_\theta(\,\mathrm{d}y|x)\, \mathbb{1}_B(\theta + \gamma H(\theta,y)) \\ &+ \delta_{\theta_c}(B) \int_A P_\theta(\,\mathrm{d}y|x)\, \mathbb{1}_{\mathbb{R}^{n_\theta}\setminus\Theta}(\theta + \gamma H(\theta,y)) \end{aligned}$$

où $A \in \mathcal{B}(\mathcal{X})$ et $B \in \mathcal{B}(\overline{\Theta})$.

Troisièmement, il est possible d'utiliser une **reprojection** de θ^*_{n+1} sur $\widetilde{\Theta}$ en posant, par exemple, θ_{n+1} comme le point le plus près de θ^*_{n+1} parmi $\widetilde{\Theta}$, où la distance considérée est celle de l'espace métrique Θ. Les différentes composantes de θ dicteront la manière d'effectuer la reprojection. Atchadé (2006), dans un algorithme MALA adaptatif, utilise cette méthode (voir l'exemple 3.5.)

Exemple 3.5 Reprojections sur un espace de paramètres compact dans un algorithme MALA adaptatif (Atchadé, 2006)

On considère un algorithme MALA avec les propositions suivantes :

$$Y_{n+1}|(x_n,\theta_n) \sim \mathcal{N}_d\left(x_n + \frac{\sigma^2}{2}\Sigma D(x_n), \sigma^2\Sigma\right),$$

où

$$D(x) = \frac{\delta}{\max(\delta, \|\nabla \log \pi(x)\|_2)} \nabla \log \pi(x)$$

est le log-gradient de π tronqué à $\delta > 0$ évalué en x de sorte à borner la dérive de l'algorithme MALA. Les paramètres de la proposition sont μ, la moyenne, σ, le paramètre d'échelle et Σ, la covariance. Notons que le paramètre μ n'est pas utilisé dans la proposition Y_{n+1}, mais bien dans la mise à jour de la covariance. L'espace des paramètres est donc

$$\Theta = \Theta_\mu \times \Theta_\sigma \times \Theta_\Sigma,$$

où $\Theta_\mu = \mathbb{R}^d$, $\Theta_\sigma = \mathbb{R}_{>0}$ et Θ_Σ est la famille des matrices $d \times d$ définies positives. Évidemment, ces trois espaces ne sont pas bornés ; on considère donc l'espace compact $\widetilde{\Theta} = \widetilde{\Theta}_\mu \times \widetilde{\Theta}_\sigma \times \widetilde{\Theta}_\Sigma$ où

$$\begin{aligned}
\widetilde{\Theta}_\mu &= \left\{\mu \in \mathbb{R}^d : \|\mu\|_2 \leqslant A_\mu\right\}; \\
\widetilde{\Theta}_\sigma &= \left\{\sigma \in \mathbb{R}_{>0} : \varepsilon_\sigma \leqslant \sigma \leqslant A_\sigma\right\}; \\
\widetilde{\Theta}_\Sigma &= \left\{\Sigma \in \Theta_\Sigma : \|\Sigma\|_F \leqslant A_\Sigma\right\}.
\end{aligned}$$

(Il pourrait également être intéressant de borner inférieurement $\|\Sigma\|_F$ afin d'éviter la dégénérescence de Σ vers la matrice nulle.) Pour $\theta = (\mu,\sigma,\Sigma)$, la mise-à-jour consiste au calcul de

$$\theta^*_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1})$$

pour une certaine fonction H représentant la mise-à-jour dans l'Algorithme ASWAM (3.10) puis aux reprojections suivantes selon l'appartenance à $\widetilde{\Theta}$ de chacune des composantes de θ^*_{n+1} :

$$\mu_{n+1} = \begin{cases} \mu^*_{n+1}, & \|\mu^*_{n+1}\|_2 \leqslant A_\mu; \\ \frac{A_\mu \mu^*_{n+1}}{\left\|\mu^*_{n+1}\right\|_2}, & \text{sinon.} \end{cases}$$

$$\sigma_{n+1} = \begin{cases} \varepsilon_\sigma, & \sigma^*_{n+1} < \varepsilon_\sigma; \\ \sigma^*_{n+1}, & \varepsilon_\sigma \leqslant \sigma^*_{n+1} \leqslant A_\sigma; \\ A_\sigma, & \sigma^*_{n+1} > A_\sigma. \end{cases}$$

$$\Sigma_{n+1} = \begin{cases} \Sigma^*_{n+1}, & \|\Sigma^*_{n+1}\|_F \leqslant A_\Sigma; \\ \frac{A_\Sigma \Sigma^*_{n+1}}{\left\|\Sigma^*_{n+1}\right\|_F}, & \text{sinon.} \end{cases}$$

Si l'adaptation diminuante tient sur l'ensemble de Θ, alors elle sera également valide avec ses reprojections sur $\widetilde{\Theta}$ puisqu'on aura $\text{dist}(\theta_n,\theta_{n+1}) \leqslant \text{dist}(\theta_n,\theta^*_{n+1})$ par construction. En restreignant ainsi θ à un espace compact, les auteurs montrent l'ergodicité géométrique apériodique forte simultanée (condition 3.8) de l'algorithme, ce qui assure la convergence bornée.

Pour ce qui est de forcer $x \in \widetilde{\mathcal{X}}$, une solution couramment utilisée est d'échantillonner par rapport à $\pi|_{\widetilde{\mathcal{X}}}$ plutôt que par rapport à π. Par exemple, les propositions Metropolis-Hastings à l'extérieur de

$\widetilde{\mathcal{X}}$ seront systématiquement refusées : si $y \notin \widetilde{\mathcal{X}}$, alors $\pi|_{\widetilde{\mathcal{X}}}(y) = 0$ et donc la probabilité d'acceptation sera nulle.

Plusieurs chercheurs ont développé des algorithmes adaptatifs utilisant ces méthodes et rapportent que la restriction de $\mathcal{X} \times \Theta$ à un sous-ensemble compact n'est pas problématique empiriquement puisque l'algorithme ne tente jamais de sortir de la région bornée. En fait, il est possible de spécifier le sous-ensemble compact d'une manière arbitrairement large de sorte à facilement inclure le support effectif de π et les valeurs plausibles de θ. Dans l'exemple 3.5, les valeurs A_μ, A_σ et A_Σ peuvent être choisies arbitrairement grandes (ou petite pour ε_σ) sans influencer les propriétés de l'algorithme. Pour cette raison, supposer $\mathcal{X}$ et Θ compacts s'avère des conditions peu restreignante en pratique dans la mesure où on peut altérer le problème—et ce, d'une manière aussi insignifiante que souhaitée—pour se placer dans un contexte où ses suppositions sont satisfaites.

3.2.3.2 Troncation et recouvrement compact

La condition $\sum_{n \geqslant 0} \gamma_n = +\infty$ est requise afin d'assurer que l'approximation stochastique puisse se rendre à une valeur optimale de θ. Cependant, cette condition permet également au processus θ_n de diverger vers l'infini lorsque Θ est non-borné. Dans leur introduction aux approximations stochastiques pour les algorithmes MCMC adaptatifs, Andrieu et Robert (2001) proposent d'utiliser une construction due à Chen et collab. (1988), appelée **Troncation à des bornes variant aléatoirement**, où Θ est successivement restreint à un sous-ensemble compact. Depuis, cette technique à été reprise par Andrieu et Moulines (2006) où l'étude dans le cas d'algorithmes MCMC adaptatifs est considérée.

L'argument et l'algorithme est généralisé dans Andrieu et Vihola (2014) où les pas d'adaptation peuvent être aléatoires. Voir également Atchadé et Rosenthal (2005, algorithme 4.1) et Saksman et Vihola (2010, section 2) pour une application à l'algorithme AM ainsi que Vihola (2011b, section 5) pour une application à l'algorithme ASWAM.

Cette construction permet non seulement de borner θ, mais surtout de garder θ à distance de $\partial\Theta$, la frontière de l'espace paramétrique ou plus généralement un sous-ensemble de la fermeture de Θ pouvant causer problème. Alors que le cas où θ tend vers l'infini est évidemment souvent problématique, des situations où θ tend vers une extrémité (finie) de Θ peuvent également empêcher l'ergodicité : on pense par exemple à la situation où θ contient une composante de variance qui pourrait tendre vers 0. La troncation de Θ se fera d'une part pour empêcher les variances infinies et d'autre part pour empêcher les variances nulles.

On considère la situation suivante. Soit $\{P_\theta\}_{\theta \in \Theta}$ une famille de transitions de Markov sur $\mathcal{X}$, toutes π-irréductibles, et admettant π comme distribution invariante. Soit $\{H(\theta,X) : \Theta \times \mathcal{X} \to \mathbb{R}^{n_\theta}\}$ une famille de fonctions de mise à jour, où n_θ est la dimension de l'espace des paramètres $\Theta \subseteq \mathbb{R}^{n_\theta}$. L'espace des paramètres est augmenté d'un point cimetière $\theta_c \notin \Theta$ afin d'éventuellement tenir compte des cas où $\theta + \gamma H(\theta,X) \notin \Theta$ et on définit $\overline{\Theta} = \Theta \cup \{\theta_c\}$.

On définit ensuite la notion de **recouvrement compact**. Il s'agit d'une suite croissante d'ensembles compacts recouvrant Θ :

Définition 3.2 *Une famille $\{\mathcal{K}_r\}_{r \geqslant 0}$ de sous-ensembles compacts de Θ est un recouvrement compact de Θ si*

$$\bigcup_{r \geqslant 0} \mathcal{K}_r = \Theta, \qquad \mathcal{K}_r \subset \mathring{\mathcal{K}}_{r+1},$$

où $\mathring{A}$ dénote l'intérieur de l'ensemble A.

La troncation de Θ s'effectue de la manière suivante. L'algorithme est initialisé avec $\mathcal{K}_0$ comme espace paramétrique et, à chaque fois que le paramètre d'adaptation sort de $\mathcal{K}_r$, le paramètre est reprojeté sur $\mathcal{K}_0$ et l'espace paramétrique est augmenté à $\mathcal{K}_{r+1}$. L'indice du recouvrement compact est initialisé à $r_0 = 0$ et les récursions de l'approximation stochastique sont désormais

$$\begin{aligned} X_{n+1}|\theta_n,x_n &\sim P_{\theta_n}(\cdot|x_n); \\ \theta^*_{n+1} &= \theta_n + \gamma_{n+1}H(\theta_n,x_{n+1}); \\ \theta_{n+1} &= \begin{cases} \theta^*_{n+1}, & \theta^*_{n+1} \in \mathcal{K}_{r_n}, \\ \Pi_{\Theta\to\mathcal{K}_0}(\theta^*_{n+1}), & \theta^*_{n+1} \notin \mathcal{K}_{r_n}; \end{cases} \\ r_{n+1} &= \begin{cases} r_n, & \theta^*_{n+1} \in \mathcal{K}_{r_n}, \\ r_n + 1, & \theta^*_{n+1} \notin \mathcal{K}_{r_n}, \end{cases} \end{aligned}$$

où $\Pi_{\Theta\to\mathcal{K}_0}$ décrit une reprojection de $\Theta/\mathcal{K}_0$ vers $\mathcal{K}_0$. L'algorithme 3.12 détaille la procédure. Il sera parfois nécessaire de réinitialiser la chaîne $\{X_n\}_{n\geqslant 0}$ lorsque la troncation est agrandie. On considère un ensemble compact $\mathcal{X}_0 \subset \mathcal{X}$ et au moment de générer $X_{n+1}|\theta_n,x_n$ on considère plutôt

$$x^*_n = \begin{cases} x_n, & \theta^*_n \in \mathcal{K}_{r_{n-1}}, \\ \Pi_{\mathcal{X}\to\mathcal{X}_0}(x_n), & \theta^*_n \notin \mathcal{K}_{r_{n-1}}, \end{cases} \qquad X_{n+1}|\theta_n,x_n \sim P_{\theta_n}(\cdot|x^*_n).$$

L'intérêt de cette construction est de trouver un sous-ensemble compact de Θ tel que l'adaptation reste dans cet ensemble. La spécification de $\widetilde{\Theta} \subset \Theta$ à la section précédente n'est donc plus requise ; ici, la suite croissante d'ensembles compacts effectue automatiquement le choix de $\widetilde{\Theta}$ de sorte que l'algorithme n'en sorte jamais.

Afin de montrer l'ergodicité de cet algorithme, une condition principale sera que le nombre de réinitialisations, c.-à-d., le r maximal requis, soit fini, impliquant alors que θ soit restreint à $\mathcal{K}_{r_{\max}}$. Dans ce cas, l'adaptation est effectivement restreinte à un ensemble évitant les valeurs problématiques de $\partial\Theta$, ce qui assurera une ergodicité uniforme par rapport à $\theta \in \Theta$. On supposera donc pour l'instant que le nombre de réinitialisations est borné presque sûrement ; il sera question de conditions sous lesquelles cette supposition est valide à l'annexe 3.4.2.

On considère le corollaire 3.4 qui suggère des conditions suffisantes pour assurer la convergence bornée. Plutôt que d'exiger que le temps de convergence soit borné en probabilité uniformément par rapport à $\theta \in \Theta$, on requiert que θ soit borné en probabilité (ce qui découle du nombre de réinitialisations borné presque sûrement) et que sur tout sous-ensemble compact de Θ le temps de convergence soit uniformément borné.

Corollaire 3.4 (Bai, 2009b, corollaire 3.2) *Soit Θ un espace métrique tel que θ_n est borné en probabilité. Supposons que pour tout $\mathcal{K} \subset \Theta$ compact et pour tout $\varepsilon > 0$ le temps de convergence-ε sur $\mathcal{K}$ est borné en probabilité, c.-à-d.,*

$$\forall\, \delta > 0, \exists M \in \mathbb{N} \ \textit{tel que} \ \forall\, n \in \mathbb{N}, \ \mathbb{P}_{x_0,\theta_0}\left(\widetilde{M}_{\varepsilon,\mathcal{K}}(X_n) \leqslant M\right) \geqslant 1 - \delta,$$

où

$$\widetilde{M}_{\varepsilon,\mathcal{K}}(x) = \min\left\{j \geqslant 1 : \sup_{\theta\in\mathcal{K}} \left\|P^j_\theta(\cdot|x) - \pi(\cdot)\right\|_{\mathrm{TV}} < \varepsilon\right\}.$$

Alors, l'algorithme satisfait la convergence bornée (condition 3.2.)

Lorsque θ n'est pas borné, il est souvent difficile, voire impossible, d'établir des conditions

Algorithme 3.12 MCMC par approximation stochastique avec recouvrement compact, Andrieu et collab. (2005, algorithme 1)

Données Densité cible π, famille de densité de transition $\{P_\theta\}_{\theta\in\Theta}$, suite de pas d'adaptation $\{\gamma_n\}_{n\geqslant 1}$ décroissante telle que $\sum_{n\geqslant 1}\gamma_n = \infty$, $\{\mathcal{K}_q\}_{q\geqslant 0}$ un recouvrement compact de Θ, $\mathcal{X}_0$ un sous-ensemble compact de $\mathcal{X}$, $\{H(\theta,x)\}$ une famille de fonctions de mise à jour et fonctions de reprojections $\Pi_{\Theta\to\mathcal{K}_0}$ et $\Pi_{\mathcal{X}\to\mathcal{X}_0}$.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne $x_0 \in \mathcal{X}$, paramètre initial $\theta_0 \in \mathcal{K}_0$ et $r_0 = 0$ l'index de la troncation actuelle.
2. Pour $n = 0,\ldots,N-1$,
   (a) *Échantillonage.* Génération du nouvel état $X_{n+1} \sim P_{\theta_n}(\cdot|x_n^*)$;
   (b) *Adaptation.*
   — Calcul du nouveau paramètre proposé
   $$\theta_{n+1}^* = \theta_n + \gamma_{n+1}H(\theta_n, x_{n+1}).$$
   — Si $\theta_{n+1}^* \in \mathcal{K}_{r_n}$, alors
   $$\theta_{n+1} = \theta_{n+1}^*, \qquad r_{n+1} = r_n, \qquad x_{n+1}^* = x_{n+1},$$
   — Sinon,
   $$\begin{aligned}\theta_{n+1} &= \Pi_{\Theta\to\mathcal{K}_0}(\theta_{n+1}^*),\\ r_{n+1} &= r_n + 1,\\ x_{n+1}^* &= \Pi_{\mathcal{X}\to\mathcal{X}_0}(x_{n+1}).\end{aligned}$$

Sortie L'échantillon $x_{0:N}$ et l'approximation stochastique θ_N à la solution à l'équation $h(\theta) = 0$.

d'ergodicité uniforme. Lorsque θ est restreint à un ensemble compact, il est plus aisé de vérifier une ergodicité uniforme, voir e.g. l'exemple 3.7. Ainsi, la majeure partie du travail sera de montrer que le processus $\{\theta_n\}_{n\geqslant 0}$ est borné en probabilité ; le temps de convergence borné sur tout compact pourra être vérifié par une ergodicité uniforme sur tout compact.

Andrieu et Moulines (2006), étant contemporains à Roberts et Rosenthal (2007), ne montrent pas l'ergodicité de leur algorithme en passant par les conditions d'adaptation diminuante (condition 3.1) et de convergence bornée (condition 3.2.) Cependant, les conditions qu'ils supposent peuvent être réinterprétées dans ce contexte. D'abord, ils supposent une ergodicité géométrique uniforme sur tout sous-ensemble compact de Θ, condition plus faible que l'ergodicité géométrique apériodique forte simultanée (condition 3.8.)

Condition 3.3 (Ergodicité géométrique uniforme sur compacts, Andrieu et Moulines, 2006, condition A1) *Pour chaque $\theta \in \Theta$, P_θ admet π comme distribution invariante. Il existe une fonction test $V : \mathcal{X} \to [1,\infty)$ telle que $\sup_{\mathcal{X}_0} V < \infty$ et, pour tout ensemble compact $\mathcal{K} \subset \Theta$, on a les propriétés suivantes :*

(i) Minorisation. Il existe $C \in \mathcal{B}(\mathcal{X})$, $\delta > 0$ et une mesure de probabilité ν avec $\nu(C) > 0$ tels que

$$P_\theta(A|x) \geqslant \delta\nu(A), \qquad \forall\, A \in \mathcal{B}(\mathcal{X}), \theta \in \mathcal{K}, x \in C.$$

(ii) Dérive géométrique. Il existe $\lambda \in [0,1)$ et $b \in (0,\infty)$ tels que

$$P_\theta V(x) \leqslant \begin{cases}\lambda V(x), & x \notin C,\\ b, & x \in C,\end{cases} \qquad \forall\, \theta \in \mathcal{K}.$$

Conjointement avec la supposition que le processus $\{\theta_n\}_{n\geqslant 0}$ est borné en probabilité, cette

condition permet de vérifier la Convergence bornée. En effet, cette condition suffit pour vérifier que le temps de convergence-ε sur tout compact (cf. corollaire 3.4) est borné en probabilité.

Proposition 3.5 *Supposons que l'algorithme adaptatif à recouvrement compact satisfait la condition 3.3. Alors, $\{\widetilde{M}_{\varepsilon,\mathcal{K}}(X_n)\}_{n\geqslant 0}$ est borné en probabilité pour tout $\mathcal{K}\subset\Theta$ compact.*

Démonstration. Pour tout $\theta\in\mathcal{K}$, on a que P_θ satisfait la condition de minorisation et de dérive géométrique. Par la proposition 3.19, il existe $\rho<1$ et $K<\infty$ ne dépendant que de δ, λ, b et v tels que

$$\left\|P_\theta^j(\cdot|x)-\pi(\cdot)\right\|_{\mathrm{TV}}\leqslant KV(x)\rho^j.$$

Puisque le choix de δ, λ, b et v est uniforme pour tout $\theta\in\mathcal{K}$, on a alors

$$\sup_{\theta\in\mathcal{K}}\left\|P_\theta^j(\cdot|x)-\pi(\cdot)\right\|_{\mathrm{TV}}\leqslant KV(x)\rho^j.$$

On montre que $\{V(X_n)\}_{n\geqslant 0}$ est borné en probabilité. En effet, pour un $\mathcal{K}$ fixé, on a la condition de dérive géométrique uniforme

$$P_\theta V\leqslant\lambda V+b\,\mathbb{1}_C\,,\qquad\forall\,\theta\in\mathcal{K},$$

pour un certain choix de $\lambda\in(0,1)$, de $b\in[0,\infty)$ et de $C\subset\mathcal{X}$ compact. Lorsque la chaîne se trouve en (x_n,θ_n), ceci correspond à

$$\mathbb{E}_{X_{n+1}}\left\{V(X_{n+1})\middle|x_n,\theta_n\right\}\leqslant\lambda V(x_n)+b\,\mathbb{1}_C(x_n).$$

En prenant l'espérance par rapport à θ_n, on obtient

$$\mathbb{E}_{X_{n+1}}\left\{V(X_{n+1})\middle|x_n\right\}\leqslant\lambda V(x_n)+b\,\mathbb{1}_C(x_n).$$

Puis, l'espérance par rapport à X_n donne

$$\mathbb{E}_{X_{n+1}}\left\{V(X_{n+1})\right\}=\mathbb{E}_{X_n}\,\mathbb{E}_{X_{n+1}}\left\{V(X_{n+1})\middle|X_n\right\}\leqslant\lambda\,\mathbb{E}_{X_n}\,V(X_n)+b.$$

Par Roberts et Rosenthal (2007, lemme 14), on trouve

$$\sup_n\mathbb{E}\,V(X_n)\leqslant\max\left\{\mathbb{E}\,V(X_0),\frac{b}{1-\lambda}\right\}.$$

On a donc la suite de nombres réels $a_n=\mathbb{E}\,V(X_n)$ qui satisfait $a_{n+1}\leqslant\lambda a_n+b$ avec $\lambda\in(0,1)$ et $b\in(0,\infty)$. Par induction, on trouve que $a_n\leqslant\lambda^n a_0+b\sum_{i=0}^{n-1}\lambda^i$. En effet, $a_0\leqslant 1\cdot a_0+0$ et

$$a_{n+1}\leqslant\lambda(a_n)+b\leqslant\lambda\left(\lambda^n a_0+b\sum_{i=0}^{n-1}\lambda^i\right)+b=\lambda^{n+1}a_0+b\sum_{i=0}^{n}\lambda^i.$$

Ainsi, pour $\lambda\in(0,1)$ on trouve

$$a_n\leqslant\lambda^n a_0+b\frac{1-\lambda^n}{1-\lambda}\leqslant a_0+\frac{b}{1-\lambda},\qquad\forall\,n\geqslant 0.$$

On note que $\mathbb{E}\,V(X_0)<\infty$ puisque $x_0\in\mathcal{X}_0$ par construction de l'algorithme et par $\sup_{\mathcal{X}_0}V(x)<\infty$ étant donné la condition 3.3. C'est donc dire que

$$\sup_n a_n=\sup_n\mathbb{E}\,V(X_n)\leqslant\mathbb{E}\,V(X_0)+\frac{b}{1-\lambda}<\infty.$$

Ainsi, $\sup_n\mathbb{E}\,V(X_n)<\infty$. Puis, par l'inégalité de Markov, on trouve $\{V(X_n)\}_{n\geqslant 0}$ borné en probabilité :

$$\mathbb{P}_{x_0,\theta_0}\left(V(X_n)\geqslant M\right)\leqslant\frac{1}{M}\,\mathbb{E}\left\{V(X_n)\right\}\leqslant\frac{1}{M}\sup_m\mathbb{E}\left\{V(X_m)\right\}\to 0,\qquad\forall\,n\geqslant 0.$$

Donc, pour tout $\delta>0$, il existe $\widetilde{V}=\widetilde{V}(\delta)<\infty$ tel que

$$\mathbb{P}_{x_0,\theta_0}\left(V(X_n)\leqslant\widetilde{V}\right)\geqslant 1-\delta,\qquad\forall\,n\geqslant 1$$

C'est donc dire que

$$\mathbb{P}_{x_0,\theta_0}\left(\sup_{\theta\in\mathcal{K}}\left\|P_\theta^j(\cdot|X_n)-\pi(\cdot)\right\|_{\mathrm{TV}}\leqslant K\widetilde{V}\rho^j\right)\geqslant 1-\delta,\qquad\forall\,n\geqslant 1$$

Soit $\varepsilon > 0$ alors $K\widetilde{V}\rho^j < \varepsilon$ implique

$$j \log \rho < \log \varepsilon - \log K - \log \widetilde{V}.$$

Pour $\rho \in (0,1)$, on trouve

$$j > \frac{1}{\log \rho} \left(\log \varepsilon - \log K - \log \widetilde{V} \right),$$

ce qui implique alors que

$$\widetilde{M}_{\varepsilon,\mathcal{K}}(X_n) \leqslant \frac{1}{\log \rho} \left(\log \varepsilon - \log K - \log \widetilde{V} \right).$$

On peut donc borner $\{\widetilde{M}_{\varepsilon,\mathcal{K}}(X_n)\}_{n\geqslant 0}$ en probabilité :

$$\begin{aligned}
&\mathbb{P}_{x_0,\theta_0}\left(\widetilde{M}_{\varepsilon,\mathcal{K}}(X_n) \leqslant \frac{1}{\log \rho} \left(\log \varepsilon - \log K - \log \widetilde{V} \right) \right) \\
&\geqslant \mathbb{P}_{x_0,\theta_0}\left(\sup_{\theta\in\mathcal{K}} \left\| P_\theta^j(\cdot|X_n) - \pi(\cdot) \right\|_{\mathrm{TV}} \leqslant K\widetilde{V}\rho^j \right) \\
&\geqslant 1-\delta, \qquad \forall\, n \geqslant 1,
\end{aligned}$$

où $\mathcal{K}$ est arbitraire, ce qui conclut la preuve en choisissant $M = \frac{1}{\log \rho} \left(\log \varepsilon - \log K - \log \widetilde{V} \right)$ dont le choix ne dépends que de δ (à travers $\widetilde{V}$) pour ε et $\mathcal{K}$ fixés. □

Corollaire 3.6 *Supposons que l'algorithme adaptatif à recouvrement compact satisfait la condition 3.3 et que $\{\theta_n\}_{n\geqslant 0}$ est borné en probabilité, alors l'algorithme satisfait la convergence bornée (condition 3.2.)*

Démonstration. Conséquence directe du corollaire 3.4 et de la proposition 3.5. □

3.2.3.3 Sur l'efficacité de l'algorithme

Tel qu'observé à l'annexe 3.4.3, le théorème 3.31 implique la convergence de l'approximation stochastique vers un des points stationnaires du champ moyen. Lorsque le champ moyen de l'approximation stochastique est défini tel que ses points stationnaires correspondent à un algorithme MCMC optimal en terme d'efficacité, la convergence de l'algorithme vers un de ces points stationnaire est donc souhaitable (le but principal des MCMC adaptatifs est d'éviter la mise au point manuelle de transitions pour estimer $\pi(f)$ efficacement.)

Andrieu et Atchadé (2007) proposent une étude de tels algorithmes en supposant en plus la convergence du paramètre d'adaptation vers une limite fixe et unique $\theta_n \to \theta^*$. En comparant une chaîne de Markov ayant comme densité de transition P_{θ^*} à la chaîne produite par l'algorithme adaptatif, ils arrivent à produire des conditions sous lesquelles l'algorithme adaptatif montre un comportement asymptotique semblable à celui de la chaîne avec P_{θ^*}.

Informellement, le résultat est le suivant. On requiert que la famille $\{P_\theta\}_{\theta\in\Theta}$ soit géométriquement simultanément ergodique (condition 3.8) et lipschitzienne (condition 3.24), et que $\theta_n \xrightarrow{\mathcal{L}_2} \theta_*$ pour un certain $\theta_* \in \Theta$ fixe, c.-à-d. qu'il existe une suite $\{\alpha_n\}_{n\geqslant 0}$ avec $\alpha_n \to 0$ telle que

$$\sqrt{\mathbb{E}\left\{ \|\theta_n - \theta_*\|_\Theta^2 \right\}} = \mathcal{O}(\alpha_n). \tag{3.21}$$

On rappelle la notation $A^{(n)}((x,\theta),\cdot)$ qui désigne la distribution de X_n dans l'algorithme adaptatif à partir des conditions initiales $(X_0,\theta_0) = (x,\theta)$ et $P_\theta^{(n)}(\cdot|x)$ qui désigne la distribution de X_n dans l'algorithme avec transition fixe P_θ à partir de $X_0 = x$. Alors, la distribution marginale de X_n dans l'algorithme adaptatif converge en distribution vers la distribution marginale dans l'algorithme à transition fixe optimale, c.-à-d. $A^{(n)}((x,\theta),\cdot) \xrightarrow{\mathcal{D}} P_{\theta_*}^{(n)}(\cdot|x)$ avec $n \to \infty$ (Andrieu et Atchadé, 2007,

corollaire 2.1). On dit alors que $\{X_n\}_{n\geqslant 0}$ est **faiblement efficace.** Si le rythme de convergence de θ_n est sommable, c.-à-d. $\sum_{n\geqslant 0}\alpha_n < \infty$, alors (Andrieu et Atchadé, 2007, corollaire 2.2) la convergence est en variation totale, c.-à-d.,

$$\lim_{n\to\infty}\left\|A^{(n)}((x,\theta),\cdot) - P_{\theta_*}^{(n)}(\cdot|x)\right\|_{\mathrm{TV}} = 0.$$

On dit alors que $\{X_n\}_{n\geqslant 0}$ est **fortement efficace.**

Dans le contexte des algorithmes adaptatifs par approximations stochastiques, Andrieu et Atchadé (2007, théorème 3.1) montrent que le ryhtme de convergence est donné par le pas d'adaptation, c.-à-d., $\alpha_n = \gamma_n$. Ainsi, dans le cas de l'algorithme AM 3.1 que l'on sait converger vers un unique point stationnaire, on trouve $\alpha_n = n^{-1}$ et donc l'algorithme est faiblement efficace. Similairement, pour toute suite $\gamma_n = n^{-\gamma}$ où $\gamma \in (1/2,1]$, l'algorithme sera faiblement efficace. Maintenant, le point stationnaire correspond à choisir $\Sigma = \Sigma_\pi$ comme covariance de la densité de proposition gaussienne. En se référant ensuite à des résultats tels que ceux de la section 2.5, on trouve que l'algorithme est optimal en terme d'efficacité d'estimation de $\pi(f)$, ce qui justifie l'utilisation d'algorithmes adaptatifs.

La question de l'efficacité sera approfondie lorsqu'il sera question de théorèmes centraux limites (sous-section 3.3.2), où il sera possible de comparer la variance asymptotique de l'estimation par un algorithme adaptatif à la variance asymptotique d'un algorithme non-adaptatif, mais à transition idéale.

3.3 Propriétés de l'estimation dans les algorithmes adaptatifs

L'ergodicité des algorithmes adaptatifs, étudiée à la Section 3.2, ne renseigne que sur la distribution marginale de la chaîne produite par l'algorithme. Cette propriété est indépendante de la fonction $f : \mathcal{X} \to \mathbb{R}^p$ dont l'espérence sous π est recherchée. Il n'est donc pas question de l'estimation Monte Carlo de $\pi(f) = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x)$ donnée par l'intégrale par rapport à la distribution empirique $\hat{\pi}_N = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}$ définie par l'échantillon produit $x_{1:N}$, c.-à-d.

$$\hat{\pi}_N(f) = \int_{\mathcal{X}} f(x)\hat{\pi}_N(\mathrm{d}x) = \frac{1}{N}\sum_{n=1}^{N} f(x_n).$$

Principalement, deux types de propriétés seront recherchées de l'estimateur Monte Carlo. La première, connue sous le nom de **loi des grands nombres** considère la convergence de $\hat{\pi}_N(f)$ vers $\pi(f)$ lorsque $N \to \infty$, assurant alors que l'estimation se rapproche de plus en plus de la vraie valeur lorsque la taille de l'échantillon augmente. La seconde, connue sous le nom de **théorème central limite**, renseigne plutôt sur la distribution de $\hat{\pi}_N(f)$ lorsque $N \to \infty$. Cette distribution asymptotique peut alors être utilisée afin de produire des erreurs standards ou bien des régions de confiance sur l'estimation ou même effectuer de l'inférence.

3.3.1 Lois des grands nombres

On dit qu'un algorithme MCMC adaptatif satisfait une **loi des grands nombres** si $\hat{\pi}_N(f)$ converge vers $\pi(f)$ lorsque $N \to \infty$. Si la convergence est en probabilité, c.-à-d. si

$$\forall\, \varepsilon > 0, \quad \lim_{N\to\infty} \mathbb{P}\left(|\hat{\pi}_N(f) - \pi(f)| > \varepsilon\right) = 0,$$

alors l'algorithme satisfait la loi **faible** des grands nombres. Si la convergence est également presque sûrement, c.-à-d. si

$$\mathbb{P}\left(\lim_{N\to\infty} \hat{\pi}_N(f) = \pi(f)\right) = 1,$$

alors l'algorithme satisfait la loi **forte** des grands nombres. Il est clair que la loi forte des grands nombres est une propriété plus forte que la loi faible ; il y aura donc des situations ou seule la loi faible sera satisfaite et des conditions pour établir l'une ou les deux lois sont nécessaires.

On note d'abord qu'il existe un certain lien entre la loi faible des grands nombres et l'ergodicité ; on verra éventuellement que les conditions suffisantes à la loi faible des grands nombres sont similaires à celles de l'ergodicité pour cette raison. En effet, un algorithme adaptatif V-ergodique pour π est tel que

$$\lim_{N\to\infty} \left|\!\left|\!\left| A^{(N)}((\cdot,\gamma),\cdot) - \pi(\cdot) \right|\!\right|\!\right|_V = 0, \qquad \forall\, \gamma \in \Gamma,$$

c'est-à-dire tel que

$$\lim_{N\to\infty} \sup_{\{f:\|f\|_V \leqslant 1\}} \left\| A^{(N)}((\cdot,\gamma),f) - \pi(f) \right\|_V = 0, \qquad \forall\, \gamma \in \Gamma.$$

Ainsi, pour toute fonction f de V-norme finie (après mise à l'échelle), on a que l'espérance de f par rapport à la distribution marginale converge vers $\pi(f)$. On voit alors la similitude entre les deux propriétés en remplaçant la distribution marginale théorique par la distribution empirique de la chaîne, puis en intégrant par rapport à la distribution marginale. En fait, plusieurs auteurs ne s'attardent pas à démontrer l'ergodicité de leur algorithme et ne considèrent que la loi des grands nombres étant donné cette grande similarité entre les deux concepts, surtout que la loi des grands nombres est une propriété plus pratique considérant le but d'estimer $\pi(f)$ par $\hat{\pi}_N(f)$ et non par $A^{(N)}((x,\gamma),f)$.

Évidemment, vérifier une loi des grands nombres dépend des propriétés de l'algorithme, celles de la distribution cible π ainsi que de celles de la fonction f. Cette sous-section considère donc différentes conditions permettant de vérifier d'abord la loi faible puis la loi forte des grands nombres.

3.3.1.1 Loi faible des grands nombres

En supposant que f soit bornée sur $\mathcal{X}$, Roberts et Rosenthal (2007) montrent que l'adaptation diminuante (condition 3.1) et la convergence bornée (condition 3.2) sont des conditions suffisantes pour vérifier la loi faible des grands nombres d'un algorithme. On voit donc à nouveau le lien entre ergodicité et loi des grands nombres puisque ces deux conditions assurent l'ergodicité ainsi que la loi faible des grands nombres pour toute fonction bornée.

Théorème 3.7 (Roberts et Rosenthal, 2007, théorème 23) *Soit un algorithme MCMC adaptatif satisfaisant la l'adaptation diminuante (condition 3.1) et la convergence bornée (condition 3.2) et soit $f : \mathcal{X} \to \mathbb{R}$ une fonction bornée mesurable. Alors, pour toutes valeurs initiales $(X_0,\Gamma_0) = (x_0,\gamma_0)$, la loi faible des grands nombres est vérifiée par l'algorithme, conditionnellement aux valeurs initiales.*

La supposition f bornée est généralement trop forte : par exemple, si l'on veut estimer le premier moment de π alors $f(x) = x$ est non-borné dès que $\mathcal{X}$ est non-borné. Yang (2008a) propose deux ensembles de conditions permettant de généraliser le résultat précédent au cas où f est non-borné.

Une première possibilité est de renforcer l'adaptation diminuante. Plutôt que d'exiger que l'adaptation entre deux transitions consécutives tende vers 0, on requiert que l'adaptation soit sommable.

Condition 3.4 (Adaptation sommable, Yang, 2008a, section 3) *L'adaptation entre deux transitions est sommable en probabilité conditionnellement aux valeurs initiales $(X_0,\Gamma_0) = (x_0,\gamma_0)$, c.-à-d., pour tout $\delta > 0$ il existe $M \in \mathbb{N}$ tel que*

$$\mathbb{P}\left(\sum_{n\geqslant 0} \left\|P_{\Gamma_{n+1}} - P_{\Gamma_n}\right\|_{\mathrm{TV}} \leqslant M \middle| X_0 = x_0, \Gamma_0 = \gamma_0\right) \geqslant 1-\delta.$$

Sous cette condition, il est possible de montrer la loi faible des grands nombres pour toute fonction π-intégrable.

Théorème 3.8 (Yang, 2008a, théorème 3.1) *Soit un algorithme MCMC adaptatif satisfaisant l'adaptation sommable (condition 3.4) et la convergence bornée (condition 3.2) et soit $f : \mathcal{X} \to \mathbb{R}$ une fonction mesurable telle que $\pi(|f|) < \infty$. Alors, pour toutes valeurs initiales $(X_0,\Gamma_0) = (x_0,\gamma_0)$, la loi faible des grands nombres est vérifiée par l'algorithme, conditionnellement aux valeurs initiales.*

Une seconde possibilité est de considérer un algorithme adaptatif à transitions Metropolis-Hastings satisfaisant les conditions du corollaire 3.15, renforçant plutôt la convergence bornée. La fonction f doit être π-intégrable ainsi qu'intégrable par rapport à la mesure $\lambda(\cdot)$ par rapport à laquelle les densités de proposition et la densité cible sont supposées absolument continues. Notons qu'un résultat similaire peut être établi sous les conditions du corollaire 3.14 lorsque les transitions sont des densités (cf. Yang, 2008a, corollaire 4.11.)

Théorème 3.9 (Yang, 2008a, théorème 4.1) *Soit un algorithme MCMC adaptatif à transitions Metropolis-Hastings satisfaisant l'adaptation diminuante (condition 3.1) et les conditions du corollaire 3.15 et soit $f : \mathcal{X} \to \mathbb{R}$ une fonction mesurable telle que $\pi(|f|) < \infty$ ainsi que $\lambda(|f|) < \infty$. Alors, pour toutes valeurs initiales $(X_0,\Gamma_0) = (x_0,\gamma_0)$, la loi faible des grands nombres est vérifiée par l'algorithme, conditionnellement aux valeurs initiales.*

La condition de compacité de l'ensemble produit $\mathcal{X} \times \mathcal{Y}$, dans les conditions du corollaire 3.15, est souvent problématique. Dans le contexte des approximations stochastiques (sous-section 3.2.3), il a été possible de relâcher cette supposition à l'aide de recouvrements compacts. On peut donc s'attendre à ce que le théorème 3.9 puisse être généralisé au cas où $\mathcal{X} \times \mathcal{Y}$ n'est pas compact. En fait, la loi forte des grands nombres pourra être vérifiée dans cette situation en imposant certaines conditions dans le comportement de f par rapport à celui de π.

3.3.1.2 Loi forte des grands nombres

Initialement, Haario et collab. (2001) montrent que leur algorithme AM 3.1 peut satisfaire à la loi forte des grands nombres selon des conditions assez fortes. Ils exigent, entre autres, que f soit borné et que le support $\mathcal{X}$ de π soit compact. De plus, bien que le paramètre Σ soit lui-même non-borné par le bas, la covariance $s_d\Sigma + \varepsilon I_d$ est elle bornée par le bas ce qui permet d'établir certaines propriétés uniformément sur la famille de transition.

Évidemment, il est possible de faire mieux sur les conditions sur π, sur Σ et sur f. Lors de la comparaison avec la V-ergodicité en introduction, on a vu que la loi des grands nombres y ressemblait lorsque la V-norme de f est finie : il s'agira en fait de la principale condition à supposer pour vérifier la loi forte des grands nombres pour les algorithmes adaptatifs.

D'abord, un résultat basé seulement sur l'adaptation diminuante et la convergence bornée, semblable au théorème 3.7, n'est pas possible. En effet, même pour une fonction f borné, cet ensemble de conditions n'est pas suffisant à vérifier la loi forte, tel que le montrent Roberts et Rosenthal (2007, exemple 24).

Atchadé et Rosenthal (2005, théorème 3.2) vérifient la loi forte des grands nombres dans des conditions assez générales. Les conditions utilisées sont alors relativement complexes, mais il est possible de leur trouver des conditions suffisantes plus compréhensibles. En effet, leur condition 3.1 peut être assurée par l'ergodicité géométrique simultanée (condition 3.8) et par l'adaptation diminuante simultanément. Dans ce cas, on doit supposer que la fonction f est telle que $\|f\|_V < \infty$, où V provient de la condition de dérive géométrique. Atchadé et Fort (2010, théorème 2.5) considèrent un résultat similaire en supposant plutôt une dérive polynomiale uniforme du type $P_\gamma V \leqslant V - bV^{1-\alpha} + c\,\mathbb{1}_C$ (c.-à-d.„ la condition 3.10.) Dans ce cas, on doit supposer $\|f\|_{V^\beta} < \infty$ pour un certain $\beta \in [0,1-\alpha)$.

Dans ces deux cas, l'ergodicité uniforme (géométrique ou polynomiale) peut être difficile à établir ou même fausse lorsque l'espace $\mathcal{Y}$ n'est pas compact par rapport à une certaine topologie. Dans le contexte des algorithmes adaptatifs par approximation stochastique, il a été possible, en introduisant un recouvrement compact de l'espace des paramètres $\mathcal{Y} = \Theta$, de contourner ce problème afin de dériver des conditions suffisantes à l'ergodicité. Il est possible d'employer cette même astuce pour trouver des conditions assurant la loi forte des grands nombres lorsque Θ n'est pas compact. En effet, Andrieu et Moulines (2006) démontrent le résultat suivant.

Théorème 3.10 (Andrieu et Moulines, 2006, théorème 8) *Soit $\{\mathcal{K}_r\}_{r\geqslant 0}$ un recouvrement compact de Θ et soit $\{\gamma_n\}_{n\geqslant 0} \subset [0,\infty)$ une suite non-croissante telle que $\sum_{n\geqslant 1} n^{-1}\gamma_n < \infty$.*

On considère la chaîne conjointe $\{(X_n,\theta_n,R_n)\}_{n\geqslant 0}$ donnée par l'algorithme 3.12. Supposons que l'algorithme satisfasse aux conditions 3.3 , 3.24 et 3.25 et que le nombre de réinitialisations soit borné presque sûrement, c.-à-d., $\mathbb{P}\left(\lim_{n\to\infty} R_n < \infty\right) = 1$. Alors, pour toute fonction $f : \mathcal{X} \to \mathbb{R}$ telle que $\|f\|_{V^\alpha} < \infty$ pour un $\alpha \in [0,1-\beta)$ où β provient de la condition 3.25 et V de la condition 3.3, la loi forte des grands nombres est satisfaite.

Par la discussion à la sous-section 3.2.3, on sait que les conditions 3.3, 3.24 et 3.25 ainsi qu'un nombre de réinitialisations borné en probabilité est suffisant afin de montrer l'ergodicité de ce type d'algorithme. Le théorème 3.10 montre à nouveau la correspondance entre ces deux concepts puisqu'on ne requiert seulement qu'une condition supplémentaire sur $\{\gamma_n\}_{n\geqslant 0}$, qui est nécessaire à la condition 3.27 utilisée pour montrer un nombre de réinitialisations borné, ainsi qu'une condition sur la fonction f.

Étant donné que les algorithmes adaptatifs par approximations stochastiques sont particulièrement pratiques et répandus, le théorème 3.10 trouve de multiples applications.

Andrieu et Moulines (2006, théorème 1) vérifient la loi des grands nombres pour l'algorithme AM 3.1 lorsque la covariance est non-bornée pour toute fonction f telle que $\|f\|_{V^{\alpha/2}} < \infty$ où $V \propto \pi^{-1}$ et $\alpha \in [0,1)$. Similairement, Saksman et Vihola (2010, théorème 10) arrivent au même résultat avec $V \propto \pi^{-1/2}$ et $\alpha = 1$.

En fait, l'annexe 3.4.3 montre (par le théorème 3.10) que l'algorithme AM vérifie la loi forte dès que π est régulière (condition 3.13), admet des contours réguliers (condition 3.14) et des ailes super-exponentielles (condition 3.18) et que la suite des pas d'adaptation est de l'ordre de $\gamma_n \sim n^{-\gamma}$, $\gamma \in (1/2,1]$. Dans l'exemple, on utilise $V \propto \pi^{-\eta}$ pour un $\eta \in (0,1)$ et, donc, la loi forte est satisfaite pour toute fonction f telle que $\|f\|_{V^\alpha} < \infty$ pour $\alpha \in [0,1/2]$ étant donné $\beta \in [0,1/2)$ dans la conditions 3.25.

De plus, Atchadé et Fort (2010, proposition 5) vérifient la loi des grands nombres pour l'algorithme ASWAM 3.10 en supposant seulement la dérive polynomiale plutôt que géométrique ; dans ce cas, la fonction f doit être telle que $\|f\|_V < \infty$ où $V = 1 + \pi^{-r}$ pour un certain $r \geqslant 0$. Enfin, Atchadé (2006, théorème 2.1) vérifie la loi des grands nombres pour un algorithme MALA adaptatif (voir exemple 3.5) ; dans ce cas, on requiert $\|f\|_V < \infty$ où $V \propto \pi^{-1/4}$.

La famille des algorithme AirMCMC 3.6 permet une très grande liberté sur la manière dont le nouveau paramètre de la transition est calculé. En imposant seulement que la fréquence de mise-à-jour du paramètre tende vers 0, il est possible de produire un algorithme qui soit ergodique. Similairement, l'algorithme peut vérifier la loi forte des grands nombres. Par exemple, si la famille de transitions satisfait une condition d'ergodicité géométrique uniforme alors (Chimisov et collab., 2018, théorème 2) la loi forte est satisfaite pour toute fonction f telle que $\|f\|_{V^\beta} < \infty$ où $\beta > 1/2$, V provient de la dérive géométrique et le temps entre deux adaptations augmente selon $n_k \sim k^\beta$.

On conclut cette section par une discussion sur les conditions du type $\|f\|_V < \infty$:

$$\sup_{x\in\mathcal{X}} \frac{|f(x)|}{V(x)} = \|f\|_V < \infty,$$

que l'on peut réécrire selon

$$\frac{|f(x)|}{\|f\|_V} \leqslant V(x), \qquad \forall\, x \in \mathcal{X}\,.$$

Sans perte de généralité (après mise à l'échelle), on peut supposer $\|f\|_V = 1$: cette condition correspond alors à $|f| \leqslant V$ sur $\mathcal{X}$. La fonction f sera souvent bornée sur les compacts de $\mathcal{X}$ (e.g. f continue) et c'est donc le comportement de f pour de grands $\|x\|_2$ qui importe pour vérifier cette condition.

Dans le théorème 3.10, la fonction V correspond à la fonction utilisée dans la dérive géométrique et, dans bien des exemples, on a $V \propto \pi^{-r}$ pour un certain $r \geqslant 0$. Lorsque la densité cible π admet des ailes super-exponentielles (condition 3.18), on a que le comportement des ailes de la densité cible correspond à celui de $\exp(-\|x\|_2^s)$ pour un $s > 1$. Ainsi, le comportement de V pour de grands $\|x\|_2$ sera celui de $\exp(r\|x\|_2^s)$. C'est donc dire que la loi forte sera satisfaite pour toute fonction dont les ailes n'augmentent pas plus rapidement qu'une exponentielle d'ordre $s > 1$, ce qui définit une large classe de fonctions. En particulier, toute fonction polynomiale peut être estimée dans ces conditions, ce qui permet d'estimer tout moment $\mathbb{E}\{|X|^p\}$, $0 \leqslant p < \infty$, de π en ayant les garanties de la loi forte des grands nombres.

3.3.2 Théorèmes limites centraux

La loi des grands nombres est souvent satisfaite dans l'estimation de $\pi(f)$ par son équivalent empirique $\hat{\pi}_N(f)$. Cependant, cette propriété assure seulement que l'estimateur Monte Carlo converge vers l'espérance et ne renseigne donc pas sur la variabilité de cette estimateur. Quant à elle, l'estimation par région de confiance permet de décrire une région contenant un ensemble d'estimations ponctuelles plausibles. Ensemble, l'estimation ponctuelle et l'estimation par région de confiance permettent donc une estimation directe de $\pi(f)$, accompagnée d'une certaine région près de l'estimé contenant des valeurs tout aussi plausibles, mais qui n'ont pas été obtenues simplement par la nature aléatoire du processus.

Afin d'obtenir une région de confiance autour d'un estimé, il est nécessaire d'avoir la distribution de l'estimateur produisant l'estimé. En général, une telle distribution est difficile à trouver puisqu'il s'agit d'une somme de variables aléatoires (fortement auto-corrélés dans le cas MCMC.) Par contre, des résultats asymptotiques peuvent fournir une bonne approximation de cette distribution. En effet, la classe des **théorèmes centraux limites** montrent que la distribution d'une moyenne converge en distribution, évidemment sous certaines conditions, vers la distribution d'une loi normale centrée en l'espérance de la moyenne et d'une certaine variance asymptotique. Ainsi, il sera parfois possible d'approcher la distribution de $\hat{\pi}_N(f)$ par celle d'une loi normale ; il sera ensuite possible d'utiliser cette distribution afin d'estimer une certaine région de confiance autour de $\hat{\pi}_N(f)$.

D'une manière générale, un théorème central limite énonce un résultat tel que

$$N^{-1/2} \sum_{n=1}^{N} [f(X_n) - \pi(f)] \quad \xrightarrow{\mathcal{D}} \quad Z, \qquad N \to \infty,$$

où $Z \sim \mathcal{N}(0, \sigma_\infty^2)$ et $0 \leqslant \sigma_\infty^2 < \infty$ est appelée la **variance asymptotique**. Généralement, on aura $\sigma_\infty^2 > 0$, mais certains cas dégénérés sont parfois possibles où $Z \sim \delta_0$. Similairement, dans le cas où f est multivarié, un tel théorème énonce

$$N^{-1/2} \sum_{n=1}^{N} [f(X_n) - \pi(f)] \quad \xrightarrow{\mathcal{D}} \quad Z, \qquad N \to \infty,$$

où $Z \sim \mathcal{N}_q(\mathbf{0}, \Sigma_\infty)$ et Σ_∞ est une matrice symétrique définie positive appelée la **covariance asymptotique**. À nouveau, il est possible que Σ_∞ ne corresponde pas à une covariance dans certains cas. On doit alors considérer la distribution normale généralisée où la covariance peut n'être que définie semi-positive, constituant alors une densité dans un sous-espace de $\mathbb{R}^q$. On se limitera ici au cas où f est univarié

étant donné qu'aucun résultat n'existe par rapport à f multivarié pour les algorithmes adaptatifs ; la section 2.4 contient certaines remarques sur le cas multivarié pour les chaînes homogènes.

Cette section sera consacrée à l'étude des conditions sous lesquelles un algorithme MCMC adaptatif vérifie un tel théorème ainsi qu'au calcul de Σ_∞ et de son estimation. Enfin, le tout sera relié à la question de l'efficacité de l'algorithme, ce qui constituait le motif principal pour introduire l'adaptation dans les méthodes MCMC.

3.3.2.1 Conditions suffisantes

Comme ce fut le cas pour les lois des grands nombres, une condition sur la fonction f ainsi que des conditions assurant l'ergodicité de l'algorithme seront requises. Puis, comme il s'agit d'un type de résultat plus fort qu'une loi des grands nombres, une hypothèse additionnelle sera nécessaire. Afin d'espérer vérifier un théorème central limite, il sera primordial de supposer la convergence en distribution de l'index d'adaptation vers une certaine limite possiblement aléatoire, c.-à-d., $\Gamma_n \xrightarrow{\mathcal{D}} \Gamma_\infty$ pour une certaine variable aléatoire Γ_∞ à support dans $\mathcal{Y}$. En effet, si la distribution de Γ_n ne converge pas vers une certaine limite, le comportement du processus $\{(X_n,\Gamma_n)\}_{n\geqslant 0}$ sera généralement trop difficile à contrôler pour prouver un tel résultat. La plupart des résultats dans la littérature exigent même que la limite Γ_∞ soit une variable aléatoire dégénérée (et donc $\Gamma_n \xrightarrow{\mathbb{P}} \gamma_\infty \in \mathcal{Y}$) ; on trouvera alors que la distribution de transition se stabilise asymptotiquement vers une transition fixe. Dans ce cas, il n'est pas surprenant de trouver un théorème central limite puisque l'algorithme se réduit effectivement à un algorithme MCMC non-adaptatif pour lequel divers théorèmes limites centraux existent (section 2.2.3.2).

Le contexte général des algorithmes adaptatifs introduit par Roberts et Rosenthal (2007), dans lequel les conditions générales d'adaptation diminuante et de convergence bornée permettaient de vérifier l'ergodicité et certaines lois des grands nombre, n'est pas suffisant afin de vérifier un théorème central limite. En effet, l'adaptation diminuante n'implique pas nécessairement une convergence de l'index de la transition et il donc possible de construire un contre-exemple : Roberts et Rosenthal (2007, exemple 24) considèrent un algorithme où les deux conditions d'ergodicité sont satisfaites, mais où une loi forte n'est pas satisfaite pour une fonction bornée. Par conséquent, le théorème central limite ne peut être satisfait pour cette même fonction bornée.

Dans l'étude de l'ergodicité des algorithmes adaptatifs par approximations stochastiques (section 3.2.3), le résultat principal dans le cas général à recouvrements compacts (théorème 3.31) implique une convergence du paramètre de la transition $\theta_n = \Gamma_n$ vers une limite fixe qui est un point stationnaire du champ moyen. Ce contexte se prête donc particulièrement bien à un possible théorème central limite puisqu'on a l'ergodicité, la loi forte des grands nombres pour une grande classe de fonctions (théorème 3.10) et la convergence du paramètre. Le résultat suivant n'est donc pas surprenant :

Théorème 3.11 (Andrieu et Moulines, 2006, théorème 9) *Soit un algorithme adaptatif à recouvrement compact $\{\mathcal{K}_r\}_{r\geqslant 0}$ de $\Theta \subseteq \mathbb{R}^p$ et soit $\{\gamma_n\}_{n\geqslant 1}$ une suite de pas d'adaptation décroissante telle que $\sum_{n\geqslant 1} n^{-1/2}\gamma_n < \infty$. supposons que l'algorithme satisfait les conditions 3.3, 3.24 et 3.25 pour un certain choix de V, α et β. Pour toute fonction $f : \mathcal{X} \to \mathbb{R}$ et $\theta \in \Theta$, on définit*

$$\sigma^2(\theta,f) := \pi\left((\hat{f}_\theta - P_\theta \hat{f}_\theta)^2\right), \qquad \hat{f}_\theta(\cdot) := \sum_{n\geqslant 0} (P_\theta^n f(\cdot) - \pi(f)). \tag{3.22}$$

Supposons, de plus, qu'il existe une variable aléatoire θ_∞ à support dans Θ telle que $\pi(\hat{f}^2_{\theta_\infty})$ et $\pi((P_{\theta_\infty}\hat{f}_{\theta_\infty})^2)$ soient bornés presque sûrement et telle que

$$\limsup_{n\to\infty} |\theta_n - \theta_\infty| = 0, \qquad p.s.$$

Alors, si $f \in \mathcal{L}_{V^\alpha}$, le théorème central limite est satisfait pour f :

$$n^{-1/2} \sum_{n=1}^{N} [f(X_n) - \pi(f)] \quad \xrightarrow{\mathcal{D}} \quad Z, \qquad N \to \infty,$$

où Z a la function caractéristique $\mathbb{E}\{\exp(-\frac{1}{2}\sigma^2(\theta_\infty,f)t^2)\}$. En particulier, si θ_∞ est constant avec $\sigma^2(\theta_\infty,f) > 0$, alors $Z \sim \mathcal{N}(0,\sigma^2(\theta_\infty,f))$.

On note que les conditions 3.26 et 3.27 nécessaires à démontrer la convergence de l'adaptation vers une valeur fixe dans le théorème 3.31 ne sont pas requises. On requiert plutôt la supposition plus faible que θ_n converge vers une variable aléatoire fixe.

Dans le cas plus restreint où Θ est supposé compact et que θ_n converge vers une valeur fixe $\theta_\infty \in \theta$, Andrieu et Atchadé (2007, théorème 2.3) arrivent à un résultat similaire pour un ensemble de conditions pratiquement identiques.

Andrieu et Moulines (2006, théorème 15) et Saksman et Vihola (2010, théorème 18) utilisent le théorème 3.11 afin de démontrer un théorème central limite pour l'algorithme AM 3.1 pour toute fonction $f \in \mathcal{L}_{V^\alpha}$ où $V = \pi^{-1/2} \sup^{1/2} \pi$ et $0 \leqslant \alpha < 1/2$. Par la convergence de l'adaptation vers une limite fixe, ils obtiennent $Z \sim \mathcal{N}(0,\sigma^2)$ pour une certaine variance asymptotique $\sigma^2 \in [0,\infty)$ qui peut être calculée par (3.22).

Le cas plus général (pas nécessairement par approximations stochastiques) est plus complexe ; voici quelques cas spécifiques où un théorème central limite est satisfait. D'abord, un théorème central limite a été établi pour les algorithmes adaptatifs par régénération (section 3.1.3.1) par Gilks et collab. (1998, théorème 2). De plus, Atchadé (2010, théorème 3.3) démontrent un théorème central limite pour l'échantillonneur Équi-énergie où les transitions n'admettent pas nécessairement π comme distribution invariante, mais bien π_n qui doivent satisfaire $\pi_n(f) \xrightarrow{\text{p.s.}} \pi(f)$. Fort et collab. (2014) généralisent ce résultat à tout algorithme adaptatif où la distribution invariante de P_θ peut dépendre de θ. Les conditions sont assez techniques, mais des applications à l'algorithme AM et à l'algorithme à tempérage en parallèle sont considérées. Enfin, tout comme il est possible de vérifier la loi forte des grands nombres pour les algorithmes AirMCMC 3.6, un théorème central limite (Chimisov et collab., 2018, théorème 1) est également satisfait en supposant que $\mathcal{Y}$ est un espace métrique, que Γ_n converge presque sûrement vers une variable aléatoire fixe et que P_γ est une fonction continue de γ. Notons que l'ergodicité géométrique de $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ requise peut même être affaiblie à une ergodicité polynomiale (Chimisov et collab., 2018, théorème 3).

3.3.2.2 Variance asymptotique et estimation

Lorsque θ_∞ est une constante, on trouve bel et bien une distribution asymptotique normale pour l'estimateur $\hat{\pi}(f)$. Dans ce cas, on s'intéresse souvent au calcul de la variance asymptotique donnée par (3.22) et de son estimation en utilisant l'échantillon produit par l'algorithme.

Dans le cas de chaînes de Markov homogènes, il existe plusieurs méthodes afin d'estimer la variance asymptotique $\sigma^2(\theta,f)$ et la sous-section 2.4.2 en présente quelques-unes. Les algorithmes adaptatifs ne jouissent pas de cette homogénéité et la question de l'estimation de $\sigma^2(\theta,f)$ ne pourra pas avoir les mêmes garanties en utilisant les mêmes techniques. Cependant, puisque la chaîne tend à devenir homogène par la convergence de θ_n, on s'attend à ce que ces techniques d'estimations demeurent relativement valides : en écartant une portion initiale de la chaîne, on obtient une chaîne qui est alors pratiquement homogène et les propriétés de l'estimation dans cette situation sont mieux connues.

Atchadé (2011) considèrent l'estimateur spectral de la variance asymptotique,

$$\hat{\sigma}_N^2(\theta,f) = \sum_{n=-N}^{N} w(nb_N)\hat{\gamma}_N(n),$$

où $\hat{\gamma}_N(n)$ est l'autocovariance empirique d'ordre n de $\{f(X_n)\}_{n=0}^N$, $w : \mathbb{R} \to \mathbb{R}$ est un noyau dont le support est $[-1,1]$ et $b_N > 0$ est la fenêtre d'estimation. Sous certaines conditions sur w, sur b_N et sur l'algorithme adaptatif (une ergodicité géométrique et une adaptation diminuante), Atchadé (2011, théorèmes 4.1-3) montrent que cet estimateur est $\mathcal{L}^p$-convergent. Pour ce qui est de l'estimateur par moyenne par lot (section 2.4.2.3), les propriétés théoriques dans le cas adaptatif sont inconnues pour le moment. Enfin, lorsque f est multivarié, aucun résultat théorique quant à l'estimation de la variance asymptotique n'existe actuellement pour les algorithmes adaptatifs. En pratique, deux options sont alors possibles. D'une part, il est possible de construire des intervalles de confiance pour chacune des composantes de f puis de les ajuster par une méthode telle que celle de Bonferroni. D'autre part, en se fiant à l'extension des résultats univariés à multivariés dans le cas homogène, on peut s'attendre à ce qu'une même extension soit possible pour les algorithmes adaptatifs ; on pourrait donc utiliser les estimateurs multivariés considérés à la section 2.4.2.5.

3.3.2.3 *Efficacité de l'algorithme*

La motivation principale d'introduire une composante d'adaptation dans un algorithme MCMC est la mise-au-point automatique des différents paramètres de la distribution de transition. L'automatisation sera construite de sorte à rendre l'algorithme optimal par rapport à une certaine mesure (voir sous-section 3.1.4 pour quelques-uns de ces critères.) La plupart des critères sont en fait des proxys pour l'efficacité de l'estimation définie en fonction de la variance asymptotique de l'estimation de $\pi(f)$,

$$\mathrm{eff}_{\mathcal{Y}}(\gamma,f) = \frac{\sigma^2(\gamma,f)}{\sigma_*^2(f)}, \qquad \sigma_*^2(f) = \inf_{\gamma\in\mathcal{Y}} \sigma^2(\gamma,f).$$

Ainsi, l'algorithme adaptatif cherchera généralement à trouver γ tel que la variance asymptotique $\sigma^2(\gamma,f)$ soit minimale au sein de la famille $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$. Dans certaines situations, la transition optimale est en fait connue en relation avec π et l'algorithme estime alors les propriétés de π requise à la définition de P_{γ_*} où $\sigma^2(\gamma_*,f) = \sigma_*^2(f)$. Par exemple, un algorithme Metropolis à propositions gaussiennes est optimal lorsque la covariance de la proposition est un certain multiple de la covariance de π : l'algorithme AM estime donc Σ_π pour ainsi approcher une transition optimale.

Les théorèmes centraux limites permettent ce genre de conclusions. En effet, ils garantissent une variance asymptotique finie et fournissent une manière de la calculer puis de l'estimer. De plus, lorsque le critère d'optimisation utilisé dans l'algorithme n'est pas directement l'efficacité de l'estimation, il est possible de comparer différents algorithmes adaptatifs en comparant leurs variances asymptotiques tout comme cette mesure est utilisée pour comparer l'efficacité des algorithme à transitions fixes.

3.4 Annexes au chapitre 3

3.4.1 Sur la convergence bornée

Cette section considère donc des conditions suffisantes à la convergence bornée (condition 3.2). Par le théorème 3.1, l'ergodicité peut ensuite être vérifiée en supposant la condition d'adaptation diminuante 3.1.

3.4.1.1 Ergodicité uniforme simultanée

On considère la condition suivante, telle qu'énoncée par Roberts et Rosenthal (2007) :

Condition 3.5 (Ergodicité uniforme simultanée) *Pour tout $\varepsilon > 0$, il existe $N = N(\varepsilon) \in \mathbb{N}$ tel que*

$$\|P_\gamma^N(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}} \leqslant \varepsilon, \qquad \forall x \in \mathcal{X}\,, \gamma \in \mathcal{Y}\,.$$

Cette condition, beaucoup plus forte que la Convergence bornée, est suffisante pour assurer l'ergodicité d'un algorithme adaptatif. Elle requiert en fait que chaque P_γ soit ergodique à π (dans le sens des chaînes de Markov homogènes), et ce, à un rythme uniforme par rapport aux valeurs initiales.

Il est simple de voir que l'ergodicité uniforme simultanée (condition 3.5) implique la convergence bornée puisqu'on obtient $M_\varepsilon(x,\gamma) \leq M_* := N(\varepsilon)$ pour tout $x \in \mathcal{X}$ et $\gamma \in \mathcal{Y}$: $M_\varepsilon(x,\gamma)$ est donc borné en probabilité. La convergence bornée ne requiert pas, quant à elle, que chaque P_γ soit ergodique d'une telle manière uniforme, bien que ce serait suffisant. Le résultat suivant est donc une conséquence triviale du théorème 3.1.

Théorème 3.12 (Roberts et Rosenthal, 2007, théorème 1) *Soit $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ une famille de transitions sur $\mathcal{X}$ telle que P_γ admette π comme distribution stationnaire pour tout $\gamma \in \mathcal{Y}$. Sous les conditions 3.1 et 3.5, l'algorithme adaptatif est ergodique.*

Maintenant, l'ergodicité uniforme simultanée n'est pas facile à vérifier en pratique et des conditions plus accessibles sont requises. En supposant seulement que chaque P_γ soit ergodique, l'uniformité par rapport à $\gamma \in \mathcal{Y}$ et $x \in \mathcal{X}$ sera généralement violée lorsque γ se dirige vers certaines régions de $\mathcal{Y}$ où le temps de convergence augmente sans borne. Il faut alors ajouter une condition sur $\mathcal{X} \times \mathcal{Y}$ ainsi que sur la relation entre les transitions P_γ. Le corollaire 3.13 impose que $\mathcal{X} \times \mathcal{Y}$ soit un espace compact et que la différence entre la distribution marginale et la distribution cible, $T(x,\gamma,n)$, varie d'une manière continue par rapport à x et γ.

Corollaire 3.13 (Roberts et Rosenthal, 2007, corollaire 3) *Soit un algorithme MCMC adaptatif satisfaisant la condition d'adaptation diminuante 3.1 et où chaque P_γ est ergodique à π. Si $\mathcal{X} \times \mathcal{Y}$ est un espace compact dans une certaine topologie par rapport à laquelle l'application $(x,\gamma) \mapsto T(x,\gamma,n)$ est continue pour chaque $n \in \mathbb{N}$ fixé, alors l'algorithme est ergodique. En fait, la condition d'ergodicité uniforme simultanée 3.5 est satisfaite par l'algorithme.*

Le corollaire 3.14 exige également la compacité de $\mathcal{X} \times \mathcal{Y}$, mais requiert, quant à lui, que chaque $P_\gamma(\cdot|x)$ admette une densité bornée uniformément par rapport à γ et que ces densités soient des fonctions continues par rapport à x et à γ.

Corollaire 3.14 (Roberts et Rosenthal, 2007, corollaire 4) *Soit un algorithme MCMC adaptatif satisfaisant la condition d'adaptation diminuante 3.1 et où chaque P_γ est ergodique à π. Supposons que, pour chaque $\gamma \in \mathcal{Y}$, $P_\gamma(\mathrm{d}z|x) = f_\gamma(z|x)\lambda(\mathrm{d}z)$ a une densité $f_\gamma(\cdot|x)$ par rapport à une mesure λ finie sur $\mathcal{X}$. Supposons aussi que les $f_\gamma(z|x)$ sont uniformément bornées et que, pour chaque $z \in \mathcal{X}$ fixé, l'application $(x,\gamma) \mapsto f_\gamma(z|x)$ est continue par rapport à une topologie d'espace métrique par rapport à laquelle $\mathcal{X} \times \mathcal{Y}$ est compacte. Alors, l'algorithme adaptatif est ergodique à π.*

Les transitions M.-H. n'admettent pas de densité étant donné la probabilité positive de demeurer au même endroit ; on ne peut donc pas appliquer le corollaire 3.14 directement. Cependant, la forme très régulière de ces transition – un mélange d'une proposition et d'une masse discrète en un point – permet tout de même d'obtenir un résultat similaire en supposant que les distributions de proposition admettent toutes une densité.

Corollaire 3.15 (Roberts et Rosenthal, 2007, corollaire 5) *Soit un algorithme MCMC adaptatif satisfaisant la condition d'adaptation diminuante 3.1 et où chaque P_γ est ergodique à π. Supposons que, pour chaque $\gamma \in \mathcal{Y}$, P_γ est la transition d'un algorithme Metropolis-Hastings avec proposition $Q_\gamma(\mathrm{d}z|x) = f_\gamma(z|x)\lambda(\mathrm{d}z)$ ayant une densité $f_\gamma(\cdot|x)$ par rapport à une mesure finie λ sur $\mathcal{X}$. Supposons aussi que π admet une densité g par rapport à λ : $\pi(\mathrm{d}y) = g(y)\lambda(\mathrm{d}y)$. Supposons enfin que les $f_\gamma(z|x)$ sont uniformément bornées et que, pour chaque $z \in \mathcal{X}$ fixé, l'application $(x,\gamma) \mapsto f_\gamma(z|x)$ est continue par rapport à une topologie d'espace métrique par rapport à laquelle $\mathcal{X} \times \mathcal{Y}$ est compacte. Alors, l'algorithme adaptatif est ergodique à π.*

3.4.1.2 Conditions de pas bornés

Que ce soit dans le corollaire 3.14 ou dans le corollaire 3.15, les densités de transition ou de proposition doivent être uniformément bornées par rapport à $x \in \mathcal{X}$ et à $\gamma \in \mathcal{Y}$. Cette propriété peut être difficile à établir ou même fausse pour certains algorithmes, particulièrement lorsque les espaces $\mathcal{X}$ et $\mathcal{Y}$ ne sont pas bornés.

Le résultat suivant, dû à Craiu et collab. (2015), impose des conditions supplémentaires, quoique plus réalistes à vérifier, assurant la convergence bornée. Principalement, on suppose que le pas entre deux itérations X_n et X_{n+1} est borné et que l'espace $\mathcal{Y}$ est compact. Les pas bornés se prêtent très bien aux algorithmes Metropolis de type marche aléatoire, puisque les propositions seront toujours effectuées par rapport à l'état actuel de la chaîne et il sera donc facile d'imposer une borne sur le pas d'un algorithme. Quant à la supposition $\mathcal{Y}$ compacte, la borne sur $\mathcal{Y}$ imposant la compacité peut être arbitrairement grande, ce qui ne pose pas de problème en pratique. À la sous-section 3.2.3, une construction particulière permettra même de relâcher cette contrainte dans le cadre des algorithmes adaptatifs par approximations stochastiques.

Théorème 3.16 (Craiu et collab., 2015, théorème 21) *Soit $\mathcal{X}$ un sous-ensemble ouvert de $\mathbb{R}^d$, K un sous-ensemble borné de $\mathcal{X}$ et $\mathcal{Y}$ un ensemble compact indexant la famille de transitions $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ telle que chaque P_γ admette π comme distribution invariante et soit (Harris-)ergodique à π. On définit K_r comme le sous-ensemble de $\mathcal{X}$ de tous les points à une distance d'au plus r de K, qui est donc borné pour tout $r < \infty$. Dans le cas où les P_γ représentent une transition Metropolis-Hastings, on peut remplacer P_γ par Q_γ, la proposition associée, dans chacune des conditions qui suivent. On suppose que π et chaque P_γ admettent tous une densité positive et qu'elles sont continues par rapport à $x \in \mathcal{X}$, à $y \in \mathcal{X}$ et à $\gamma \in \mathcal{Y}$.*

Supposons que chaque P_γ satisfait à la condition de sauts bornés par $D < \infty$ en probabilité,

$$P_\gamma\left(\{y \in \mathcal{X} : \|y - x\|_2 \leqslant D\}|x\right) = 1, \qquad \forall\, x \in \mathcal{X}, \gamma \in \mathcal{Y}, \tag{3.23}$$

où $\|\cdot\|_2$ dénote la norme euclidienne usuelle. Supposons de plus qu'aucune adaptation n'est effectuée à l'extérieur de K, i.e. la chaîne suit une transition fixe P lorsque $x \notin K$

$$P_\gamma(A|x) = P(A|x), \qquad A \in \mathcal{F}, x \in \mathcal{X}\setminus K. \tag{3.24}$$

Supposons aussi que la transition P est supposée bornée par le haut,

$$\exists\, M < \infty, \text{ tel que } P(\,\mathrm{d}y|x) \leqslant M\lambda(\,\mathrm{d}y), \qquad \forall\, x \in K_D \setminus K,\, y \in K_{2D} \setminus K_D, \tag{3.25}$$

où λ dénote la mesure de Lebesgue, et que P satisfait la condition ε,δ suivante pour un rectangle J contenant $K_{2D} \setminus K_D$

$$\forall\, x,y \in J : \|y - x\|_2 < \delta \quad \Rightarrow \quad P(\,\mathrm{d}y|z) \geqslant \varepsilon\lambda(\,\mathrm{d}y). \tag{3.26}$$

Alors, la convergence bornée est satisfaite par l'algorithme qui est donc ergodique si la l'adaptation diminuante est vérifiée.

Bien que le théorème 3.16 requiert la vérification de plusieurs conditions technique, il reste néanmoins que ce résultat fournit une « recette » pour construire des algorithmes adaptatifs ergodiques. En fait, les différentes conditions sont relativement faciles à imposer à un algorithme en pratique. Aucune restriction sur la méthode d'adaptation n'est imposée à l'exception que γ doive rester dans un espace compact $\mathcal{Y}$. On choisit $K \subseteq \mathcal{X}$ borné (généralement arbitrairement large), une transition fixe P et $D < \infty$. Puis, on doit définir les transitions de sorte à satisfaire les conditions. L'exemple 3.6 montre que l'algorithme BAM (*Bounded adaptation Metropolis*, 3.13, simplification due à Rosenthal et Yang, 2018) satisfait toutes les conditions du théorème 3.16.

Les conditions du théorème 3.16 sont en fait utilisées pour vérifier les conditions de la proposition suivante qui est un peu plus générale.

Proposition 3.17 (Craiu et collab., 2015, proposition 23) *Soit un algorithme MCMC adaptatif à famille de transition $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ compacte telle que chaque P_γ admette π comme distribution invariante et soit Harris-ergodique à π. Si l'application $(x,\gamma) \mapsto \Delta(x,\gamma,n) := \left\|P_\gamma^n(\cdot|x) - \pi(\cdot)\right\|_{\mathrm{TV}}$ est continue conjointement pour tout $(x,\gamma) \in \mathcal{X}\times\mathcal{Y}$ et si $\{X_n\}_{n\geqslant 1}$ est borné en probabilité, alors la convergence bornée est satisfaite par l'algorithme qui est donc ergodique si l'adaptation diminuante est vérifiée.*

Algorithme 3.13 Metropolis à adaptation bornée (BAM, Rosenthal et Yang, 2018)

Données Densité cible π à support $\mathcal{X} = \mathbb{R}^d$, $K \subseteq \mathcal{X}$ borné, $D > 0$, $\mathcal{Y}$ un ensemble compact de matrices $d \times d$ définies positives et $\Sigma_* \in \mathcal{Y}$ fixé.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne $x_0 \in K$ et covariance initiale $\Sigma_0 \in \mathcal{Y}$.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonage Metropolis.*
      i. Proposition :
         — Si $x_n \notin K$, alors $Y \sim \mathcal{N}_d(x_n, \Sigma_*)$;
         — Si $x_n \in K$, alors $Y \sim \mathcal{N}_d(x_n, \Sigma_n)$;
      ii. Si $\|y - x_n\|_2 > D$, la proposition est automatiquement rejetée ;
      iii. Sinon :
         A. Calcul de la probabilité d'acceptation
         $$\alpha(x_n, y) = \min\left\{1, \frac{\pi(y)}{\pi(x_n)}\right\};$$
         B. Avec probabilité $\alpha(x_n, y)$, acceptation de la proposition $(x_{n+1} = y)$; sinon rejet $(x_{n+1} = x_n.)$
   (b) *Adaptation.* Si $x_{n+1} \in K$, alors mise à jour de la covariance Σ_{n+1}. Par exemple,
   $$\Sigma_{n+1} = (2.38)^2 \operatorname{Cov}(x_{0:n})/d,$$
   en s'assurant que $\Sigma_{n+1} \in \mathcal{Y}$.

Sortie L'échantillon $x_{1:N}$.

Exemple 3.6 Vérification de la convergence bornée de l'algorithme BAM 3.13 (Craiu et collab., 2015, proposition 22)

On montre que l'algorithme satisfait à toutes les conditions du théorème 3.16.

Pour $\Sigma \in \mathcal{Y}$ une matrice définie positive, où $\mathcal{Y}$ est compacte (par exemple en bornant le déterminant Σ par le haut et par le bas), les propositions Metropolis $Q_\Sigma(y|x) = \mathcal{N}_d(y|x, \Sigma)$ admettent des densités positives sur $\mathcal{X} = \mathbb{R}^d$ et continues par rapport à x, à y et à Σ. En effet, les densités gaussiennes dépendent de x et de y qu'à travers $\exp(-\frac{1}{2}(x-y)^\top \Sigma^- 1(x-y)$ qui est continu par rapport à x et à y pour Σ fixé. Ces mêmes densités dépendent de Σ à travers ce même terme exponentiel ainsi qu'à travers $\det(\Sigma)^{-1/2}$. Le déterminant est une fonction continue de la matrice puisqu'il s'agit d'une somme de produits des composantes. Puis, par la définition de $\mathcal{Y}$, le déterminant est compris dans un certain intervalle fermé sous-ensemble de $(0,\infty)$. On note enfin que $x \mapsto x^{-1/2}$ est continu sur $(0,\infty)$. Ensuite, l'inverse d'une matrice peut être écrit en fonction de son adjointe (fonction continue) : parmi la classe des matrices à déterminant strictement positif, l'inverse d'une matrice est une fonction continue. Puis, la forme quadratique est une somme de produits des composantes de l'inverse, ce qui est à nouveau continu. Finalement, l'exponentielle est continue de sorte que la densité est continue par rapport à Σ.

Les transitions M.-H. de type marche aléatoire sont invariantes pour π (proposition 2.24 et théorème 2.7) et sont ergodiques à π sous certaines conditions satisfaites par les densités gaussiennes tronquées. En effet, la condition ε,δ (3.26) est suffisantes à la π-irréductibilité (proposition 2.25). Puis, le noyau est apériodique par la proposition 2.26, ce qui permet de conclure à l'ergodicité de la chaîne via le théorème 2.10. Notons que les propositions M.-H. sont des lois gaussiennes tronquées à D ; les probabilités d'acceptation peuvent tout de même être calculées avec les densités gaussiennes complètes puisque la constante d'intégration s'annulera dans le quotient.

Les propositions M.-H. étant automatiquement refusée pour $\|y-x\|_2 > D$, la condition de sauts bornés (3.23) est satisfaite par construction. Également par construction, la transition est fixée à l'extérieur de K à $Q_{\Sigma_*}(y|x) = \mathcal{N}_D(y|x,\Sigma_*)$ et est bornée

$$\begin{aligned} Q_{\Sigma_*}(y|x) &= (2\pi)^{-d/2} \det(\Sigma_*)^{-1/2} \exp\left\{-\frac{1}{2}(y-x)^\top \Sigma_*^{-1}(y-x)\right\} \\ &\leqslant (2\pi)^{-d/2} \det(\Sigma_*)^{-1/2} =: M < \infty, \qquad \forall\, x,y \in \mathcal{X}\,. \end{aligned}$$

Soit J un rectangle contenant K_{2D}. Puisque K_{2D} est borné, alors J pourra être choisi borné également. Puisque $Q = Q_{\Sigma_*}$ est continue sur $\mathcal{X}$, Q sera continue sur J qui est compact. Ainsi, Q est borné par le bas sur J et la condition ε,δ (3.26) est satisfaite.

Toutes les conditions sont ainsi satisfaites, ce qui implique que l'algorithme BAM satisfait à la convergence bornée, et ce, peu importe le choix de la méthode d'adaptation dans K. En choisissant judicieusement la méthode d'adaptation, il sera aisé d'imposer l'adaptation diminuante à l'algorithme, ce qui assurera l'ergodicité.

Rosenthal et Yang (2018) généralisent légèrement le théorème 3.16 en affaiblissant d'abord la condition ε,δ (3.26), mais surtout en relâchant la condition de continuité des transitions par rapport à l'état de la chaîne $x \in \mathcal{X}$ au profit d'une condition de **combocontinuité** et en retirant complètement la condition de continuité par rapport au nouvel état de la chaîne $y \in \mathcal{X}$. Ceci permet donc une plus grande flexibilité dans le choix de la famille de transitions $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$.

Définition 3.3 (Combocontinuité, Rosenthal et Yang, 2018) *Une fonction f sur un espace S est dite combocontinue si elle peut être écrite comme une combinaison finie de fonctions continues. Posons $g_1, \ldots, g_m$ des fonctions continues sur S et $I : S \to \{1, \ldots, m\}$ une fonction déterminant l'index de la fonction à utiliser au point $x \in S$, alors la fonction f satisfait*

$$f(x) = g_{I(x)}(x) = \sum_{i=1}^{m} g_i(x)\, \mathbb{1}_i(I(x)).$$

La combocontinuité généralise en quelque sorte la notion de continuité par morceaux. En effet, si la fonction I est constante par morceaux, alors f est continue par morceaux. Cependant, la définition de combocontinuité est bien plus générale étant donné que la seule supposition sur I est qu'elle prend un nombre fini de valeurs ; la préimage de chaque $i \in \{1, \ldots, m\}$ peut être n'importe quel sous-ensemble du domaine. Cette définition permet la généralisation suivante du théorème 3.16.

Théorème 3.18 (Rosenthal et Yang, 2018, théorème 2) *Soit $\mathcal{X}$ un espace d'états avec une σ-algèbre de Borel $\mathcal{F}$, ayant une distance* dist *et une origine $\mathbf{0} \in \mathcal{X}$. Soit P une transition de Markov sur $\mathcal{X}$ admettant π comme distribution invariante et (Harris-)ergodique à π. On considère le processus stochastique $\{X_n\}_{n\geqslant 0}$ sur $\mathcal{X}$ tel qu'il satisfait aux conditions suivantes :*

(i) Le processus $\{X_n\}_{n\geqslant 0}$ ne bouge au maximum que d'une distance $D > 0$, i.e.

$$\mathbb{P}\left(\|X_{n+1} - X_n\|_2 > D\right) = 0;$$

en particulier, la transition P satisfait

$$P\left(\{y \in \mathcal{X} : \mathrm{dist}(x,y) \leqslant D\}|x\right) = 1, \qquad \forall\, x \in \mathcal{X};$$

(ii) Le processus $\{X_n\}_{n\geqslant 0}$ se déplace selon une transition fixe P lorsque la chaîne se trouve à l'extérieur d'un sous-ensemble compact $K \subset \mathcal{X}$. Sur K, la chaîne peut se déplacer selon n'importe quelle transition à n'importe quel point de K_D ;

(iii) La transition fixe P est bornée par le haut selon

$$P(\mathrm{d}y|x) \leqslant M\mu_*(\mathrm{d}y), \qquad \forall\, x \in K_D \setminus K, y \in K_{2D} \setminus K_D,$$

pour une constante finie $M < \infty$ et n'importe quelle mesure de probabilité μ_* concentrée sur $K_{2D} \setminus K_D$;

(iv) La transition fixe P est bornée par le bas selon

$$P^{n_0}(A|x) \geqslant \varepsilon \nu_*(A), \qquad \forall\, x \in K_{2D} \setminus K_D, A \in \mathcal{F},$$

pour un $n_0 \in \mathbb{N}$, une constante $\varepsilon > 0$ et n'importe quelle mesure de probabilité ν_* sur $\mathcal{X}$ telle que $\nu_* = \mu_*$ ou bien telle que P est réversible par rapport à π et $\mu_* = \pi|_{K_{2D}\setminus K_D}$, où $\mu|_A$ dénote la restriction de la mesure μ à l'ensemble A ;

(v) Pour une certaine mesure de référence λ sur $\mathcal{X}$, π admet une densité g par rapport à λ et, pour la famille de transitions $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ où $\mathcal{Y}$ est compact, on a une des conditions suivantes :

(v') Chaque P_γ admet une densité par rapport à λ, l'application $\gamma \mapsto p_\gamma(y|x)$ est continue en γ pour toute paire (x,y) et p_γ est combocontinue par rapport à x pour chaque y ;

(v") Chaque P_γ représente une transition Metropolis-Hastings avec proposition Q_γ admettant une densité q_γ par rapport à λ telle que l'application $\gamma \mapsto q_\gamma(y|x)$ est continue en γ pour toute paire (x,y) et q_γ est combocontinue par rapport à x pour chaque y.

Alors, la convergence bornée est satisfaite par l'algorithme qui est donc ergodique si l'adaptation diminuante est aussi vérifiée.

Alors que $\mathcal{X}$ est en général non-borné, les conditions (i) à (iv) du théorème 3.18 impliquent que le processus $\{X_n\}_{n\geqslant 0}$ est borné en probabilité (cf. Craiu et collab., 2015, théorème 5), i.e.

$$\lim_{L\to\infty} \sup_{n\in\mathbb{N}} \mathbb{P}\left(\mathrm{dist}(X_n,\mathbf{0}) > L | X_0 = x_0\right) = 0.$$

De plus, par Craiu et collab. (2015, proposition 5), la condition ε,δ (3.26) est un cas particulier de la condition (iv) avec $\mathcal{X}$ un sous-ensemble ouvert de $\mathbb{R}^d$ et $\nu_* = \text{uniforme}(K_{2D} \setminus K_D)$. On voit bien que la condition (v) est en fait exactement la condition de continuité dans le théorème 3.18 où λ est la mesure de Lebesgue, la continuité par rapport à y est omise et la continuité par rapport à x est remplacée par la combocontinuité. Ainsi, la « recette » pour définir des algorithmes adaptatifs satisfaisant la convergence bornée permet donc plus de liberté dans le choix des transitions P_γ (ou des propositions Q_γ dans le cas d'algorithmes M.-H) que ce que le théorème 3.16 prévoyait initialement.

3.4.1.3 Conditions de dérive géométrique

Pour une transition fixée P, l'ergodicité de la chaîne de Markov homogène résultante peut être assurée par deux conditions qui font en sorte que le rythme de convergence de la chaîne est borné géométriquement.

D'abord, la condition de **minorisation** (ou de $(1,\delta,\nu)$**-minorisation**) requiert l'existence d'un ensemble $C \in \mathcal{F}$ de probabilité positive telle que, pour tout $x \in C$, la transition $P(\cdot|x)$ soit bornée par le bas sur C. Cette condition est également parfois appelée condition de **récurrence positive vers un petit ensemble** C, puisque cette condition implique que la chaîne a une probabilité positive de demeurer dans C lorsqu'elle s'y trouve. Spécifiquement, on définit

Condition 3.6 (Minorisation apériodique forte) *Soit P une transition sur un espace d'états $(\mathcal{X},\mathcal{F})$. Il existe $C \in \mathcal{F}$, $\delta > 0$ et ν une mesure de probabilité concentrée sur C tels que*

$$P(\cdot|x) \geqslant \delta\nu(\cdot), \qquad \forall x \in C.$$

Ensuite, la condition de **dérive géométrique** requiert l'existence d'une fonction test $V \geqslant 1$ bornée sur un ensemble et qui contrôle P à l'extérieur de cet ensemble.

Condition 3.7 (Dérive géométrique) *Soit P une transition sur un espace d'états $(\mathcal{X},\mathcal{F})$ admettant π comme distribution stationnaire. Il existe $C \in \mathcal{F}$, $\lambda < 1$, $b < \infty$ et $V : \mathcal{X} \to [1,\infty)$ tels que $\sup_C V = v < \infty$ et $PV \leqslant \lambda V + b\,\mathbb{1}_C$, c.-à-d.,*

$$(PV)(x) \leqslant \lambda V(x) + b\,\mathbb{1}_C(x), \qquad \forall x \in C,$$

où

$$(PV)(x) = \mathbb{E}\{V(X_1)|X_0 = x\}$$

est l'espérance de $V(X_1)$ *pour* $X_1 \sim P(\cdot|x_0)$.

Suivant Meyn et Tweedie (1994, théorème 2.3, voir aussi Baxendale, 2005, théorème 1.1), ces deux conditions impliquent une convergence au moins géométrique entre la distribution marginale de la chaîne et la distribution stationnaire.

Proposition 3.19 (Roberts et Rosenthal, 2007, proposition 3) *Soit* P *une transition sur un espace d'états* $(\mathcal{X}, \mathcal{F})$ *admettant* π *comme distribution stationnaire telle que les conditions de minorisation apériodique forte 3.6 et de dérive géométrique 3.7 soient satisfaites pour* $C \in \mathcal{F}$, $V : \mathcal{X} \to [1,\infty)$, $\delta > 0$, $\lambda < 1$ *et* $b < \infty$ *tels que* $\sup_C V = v < \infty$. *Alors, il existe* $K < \infty$ *et* $\rho < 1$, *dépendant seulement de* δ, λ, b *et* v *tels que*

$$\left|\left|\left| P^n(\cdot|x) - \pi(\cdot) \right|\right|\right|_V \leqslant K\rho^n. \tag{3.27}$$

D'une manière équivalente,

$$\|P^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}} \leqslant KV(x)\rho^n, \qquad \forall\, x \in \mathcal{X}\,. \tag{3.28}$$

Maintenant, si un même choix de fonction test V, d'ensemble C et de constantes δ, λ, b permet d'établir l'ergodicité de tous les P_γ, alors le rythme de convergence pourra être borné de la même façon à travers $\mathcal{Y}$, tel qu'indiqué par (3.27) où K et ρ ne dependent de P qu'à travers ces choix. Ceci mène à la condition suivante qui est suffisante pour assurer l'ergodicité d'un algorithme MCMC adaptatif.

Condition 3.8 (Ergodicité géométrique apériodique forte simultanée, Roberts et collab., 1998) *Soit* $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ *une famille de transitions de Markov. Il existe* $C \in \mathcal{F}$, $V : \mathcal{X} \to [1,\infty)$, $\delta > 0$, $\lambda < 1$ *et* $b < \infty$ *tels que* $\sup_C V = v < \infty$ *et*

(i) (Minorisation uniforme) Pour chaque $\gamma \in \mathcal{Y}$, *il existe une mesure de probabilité* ν_γ *concentrée sur* C *avec* $P_\gamma(\cdot|x) \geqslant \delta\nu_\gamma(\cdot)$ *pour tout* $x \in C$;

(ii) (Dérive géométrique uniforme) Pour chaque $\gamma \in \mathcal{Y}$, $P_\gamma V \leqslant \lambda V + b\,\mathbb{1}_C$.

Théorème 3.20 (Roberts et Rosenthal, 2007, théorème 3) *Soit un algorithme MCMC adaptatif satisfaisant la condition d'adaptation diminuante 3.1 et telle que la famille* $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ *satisfait aux conditions d'ergodicité géométrique apériodique forte simultanée 3.8 avec* $\pi(V) < \infty$. *Alors, l'algorithme satisfait à la condition de convergence bornée 3.2 et est donc ergodique à* π.

Bai et collab. (2011b) proposent un ensemble de conditions pour vérifier la convergence bornée qui sont en fait suffisantes (voir Jarner et Hansen, 2000, lemme 3.5) à l'ergodicité géométrique apériodique forte simultanée (condition 3.8.) On définit d'abord la notion de **petit ensemble** pour ensuite énoncer le résultat.

Définition 3.4 (Petit ensemble, Meyn et Tweedie, 2009, section 5.2) *Soit* P *une transition de Markov sur un espace* $\mathcal{X}$. *Un ensemble* $C \subseteq \mathcal{X}$ *est dit* ***petit*** *(ou* ν_m***-petit***$)$ *s'il existe* $m > 0$ *et une mesure (non-triviale)* ν_m *sur* $\mathcal{B}(\mathcal{X})$ *telle que pour tout* $x \in C$ *et* $B \in \mathcal{B}(\mathcal{X})$

$$P^m(B|x) \geqslant \nu_m(B).$$

Proposition 3.21 (Bai et collab., 2011b, proposition 2.3) *Soit* $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ *une famille de transitions de Markov sur* $\mathcal{X}$. *Supposons que tous les ensembles compacts sont petits pour* $P_\gamma, \gamma \in \mathcal{Y}$, *et qu'il existe une fonction* $V : \mathcal{X} \to (1,\infty)$ *telle que*

$$\sup_{x\in C, \gamma\in\mathcal{Y}} P_\gamma V(x) < \infty, \qquad \forall\, C \subseteq \mathcal{X} \ \text{compact}, \tag{3.29}$$

et

$$\limsup_{\|x\|\to\infty} \sup_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < 1. \tag{3.30}$$

Alors, toute stratégie d'adaptation sur $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ *satisfait la convergence bornée (condition 3.2.)*

En supposant que tous les ensembles compacts de $\mathcal{X}$ soient petits, on assure alors la minorisation tout en retirant le lien entre la condition de minorisation et celle de dérive géométrique. La dérive géométrique peut être facilement retrouvée à partir de (3.29) et de (3.30). Ceci explique donc la suffisance des conditions de la proposition 3.21.

3.4.1.4 Conditions de dérive polynomiale

La condition de dérive géométrique, comme son nom l'indique, assure qu'il est possible de borner la convergence de chaque transition P_γ d'une manière géométrique (e.g. la borne $K\rho^n$ dans (3.27)). Ceci permet ensuite de borner uniformément la convergence. Ce rythme de convergence peut s'avérer trop exigeant comme condition ; un rythme plus lent, dit **sous-géométrique**, peut s'avérer suffisant pour borner uniformément la convergence. On considère ici des conditions de ce type.

Dans le même article phare de Roberts et Rosenthal (2007), on trouve la condition d'**ergodicité polynomiale uniforme** (voir aussi Bai, 2009b), inspirée de la dérive polynomiale considérée dans Fort et Moulines (2000a). On requiert la même condition de minorisation 3.6, mais la dérive géométrique est remplacée par une nombre fini q de conditions de dérive polynomiale.

Condition 3.9 (Ergodicité polynomiale uniforme, Roberts et Rosenthal, 2007) *On dit qu'une famille de transitions de Markov $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ satisfait la condition d'ergodicité polynomiale uniforme si*

(i) Pour chaque $\gamma \in \mathcal{Y}$, P_γ est π-irréductible et admet π comme distribution stationnaire ;

(ii) Il existe $C \in \mathcal{F}$, $m \in \mathbb{N}$, $\delta > 0$ et des mesures de probabilité ν_γ (pour chaque γ) tels que $\pi(C) > 0$ et

$$P_\gamma^m(\cdot|x) \geqslant \delta\nu_\gamma(\cdot), \qquad \forall x \in C, \gamma \in \mathcal{Y};$$

(iii) Il existe $q \in \mathbb{N}$, des fonctions mesurables $V_0, V_1, \ldots, V_q : \mathcal{X} \to (0,\infty)$ et, pour $k = 0, 1, \ldots, q-1$, il existe $0 < \alpha_k < 1$, $b_k < \infty$ et $c_k > 0$ satisfaisant

$$(P_\gamma V_{k+1})(x) \leqslant V_{k+1}(x) - V_k(x) + b_k \mathbb{1}_C(x), \qquad \forall\, x \in \mathcal{X}, \gamma \in \mathcal{Y}, \tag{3.31a}$$

$$V_k(x) \geqslant c_k, \qquad \forall\, x \in \mathcal{X}, \tag{3.31b}$$

$$V_k(x) - b_k \geqslant \alpha_k V_k(x), \qquad \forall\, x \in \mathcal{X}\setminus C, \tag{3.31c}$$

$$\sup_C V_q < \infty \tag{3.31d}$$

$$\pi(V_q^\beta) < \infty, \qquad \text{pour un } 0 < \beta \leqslant 1. \tag{3.31e}$$

Fort et Moulines (2000a) étudient ce type de condition et montrent qu'ils engendrent en effet une borne polynomiale à l'ergodicité. Les détails sont fastidieux et hors de l'intérêt de ce texte, mais l'existence de ce résultat permet de borner uniformément l'ergodicité des transitions. On obtient ainsi le résultat suivant.

Théorème 3.22 (Roberts et Rosenthal, 2007, théorème 4) *Soit un algorithme MCMC adaptatif satisfaisant la condition d'adaptation diminuante 3.1 et telle que la famille $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ satisfait aux conditions de dérive polynomiale simultanée 3.9. Alors, l'algorithme satisfait à la condition de convergence bornée 3.2 et est donc ergodique à π.*

Ce type de condition est en pratique très difficile à vérifier. Cependant, plusieurs cas particuliers peuvent être étudiés lorsque q est petit. Pour $q = 1$ et en supposant que $V_1 = V^\alpha$, où $V = V_0$ et pour un certain $\alpha \in (0,1)$, on obtient la condition suivante proposée par Bai et collab. (2011b).

Condition 3.10 (Dérive polynomiale simultanée I, Bai et collab., 2011b, proposition 2.4) *Soit $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ une famille de transitions de Markov sur $\mathcal{X}$. Supposons qu'il existe $C \subset \mathcal{X}$ tel que $\pi(C) > 0$, $m \in \mathbb{N}$, $\delta > 0$ et des mesures de probabilité ν_γ sur $\mathcal{X}$ tels que*

$$P_\gamma^m(\cdot|x) \geqslant \delta\, \mathbb{1}_C(x)\nu_\gamma(\cdot), \qquad \forall\, \gamma \in \mathcal{Y}.$$

Supposons aussi qu'il existe $\alpha \in (0,1)$, $\beta \in (0,1]$, $b > 0$, $c > 0$ *et une fonction mesurable* $V : \mathcal{X} \to [1,\infty)$ *avec* $cV(x) > b$ *sur* $\mathcal{X} \setminus C$, $\sup_C V(x) < \infty$ *et* $\pi(V^\beta) < \infty$ *tels que*

$$P_\gamma V \leqslant V - cV^\alpha + b\,\mathbb{1}_C\,, \qquad \forall\,\gamma \in \mathcal{Y}\,.$$

Sous ce type de condition, Jarner et Roberts (2002, théorème 3.6) montrent que la convergence suit un rythme polynomial d'ordre $1-\beta$. Spécifiquement, pour $V_\beta = V^{1-\beta(1-\alpha)}$,

$$(n+1)^{\beta-1} |||P^n - \pi|||_{V_\beta} \to 0, \qquad n \to \infty.$$

Tout comme la proposition 3.19 le faisait avec la dérive géométrique, ce résultat permet de borner uniformément la convergence et d'assurer du coup la convergence bornée.

Proposition 3.23 (Bai et collab., 2011b, proposition 4) *Soit un algorithme MCMC adaptatif satisfaisant la Condition de dérive polynomiale simultanée II 3.11. Alors, toute stratégie d'adaptation n'utilisant que* $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ *satisfait la convergence bornée 3.2.*

Si, toujours avec $q = 1$, on choisit plutôt $V_0 \equiv 1$ et $V_1 = V$, on trouve la condition suivante, suggérée par Yang (2008b).

Condition 3.11 (Dérive polynomiale simultanée II, Yang, 2008b, théorème 3.1) *Soit un algorithme MCMC adaptatif tel qu'il existe* $C \in \mathcal{F}$, $V : \mathcal{X} \to [1,\infty)$ *mesurable avec* $\pi(V) < \infty$ *et* $\sup_C V = v < \infty$, $\delta > 0$ *et* $b < \infty$ *pour lesquels on a les propriétés suivantes :*

(i) Pour chaque $\gamma \in \mathcal{Y}$, *il existe une mesure de probabilité* ν_γ *sur* C *avec*

$$P_\gamma(\cdot|x) \geqslant \delta\nu_\gamma(\cdot), \qquad \forall x \in C;$$

(ii) Pour chaque $\gamma \in \mathcal{Y}$, *on a*

$$P_\gamma V \leqslant V - 1 + b\,\mathbb{1}_C;$$

(iii) L'ensemble Δ_C *est compact par rapport à la distance* $\mathrm{dist}(\cdot,\cdot)$, *où*

$$\Delta_C = \{\gamma \in \mathcal{Y}\,|P_\gamma V \leqslant V - 1 + b\,\mathbb{1}_C\};$$

(iv) Le processus $\{V(X_n)\}_{n\geqslant 0}$ *est borné en probabilité conditionnellement à* $X_0 = x_*$ *et à* $\Gamma_0 = \gamma_0$.

Théorème 3.24 (Yang, 2008b, théorème 3.1) *Soit* $\mathcal{Y} \subseteq \mathbb{R}^q$ *un espace métrique et un algorithme MCMC adaptatif tel que l'adaptation diminuante 3.1 et la condition 3.11 soient satisfaites. Alors, pour le choix de* x_* *et de* γ_* *dans la condition 3.11(iv), on a*

$$\lim_{n\to\infty} T(x_*,\gamma_*,n) = 0.$$

Une condition inspirée d'Atchadé et Fort (2010, condition A2) est suffisante pour montrer la convergence bornée par Atchadé et Fort (2010, proposition 2.4) en supposant une adaptation markovienne. On considère ici une reformulation légèrement plus forte de Bai et collab. (2011b).

Condition 3.12 (Dérive polynomiale simultanée III, Bai et collab., 2011b, conditions M1 à M3) *Soit un algorithme MCMC adaptatif tel que les propriétés suivantes sont satisfaites :*

(i) Il existe une mesure de probabilité ν, *une constante* $\delta > 0$ *et un ensemble* $C \in \mathcal{F}$ *tels que*

$$P_\gamma(\cdot|x) \geqslant \delta\,\mathbb{1}_C(x)\nu(\cdot);$$

(ii) Il existe une fonction mesurable $V : \mathcal{X} \to [1,\infty)$ *et* $b > 0$ *tels que*

$$(P_\gamma V)(x) - V(x) \leqslant -1 + b\,\mathbb{1}\,C(x), \qquad \forall\gamma \in \mathcal{Y};$$

(iii) Pour tout sous-ensemble de niveau l *de* V,

$$\mathcal{D}_l = \{x \in \mathcal{X} : V(x) \leqslant l\}\,,$$

on a

$$\lim_{n\to\infty} \sup_{\mathcal{D}_l\times\mathcal{Y}} \|P_\gamma^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}}.$$

Proposition 3.25 (Atchadé et Fort, 2010, proposition 2.4) *Soit un algorithme MCMC adaptatif à adaptation markovienne tel que l'adaptation diminuante 3.1 et la condition 3.12 soient satisfaites. Alors, l'algorithme est ergodique.*

3.4.1.5 Étude de cas – algorithme Metropolis adaptatif

Que ce soit via l'ergodicité géométrique uniforme (condition 3.8) ou une ergodicité polynomiale uniforme (e.g. condition 3.9), vérifier la condition de convergence bornée 3.2 requiert une étude approfondie de la distribution cible, des transitions P_γ ainsi que de la manière dont varient les transitions par rapport à $\gamma \in \mathcal{Y}$.

Dans cette section, on considère un algorithme Metropolis de type marche aléatoire et certaines classes de distributions cibles afin de montrer comment, en pratique, il est possible de vérifier la convergence bornée. La majeure partie de cette discussion est inspirée de Bai et collab. (2011b, section 5).

On considère d'abord deux conditions générales sur la distribution cible. On requiert que la distribution cible admette une densité π qui soit positive (non-nulle) et bornée sur tout ensemble compact.

Condition 3.13 (Régularité de la distribution cible) *La distribution cible admet une densité π par rapport à la mesure de Lebesgue qui est positive et bornée sur tout sous-ensemble compact mesurable de $\mathcal{X}$.*

On requiert également que la densité π soit continuellement différentiable et que sa décroissance soit rapide pour de grands x.

Condition 3.14 (Décroissance forte de la densité cible) *La densité cible π admet des dérivés premières continues $\nabla\pi$ et satisfait*

$$\limsup_{\|x\|_2\to\infty} \left(\frac{x}{\|x\|_2}\right)^\top \left(\frac{\nabla\pi(x)}{\|\nabla\pi(x)\|_2}\right) =: \limsup_{\|x\|_2\to\infty} n(x)^\top m(x) < 0.$$

La décroissance forte (condition 3.14) correspond à la définition que la densité π admette des **contours réguliers**. Le vecteur $x/\|x\|_2$ pointe vers l'origine, alors que le vecteur $\nabla\pi(x)/\|\nabla\pi(x)\|_2$ pointe dans la direction de la plus grande croissance de π. Le produit scalaire entre ces deux vecteurs est donc négatif si π a tendance à décroitre vers les grands x. En considérant la $\limsup_{\|x\|_2\to\infty}$ on trouve que, pour les grands x, la densité π décroit dans toutes les directions, d'où le nom de contours réguliers.

La famille de transitions $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ est définie par la transition Metropolis de type marche aléatoire

$$P_\gamma(\,\mathrm{d}y|x) = \alpha(y|x)Q_\gamma(\,\mathrm{d}y|x) + (1-\alpha(y|x))\,\delta_x(y),$$

où Q_γ est la distribution de transition qui admet une densité par rapport à la mesure de Lebesgue, i.e.

$$Q_\gamma(\,\mathrm{d}y|x) = q_\gamma(y|x)\lambda(\,\mathrm{d}y),$$

et $\alpha(y|x)$ est la probabilité d'acceptation Metropolis

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}.$$

Par construction, les transitions de Metropolis-Hastings admettent la densité cible π comme distribution stationnaire.

La plupart des résultats d'ergodicité requièrent que la transition soit ϕ-irréductible et apériodique. En supposant que la condition suivante sur les densités de proposition,

Condition 3.15 (Positivité locale de la densité de proposition) *Pour chaque $\gamma \in \mathcal{Y}$ il existe $\delta = \delta(\gamma) > 0$ et $\varepsilon = \varepsilon(\gamma) > 0$ tels que*

$$\|x - y\|_2 \leqslant \delta \quad \Rightarrow \quad q_\gamma(y|x) \geqslant \varepsilon.$$

Roberts et Tweedie (1996b, théorème 2.2) montrent que si π satisfait la condition 3.13 alors chaque q_γ est λ-irréductible et apériodique et tous les ensembles compacts non-nuls C sont petits (définition 3.4.) On requiert également la symétrie de la densité de proposition.

Condition 3.16 (Symétrie de la densité de proposition) *Pour chaque $\gamma \in \mathcal{Y}$, la densité de proposition q_γ satisfait*

$$q_\gamma(y|x) = q_\gamma(y - x|\mathbf{0}) = q_\gamma(x|y).$$

Puisque π est une densité sur $\mathcal{X} \subseteq \mathbb{R}^d$, on a évidemment que $\lim_{\|x\|_2 \to \infty} \pi(x) = 0$ afin d'assurer l'intégrabilité à 1. Cependant, le rythme auquel la densité tend vers 0 influence le rythme de convergence d'une chaîne de Markov homogène vers la distribution stationnaire. Évidemment, cette propriété se transpose aux algorithme MCMC adaptatifs. On considère donc les conditions suivantes qui définissent différents rythmes de décroissances des ailes de π.

Une densité π dont le comportement pour grands $\|x\|_2$ est tel que $\log \pi(x) \approx -\beta \|x\|_2$ pour un $\beta > 0$ est dite à ailes **exponentielles**.

Condition 3.17 (Ailes exponentielles) *La densité π à support $\mathcal{X} \subseteq \mathbb{R}^d$ a des ailes exponentielles si elle admet des dérivés premières continues telles que*

$$\limsup_{\|x\|_2 \to \infty} n(x)^\top (\nabla \log \pi(x)) < 0,$$

où $n(x) = x/\|x\|_2$.

Une densité π dont le comportement pour grands $\|x\|_2$ est tel que $\log \pi(x) \sim -\|x\|_2^s$ pour un $s > 1$ est dite à ailes **super-exponentielles**.

Condition 3.18 (Ailes super-exponentielles) *La densité π à support $\mathcal{X} \subseteq \mathbb{R}^d$ a des ailes super-exponentielles si elle admet des dérivés premières continues telles que*

$$\limsup_{\|x\|_2 \to \infty} n(x)^\top (\nabla \log \pi(x)) = -\infty,$$

où $n(x) = x/\|x\|_2$.

Une densité π dont le comportement pour grands $\|x\|_2$ est tel que $\log \pi(x) \sim -\|x\|_2^s$ pour un certain $0 < s < 1$ est dite à ailes **hyperboliques** (ou parfois **sous-exponentielles**.) Cette définition a été initialement introduite par Fort et Moulines (2000b), où l'ergodicité d'algorithmes Metropolis-Hastings (non-adaptatifs) est démontrée sous ce genre de condition.

Condition 3.19 (Ailes hyperboliques) *La densité π à support $\mathcal{X} \subseteq \mathbb{R}^d$ a des ailes hyperboliques si elle admet des dérivés secondes $\nabla^2 \pi$ continues et qu'il existe $0 < m < 1$, d_i, D_i pour $i = 0, 1, 2$ des constantes positives finies telles que pour $\|x\|_2$ suffisamment grand*

$$0 < d_0 \|x\|_2^m \leqslant -\log \pi(x) \leqslant D_0 \|x\|^m; \tag{3.32a}$$

$$0 < d_1 \|x\|_2^{m-1} \leqslant \|\nabla \log \pi(x)\|_2 \leqslant D_1 \|x\|^{m-1}; \tag{3.32b}$$

$$0 < d_2 \|x\|_2^{m-2} \leqslant \left\|\nabla^2 \log \pi(x)\right\|_F \leqslant D_2 \|x\|^{m-2}. \tag{3.32c}$$

Additionnellement à la condition d'ailes exponentielles 3.17, la condition suivante, sur la relation entre la densité cible et les densités de transition, est requise pour assurer la convergence bornée. On considère l'ensemble suivant

$$A(\delta,\Delta,u,\varepsilon) := \left\{ z = a\xi : \delta \leqslant a \leqslant \Delta, \xi \in S^{d-1}, \|\xi - u\|_2 < \frac{\varepsilon}{3} \right\}, \qquad u \in S^{d-1},$$

où S^{d-1} dénote l'hyper-sphère en dimension d de rayon 1. L'ensemble $A(\delta,\Delta,x,\varepsilon)$ comporte les points en direction de $u \in S^{d-1}$ se trouvant à moins de $\varepsilon/3$ de u sur S^{d-1} et tels que $\|z\|_2 \in [\delta,\Delta]$. Il s'agit donc d'un cône autour de u tronqué à l'extérieur de $[\delta,\Delta]$.

Condition 3.20 (Bai et collab., 2011b, condition 5.5) *Soit π vérifiant la condition d'ailes exponentielles 3.17 et de décroissance forte 3.14 avec*

$$\eta_1 := -\limsup_{\|x\|_2 \to \infty} n(x)^\top m(x), \qquad \eta_2 := -\limsup_{\|x\|_2 \to \infty} n(x)^\top \nabla \log \pi(x),$$

où $n(x) = x/\|x\|_2$ et $m(x) = \nabla\pi(x)/\|\nabla\pi(x)\|_2$. Supposons qu'il existe $\varepsilon \in (0,\eta_1)$, $\beta \in (0,\eta_2)$, δ et Δ tels que

$$0 < \frac{3}{\beta\varepsilon} \leqslant \delta < \Delta \leqslant \infty$$

de sorte que pour n'importe qu'elle séquence $\{(x_n,\gamma_n)\}_{n\geqslant 0}$, avec $\|x_n\|_2 \to \infty$ et $\{\gamma_n\}_{n\geqslant 0} \subset \mathcal{Y}$, il existe une sous-séquence $\{(x_{n_k},\gamma_{n_k})\}_{k\geqslant 0}$, avec $\|x_{n_k}\|_2 \to \infty$, qui satisfait

$$\lim_{k\to\infty} \int_{A(\delta,\Delta,n(x_{n_k}),\varepsilon)} \|z\|_2 q_{\gamma_{n_k}}(z|\mathbf{0})\lambda(\,\mathrm{d}z) > \frac{3}{\beta\varepsilon(e-1)}.$$

D'une manière équivalente,

$$\int_{A(\delta,\Delta,u,\varepsilon)} \|z\|_2 q_\gamma(z|\mathbf{0})\lambda(\,\mathrm{d}z) > \frac{3}{\beta\varepsilon(e-1)}, \qquad \forall\,(u,\gamma) \in S^{d-1}\times\mathcal{Y}.$$

La condition 3.20, bien qu'à l'apparence technique, assure intuitivement que les ailes de la densité de proposition soient suffisamment lourdes comparativement aux ailes exponentielles de la densité cible, et ce, uniformément par rapport à la direction $u \in S^{d-1}$ et, surtout, par rapport au paramètre de la densité de proposition $\gamma \in \mathcal{Y}$. La constante $\delta > 0$ produit une borne inférieure sur le domaine d'intégration ; cette condition impose donc que la famille de propositions soit positive au-delà de δ. Ainsi, puisque les densité de proposition auront toutes des ailes assez lourdes pour visiter tout le support de π, il n'est pas surprenant que l'algorithme adaptatif résultant soit ergodique.

Proposition 3.26 (Bai et collab., 2011b, proposition 5.4 et théorème 5.5) *Soit π une densité cible régulière (condition 3.13), fortement décroissante (condition 3.14) et aux ailes exponentielles (condition 3.17.) Un algorithme Metropolis adaptatif tel que chaque proposition Q_γ admet une densité symétrique (condition 3.16) et localement positive (condition 3.15) pour chaque $\gamma \in \mathcal{Y}$ et tel que la condition 3.20 est satisfaite vérifie la convergence bornée (condition 3.2.) L'algorithme est donc ergodique à π dès que l'adaptation diminuante (condition 3.1) est satisfaite.*

Sans entrer dans les détails de la preuve, notons surtout que la convergence bornée est montrée en utilisant la proposition 3.21 avec la fonction test $V(x) = c\pi^{-s}(x)$ pour certains $c > 0$ et $s \in (0,1)$.

Maintenant, si la densité cible π admet des ailes super-exponentielles, il est clair que la densité satisfait également aux conditions d'ailes exponentielles. Dans la condition 3.20 on trouve alors $\eta_2 = \infty$, ce qui ne borne plus β qui pourra être choisi arbitrairement grand (même infini). Il sera donc possible de choisir $\delta > 0$ arbitrairement petit. Ainsi, on trouve la condition suivante qui sera suffisante pour montrer la convergence bornée.

Condition 3.21 (Bai et collab., 2011b, condition 5.6) *Soit π vérifiant la condition d'ailes*

super-exponentielles 3.18 et de décroissance forte 3.14 avec

$$\eta := -\limsup_{\|x\|_2 \to \infty} n(x)^\top m(x).$$

Supposons qu'il existe $\varepsilon \in (0,\eta)$, δ *et* Δ *avec* $0 < \delta < \Delta \leqslant \infty$ *tels que pour toute séquence* $\{(x_n,\gamma_n)\}_{n\geqslant 1} \subset (\mathcal{X}\times\mathcal{Y})^{\mathbb{N}}$ *avec* $\|x_n\|_2 \to \infty$, *il existe une sous-séquence* $\{(x_{n_k},\gamma_{n_k})\}_{k\geqslant 1}$ *avec* $\|x_{n_k}\|_2 \to \infty$ *telle que*

$$\liminf_{k\to\infty} \int_{A(\delta,\Delta,n(x_{n_k}),\varepsilon)} q_{\gamma_{n_k}}(z|\mathbf{0})\lambda(\,\mathrm{d}z) > 0.$$

D'une manière équivalente,

$$\inf_{(u,\gamma)\in S^{d-1}\times\mathcal{Y}} \int_{A(\delta,\Delta,u,\varepsilon)} q_\gamma(z|\mathbf{0})\lambda(\,\mathrm{d}z) > 0.$$

Proposition 3.27 (Bai et collab., 2011b, théorème 5.7) *Soit* π *une densité cible régulière (condition 3.13), fortement décroissante (condition 3.14)et aux ailes super-exponentielles (condition 3.18.) Un algorithme Metropolis adaptatif tel que chaque proposition* Q_γ *admet une densité symétrique (condition 3.16) et localement positive (condition 3.15) pour chaque* $\gamma \in \mathcal{Y}$ *et tel que la condition 3.21 est satisfaite vérifie la convergence bornée (condition 3.2.) L'algorithme est donc ergodique à* π *dès que l'adaptation diminuante (condition 3.1) est satisfaite.*

La condition 3.21 impose surtout que la famille de propositions ne comporte pas de densités qui tendent à se concentrer à l'intérieur d'un $\delta > 0$. Par exemple, pour γ correspondant à la variance d'une densité gaussienne unidimensionnelle, si $\gamma \to 0$ alors cette condition ne sera pas vérifiée. On peut corriger cette situation en se limitant, par exemple, à $\gamma \geqslant a$ pour un certain $a > 0$ arbitrairement petit mais fixé.

Bai et collab. (2011b) proposent également une façon de vérifier les conditions 3.20 et 3.21 en bornant par le bas les densités de proposition pour grands $\|x\|_2$ uniformément sur $\gamma \in \mathcal{Y}$. Le résultat suivant traite du cas super-exponentiel, mais un résultat similaire peut également être obtenu dans le cas exponentiel (Bai et collab., 2011b, lemme 5.3.)

Lemme 3.28 (Bai et collab., 2011b, lemme 5.6) *Soit* π *une densité cible régulière (condition 3.13), fortement décroissante (condition 3.14) et aux ailes super-exponentielles (condition 3.18) et un algorithme Metropolis adaptatif aux propositions* $\{Q_\gamma\}_{\gamma\in\mathcal{Y}}$ *admettant chacune une densité* q_γ. *Supposons qu'il existe* $M > 0$ *tel qu'il existe une fonction positive* $q^- : \mathcal{X} \to \mathbb{R}^+$ *satisfaisant*

$$q_\gamma(z) \geqslant q^-(z) > 0, \qquad \forall\, \|z\|_2 > M, \gamma \in \mathcal{Y}\,.$$

Alors, la condition 3.21 est satisfaite.

Ce genre de condition peut être artificiellement implantée dans un algorithme. On considère q_γ comme un mélange aux poids fixes entre une composante fixe à support sur $\mathcal{X}$ et une composante adaptative. Ainsi, q_γ est évidemment bornée par le bas par la composante fixe sur $\mathcal{X}$, ce qui satisfait aux conditions du lemme 3.28. Le poids de la composante fixe est choisi arbitrairement petit afin que q_γ ne soit pratiquement constituée que de la composante adaptative. Similairement, imposer certaines bornes à l'espace $\mathcal{Y}$ peut s'avérer suffisant afin de trouver la fonction q^-.

Finalement, lorsque la densité cible n'a que des ailes hyperboliques, il est tout de même possible de trouver des conditions sur la famille de densités de proposition qui seront suffisantes pour assurer l'ergodicité. Ce rythme de décroissance particulièrement lent impose cependant des conditions beaucoup plus fortes que dans le cas d'ailes exponentielles. Bai et collab. (2011b) proposent les conditions suivantes. On suppose que la famille de densités de proposition admette un support uniformément compact et que chaque densité soit uniformément bornée par le haut.

Condition 3.22 (Support uniformément compact, Bai et collab., 2011b, condition 5.7) *Il existe $M > 0$ tel que*

$$q_\gamma(z) = 0, \qquad \forall \, \|z\|_2 > M, \gamma \in \mathcal{Y}\,.$$

Condition 3.23 (Densités uniformément bornées) *Il existe une fonction $q^+ : \mathcal{X} \to \mathbb{R}^+$ telle que $\lambda(q^+) < \infty$ et*

$$q_\gamma(z) \leqslant q^+(z), \qquad \forall\, \gamma \in \mathcal{Y}\,.$$

Proposition 3.29 (Bai et collab., 2011b, théorème 5.13) *Soit π une densité cible régulière (condition 3.13), fortement décroissante (condition 3.14) et aux ailes hyperboliques (condition 3.18.) Un algorithme Metropolis adaptatif tel que chaque proposition Q_γ admet des densité symétrique (condition 3.16), localement positive (condition 3.15), uniformément bornées (condition 3.23) et au support uniformément compact (condition 3.22) vérifie la convergence bornée (condition 3.2.) L'algorithme est donc ergodique à π dès que l'adaptation diminuante (condition 3.1) est satisfaite.*

La preuve de ce résultat fait usage d'une condition de dérive polynomiale (condition 3.10, proposition 3.23) avec la fonction test $V_s(x) = (-\log \pi(x))^s$.

Un support uniformément compact peut être facilement implanté en tronquant les densités de transition au-delà de $M > 0$, comme dans l'algorithme BAM 3.13. Cependant, borner uniformément les densités par le haut est un peu plus difficile. Une manière de procéder est en considérant un espace des paramètres $\mathcal{Y}$ compact.

L'exemple 3.7 traite de l'ergodicité de l'algorithme AM 3.1 lorsque π admet des ailes super-exponentielles. Puisque le choix de la famille de densités de proposition est fixé, seul le choix des conditions sur la densité cible influenceront l'ergodicité. On doit cependant ajouter la condition que l'espace des paramètres $\mathcal{Y}$ soit borné afin de faciliter la vérification de la condition 3.21. Saksman et Vihola (2010, voir aussi Andrieu et Moulines, 2006) ne supposent pas que l'espace $\mathcal{Y}$ est borné pour montrer l'ergodicité. Pour y arriver, ils considèrent une couverture compacte de $\mathcal{Y}$; intuitivement, ils requièrent que la covariance ne soit bornée seulement qu'en probabilité. Cette construction sera étudiée à la sous-section 3.2.3.

Supposer que l'espace des paramètres est compact est souvent bien suffisant pour assurer l'ergodicité d'un algorithme Metropolis adaptatif. Par exemple, Atchadé et Fort (2010) démontrent l'ergodicité de l'algorithme ASWAM 3.10, où l'échelle s_d est aussi adaptée, en supposant des ailes seulement hyperboliques à la densité cible π. Une condition supplémentaire (Atchadé et Fort, 2010, D3) est évidemment requise : le comportement asymptotique ($\|x\|_2 \to \infty$) de la probabilité d'acceptation moyenne de l'algorithme doit être contrôlé d'une certaine manière.

Une version modifiée de l'algorithme AM a également été suggérée par Roberts et Rosenthal (2009) ; l'exemple 3.8 traite de l'ergodicité de cet algorithme. Cette reformulation montre bien qu'un design judicieux d'un algorithme adaptatif peut grandement simplifier le traitement théorique, par opposition à tenter de prouver l'ergodicité après avoir défini l'algorithme.

Exemple 3.7 Vérification de la convergence bornée de l'algorithme AM 3.1 (Haario et collab., 2001)

L'algorithme AM consiste en une marche aléatoire Metropolis à proposition gaussienne adaptée selon la covariance du passé de la chaîne. Pour $n \geqslant n_0$, on a

$$\begin{aligned} Y|x_n, \Sigma_n &\sim \mathcal{N}_d\left(x_n, s_d\Sigma_n\right); \\ X_{n+1} &= \begin{cases} y, & \text{avec prob. } \alpha(y|x_n), \\ x_n, & \text{avec prob. } 1-\alpha(y|x_n); \end{cases} \\ \Sigma_{n+1} &= \mathrm{Cov}(x_{0:n+1}) + \kappa I_d, \end{aligned}$$

où $s_d, \kappa > 0$ sont des constantes positives (e.g. $s_d = (2{,}38)^2/d$ et $\kappa = 10^{-3}$.) Dans ce cas, l'espace des paramètres de la densité de proposition $\mathcal{Y}$ est constitué de l'ensemble des matrices Σ de dimension $d \times d$ définies positives dont la diagonale est bornée par le bas par $\kappa > 0$.

On dénote $q_\Sigma(y|\mu) = \varphi(y|\mu,\Sigma)$, la densité gaussienne de dimension d, à moyenne μ et à covariance Σ donnée par

$$\varphi(y|\mu,\Sigma) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}.$$

Cette densité est positive sur $\mathbb{R}^d$ et continue ; la condition de positivité locale 3.15 est donc triviale à vérifier puisque $\|x-y\|_2 \leqslant \delta$ définit un ensemble compact. Ensuite, la densité gaussienne est évidemment symétrique (condition 3.16) puisque fonction de $(x-y)$ seulement à travers la forme quadratique $(x-y)^\top \Sigma^{-1}(x-y)$:

$$q_\Sigma(y|x) = \varphi(y|x,\Sigma) = \varphi(x|y,\Sigma) = q_\Sigma(x|y),$$
$$q_\Sigma(y-x) := \varphi(y-x|\mathbf{0},\Sigma).$$

À partir de cette définition de l'algorithme, on tente de dégager les conditions les moins fortes possibles sur π qui assureront l'ergodicité de l'algorithme. Plusieurs auteurs se sont déjà prêté à l'exercice de démontrer l'ergodicité de cet algorithme.

Dans l'article initial de Haario et collab. (2001), l'ergodicité est montrée en supposant que le support de π, donné par $\mathcal{X} \subset \mathbb{R}^d$, est borné et que π est bornée par le haut. Andrieu et Atchadé (2007) et Roberts et Rosenthal (2007) considèrent également cet algorithme et imposent les mêmes conditions afin de montrer l'ergodicité. La condition du support borné est particulièrement forte et les résultats énoncés dans cette section semblent indiquer que l'on pourrait faire mieux. Le problème principal que l'on rencontrera est que Σ devra être borné ; lorsque le support est borné, il est possible de montrer que Σ_n tel que défini dans les récursions sera lui aussi borné.

Saksman et Vihola (2010) ainsi que Andrieu et Moulines (2006) supposent quant à eux que π est bornée par le haut, bornée par le bas sur tout ensemble compact (donc régulière, condition 3.13), différentiable, fortement décroissante (condition 3.14) et admet des ailes super-exponentielles (condition 3.18.) Ces conditions sont exactement celles de la Proposition 3.27 où il ne manquerait que la condition 3.21 sur la famille de propositions pour conclure à l'ergodicité. Pour ce, on considère l'intégrale

$$\int_{A(\delta,\Delta,u,\varepsilon)} q_\Sigma(z)\lambda(\,\mathrm{d}z), \tag{3.33}$$

que l'on doit borner positivement par le bas uniformément sur $\Sigma \in \mathcal{Y}$ et $u \in S^{d-1}$ selon un certain choix de $0 < \delta \leqslant \Delta$, $\varepsilon \in (0,\eta)$, où $\eta > 0$ est donné par la condition d'ailes super-exponentielles (condition 3.18.) Ceci sera possible principalement puisque Σ est borné par le bas par κI_d.

On cherche à borner q_Σ par le bas sur $A(\delta,\Delta,u,\varepsilon)$ uniformément par rapport à u et à Σ. Dans l'expression de q_Σ, deux termes seront à minimiser, soit

$$(\det\Sigma)^{-1/2} \quad \text{et} \quad -\frac{1}{2}z^\top\Sigma^{-1}z.$$

Pour Σ définie positive, son déterminant pourra être arbitrairement élevé et, donc $(\det\Sigma)^{-1/2}$ pourra être arbitrairement petit. On est donc forcé d'admettre une borne supérieure sur Σ ; on posera que toutes les valeurs propres de Σ sont bornées par $L > 0$. Ainsi,

$$(\det\Sigma)^{-1/2} \geqslant L^{-d/2}.$$

Pour l'argument de l'exponentielle $-\frac{1}{2}z^\top\Sigma^{-1}z$, on considère la décomposition spectrale de la matrice Σ^{-1}, donnée par

$$\Sigma^{-1} = S^\top \Lambda S = \sum_{i=1}^d \lambda_i s_i s_i^\top,$$

où $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ est une matrice diagonale contenant les valeurs propres de Σ^{-1} et S est une matrice orthogonale ($S^\top S = SS^\top = I_d$) dont les colonnes $s_1, \ldots, s_d$ sont des vecteurs propres orthogonaux unitaires

de Σ^{-1}. En particulier, on trouve

$$z^\top \Sigma^{-1} z = \sum_{i=1}^{d} \lambda_i z^\top s_i s_i^\top z = \sum_{i=1}^{d} \lambda_i (z^\top s_i)^2 \leqslant \lambda_{\max}(\Sigma^{-1}) \sum_{i=1}^{d} (z^\top s_i)^2.$$

On note que S est orthogonale, d'où

$$\sum_{i=1}^{d} (z^\top s_i)^2 = (S^\top z)^\top (S^\top z) = z^\top S S^\top z = z^\top z = \|z\|_2^2.$$

Ainsi, il est possible de borner le terme exponentiel dans la densité normale

$$-\frac{1}{2} z\Sigma^{-1} z \geqslant -\frac{1}{2} \lambda_{\max}(\Sigma^{-1}) \|z\|_2^2,$$

qui est uniforme par rapport à u.

Puisque Σ est définie positive, alors toutes les valeurs propres de Σ sont positives et, donc, la plus petite valeur propre de Σ est exactement $\lambda_{\max}^{-1}(\Sigma^{-1})$. D'où

$$-\frac{1}{2} z\Sigma^{-1} z \geqslant -\frac{1}{2} \lambda_{\min}^{-1}(\Sigma) \|z\|_2^2,$$

On cherche donc la valeur maximale que peut prendre $\lambda_{\min}^{-1}(\Sigma)$, donc la valeur minimale de $\lambda_{\min}(\Sigma)$. Puisqu'on peut écrire $\Sigma = C + \kappa I_d$ où C est une matrice $d \times d$ définie positive, une valeur propre λ de Σ est telle qu'il existe un vecteur v avec $Cv + \kappa v = \lambda v$. donc $Cv = (\lambda - \kappa)v$, c'est-à-dire que $\lambda - \kappa$ est une valeur propre de C. Étant donné que C est définie positive, ses valeurs propres sont positives et on trouve alors $\lambda > \kappa$ ce qui permet de borner $\lambda_{\min} > \kappa$. Enfin, on trouve

$$-\frac{1}{2} z\Sigma^{-1} z \geqslant -\frac{1}{2} \kappa^{-1} \|z\|_2^2.$$

Cette borne et celle sur $(\det \Sigma)^{-1/2}$ permettent de vérifier le lemme 3.28 ; la condition 3.21 est cependant simple à vérifier également à partir d'ici. Sur $A(\delta,\Delta,u,\varepsilon)$, on a $\|z\|_2 \leqslant \Delta$, d'où

$$-\frac{1}{2} z\Sigma^{-1} z \geqslant -\frac{1}{2} \kappa^{-1} \Delta^2,$$

En bornant de la sorte les deux termes qui dépendent de Σ et de u, on obtient donc une borne sur l'intégrale (3.33). Si $A = A(\delta,\Delta,u,\varepsilon)$, alors

$$\int_A q_\Sigma(z) \lambda(\mathrm{d}z) \geqslant (2\pi L)^{-d/2} \exp\left\{-\frac{1}{2} \kappa^{-1} \Delta^2\right\} \lambda(A), \qquad \forall\, u \in S^{d-1}, \Sigma \in \mathcal{Y}.$$

Ainsi, en supposant que $\mathcal{Y}$ est borné, il est possible de montrer la condition 3.21 et donc la convergence bornée de l'algorithme AM lorsque π admet des ailes super-exponentielles.

Exemple 3.8 Vérification de la convergence bornée de l'algorithme AM modifié (Roberts et Rosenthal, 2009)

On considère la modification suivante de l'algorithme AM 3.1. Plutôt que d'ajouter $\kappa > 0$ sur la diagonale de $\mathrm{Cov}(x_{0:n})$ dans le calcul de Σ_n, on choisit la densité de proposition suivante :

$$q_\Sigma(y|x) = (1-\beta)\varphi(y|x, s_d\Sigma) + \beta\varphi(y|x, s_d^* I_d)$$

pour un certain choix $s_d, s_d^* > 0$ et $\beta \in (0,1)$ (typiquement, $s_d = (2{,}38)^2/d$, $s_d^* = (0{,}1)^2/d$ et β près de 0.) Le reste de l'algorithme est identique à l'algorithme AM. Lorsque β est choisi petit, l'algorithme utilisera majoritairement la proposition adaptée $\varphi(y|x, s_d\Sigma)$ tel que souhaité. La proposition fixe $\varphi(y|x, s_d^* I_d)$ sera généralement mal adaptée à la situation (covariance trop petite et donc propositions trop souvent acceptées), mais un petit β cachera cette perte d'efficacité.

Cette forme particulière de la densité de proposition permet une borne inférieure facile à identifier :

$$q_\Sigma(y|x) \geqslant \beta\varphi(y|x, s_d^* I_d), \qquad \forall\, \Sigma \in \mathcal{Y}\,.$$

On note même que cette borne est sphériquement symétrique et qu'elle constitue donc une borne uniforme par rapport à $u \in S^{d-1}$. Il sera donc possible de vérifier la condition 3.21 en utilisant le lemme 3.28 ; puis, en supposant des ailes super-exponentielles (condition 3.17), il sera donc possible d'assurer la convergence bornée.

Si on suppose plutôt que π admet des ailes qui ne sont qu'exponentielles, Bai et collab. (2011b, proposition 2.12) montrent qu'il est également possible d'assurer la convergence bornée. Pour ce faire, la densité fixe gaussienne $\varphi(y|x, s_d^* I_d)$ doit être remplacée par une densité uniforme sur une hyperboule suffisamment large.

3.4.2 Suppléments à la section 3.2.3.2

Dans cette annexe, il sera question des conditions permettant de vérifier que le paramètre d'adaptation est borné en probabilité dans le contexte d'algorithmes adaptatifs par approximations stochastiques avec troncation et couverture compacte.

On considère d'abord une condition de continuité lipschitzienne sur la variation de P_θ par rapport à θ et sur la variation de H par rapport à θ.

Condition 3.24 (Transitions lipschitziennes) *Soit V donnée dans la condition 3.3. Pour chaque $\mathcal{K} \subset \Theta$ compact et $r \in [0,1]$, il existe une constante C telle que pour toute paire $(\theta,\theta') \in \mathcal{K} \times \mathcal{K}$ et toute fonction $f \in \mathcal{L}_{V^r} := \{g : \|g\|_{V^r} < \infty\}$,*

$$\|P_\theta f - P_{\theta'} f\|_{V^r} \leqslant C\|f\|_{V^r} \left\|\theta - \theta'\right\|_2.$$

Andrieu et Moulines (2006, lemme 13) montrent, par exemple, que les transitions Metropolis à proposition $\mathcal{N}(\mathbf{0},\Sigma)$ sont lipschitzienne par rapport à Σ dès qu'on considère un ensemble compact $\mathcal{K}$ de matrices définies positives tel que la plus petite valeur propre de Σ est bornée uniformément par le bas sur $\mathcal{K}$. Ce résultat est généralisé (Andrieu et Moulines, 2006, proposition 16) aux transitions Metropolis de type marche aléatoire à propositions symétriques venant d'une famille exponentielle courbée.

Condition 3.25 (Mises à jour lipschitziennes) *Soit V donnée dans la condition 3.3. La famille des mises à jour indexées par θ, i.e. $\{H_\theta\}_{\theta\in\Theta}$ où $H_\theta(x) = H(\theta,x)$, est V^β-Lipschitzienne pour un certain $\beta \in [0,1/2]$. C'est-à-dire que pour tout $\mathcal{K} \subset \Theta$ compact on a*

$$\sup_{\theta\in\mathcal{K}} \|H_\theta\|_{V^\beta} < \infty, \qquad \sup_{\theta\neq\theta'\in\mathcal{K}\times\mathcal{K}} \left\|\theta - \theta'\right\|_2^{-1} \|H_\theta - H_{\theta'}\|_{V^\beta} < \infty.$$

En plus de contribuer à vérifier que θ_n est borné en probabilité, les conditions 3.24 et 3.25 assurent également l'adaptation diminuante. Pour θ_n borné en probabilité, alors il existe un compact contenant $\{\theta_n\}_{n\geqslant 0}$ avec probabilité arbitrairement élevée. Puis, en assumant de plus que les pas d'adaptation γ_n diminuent vers 0, on trouvera par la condition sur les mises à jour que $\gamma_{n+1}H(\theta_n,x_{n+1})$ tendra elle aussi vers 0 puisque H est borné. Ainsi, la différence $\theta_n - \theta_{n+1}$ tendra elle-aussi vers 0. Enfin, la condition de continuité lipschitzienne sur les transitions assure que la différence $P_{\theta_{n+1}} - P_{\theta_n}$ tend elle aussi vers 0, c'est-à-dire que l'adaptation diminuante est vérifiée.

Proposition 3.30 *Si les conditions 3.3 et 3.24 sont satisfaites, si θ_n est borné en probabilité et si $\sup_{\theta\in\mathcal{K}} \|H_\theta\|_{V^\beta} < \infty$, alors l'algorithme à couverture compacte satisfait l'adaptation diminuante (condition 3.1.)*

Démonstration. Par la condition sur H_θ, pour tout $\mathcal{K} \subset \Theta$ compact, on trouve $\widetilde{H} := \sup_{\theta\in\mathcal{K}} \|H_\theta\|_{V^r} < \infty$, où $\|\cdot\|_V$ est donnée par (3.14). En particulier, on a

$$\|H(\theta,x)\|_2 \leqslant \widetilde{H} V^r(x), \qquad \forall\, x \in \mathcal{X}\,, \theta \in \mathcal{K}\,.$$

Pour $r \in (0,1]$ et $V \geqslant 1$, on a $V^r(x) \leqslant V(x)$ pour tout $x \in \mathcal{X}$. La Condition 3.3, permet de borner $V(X)$ en probabilité tel que démontré à la Proposition 3.5.

Donc, pour tout $\delta > 0$, il existe $\widetilde{V} = \widetilde{V}(\delta) < \infty$ tel que $\mathbb{P}_{x_0,\theta_0}\left(V(X_n) \leqslant \widetilde{V}\right) \geqslant 1 - \frac{\delta}{4}$ pour tout $n \geqslant 1$. Alors,

$$\mathbb{P}_{x_0,\theta_0}\left(\|H(\theta,X_n)\|_2 \leqslant \widetilde{H}\widetilde{V}\right) \geqslant 1 - \frac{\delta}{4}, \qquad \forall\, \theta \in \mathcal{K}\,.$$

Puis, pour $\theta_{n+1} - \theta_n = \gamma_{n+1} H(\theta_n, X_n)$, on obtient

$$\mathbb{P}_{x_0,\theta_0}\left(\|\theta_{n+1} - \theta_n\|_2 \leqslant \gamma_{n+1}\widetilde{H}\widetilde{V} \middle| \theta_n\right) \geqslant 1 - \frac{\delta}{4}, \qquad \forall\, \theta_n \in \mathcal{K}\,.$$

Puisque $\{\theta_n\}_{n\geqslant 0}$ est borné en probabilité, il existe alors $\mathcal{K} \subset \Theta$ compact tel que

$$\mathbb{P}_{x_0,\theta_0}\left(\theta_n \in \mathcal{K}\right) \geqslant 1 - \frac{\delta}{4},$$

d'où

$$\mathbb{P}_{x_0,\theta_0}\left(\|\theta_{n+1} - \theta_n\|_2 \leqslant \gamma_{n+1}\widetilde{H}\widetilde{V}\right) \geqslant \left(1 - \frac{\delta}{4}\right)^2.$$

Ensuite, par la condition 3.24, il existe $C < \infty$ tel que

$$\left\|P_{\theta_{n+1}} f - P_{\theta_n} f\right\|_{V^r} \leqslant C\|\theta_{n+1} - \theta_n\|_2, \qquad \forall\, \|f\|_{V^r} \leqslant 1, \theta_{n+1}, \theta_n \in \mathcal{K}\,.$$

Ainsi,

$$\left|\left|\left|P_{\theta_{n+1}} - P_{\theta_n}\right|\right|\right|_{V^r} \leqslant C\|\theta_{n+1} - \theta_n\|_2, \qquad \forall\, \theta_{n+1}, \theta_n \in \mathcal{K},$$

où $\|\cdot\|_V$ est donné par (3.13) On trouve donc,

$$\mathbb{P}_{x_0,\theta_0}\left(\left|\left|\left|P_{\theta_{n+1}} - P_{\theta_n}\right|\right|\right|_{V^r} \leqslant \gamma_{n+1} C\widetilde{H}\widetilde{V} \middle| \theta_{n+1} \in \mathcal{K}\right) \geqslant \left(1 - \frac{\delta}{4}\right)^2.$$

À nouveau, on peut borner la probabilité de $\theta_{n+1} \in \mathcal{K}$:

$$\mathbb{P}_{x_0,\theta_0}\left(\left|\left|\left|P_{\theta_{n+1}} - P_{\theta_n}\right|\right|\right|_{V^r} \leqslant \gamma_{n+1} C\widetilde{H}\widetilde{V}\right) \geqslant \left(1 - \frac{\delta}{4}\right)^3.$$

Finalement, $\gamma_n \to 0$ implique qu'il existe, pour tout $\varepsilon > 0$, $M = M(\varepsilon) \in \mathbb{N}$ tel que $\gamma_{n+1} C\widetilde{H}\widetilde{V} \leqslant \varepsilon$ pour tout $n \geqslant M$. Donc, pour tout $n \geqslant M$, on a

$$\mathbb{P}_{x_0,\theta_0}\left(\left|\left|\left|P_{\theta_{n+1}} - P_{\theta_n}\right|\right|\right|_{V^r} \leqslant \varepsilon\right) \geqslant \left(1 - \frac{\delta}{4}\right)^3 \geqslant 1 - \delta,$$

i.e. $\mathrm{dist}(\theta_{n+1},\theta_n) \to 0$ en probabilité. □

On considère donc maintenant des conditions suffisantes pour assurer un nombre de réinitialisations borné. On rappelle que les réinitialisations se produisent lorsque θ tente de s'échapper de l'ensemble compact actuel $\mathcal{K}_r$; on cherche donc des conditions qui permettront de contenir θ dans un compact $\mathcal{K} \subset \Theta$. Andrieu et Moulines (2006, voir aussi Andrieu et collab., 2005 pour plus de généralité) proposent une condition sur le champ moyen $h(\theta)$ et une sur la séquence des pas d'adaptation qui assureront que θ demeure dans un ensemble compact.

Condition 3.26 (Andrieu et Moulines, 2006, condition A4) *Soit Θ un sous-ensemble ouvert de $\mathbb{R}^{n_\theta}$. Le champ moyen $h : \Theta \to \mathbb{R}^{n_\theta}$ est continu et il existe une fonction à dérivés premières continues $w : \Theta \to [0,\infty)$ (parfois appelée fonction de Lyapunov) telle que :*

(i) Pour tout $M > 0$, l'ensemble de niveau $\mathcal{W}_M := \{\theta \in \Theta : w(\theta) \leqslant M\}$ est compact ;

(ii) L'ensemble des points stationnaires de $\mathcal{L} := \{\theta \in \Theta : \nabla w(\theta)^\top h(\theta) = 0\}$ fait partie de l'intérieur de Θ ;

(iii) Pour tout $\theta \in \Theta$, on a $\nabla w(\theta)^\top h(\theta) \leqslant 0$ et la fermeture de $w(\mathcal{L})$ possède un intérieur vide.

La condition 3.26 exige l'existence d'une fonction de Lyapunov $w(\cdot)$ qui contrôle le comportement de h sur Θ. Il s'agit d'une fonction qui tend vers l'infini lorsque $\theta \to \partial\Theta$ et qui dirige en quelque sorte θ vers les racines de h.

Condition 3.27 (Andrieu et Moulines, 2006, condition A5) *La séquence des pas d'adaptation* $\{\gamma_n\}_{n\geqslant 1}$ *satisfait*

$$\sum_{n\geq 1}\gamma_n = \infty, \qquad \sum_{n\geqslant 1}\left(\gamma_n^2 + n^{-1/2}\gamma_n\right) < \infty.$$

Un exemple de séquence satisfaisant la condition 3.27 est $\gamma_n = n^{-\gamma}$ pour un certain $\gamma \in (1/2,1]$. Ensemble, les conditions 3.3, 3.24, 3.25, 3.26 et 3.27 montrent (Andrieu et Moulines, 2006, théorème 11) que la probabilité du nombre maximal de réinitialisations décroit d'une manière super-exponentielle sur $\mathbb{N}$ (et est donc borné presque sûrement) et que le paramètre θ converge vers l'ensemble des points stationnaire $\mathcal{L}$.

Théorème 3.31 (Andrieu et Moulines, 2006, théorème 11) *Soit* $\{\mathcal{K}_r\}_{r\geqslant 0}$ *une couverture compacte de* Θ, $\{\gamma_n\}_{n\geqslant 1}$ *une séquence réelle et* R_n *la variable aléatoire dénotant l'index de l'ensemble compact dans l'algorithme. En supposant les conditions 3.3, 3.24, 3.25, 3.26 et 3.27, on a*

$$\limsup_{k\to\infty}\frac{1}{k}\log\left[\sup_{(x,\theta)\in\mathcal{X}\times\Theta}\mathbb{P}_{x,\theta}\left(\sup_{n\geqslant 0}R_n \geqslant k\right)\right] = -\infty,$$

où $\mathbb{P}_{x,\theta}$ *dénote la probabilité du processus* $\{(\theta_n,X_n,R_n)\}_{n\geqslant 0}$ *avec les conditions initiales* $\theta_0 = \theta$, $x_0 = x$ *et* $r_0 = 0$. *En particulier,* $\{R_n\}_{n\geqslant 0}$ *est borné presque sûrement. De plus,*

$$\inf_{(x,\theta)\in\mathcal{X}\times\Theta}\mathbb{P}_{x,\theta}\left(\lim_{n\to\infty}\operatorname{dist}_{\mathcal{L}}(\theta_n) = 0\right) = 1,$$

où $\operatorname{dist}_{\mathcal{L}}(\theta) = \inf_{\theta'\in\mathcal{L}}\|\theta-\theta'\|_2$ *est la distance de* θ *à l'ensemble* $\mathcal{L}$.

L'annexe 3.4.3 considère l'algorithme AM 3.1 où l'ergodicité est montrée à l'aide d'une couverture compacte. Il s'agit donc d'une amélioration par rapport à l'exemple 3.7 où l'on devait supposer que l'espace des covariances devait être borné.

3.4.3 Ergodicité de l'algorithme AM avec couverture compacte

L'algorithme AM considère l'espace des paramètres $\Theta = \Theta_\mu \times \Theta_\Sigma$ où $\Theta_\mu = \mathbb{R}^d$ est l'espace des vecteurs de moyenne et où Θ_Σ est l'espace des matrices $d\times d$ symétriques et définies positives. Lorsque chacun des espaces Θ_μ et Θ_Σ n'est pas borné, on doit utiliser une couverture compacte afin d'assurer l'ergodicité. La description même de la couverture compacte n'est pas requise pour la preuve, mais on donne un exemple de construction possible. Soit $\mathcal{K}_r = \mathcal{K}_{r,\mu}\times\mathcal{K}_{r,\Sigma}$ avec

$$\begin{aligned}\mathcal{K}_{r,\mu} &= \{\mu\in\Theta_\mu : \|\mu\|_2 \leqslant r\},\\ \mathcal{K}_{r,\Sigma} &= \left\{\Sigma\in\Theta_\Sigma : \max(2,r)^{-1} \leqslant \lambda_1(\Sigma) \leqslant \ldots \leqslant \lambda_d(\Sigma) \leqslant \max(2,r)\right\},\end{aligned}$$

où $\lambda_i(\Sigma)$ dénote la i-ième valeur propre de Σ (prises en ordre croissant). Puisque les matrices définies positives admettent des valeurs propres qui soient toutes positives ; il est donc clair que $\mathcal{K}_r$ constitue une couverture compacte de Θ.

Pour voir ceci, on doit définir une norme sur Θ afin de pouvoir considérer la notion de compacité.

Soit

$$\begin{aligned}\|\theta\|_\Theta &= \sqrt{\|\mu\|_2^2 + \|\Sigma\|_F^2},\\ \|\mu\|_2^2 &= \mu^\top \mu,\\ \|\Sigma\|_F^2 &= \operatorname{tr}\left(\Sigma^\top \Sigma\right).\end{aligned}$$

La compacité dans Θ par rapport à la distance $\operatorname{dist}_\Theta(\theta,\theta') := \|\theta - \theta'\|_\Theta$ sera la même notion que dans un espace Euclidien (en identifiant $\Theta \subset \mathbb{R}^{d+d^2}$) puisque la norme de Frobenius correspond à la norme Euclidienne en vectorisant les matrices. Un ensemble compact de Θ sera donc fermé et borné. On note que Θ est non-borné vers les grands $\|\theta\|_\Theta$ et ouvert autour de $\|\theta\|_\Theta = 0$ puisque les matrices sont définies positives et auront donc une norme de Frobenius non-nulle. Ainsi, tout ensemble compact de Θ doit borner $\|\mu\|_2$ et $\|\Sigma\|_F$ par le haut ainsi que $\|\Sigma\|_F$ par le bas.

Enfin, on note que si $\|\Sigma\|_F$ est borné par le haut et par le bas, alors toutes ses valeurs propres sont contenues dans un intervalle fermé $[\lambda_{\min},\lambda_{\max}] \subset (0,\infty)$. En effet, soit λ une valeur propre de Σ et posons $m \leqslant \|\Sigma\|_F \leqslant M$. Puisque Σ est définie positive et symétrique, on a $\Sigma = S^\top \Lambda S$ où $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ est une matrice diagonale contenant les valeurs propres de Σ et $SS^\top = I_d$. Ainsi, on obtient

$$\|\Sigma\|_F^2 = \operatorname{tr}\Sigma^\top\Sigma = \operatorname{tr} S^\top \Lambda S S^\top \Lambda S = \operatorname{tr} S^\top \Lambda^2 S = \operatorname{tr}\Lambda^2 = \sum_{i=1}^d \lambda_i^2,$$

ce qui implique $m \leqslant \lambda \leqslant M$ puisque $\lambda > 0$.

On comprend donc que la couverture compacte considérée plus haut contient en quelque sorte toutes les couvertures compactes possibles en sous-séquence. En effet, soit $\mathcal{K}$ un sous-ensemble compact de Θ, alors il existe $M > 0$ et $0 < \lambda_{\min} \leqslant \lambda_{\max} < \infty$ tels que

$$\begin{aligned}&\|\mu\|_2 \leqslant M, && \forall\, \mu \in \mathcal{K},\\ &\lambda_{\min} \leqslant \lambda_i(\Sigma) \leqslant \lambda_{\max}, && \forall\, i = 1, \dots, d, \Sigma \in \mathcal{K}.\end{aligned}$$

Il existe donc $r \in \mathbb{N}$ tel que $\mathcal{K} \subseteq \mathcal{K}_r$.

Maintenant, les récursions sur $\theta = (\mu,\Sigma)$ sont données par (exemple 3.2)

$$\begin{aligned}\theta_{n+1} &= \theta_n + \gamma_{n+1} H(\theta_n, x_{n+1}),\\ \gamma_{n+1} &= \frac{1}{n+2},\\ H(\theta,x) &= \left[x - \mu, (x-\mu)(x-\mu)^\top - \Sigma\right]^\top.\end{aligned}$$

Ainsi, on trouve que le champ moyen est donnée par

$$\begin{aligned}
h(\theta) &= [h_\mu(\theta), h_\Sigma(\theta)]^\top \\
h_\mu(\theta) &= \int_{\mathcal{X}} (x-\mu)\pi(\,\mathrm{d}x) = \int_{\mathcal{X}} x\pi(\,\mathrm{d}x) - \mu =: \mu_\pi - \mu, \\
h_\Sigma(\theta) &= \int_{\mathcal{X}} \left((x-\mu)(x-\mu)^\top - \Sigma\right)\pi(\,\mathrm{d}x) \\
&= \int_{\mathcal{X}} \left((x-\mu_\pi+\mu_\pi-\mu)(x-\mu_\pi+\mu_\pi-\mu)^\top\right)\pi(\,\mathrm{d}x) - \Sigma \\
&= \int_{\mathcal{X}} \left((x-\mu_\pi)(x-\mu_\pi)^\top - 2(x-\mu_\pi)(\mu_\pi-\mu)^\top + (\mu_\pi-\mu)(\mu_\pi-\mu)^\top\right)\pi(\,\mathrm{d}x) - \Sigma \\
&=: \Sigma_\pi - 0 + (\mu_\pi-\mu)(\mu_\pi-\mu)^\top - \Sigma,
\end{aligned}$$

où l'on suppose que les espérances μ_π et Σ_π existent.

On cherche donc à vérifier les conditions 3.3, 3.24, 3.25, 3.26 et 3.27 pour les transitions Metropolis de marche aléatoire à proposition $\mathcal{N}_d(x_n,s_d\Sigma)$ (on omet le terme $+\varepsilon I_d$ qui n'est plus requis), une densité cible π à support dans $\mathcal{X}$ contenant $\mathcal{X}_0$ compact, l'espace de paramètre Θ général accompagné d'une couverture compacte, la séquence de pas d'adaptation $\{\gamma_n\}_{n\geqslant 1}$, la fonction H et son champ moyen h.

3.4.3.1 Vérification de la condition 3.3

Par construction, les transitions M.-H. admettent toutes la densité cible π comme distribution invariante. On doit trouver $V : \mathcal{K} \to [1,\infty)$ bornée sur $\mathcal{X}_0$ telle que la minorisation et la dérive géométrique sont vérifiées sur tout compact $\mathcal{K} \subset \Theta$. En supposant la condition de régularité 3.13 de la densité cible π, on aura $\sup_{\mathcal{X}} \pi < \infty$. Alors, on peut considérer

$$V_\eta(x) = \left(\frac{\sup_{\mathcal{X}} \pi}{\pi(x)}\right)^\eta \in [1,\infty),$$

pour tout $\eta > 0$. Par la régularité de π, on aura que π est bornée par le bas sur $\mathcal{X}_0$. Ainsi, on aura

$$\sup_{x\in\mathcal{X}_0} V(x) \leqslant \left(\sup_{\mathcal{X}} \pi\right)^\eta \sup_{\mathcal{X}_0}\left(\pi^{-\eta}\right) \leqslant \left(\sup_{\mathcal{X}} \pi\right)^\eta \left(\inf_{\mathcal{X}_0} \pi\right)^{-\eta} < \infty.$$

Ensuite, soit $\mathcal{K} \subset \Theta$ compact. Ainsi, tel que mentionné précédemment, tout $\theta = (\mu,\Sigma) \in \mathcal{K}$ sera tel que $\|\mu\| < M$ pour un certain $M < \infty$ et les valeurs propres λ_i de Σ seront contenues dans un intervalle $[\lambda_{\min},\lambda_{\max}] \subset (0,\infty)$, $i = 1,\ldots,d$.

Afin de montrer la condition de minorisation, on considère $\nu(A) = \lambda(A\cap C)/\lambda(C)$, où λ dénote la mesure de Lebesgue, qui constitue une mesure de probabilité sur $\mathcal{X}$, mais concentrée sur C. On doit alors trouver $C \in \mathcal{B}(\mathcal{X})$ et $\delta > 0$ tels que $P_\theta(A|x) \geqslant \delta\nu(A)$ pour tout $A \in \mathcal{B}(\mathcal{X})$, $\theta \in \mathcal{K}$ et $x \in C$. On choisit $C \subset \mathcal{X}$ compact arbitraire pour le moment. Par un argument similaire à celui utilisé à l'exemple 3.7, il sera possible de borner les densités de proposition $\varphi(\cdot|x,s_d\Sigma)$ par le bas :

$$\begin{aligned}
\varphi(z|\mathbf{0},s_d\Sigma) &= (2\pi)^{-d/2}(\det s_d\Sigma)^{-1/2}\exp\left\{-\frac{1}{2}z^\top(s_d\Sigma)^{-1}z\right\} \\
&\geqslant (2\pi s_d\lambda_{\max})^{-d/2}\exp\left\{-\frac{1}{2}(s_d\lambda_{\min})^{-1}\|z\|_2^2\right\}.
\end{aligned}$$

Donc, pour C borné, on pourra borner $\|z\|_2^2$ par le diamètre de C, définit par $\operatorname{diam} C = \sup_{x,y\in C} \|x-y\|_2$, et finalement borner $\varphi(z|\mathbf{0},s_d\Sigma)$ uniformément en Σ par le bas sur C :

$$\varphi(z|\mathbf{0},s_d\Sigma) \geqslant (2\pi s_d\lambda_{\max})^{-d/2} \exp\left\{-\frac{1}{2}(s_d\lambda_{\min})^{-1}(\operatorname{diam} C)^2\right\} > 0.$$

Puisque P_θ est une transition Metropolis à proposition $\varphi(\cdot|x,s_d\Sigma)$, il sera possible d'établir la minorisation. On rappelle que la probabilité d'acceptation Metropolis est indépendante de la densité d'une proposition symétrique :

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}.$$

Sur B compact, on aura π borné par le haut et par le bas étant donné sa Régularité. Donc,

$$\inf_{x,y\in B} \alpha(y|x) \geqslant \inf_{x,y\in B} \frac{\pi(y)}{\pi(x)} \geqslant \frac{\inf_{x,y\in B}\pi(y)}{\sup_{x,y\in B}\pi(x)} > 0.$$

Pour un ensemble $B \subset \mathcal{X}$, on introduit la notation $B - x := \{y \in \mathcal{X} : x + y \in B\}$, qui est compact dès que B est compact. Puis, on note que $B - x$ est borné par $\operatorname{diam} B$ uniformément sur $x \in B$. Enfin, on peut montrer la minorisation :

$$\begin{aligned}
P_\theta(A|x) &\geqslant \int_{y\in A} \varphi(y|x,s_d\Sigma)\alpha(y|x)\lambda(\,\mathrm{d}y) \\
&\geqslant \int_{y\in A\cap C} \varphi(y|x,s_d\Sigma)\alpha(y|x)\lambda(\,\mathrm{d}y) \\
&\geqslant \inf_{y\in A\cap C} \varphi(y|x,s_d\Sigma) \int_{y\in A} \alpha(y|x)\lambda(\,\mathrm{d}y) \\
&\geqslant \inf_{y\in C} \varphi(y|x,s_d\Sigma) \inf_{x\in C, y\in A\cap C} \alpha(y|x) \int_{y\in A\cap C} \lambda(\,\mathrm{d}y) \\
&\geqslant \nu(A) \inf_{z\in C-x} \varphi(z|\mathbf{0},s_d\Sigma) \inf_{x\in C, y\in A\cap C} \alpha(y|x) \\
&\geqslant \nu(A) \inf_{\|z\|_2 \leqslant \operatorname{diam} C} \varphi(z|\mathbf{0},s_d\Sigma) \inf_{x,y\in C} \alpha(y|x) \\
&=: \delta\nu(A), \qquad \forall\, x \in C, \theta \in \mathcal{K}\,, A \in \mathcal{B}(\mathcal{X}),
\end{aligned}$$

où

$$\delta = \inf_{\|z\|_2 \leqslant \operatorname{diam} C} \varphi(z|\mathbf{0},s_d\Sigma) \inf_{x,y\in C} \alpha(y|x) > 0.$$

Pour la dérive géométrique, on définit les deux régions suivantes :

$$\begin{aligned}
A(x) &:= \{y \in \mathcal{X} : \pi(y) \geqslant \pi(x)\} \\
R(x) &:= \{y \in \mathcal{X} : \pi(y) < \pi(x)\} = \mathcal{X} \setminus A(x).
\end{aligned}$$

L'ensemble $A(x)$ contient tous les y tels qu'ils sont acceptés automatiquement à partir de x ; l'ensemble $R(x)$ contient quant à lui les y qui sont acceptés avec une probabilité inférieure à 1. Notons, à nouveau, que les probabilité d'acceptation Metropolis sont indépendante de la densité de proposition et donc que les régions $A(x)$ et $R(x)$ sont indépendantes de θ. Jarner et Hansen (2000, théorème 4.1) montrent la dérive géométrique des algorithmes Metropolis (non-adaptatif) en supposant que π admette des ailes

super-exponentielles (condition 3.18) et que la densité de propositions q satisfassent

$$\liminf_{\|x\|_2 \to \infty} q(A(x)|x) > 0.$$

Leur théorème 4.3 montre qu'une condition suffisante est que π admette des contours réguliers (condition 3.14), ce qui relâche le lien entre les propositions et la densité cible. On ne peut pas appliquer directement le résultat à notre situation étant donné que nous cherchons une uniformité sur $\theta \in \mathcal{K}$, mais la preuve peut être adaptée pour y parvenir (Andrieu et Moulines, 2006, proposition 12). Saksman et Vihola (2010, proposition 15) montrent un résultat similaire en ne bornant les valeurs propres de Σ que par le bas. Si π satisfait la condition de régularité 3.13, admet des contours réguliers (condition 3.14) et des ailes super-exponentielles (condition 3.18), alors il existe C compact (qui peut donc être utilisé pour la Minorisation), une mesure de probabilité ν sur C et une constante $b \in [0,\infty)$ tels que pour $\Sigma \in \Theta_\Sigma$ aux valeurs propres bornées par le bas par $\kappa > 0$ on a

$$P_\theta V(x) \leqslant \lambda_\theta V(x) + b\, \mathbb{1}_C(x), \qquad \forall\, x \in \mathcal{X},$$

où $V = V_\eta$ avec $\eta = 1/2$ et $\lambda_\theta \in (0,1)$ satisfait

$$(1 - \lambda_\theta)^{-1} \leqslant c(\det \Sigma)^{1/2},$$

pour un certain $c \geqslant 1$. En bornant également les valeurs propres par le haut par $\infty > K > \kappa$, on aura $\det \Sigma \leqslant K^d$ et donc $(1-\lambda)^{-1} \leqslant cK^{d/2}$, ce qui permet de borner uniformément λ_Σ :

$$\lambda_\Sigma \leqslant 1 - c^{-1}K^{-d/2} =: \lambda \in (0,1).$$

Ceci montre donc la condition 3.3. Andrieu et Moulines (2006, proposition 12) montrent – sous peu de détails – un résultat plus général en considérant la famille de densités de proposition suivante

$$\mathcal{Q}_{a,b}(\mathcal{X}) := \left\{ q = \frac{\mathrm{d}Q}{\mathrm{d}\lambda} : q(x) = q(-x), \inf_{\|x\|_2 \leqslant a} q(x) \geqslant b \right\},$$

correspondant à l'ensemble de toutes les densités (par rapport à la mesure de Lebesgue) sur $\mathcal{X}$ qui soient symétriques et bornées uniformément par le bas sur $\|x\|_2 \leq a$. Évidemment, la famille de propositions gaussiennes $q_\theta(z) = \varphi(z|\mathbf{0}, s_d\Sigma)$ avec $\theta \in \mathcal{K}$ est un sous-ensemble de cette famille. Sans répéter l'argument complet, on considère ici les grandes lignes. On a

$$P_\theta V(x) = \int_{\mathcal{X}} V(y) P_\theta(y|x) \lambda(\mathrm{d}y).$$

Pour notre choix de V, on trouve alors

$$\frac{P_\theta V(x)}{V(x)} = \left(\frac{\pi(x)}{\sup_{\mathcal{X}} \pi}\right)^\eta \int_{\mathcal{X}} \left(\frac{\sup_{\mathcal{X}} \pi}{\pi(y)}\right)^\eta P_\theta(y|x)\lambda(\mathrm{d}y) \tag{3.34}$$

$$= \int_{\mathcal{X}} \left(\frac{\pi(x)}{\pi(y)}\right)^\eta P_\theta(y|x)\lambda(\mathrm{d}y). \tag{3.35}$$

Dans la preuve de la minorisation, on exigeait que C soit borné, mais arbitraire. La condition $P_\theta V(x) \leqslant \lambda V(x)$ sur $x \notin C$ correspond alors à borner l'expression (3.35) pour grands x par un certain $\lambda \in [0,1)$.

Pour $y \in A(x)$, on trouve

$$P_\theta(y|x) = q_\theta(y-x) + \delta_x(y) \int_{R(x)} \left(1 - \frac{\pi(z)}{\pi(x)}\right) q_\theta(y-x)\lambda(\,\mathrm{d}z),$$

et, pour $y \in R(x)$, on a

$$P_\theta(y|x) = \frac{\pi(y)}{\pi(x)} q_\theta(y-x).$$

Ainsi,

$$\begin{aligned}
\frac{P_\theta V(x)}{V(x)} &= \int_{A(x)} \left(\frac{\pi(x)}{\pi(y)}\right)^\eta P_\theta(y|x)\lambda(\,\mathrm{d}y) + \int_{R(x)} \left(\frac{\pi(x)}{\pi(y)}\right)^\eta P_\theta(y|x)\lambda(\,\mathrm{d}y) \\
&= \int_{A(x)} \left(\frac{\pi(x)}{\pi(y)}\right)^\eta q_\theta(y-x)\lambda(\,\mathrm{d}y) \\
&\quad + \int_{A(x)} \left(\frac{\pi(x)}{\pi(y)}\right)^\eta \delta_x(y) \int_{R(x)} \left(1 - \frac{\pi(z)}{\pi(x)}\right) q_\theta(z-x)\lambda(\,\mathrm{d}z)\lambda(\,\mathrm{d}y) \\
&\quad + \int_{R(x)} \left(\frac{\pi(x)}{\pi(y)}\right)^{\eta-1} q_\theta(y-x)\lambda(\,\mathrm{d}y) \\
&= \int_{A(x)} \left(\frac{\pi(y)}{\pi(x)}\right)^{-\eta} q_\theta(y-x)\lambda(\,\mathrm{d}y) \\
&\quad + \int_{R(x)} \left(1 - \frac{\pi(y)}{\pi(x)} + \left(\frac{\pi(y)}{\pi(x)}\right)^{1-\eta}\right) q_\theta(y-x)\lambda(\,\mathrm{d}y).
\end{aligned}$$

En choisissant C arbitrairement grand, le comportement de $P_\theta V/V$ pour $x \notin C$ correspondra à celui de $\|x\|_2 \to \infty$. Il est alors possible d'arguer que les ratios

$$\left(\frac{\pi(y)}{\pi(x)}\right)^{-\eta}, \frac{\pi(y)}{\pi(x)}, \left(\frac{\pi(y)}{\pi(x)}\right)^{1-\eta}$$

tendent tous vers 0 lorsque $\|x\|_2 \to \infty$ et tout y avec $\pi(y) \neq \pi(x)$ dans les régions respectives $A(x)$ et $R(x)$. Ainsi, on trouve

$$\begin{aligned}
\limsup_{\|x\|_2 \to \infty} \frac{P_\theta V(x)}{V(x)} &= \limsup_{\|x\|_2 \to \infty} \int_{R(x)} q_\theta(y-x)\lambda(\,\mathrm{d}y) \\
&= \limsup_{\|x\|_2 \to \infty} q_\theta(R(x)-x) \\
&= 1 - \liminf_{\|x\|_2 \to \infty} q_\theta(A(x)-x).
\end{aligned}$$

Enfin, puisque $\mathcal{K}$ est compact et que les contours réguliers de π font en sorte que la région $A(x)$ contient toujours une partie d'un voisinage de x, il sera possible de borner

$$\inf_{\theta \in \mathcal{K}} \liminf_{\|x\|_2 \to \infty} q_\theta(A(x)-x) > 0.$$

Ainsi,

$$\sup_{\theta \in \mathcal{K}} \limsup_{\|x\|_2 \to \infty} \frac{P_\theta V(x)}{V(x)} < 1.$$

Donc, pour un C suffisamment grand, il existe $\lambda \in (0,1)$ tel que

$$\sup_{\theta\in\mathcal{K}} \sup_{x\notin C} \frac{P_\theta V(x)}{V(x)} \leqslant \lambda.$$

Pour la borne $P_\theta V \leq b$ sur C, on considère d'abord $y \in A(x)$ dans ce cas, $\pi(x)/\pi(y) \in [0,1]$ et donc $(\pi(y)/\pi(x))^{-\eta} \leqslant 1$. Ainsi, on trouve

$$\begin{aligned}\int_{A(x)} \left(\frac{\pi(y)}{\pi(x)}\right)^{-\eta} q_\theta(y-x)\lambda(\,\mathrm{d}y) &\leqslant \int_{A(x)} q_\theta(y-x)\lambda(\,\mathrm{d}y)\\ &= q_\theta(A(x)-x)\end{aligned}$$

Puis, pour $y \in R(x)$ on a $\pi(y)/\pi(x) \in [0,1]$ et donc

$$\begin{aligned}&\int_{R(x)} \left(1 - \frac{\pi(y)}{\pi(x)} + \left(\frac{\pi(y)}{\pi(x)}\right)^{1-\eta}\right) q_\theta(y-x)\lambda(\,\mathrm{d}y)\\ &\leqslant \int_{R(x)} \sup_{u\in[0,1]} \left(1-u+u^{1-\eta}\right) q_\theta(y-x)\lambda(\,\mathrm{d}y)\\ &= q_\theta(R(x)-x) \sup_{u\in[0,1]} \left(1-u+u^{1-\eta}\right).\end{aligned}$$

Finalement, on obtient

$$\begin{aligned}\frac{P_\theta V(x)}{V(x)} &\leqslant q_\theta(A(x)-x) + q_\theta(R(x)-x) \sup_{u\in[0,1]} \left(1-u+u^{1-\eta}\right)\\ &\leqslant \max\left\{1, \sup_{u\in[0,1]} \left(1-u+u^{1-\eta}\right)\right\},\end{aligned}$$

d'où l'on conclut

$$P_\theta V(x) \leqslant \sup_{\mathcal{X}} V \cdot \max\left\{1, \sup_{u\in[0,1]} \left(1-u+u^{1-\eta}\right)\right\} < \infty.$$

3.4.3.2 Vérification de la condition 3.24

Soit $\mathcal{K} \subset \Theta$ compact et soit $V = V_\eta$ pour un $\eta \in (0,1)$. On doit montrer qu'il existe $C > 0$ et un $r \in [0,1]$ tels que

$$\|P_\theta f - P_{\theta'} f\|_{V^r} \leqslant C\|f\|_{V^r}\|\theta-\theta'\|_2, \qquad \forall\, (\theta,\theta') \in \mathcal{K}\times\mathcal{K}\,, f \in \mathcal{L}_{V^r}\,.$$

D'abord, on a

$$\|P_\theta f - P_{\theta'} f\|_{V^r} = \sup_{x\in\mathcal{X}} \frac{|P_\theta f(x) - P_{\theta'} f(x)|}{V^r(x)}.$$

On débute donc par démontrer un résultat intermédiaire afin de borner cette norme en supposant seulement la symétrie des densités de proposition $q_\theta(z) = q_\theta(-z)$. On écrit d'abord

$$\begin{aligned}P_\theta f(x) &= \int_{\mathcal{X}} f(y) P_\theta(y|x)\lambda(\,\mathrm{d}y)\\ &= \int_{\mathcal{X}} f(y) q_\theta(y|x)\alpha(y|x)\lambda(\,\mathrm{d}y) + f(x)\int_{\mathcal{X}} q_\theta(y|x)\left[1-\alpha(y|x)\right]\lambda(\,\mathrm{d}y).\end{aligned}$$

Puis, en rappelant que $\alpha(y|x)$ est indépendant de θ dans un algorithme Metropolis,

$$P_\theta f(x) - P_{\theta'} f(x) = \int_{\mathcal{X}} f(y) \left[q_\theta(y|x) - q_{\theta'}(y|x)\right] \alpha(y|x) \lambda(\,\mathrm{d}y) + f(x) \int_{\mathcal{X}} \left[q_{\theta'}(y|x) - q_\theta(y|x)\right] \alpha(y|x) \lambda(\,\mathrm{d}y).$$

Ensuite, par l'inégalité du triangle on trouve

$$\frac{|P_\theta f(x) - P_{\theta'} f(x)|}{\|f\|_{V^r} V^r(x)} \leqslant \frac{\left|\int_{\mathcal{X}} f(y) \left[q_\theta(y|x) - q_{\theta'}(y|x)\right] \alpha(y|x) \lambda(\,\mathrm{d}y)\right|}{\|f\|_{V^r} V^r(x)} + \frac{\left|f(x) \int_{\mathcal{X}} \left[q_{\theta'}(y|x) - q_\theta(y|x)\right] \alpha(y|x) \lambda(\,\mathrm{d}y)\right|}{\|f\|_{V^r} V^r(x)}.$$

On cherche à borner ces deux termes par $\int_{\mathcal{X}} |q_\theta(y|x) - q_{\theta'}(y|x)| \lambda(\,\mathrm{d}y)$. On considère d'abord le premier terme. Puisque $f \in \mathcal{L}_{V^r}$, alors

$$\frac{|f(z)|}{\|f\|_{V^r}} \leqslant V^r(z), \qquad \forall\, z \in \mathcal{X}\,.$$

On trouve donc

$$\begin{aligned} &\frac{\left|\int_{\mathcal{X}} f(y) \left[q_\theta(y|x) - q_{\theta'}(y|x)\right] \alpha(y|x) \lambda(\,\mathrm{d}y)\right|}{\|f\|_{V^r} V^r(x)} \\ &\leqslant \int_{\mathcal{X}} \frac{V^r(y)}{V^r(x)} |q_\theta(y|x) - q_{\theta'}(y|x)| \alpha(y|x) \lambda(\,\mathrm{d}y) \\ &= \int_{\mathcal{X}} |q_\theta(y|x) - q_{\theta'}(y|x)| \frac{\pi(y)^{-r\eta}}{\pi(x)^{-r\eta}} \alpha(y|x) \lambda(\,\mathrm{d}y). \end{aligned}$$

Puis, on sépare $\mathcal{X}$ selon les régions $A(x)$ et $R(x)$ où le comportement de $\alpha(y|x)$ est constant. Sur $A(x)$, on a $\alpha(y|x) = 1$ et $\pi(y)/\pi(x) \geqslant 1$; on trouve alors, pour $r\eta \in (0,1)$,

$$\frac{\pi(y)^{-r\eta}}{\pi(x)^{-r\eta}} \alpha(y|x) = \frac{\pi(y)^{-r\eta}}{\pi(x)^{-r\eta}} = \left(\frac{\pi(x)}{\pi(y)}\right)^{r\eta} \leqslant 1.$$

Sur $R(x)$, on a $\alpha(y|x) = \pi(y)/\pi(x) \leqslant 1$. On trouve également, en notant que $1 - r\eta \in (0,1)$,

$$\frac{\pi(y)^{-r\eta}}{\pi(x)^{-r\eta}} \alpha(y|x) = \frac{\pi(y)^{-r\eta}}{\pi(x)^{-r\eta}} \frac{\pi(y)}{\pi(x)} = \left(\frac{\pi(y)}{\pi(x)}\right)^{1-r\eta} \leqslant 1.$$

Donc, on obtient

$$\frac{\left|\int_{\mathcal{X}} f(y) \left[q_\theta(y|x) - q_{\theta'}(y|x)\right] \alpha(y|x) \lambda(\,\mathrm{d}y)\right|}{\|f\|_{V^r} V^r(x)} \leqslant \int_{\mathcal{X}} |q_\theta(y|x) - q_{\theta'}(y|x)| \lambda(\,\mathrm{d}y).$$

Le second terme est un peu plus direct puisque $f(x)$ est fixe et que $\alpha(y|x) \leq 1$:

$$\begin{aligned} &\frac{\left|f(x) \int_{\mathcal{X}} \left[q_{\theta'}(y|x) - q_\theta(y|x)\right] \alpha(y|x) \lambda(\,\mathrm{d}y)\right|}{\|f\|_{V^r} V^r(x)} \\ &\leqslant \int_{\mathcal{X}} |q_{\theta'}(y|x) - q_\theta(y|x)| \alpha(y|x) \lambda(\,\mathrm{d}y) \\ &\leqslant \int_{\mathcal{X}} |q_{\theta'}(y|x) - q_\theta(y|x)| \lambda(\,\mathrm{d}y). \end{aligned}$$

Finalement, ceci permet de borner

$$\frac{|P_\theta f(x) - P_{\theta'} f(x)|}{V^r(x)} \leqslant 2\|f\|_{V^r} \int_{\mathcal{X}} |q_{\theta'}(y|x) - q_\theta(y|x)| \lambda(\,\mathrm{d}y),$$

pour tout $x \in \mathcal{X}, (\theta,\theta') \in \mathcal{K}\times\mathcal{K}$ et $f \in \mathcal{L}_{V^r}$. Ainsi, afin d'établir la condition lipschitzienne sur $\|P_\theta f - P_{\theta'} f\|_{V^r}$ par rapport à θ, on doit borner $2\int_{\mathcal{X}} |q_{\theta'}(y|x) - q_\theta(y|x)|\lambda(\,\mathrm{d}y)$ par $C\|\theta - \theta'\|_2$ uniformément sur $x \in \mathcal{X}$. Ceci requiert évidemment une connaissance de q_θ ; dans notre cas, on a $q_\theta(z) = \varphi(z|\mathbf{0},s_d\Sigma) =: \varphi_\Sigma(z)$. Puisque $\mathcal{K}$ est supposé compact, il existe $\lambda_{\min} < \lambda_{\max} \in (0,\infty)$ tels que les valeurs de propres de $\overline{\Sigma} := s_d\Sigma$ soient toutes contenues dans $[\lambda_{\min},\lambda_{\max}]$ pour tout $\theta = (\mu,\Sigma) \in \mathcal{K}$.

On considère ensuite la combinaison convexe $\Sigma_t = \Sigma + t(\Sigma' - \Sigma) = (1-t)\Sigma + t\Sigma'$ (Haario et collab., 2001, théorème 1) qui vaut Σ lorsque $t = 0$, Σ' lorsque $t = 1$ et qui définit une matrice symétrique définie positive pour tout $t \in [0,1]$. En effet, Σ_t est évidemment symétrique puisque combinaison linéaire de matrices symétriques et définie positive : pour tout $t \in [0,1]$ et tout $x \in \mathbb{R}^d$,

$$x^\top \Sigma_t x = \underbrace{(1-t)}_{\geqslant 0}\underbrace{x^\top \Sigma x}_{\geqslant 0} + \underbrace{t}_{\geqslant 0}\underbrace{x^\top \Sigma' x}_{\geqslant 0} \geqslant 0.$$

Ainsi, φ_{Σ_t} est une densité gaussienne bien définie. L'intérêt de cette définition est l'identité suivante

$$\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\varphi_{\Sigma_t}(z)\,\mathrm{d}t = \varphi_{\Sigma_t}(z)|_{t=0}^{t=1} = \varphi_\Sigma(z) - \varphi_{\Sigma'}(z),$$

à condition que $\varphi_{\Sigma_t}(z)$ soit dérivable par rapport à t. Il sera alors possible de relier plus aisément la différence $\varphi_\Sigma(z) - \varphi_{\Sigma'}(z)$ à la différence $\Sigma - \Sigma'$ étant donnée que $\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t = \Sigma' - \Sigma$. En effet, on procède à la dérivation logarithmique pour trouver

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\varphi_{\Sigma_t}(z) &= \varphi_{\Sigma_t}(z)\frac{\mathrm{d}}{\mathrm{d}t}\log\varphi_{\Sigma_t}(z)\\ &= -\frac{1}{2}\varphi_{\Sigma_t}(z)\frac{\mathrm{d}}{\mathrm{d}t}\left[d\log(2\pi) + \log\det(\Sigma_t) + z^\top\Sigma_t^{-1}z\right],\end{aligned}$$

où (Petersen et Pedersen, 2008)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\log\det(\Sigma_t) &= \operatorname{tr}\left(\Sigma_t^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t\right) = \operatorname{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma)\right),\\ \frac{\mathrm{d}}{\mathrm{d}t}z^\top\Sigma_t^{-1}z &= -z^\top\left(\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t^{-1}\right)z\\ &= -z^\top\Sigma_t^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t\right)\Sigma_t^{-1}z\\ &= -z^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\Sigma_t^{-1}z\\ &= \operatorname{tr}\left(-z^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\Sigma_t^{-1}z\right)\\ &= \operatorname{tr}\left(-\Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right).\end{aligned}$$

La dépendance en la différence $\Sigma - \Sigma'$ est désormais évidente :

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_{\Sigma_t}(z) = -\frac{1}{2}\varphi_{\Sigma_t}(z)\operatorname{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma) - \Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right)$$

Par l'inégalité du triangle, on trouve

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\log\varphi_{\Sigma_t}(z)\right| \leqslant \left|\mathrm{tr}\left(\Sigma_t^{-1}(\Sigma'-\Sigma)\right)\right| + \left|\mathrm{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma'-\Sigma)\right)\right|.$$

Par l'inégalité de Hölder sur les normes matricielles de Schatten, i.e. $|\mathrm{tr}(AB)| \leqslant \|A\|_F\|B\|_F$ où la norme de Frobenius correspond à la norme Schatten d'ordre 2, on peut borner

$$\left|\mathrm{tr}\left(\Sigma_t^{-1}(\Sigma'-\Sigma)\right)\right| \leqslant \left\|\Sigma_t^{-1}\right\|_F\|\Sigma'-\Sigma\|_F.$$

Similairement, on peut borner

$$\begin{aligned}
\left|\mathrm{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma'-\Sigma)\right)\right| &\leqslant \left\|\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\right\|_F\|\Sigma'-\Sigma\|_F \\
&\leqslant \sqrt{\mathrm{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\right)}\|\Sigma'-\Sigma\|_F \\
&\leqslant \sqrt{\mathrm{tr}\left(\Sigma_t^{-1}z(z^\top\Sigma_t^{-2}z)z^\top\Sigma_t^{-1}\right)}\|\Sigma'-\Sigma\|_F \\
&\leqslant \sqrt{(z^\top\Sigma_t^{-2}z)\,\mathrm{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\right)}\|\Sigma'-\Sigma\|_F \\
&\leqslant \sqrt{(z^\top\Sigma_t^{-2}z)\,\mathrm{tr}\left(z^\top\Sigma_t^{-2}z\right)}\|\Sigma'-\Sigma\|_F \\
&\leqslant \sqrt{(z^\top\Sigma_t^{-2}z)^2}\|\Sigma'-\Sigma\|_F \\
&\leqslant \left|z^\top\Sigma_t^{-2}z\right|\|\Sigma'-\Sigma\|_F
\end{aligned}$$

Notons ensuite que, si A est symétrique, alors

$$x^\top A^2x = x^\top AAx = x^\top A^\top Ax = (Ax)^\top Ax = \|Ax\|_2 \geqslant 0.$$

Ainsi, $z^\top\Sigma_t^{-2}z > 0$ pour tout z, ce qui implique

$$\left|z^\top\Sigma_t^{-2}z\right| = z^\top\Sigma_t^{-2}z.$$

On peut donc borner

$$\begin{aligned}
\left|\frac{\mathrm{d}}{\mathrm{d}t}\log\varphi_{\Sigma_t}(z)\right| &\leqslant \left\|\Sigma_t^{-1}\right\|_F\|\Sigma'-\Sigma\|_F + \|\Sigma'-\Sigma\|_Fz^\top\Sigma_t^{-2}z \\
&= \|\Sigma'-\Sigma\|_F\left(\left\|\Sigma_t^{-1}\right\|_F + z^\top\Sigma_t^{-2}z\right).
\end{aligned}$$

Par la théorie des formes quadratiques, on note que si $Z \sim \mathcal{N}_d(\mathbf{0},\Sigma)$, alors $\mathbb{E}\{Z^\top AZ\} = \mathrm{tr}\,(A\Sigma)$, et donc

$$\int_{\mathcal{X}}(z^\top\Sigma_t^{-2}z)\varphi_{\Sigma_t}(z)\lambda(\,\mathrm{d}z) = \mathrm{tr}\left(\Sigma_t^{-2}\Sigma_t\right) = \mathrm{tr}\left(\Sigma_t^{-1}\right).$$

Ainsi, on trouve

$$\int_{\mathcal{X}}\left(\left\|\Sigma_t^{-1}\right\|_F + z^\top\Sigma_t^{-2}z\right)\varphi_{\Sigma_t}(z)\lambda(dz) = \left\|\Sigma_t^{-1}\right\|_F + \mathrm{tr}\left(\Sigma_t^{-1}\right).$$

Maintenant, la norme $\left\|\Sigma_t^{-1}\right\|_F$ ainsi que $\mathrm{tr}\left(\Sigma_t^{-1}\right)$ peuvent être bornées uniformément. En effet, puisque Σ_t est une combinaison convexe de Σ et de Σ', toutes deux aux valeurs propres dans $[\lambda_{\min},\lambda_{\max}]$, alors les valeurs propres de Σ_t seront également dans le même intervalle (conséquence de Bhatia, 1997,

théorème III.2.1). Ainsi,

$$\left\|\Sigma_t^{-1}\right\|_F^2 = \sum_{i=1}^d \lambda_i^2(\Sigma_t^{-1}) = \sum_{i=1}^d \lambda_i^{-2}(\Sigma_t) \leqslant d\lambda_{\min}^{-2},$$
$$\operatorname{tr}\left(\Sigma_t^{-1}\right) = \sum_{i=1}^d \lambda_i(\Sigma_t^{-1}) \leqslant d\lambda_{\min}^{-1}.$$

Enfin, on trouve la borne

$$\int_{\mathcal{X}} \left(\left\|\Sigma_t^{-1}\right\|_F + z^\top \Sigma_t^{-2} z\right) \varphi_{\Sigma_t}(z)\lambda(dz) \leqslant \sqrt{d}\lambda_{\min}^{-1} + d\lambda_{\min}^{-1} \leqslant 2d\lambda_{\min}^{-1}.$$

Ceci qui permet de conclure

$$\begin{aligned}
\int_{\mathcal{X}} |\varphi_\Sigma(z) - \varphi_{\Sigma'}(z)|\lambda(\,\mathrm{d}z) &= \int_{\mathcal{X}} \left|\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\varphi_{\Sigma_t}(z)\,\mathrm{d}t\right|\lambda(\,\mathrm{d}y) \\
&\leqslant \int_{\mathcal{X}}\int_0^1 \frac{1}{2}\varphi_{\Sigma_t}(z)\left|\frac{\mathrm{d}}{\mathrm{d}t}\log\varphi_{\Sigma_t}(z)\right|\mathrm{d}t\lambda(\,\mathrm{d}y) \\
&= \frac{1}{2}\int_0^1 \int_{\mathcal{X}} \left|\frac{\mathrm{d}}{\mathrm{d}t}\log\varphi_{\Sigma_t}(z)\right|\varphi_{\Sigma_t}(z)\lambda(\,\mathrm{d}y)\,\mathrm{d}t \\
&\leqslant \frac{1}{2}\int_0^1 2d\lambda_{\min}^{-1}\|\Sigma' - \Sigma\|_F\,\mathrm{d}t \\
&= \frac{d}{\lambda_{\min}}\|\Sigma' - \Sigma\|_F.
\end{aligned}$$

Finalement, on obtient bien la condition lipshitzienne sur les transitions

$$\begin{aligned}
\frac{|P_\theta f(x) - P_{\theta'} f(x)|}{V^r(x)} &\leqslant 2\|f\|_{V^r}\int_{\mathcal{X}} |q_{\theta'}(y|x) - q_\theta(y|x)|\lambda(\,\mathrm{d}y) \\
&\leqslant \frac{2d}{\lambda_{\min}}\|f\|_{V^r}\|\Sigma' - \Sigma\|_F, \qquad \forall\, x \in \mathcal{X},
\end{aligned}$$

c'est-à-dire

$$\|P_\theta f - P_{\theta'} f\|_{V^r} \leqslant \frac{2d}{\lambda_{\min}}\|f\|_{V^r}\|\Sigma' - \Sigma\|_F \leqslant \frac{2d}{\lambda_{\min}}\|f\|_{V^r}\|\theta' - \theta\|.$$

3.4.3.3 Vérification de la condition 3.25

On a

$$H_\theta(x) = \begin{pmatrix} x - \mu \\ (x-\mu)(x-\mu)^\top - \Sigma \end{pmatrix}$$

et la fonction $V(x) = (\sup_{\mathcal{X}} \pi/\pi(x))^\eta$ pour un $\eta \in (0,1)$. On veut montrer qu'il existe $\beta \in [0,1/2]$ tel que pour tout $\mathcal{K} \subset \Theta$ compact on a

$$\sup_{\theta\in\mathcal{K}} \|H_\theta\|_{V^\beta} < \infty, \qquad \sup_{\theta\neq\theta'\in\mathcal{K}\times\mathcal{K}} \|\theta - \theta'\|_2^{-1}\|H_\theta - H_{\theta'}\|_{V^\beta} < \infty.$$

Directement, par l'inégalité du triangle,

$$\begin{aligned}\frac{\|H_\theta(x)\|_2}{V^\beta(x)} &\leqslant \frac{\|x-\mu\|_2 + \left\|(x-\mu)(x-\mu)^\top - \Sigma\right\|_F}{V^\beta(x)} \\ &\leqslant \frac{\|x-\mu\|_2 + \left\|(x-\mu)(x-\mu)^\top\right\|_F + \|\Sigma\|_F}{V^\beta(x)} \\ &\leqslant \frac{\|x\|_2 + \|\mu\|_2 + \left\|xx^\top\right\|_F + 2\left\|x\mu^\top\right\|_F + \left\|\mu\mu^\top\right\|_F + \|\Sigma\|_F}{V^\beta(x)}.\end{aligned}$$

Puisque $V \geqslant 1$, les termes qui ne dépendent pas de x sont bornés :

$$\frac{\|H_\theta(x)\|_2}{V^\beta(x)} \leqslant \frac{\|x\|_2 + \left\|xx^\top\right\|_F + 2\left\|x\mu^\top\right\|_F}{V^\beta(x)} + \|\mu\|_2 + \left\|\mu\mu^\top\right\|_F + \|\Sigma\|_F,$$

où $\|\mu\|_2$, $\left\|\mu\mu^\top\right\|_F$ et $\|\Sigma\|_F$ sont bornés pour $\theta \in \mathcal{K}$ compact. Maintenant, $\|x\|_2$ est non-borné en général pour $x \in \mathcal{X}$: des conditions sur V et donc sur π sont requises pour assurer que cette expression soit bornée. D'abord, on note les identités suivantes

$$\begin{aligned}\left\|xx^\top\right\|_F^2 &= \operatorname{tr}\left(xx^\top xx^\top\right) = x^\top x \operatorname{tr}\left(xx^\top\right) = \|x\|_2^4, \\ \left\|x\mu^\top\right\|_F^2 &= \operatorname{tr}\left(x\mu^\top x\mu^\top\right) = \mu^\top x \operatorname{tr}\left(x\mu^\top\right) = (\mu^\top x)^2 \leqslant \|x\|_2^2\|\mu\|_2^2.\end{aligned}$$

Ainsi, on trouve

$$\frac{\|H_\theta(x)\|_2}{V^\beta(x)} \leqslant \frac{\|x\|_2 + \|x\|_2^2 + 2\|x\|_2\|\mu\|_2}{V^\beta(x)} + \|\mu\|_2 + \left\|\mu\mu^\top\right\|_F + \|\Sigma\|_F.$$

Pour $V^\beta(x) = (\sup_{\mathcal{X}} \pi)^{\beta\eta}\pi^{-\beta\eta}(x)$, on requiert donc

$$\sup_{x\in\mathcal{X}} \|x\|_2^2 \pi^{\beta\eta}(x) < \infty.$$

Sur tout compact $\mathcal{X}_0 \subset \mathcal{X}$ on a que $\|x\|_2^2\pi^{\beta\eta}(x)$ est borné étant donné la régularité de π. Donc, on ne requiert que la décroissance de π soit suffisamment rapide pour grands x, i.e.

$$\limsup_{\|x\|_2\to\infty} \|x\|_2^2 \pi^{\beta\eta}(x) < \infty.$$

Cette condition est cependant vérifiée pour tout $\beta > 0$ dès que l'on suppose une des conditions sur les ailes de π (des ailes hyperboliques sont suffisantes bien qu'on a supposé des ailes super-exponentielles pour vérifier la condition 3.3.) On conclut donc que $\sup_{\theta\in\mathcal{K}} \|H_\theta\|_{V^\beta} < \infty$ pour tout $\mathcal{K}$ compact.

Ensuite, on écrit

$$\begin{aligned}
\|H_\theta(x) - H_{\theta'}(x)\|_2^2 &= \|(x-\mu) - (x-\mu')\|_2^2 \\
&\quad + \left\| \left[(x-\mu)(x-\mu)^\top - \Sigma\right] - \left[(x-\mu')(x-\mu')^\top - \Sigma'\right] \right\|_F^2 \\
&= \|\mu - \mu'\|_2^2 \\
&\quad + \left\| xx^\top - 2x\mu^\top + \mu\mu^\top - \Sigma - xx^\top + 2x\mu^\top - \mu\mu^\top + \Sigma \right\|_F^2 \\
&= \|\mu - \mu'\|_2^2 + \left\| 2x(\mu'-\mu)^\top + (\mu'-\mu)(\mu'+\mu)^\top + (\Sigma' - \Sigma) \right\|_F^2 \\
&\leqslant \|\mu - \mu'\|_2^2 \left(1 + 2\|x\|_2^2 + \|\mu' + \mu\|_2^2\right) + \|\Sigma' - \Sigma\|_F^2. \\
&\leqslant \|\theta - \theta'\|_2^2 \left(1 + 2\|x\|_2^2 + \|\mu' + \mu\|_2^2\right),
\end{aligned}$$

c'est-à-dire

$$\|H_\theta(x) - H_{\theta'}(x)\|_2 \leqslant \|\theta - \theta'\|_2 \sqrt{1 + 2\|x\|_2^2 + \|\mu' + \mu\|_2^2}$$

Maintenant, pour $\mathcal{K}$ compact, $\|\mu' + \mu\|_2$ est borné uniformément. Puis, pour les mêmes raisons que précédemment, $\sqrt{1 + 2\|x\|_2^2 + \|\mu' + \mu\|_2^2}$ est borné sur tout compact de $\mathcal{X}$ et le comportement asymptotique est celui de $\|x\|_2$, qui est tel que

$$\sup_{x \in \mathcal{X}} \|x\|_2 \pi^{\beta\eta}(x) < \infty.$$

C'est donc dire que

$$\sup_{x \in \mathcal{X}} \frac{\sqrt{1 + 2\|x\|_2^2 + \|\mu' + \mu\|_2^2}}{V^\beta(x)} < \infty,$$

uniformément par rapport à $\theta, \theta' \in \mathcal{K}$. On conclut donc que

$$\begin{aligned}
&\sup_{\theta \neq \theta' \in \mathcal{K} \times \mathcal{K}} \|\theta - \theta'\|_2^{-1} \|H_\theta - H_{\theta'}\|_{V^\beta} \\
&= \sup_{\theta \neq \theta' \in \mathcal{K} \times \mathcal{K}} \sup_{x \in \mathcal{X}} \|\theta - \theta'\|_2^{-1} \frac{\|H_\theta(x) - H_{\theta'}(x)\|_2}{V^\beta(x)} \\
&\leqslant \sup_{\theta \neq \theta' \in \mathcal{K} \times \mathcal{K}} \sup_{x \in \mathcal{X}} \frac{\sqrt{1 + 2\|x\|_2^2 + \|\mu' + \mu\|_2^2}}{V^\beta(x)} \\
&< \infty.
\end{aligned}$$

3.4.3.4 Vérification de la condition 3.26

Les conditions sur le champ moyen h requièrent une fonction de Lyapunov $w : \Theta \to [0,\infty)$ telle que (i) ses ensembles de niveaux soient compacts (donc que w ne soit grande que pour de grands θ), (ii) que les points stationnaires de w soient dans l'intérieur de Θ (dans notre cas, on veut que μ et Σ soient bornés par le haut et Σ borné par le bas) et (iii) que la fermeture de l'ensemble des points stationnaires possède un intérieur vide (par exemple, il ne faut pas que les points stationnaires définissent un ouvert) et que $\nabla w^\top h \leqslant 0$, i.e. que les points stationnaires correspondent à des minimums.

Pour ce, on considère $w(\theta) = \frac{1}{2}\|\theta - \theta_\pi\|_\Theta^2$, où $\theta_\pi = (\mu_\pi, \Sigma_\pi)$. En choisissant la distance de l'ensemble Θ, les ensembles de niveaux sont directement compact. En effet, $w(\theta)$ est une fonction continue de θ et donc $\mathcal{W}_M = w^{-1}([0,M])$ est fermé puisqu'il s'agit de la pré-image d'une ensemble fermé. Ensuite, si $w(\theta) \leqslant M$ alors $\|\theta - \theta_\pi\|_\Theta \leqslant \sqrt{2M}$: l'ensemble $\mathcal{W}_M$ est donc une boule de rayon $\sqrt{2M}$ centrée en θ_π et est donc borné. Donc $\mathcal{W}_M$ est borné et fermé dans $\Theta \subset \mathbb{R}^{d+d\times d}$ sous la distance

euclidienne, ce qui implique que $\mathcal{W}_M$ est compact.

On calcule ensuite $\nabla w(\theta) \cdot h(\theta)$ afin de vérifier les conditions (ii) et (iii). On note que Σ est symétrique et donc que la dérivation par rapport à Σ se fait en deux étapes (Petersen et Pedersen, 2008). On dérive d'abord l'expression par rapport à Σ comme si les entrées de Σ étaient indépendantes (que l'on dénotera par $\mathrm{d}/\mathrm{d}\Sigma$), puis on ajuste la dérivée pour tenir compte de la dépendance induite par la symétrie (que l'on dénotera par $\partial/\partial\Sigma$.) On obtient alors

$$\begin{aligned}
\frac{\partial w(\theta)}{\partial \mu} &= \frac{1}{2}\frac{\partial}{\partial \mu}\left((\mu-\mu_\pi)^\top(\mu-\mu_\pi)\right) = \mu-\mu_\pi;\\
\frac{\mathrm{d}w(\theta)}{\mathrm{d}\Sigma} &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\Sigma}\left(\mathrm{tr}\left[(\Sigma-\Sigma_\pi)^\top(\Sigma-\Sigma_\pi)\right]\right) = \Sigma-\Sigma_\pi\\
\frac{\partial w(\theta)}{\partial \Sigma} &= \frac{\mathrm{d}w(\theta)}{\mathrm{d}\Sigma} + \left[\frac{\mathrm{d}w(\theta)}{\mathrm{d}\Sigma}\right]^\top - \mathrm{diag}\left[\frac{\mathrm{d}w(\theta)}{\mathrm{d}\Sigma}\right]\\
&= 2(\Sigma-\Sigma_\pi) - \mathrm{diag}(\Sigma-\Sigma_\pi).
\end{aligned}$$

Puis, on trouve

$$\begin{aligned}
\nabla w(\theta)\cdot h(\theta) &= \left[\frac{\partial w(\theta)}{\partial \mu}\right]^\top h_\mu(\theta) + \mathrm{tr}\left(\left[\frac{\partial w(\theta)}{\partial \Sigma}\right]^\top h_\Sigma(\theta)\right)\\
&= (\mu-\mu_\pi)^\top(\mu_\pi-\mu) + 2\,\mathrm{tr}\left((\Sigma-\Sigma_\pi)^\top(\Sigma_\pi-\Sigma)\right)\\
&\qquad - \mathrm{tr}\left(\mathrm{diag}(\Sigma-\Sigma_\pi)^\top(\Sigma_\pi-\Sigma)\right)\\
&= -\|\mu-\mu_\pi\|_2^2 - 2\|\Sigma-\Sigma_\pi\|_F^2 + \mathrm{tr}\left(\mathrm{diag}(\Sigma-\Sigma_\pi)(\Sigma-\Sigma_\pi)\right).
\end{aligned}$$

On note ensuite que

$$\begin{aligned}
&2\|\Sigma-\Sigma_\pi\|_F^2 - \mathrm{tr}\left(\mathrm{diag}(\Sigma-\Sigma_\pi)(\Sigma-\Sigma_\pi)\right)\\
&= \sum_{i,j} 2\left[\Sigma-\Sigma_\pi\right]_{ij}^2 - \sum_i \left[\Sigma-\Sigma_\pi\right]_{ii}^2\\
&= \sum_{i,j}\left[\Sigma-\Sigma_\pi\right]_{ij}^2 + \sum_{i\neq j}\left[\Sigma-\Sigma_\pi\right]_{ij}^2 \geqslant 0.
\end{aligned}$$

Ainsi, on conclut que $\nabla w(\theta)\cdot h(\theta) \leqslant 0$ pour tout $\theta \in \Theta$. On remarque alors que $\nabla w(\theta)\cdot h(\theta) = 0$ si et seulement si

$$\left[\frac{\partial w(\theta)}{\partial \mu}\right]^\top h_\mu(\theta) = 0 \qquad \text{et} \qquad \mathrm{tr}\left(\left[\frac{\partial w(\theta)}{\partial \Sigma}\right]^\top h_\Sigma(\theta)\right) = 0,$$

étant donné que ces deux quantités sont non-positives. En inspectant les calculs, on trouve que $\|\mu-\mu_\pi\|_2 = 0$ et $[\Sigma-\Sigma_\pi]_{ij} = 0$ pour tout i,j. Ainsi, le seul point stationnaire est (μ_π,Σ_π), ce qui montre que $\mathcal{L} = \{\theta\in\Theta : \nabla w(\theta)\cdot h(\theta) = 0\}$ est un singleton qui est donc dans l'intérieur de Θ qui doit être supposé ouvert. Enfin, $w(\mathcal{L}) = w\{(\mu_\pi,\Sigma_\pi)\} = \{0\}$ possède évidemment un intérieur vide.

3.4.3.5 Vérification de la condition 3.27

La séquence de pas d'adaptation est donnée par $\gamma_n = (n+1)^{-1}$. Directement,

$$\sum_{n\geqslant 1} \gamma_n = \sum_{n\geqslant 2} n^{-1} = \infty$$
$$\sum_{n\geqslant 1} \gamma_n^2 + n^{-1/2}\gamma_n = \sum_{n\geqslant 1} \frac{1}{(n+1)^2} + \frac{1}{n^{3/2}+n^{1/2}} \leqslant \sum_{n\geqslant 1} \frac{2}{n^{3/2}} < \infty.$$

Notons que toute séquence $\gamma_n = n^\gamma$ pour $\gamma \in (1/2,1]$ satisfait à cette condition.

3.4.3.6 Conclusion

Afin de vérifier les conditions 3.3, 3.24, 3.25, 3.26 et 3.27 pour l'algorithme AM, il a été nécessaire de supposer les conditions suivantes :

— La distribution cible admet une densité π régulière (condition 3.13), aux contours réguliers (condition 3.14) et aux ailes super-exponentielles (condition 3.18). Notons qu'une démonstration similaire par Saksman et Vihola (2010, section 5.2.2) ne suppose que des ailes hyperboliques (condition 3.19).

— La distribution cible admet un second moment fini de sorte que μ_π et Σ_π existent. Cependant, la condition sur les ailes de la densité implique cette propriété.

— L'espace Θ est ouvert. On peut donc choisir systématiquement l'espace maximal $\Theta = \mathbb{R}^d \times \mathcal{C}_+^d$ où $\mathcal{C}_+^d$ dénote le cône de matrices $d \times d$ symétriques définies positives.

Ainsi, au moment d'appliquer l'algorithme AM, seules des conditions sur la distribution cible sont à vérifier.

Par le théorème 3.31, ces conditions sont suffisantes pour assurer que le nombre de réinitialisations dans l'algorithme AM avec couverture compacte (algorithme 3.12) soit borné presque sûrement. Ceci implique que le processus $\{\theta_n\}_{n\geqslant 1}$ restera dans un sous-ensemble compact de Θ avec probabilité 1 ; ainsi, θ_n demeurera à distance de $\partial\Theta$ où les densités de propositions perdent leur ergodicité.

Par la proposition 3.30, l'algorithme satisfait l'adaptation dimunuante ; par le corollaire 3.6, l'algorithme satisfait la convergence bornée. Ainsi, l'algorithme adaptatif est ergodique suivant le théorème 3.1.

De plus, le théorème 3.31 indique que l'approximation stochastique θ_n converge presque sûrement vers un point stationnaire. Dans notre cas, ceci implique que $\theta_n \to \theta_\pi$ puisque θ_π est l'unique point stationnaire ; la densité de proposition convergera donc vers la distribution normale aux deux premiers moments identiques à ceux de la distribution cible. On peux ainsi comprendre que l'algorithme atteint l'optimalité de l'efficacité d'estimation dans le sens du théorème 2.34, ce qui constitue l'intérêt des algorithmes adaptatifs.

Chapitre 4

MCMC à essais multiples

L'algorithme Metropolis-Hastings (sous-section 2.3.1) génère le nouvel état de la chaîne de Markov en deux étapes : une proposition est d'abord générée, puis le nouvel état est sélectionné au hasard entre cette **proposition** et l'état précédent de la chaîne selon la probabilité M.-H. (2.20). Une généralisation de cette procédure est de considérer un ensemble d'états (les **candidats**) plutôt qu'un seul afin d'obtenir une proposition de meilleure qualité, c'est-à-dire qui a une plus grande probabilité d'être acceptée comme nouvel état. C'est l'idée derrière l'algorithme Metropolis à essais multiples (MTM : *Multiple-try Metropolis*) initialement proposé par Liu et collab. (2000).

Les résultats d'échelle optimale pour l'algorithme M.-H. montrent qu'une grande proportion des candidats générés doivent être rejetés : par exemple, la règle du 0,234 (e.g. théorème 2.34) impose que 76,6% des candidats soient rejetés afin d'obtenir un algorithme optimal en terme d'efficacité. Ce genre de résultat n'est valide que sous certaines conditions de régularité sur la densité cible et ces conditions sont généralement non-négligeables (composantes i.i.d., par exemple.) Néanmoins, la probabilité d'acceptation optimale est relativement robuste par rapport à ces conditions de sorte que l'optimalité est souvent atteinte pour un taux d'acceptation aux alentours de 23,4%. Ainsi, dans la majorité des itérations de l'algorithme, la chaîne générée ne change pas d'état : un échantillon de taille $N = 1{,}000$ ne contiendra en moyenne que 234 états distincts. De plus, à chaque itération, un seul nouvel état du support de la distribution cible π est considéré comme nouvel état de la chaîne et donc seulement 1,000 points du support $\mathcal{X}$ de π seront évalués globalement. Il semble donc que l'exploration du support de π ne soit pas particulièrement efficace en raison de ces remarques.

En considérant plutôt un ensemble de $K > 1$ candidats à chaque itération, l'algorithme génère donc $1{,}000 \times K$ nouveaux états potentiels pour la chaîne et visite donc le support de π plus rapidement et d'une manière plus fine en terme de granularité. De plus, en sélectionnant judicieusement un des candidats, appelé la proposition, parmi les K considérés afin de mettre à jour la chaîne par comparaison avec l'état précédent, on peut s'attendre à ce que les candidats soient acceptés plus souvent et que l'échantillon contienne alors en moyenne plus de 234 points distincts. Théoriquement, l'optimalité de l'algorithme sera donc atteinte avec une probabilité d'acceptation supérieure à celle de l'algorithme M.-H. à un seul candidat. Intuitivement, plus le nombre de candidats K sera élevé, plus le candidat retenu comme nouvel état potentiel sera de qualité, augmentant du coup l'efficacité de chaque itération.

Évidemment, la génération des $K - 1$ candidats supplémentaires et l'évaluation de π à ces états exige un coût computationnel supplémentaire qui doit être considéré. Bien que cet algorithme améliore théoriquement l'efficacité de chaque itération, le calcul nécessaire à chaque itération s'en voit cependant

augmenté. Ainsi, il est primordial de considérer l'augmentation de l'efficacité par rapport au temps de calcul et donc par rapport à K.

À la section 4.1, il sera question de l'algorithme MTM ainsi que de ses généralisations et de ses variations. À la section 4.2, les propriétés théoriques de ce type d'algorithmes seront explorées. À la section 4.3, l'efficacité de l'algorithme sera étudiée dans le contexte de coût computationnel additionnel.

4.1 Définition

L'algorithme MTM se voit comme une un généralisation de l'algorithme M.-H. Le candidat généré par l'algorithme M.-H. est automatiquement choisi comme proposition pour le nouvel état de la chaîne et est alors soumis au test donné par la probabilité d'acceptation M.-H. L'algorithme MTM insère une étape supplémentaire entre la génération des K candidats et le test d'acceptation : une proposition doit être sélectionnée parmi l'ensemble des candidats. Pour ce, un échantillonnage est effectué parmi les candidats où les probabilités de sélection doivent refléter, d'une certaine manière, la qualité du candidat. Plusieurs choix de poids sont possibles, mais la condition principale dans la formulation des poids sera que la transition de Markov induite par l'algorithme admette π comme distribution stationnaire.

Dans la dérivation de la probabilité d'acceptation de l'algorithme M.-H., une des conditions permettant d'assurer la stationnarité était que la chaîne soit réversible par rapport à π. La réversibilité peut quant à elle être vérifiée si la transition de Markov P satisfait la condition d'équilibre, c.-à-d.,

$$P(x|y)\pi(y) = P(y|x)\pi(x), \qquad \forall\, x,y \in \mathcal{X}^2\,. \tag{4.1}$$

Soit x_n, l'état actuel de la chaîne. On dénote par $y^{(1)}\,,\ldots,y^{(K)}$ les K candidats et par $w^{(k)}(y^{(k)}|x_n)$, $k = 1\,,\ldots,K$, le poids de $y^{(k)}$ pour l'étape de sélection. Lorsque $y = y^{(k)}$ est choisi comme proposition, celle-ci est ensuite soumise à un test d'acceptation M.-H. : y est alors choisi comme nouvel état avec probabilité α et x_n avec probabilité $1-\alpha$. Les expressions de w et de α seront alors choisies de sorte à respecter la condition d'équilibre (4.1).

La condition d'équilibre exige que le passage de x à y s'effectue selon la même distribution que le passage de y à x. Dans le cas de l'algorithme MTM, le passage de x_n à x_{n+1} s'effectue en plusieurs étapes qui ne sont pas symétriques en apparence. L'expression de α devra être donc être construite de sorte à retrouver une certaine symétrie. Puisque la première étape consiste en la production des candidats $y^{(1)}\,,\ldots,y^{(K)}$ à partir de x_n, il sera nécessaire de produire des **points de référence** $x_*^{(1)}\,,\ldots,x_*^{(K)}$ à partir de la proposition y afin d'assurer la réversibilité du processus. De plus, pour que la transition puisse s'effectuer dans les deux directions, on doit poser $x_*^{(k)} = x_n$ où k est l'indice

du candidat choisi comme proposition. Ainsi, on définit la **probabilité M.-H. généralisée**

$$\alpha(y,y^{(-k)}|x_n,x_*^{(-k)}) := \min\left\{1,\frac{\sum_{j=1}^K w^{(j)}(y^{(j)}|x_n)}{\sum_{j=1}^K w^{(j)}(x_*^{(j)}|y)}\right\} \tag{4.2}$$

$$= \min\left\{1,\frac{w^{(k)}(y^{(k)}|x_n)+\sum_{j\neq k} w^{(j)}(y^{(j)}|x_n)}{w^{(k)}(x_*^{(k)}|y)+\sum_{j\neq k} w^{(j)}(x_*^{(j)}|y)}\right\} \tag{4.3}$$

qui satisfait alors la condition d'équilibre moyennant une condition sur l'expression de $w^{(k)}(y^{(k)}|x_n)$.

Proposition 4.1 (Casarin et collab., 2013, en annexe) *Soient $Q(\cdot|x)$ la densité instrumentale servant à générer les candidats $y^{(1:K)}$, $Q^{(-k)}(\cdot|y^{(k)},x)$, la densité conditionnelle de $y^{(-k)}$ sachant $y^{(k)}$, et $Q^{(k)}(\cdot|x)$, la densité marginale de $y^{(k)}$. La probabilité d'acceptation généralisée définie dans (4.2) satisfait la condition d'équilibre (4.1) si w est de la forme*

$$w^{(k)}(y|x) = \pi(y)Q^{(k)}(x|y)s(x,y), \tag{4.4}$$

où $s(\cdot,\cdot) > 0$ est une fonction symétrique de ses arguments.

Démonstration. Si $x = y$, alors l'égalité (4.1) est évidemment respectée. On suppose donc que $x \neq y$. Dans ce cas, la proposition a été acceptée puisque la chaîne change d'état. On introduit la variable aléatoire $I \in \{1,\ldots,K\}$ avec probabilités $w(y^{(k)}|x)$, $k = 1,\ldots,K$. Notons la transition de x à y restreinte à l'acceptation M.-H. par $A(y|x)$. Alors

$$\begin{aligned}\pi(x)A(y|x) &= \pi(x)\sum_{k=1}^K \int_{\mathcal{X}^{K-1}}\int_{\mathcal{X}^{K-1}} Q(y,y^{(-k)}|x)\bar{w}^{(k)}(y;y^{(-k)}|x)\alpha(y,y^{(-k)}|x,x_*^{(-k)}) \\ &\qquad \times Q^{(-k)}(x_*^{(-k)}|y,x)\,\mathrm{d}y^{(-k)}\,\mathrm{d}x_*^{(-k)} \\ &= \sum_{k=1}^K \pi(x)\int_{\mathcal{X}^{K-1}}\int_{\mathcal{X}^{K-1}} Q(y,y^{(-k)}|x)\bar{w}^{(k)}(y;y^{(-k)}|x)\alpha(y,y^{(-k)}|x,x_*^{(-k)}) \\ &\qquad \times Q^{(-k)}(x_*^{(-k)}|y,x)\,\mathrm{d}y^{(-k)}\,\mathrm{d}x_*^{(-k)},\end{aligned}$$

où

$$\bar{w}^{(k)}(y^{(k)};y^{(-k)}|x) = \mathbb{P}\left(I=k\middle|x,y^{(1:K)}\right) = \frac{w^{(k)}(y^{(k)}|x)}{\sum_{j=1}^K w^{(j)}(y^{(j)}|x)}$$

correspond à la probabilité de sélection du candidat $y^{(k)}$. On note alors la décomposition par densités conditionnelles

$$Q(y^{(1:K)}|x) = Q^{(-k)}(y^{(-k)}|y^{(k)},x)Q^{(k)}(y^{(k)}|x),$$

ainsi que la ré-expression suivante de la probabilité d'acceptation

$$\alpha(y,y^{(-k)}|x,x_*^{(-k)}) = \left(\sum_{j=1}^K w^{(j)}(y^{(j)}|x)\right)\min\left\{\frac{1}{\sum_{j=1}^K w^{(j)}(y^{(j)}|x)},\frac{1}{\sum_{j=1}^K w^{(j)}(x_*^{(j)}|y)}\right\}.$$

Enfin, lorsque $y^{(k)}$ est sélectionné comme proposition, on note que $w^{(k)}(y^{(k)}|x) = w^{(k)}(y|x)$ et

$Q^{(k)}(y^{(k)}|x) = Q^{(k)}(y|x)$ sont indépendants des variables d'intégration. On obtient alors

$$\begin{aligned}
\pi(x)A(y|x) &= \sum_{k=1}^{K} \pi(x) \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} Q^{(-k)}(y^{(-k)}|y,x) Q^{(k)}(y|x) \frac{w^{(k)}(y^{(k)}|x)}{\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x)} \\
&\qquad \times \left(\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x) \right) \min \left\{ \frac{1}{\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x)}, \frac{1}{\sum_{j=1}^{K} w^{(j)}(x_*^{(j)}|y)} \right\} \\
&\qquad \times Q^{(-k)}(x_*^{(-k)}|y,x)\, \mathrm{d}y^{(-k)}\, \mathrm{d}x_*^{(-k)} \\
&= \sum_{k=1}^{K} \pi(x) Q^{(k)}(y|x) w^{(k)}(y|x) \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} Q^{(-k)}(y^{(-k)}|x,y) Q^{(-k)}(x_*^{(-k)}|y,x) \\
&\qquad \times \min \left\{ \frac{1}{\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x)}, \frac{1}{\sum_{j=1}^{K} w^{(j)}(x_*^{(j)}|y)} \right\} \mathrm{d}y^{(-k)}\, \mathrm{d}x_*^{(-k)}.
\end{aligned}$$

L'expression sous l'intégrale est complètement symétrique en (x,y). On note ensuite que

$$\pi(x)Q^{(k)}(y|x)w^{(k)}(y|x) = \frac{w^{(k)}(x|y)}{s(y,x)} w^{(k)}(y|x)$$

est aussi symétrique en (x,y). On conclut donc que ce choix de w et de α donne une transition satisfaisant la condition d'équilibre. □

Ensemble, les poids (4.4) et la probabilité d'acceptation (4.2) constituent une généralisation de l'étape d'acceptation M.-H. En effet, lorsque $K = 1$, l'unique candidat est automatiquement choisi comme proposition et, pour tout $s(\cdot,\cdot)$ symétrique, on obtient alors la probabilité d'acceptation M.-H. qui définit alors entièrement l'algorithme MTM 4.1 :

$$\alpha(y|x) = \min \left\{ 1, \frac{w^{(1)}(y|x)}{w^{(1)}(x|y)} \right\} = \min \left\{ 1, \frac{\pi(y)Q^{(1)}(x|y)s(y,x)}{\pi(x)Q^{(1)}(y|x)s(x,y)} \right\} = \min \left\{ 1, \frac{\pi(y)Q^{(1)}(x|y)}{\pi(x)Q^{(1)}(y|x)} \right\}.$$

Dans la preuve de la proposition 4.1, la forme explicite des poids de sélection n'est utilisée qu'à la toute dernière étape. Ceci suggère qu'une généralisation est possible. Considérons donc la réécriture suivante du rapport des poids dans la probabilité d'acceptation généralisée (4.2) :

$$\begin{aligned}
\frac{\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x)}{\sum_{j=1}^{K} w^{(j)}(x_*^{(j)}|y)} &= \frac{\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x)}{\sum_{j=1}^{K} w^{(j)}(x_*^{(j)}|y)} \times \frac{\pi(y)Q^{(k)}(x|y)s(y,x)}{w^{(k)}(y|x)} \times \frac{w^{(k)}(x|y)}{\pi(x)Q^{(k)}(y|x)s(x,y)} \\
&= \frac{\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x)}{w^{(k)}(y|x)} \times \frac{w^{(k)}(x|y)}{\sum_{j=1}^{K} w^{(j)}(x_*^{(j)}|y)} \times \frac{\pi(y)Q^{(k)}(x|y)}{\pi(x)Q^{(k)}(y|x)} \\
&= \frac{\pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x; x_*^{(-k)}|y)}{\pi(x)Q^{(k)}(y|x)\bar{w}^{(k)}(y; y^{(-k)}|x)}.
\end{aligned} \tag{4.5}$$

Cette seconde expression du rapport des poids est en fait plus générale que (4.2) puisqu'il est possible de montrer qu'une telle probabilité d'acceptation satisfait la condition d'équilibre (4.1) pour n'importe quel choix de poids. Ce résultat est démontré par Pandolfi et collab. (2010, Théorème 5.1) dans le cas où les candidats sont i.i.d., mais il est possible d'obtenir le même résultat pour un ensemble de candidats plus général en imitant la preuve de Casarin et collab. (2013, en annexe) pour les poids de la forme (4.4).

Proposition 4.2 *Soient $Q(\cdot|x)$ la densité instrumentale servant à générer les candidats $y^{(1:K)}$, $Q^{(-k)}(\cdot|y^{(k)},x)$, la densité conditionnelle de $y^{(-k)}$ sachant $y^{(k)}$, et $Q^{(k)}(\cdot|x)$, la densité marginale de $y^{(k)}$. Supposons que les densités marginales satisfassent à*

$$Q^{(k)}(x|y) > 0 \quad \Leftrightarrow \quad Q^{(k)}(y|x) > 0, \qquad k = 1, \ldots, K.$$

La probabilité d'acceptation M.-H. généralisée définie à l'aide du rapport (4.5) satisfait la condition d'équilibre (4.1) pour tout choix de $w^{(k)} > 0$, $k = 1, \dots, K$.

Démonstration. L'argument est similaire à la preuve de la proposition 4.1. On note donc la décomposition par densité conditionnelles suivante : pour $x_*^{(k)} = x$, on a

$$Q(x_*^{(1:K)}|y) = Q^{(k)}(x|y)Q^{(-k)}(x_*^{(-k)}|x,y).$$

De plus, on réécrit la probabilité d'acceptation ainsi :

$$\begin{aligned}\alpha(y,y^{(-k)}|x,x_*^{(-k)}) &= \min\left\{1, \frac{\pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x;x_*^{(-k)}|y)}{\pi(x)Q^{(k)}(y|x)\bar{w}^{(k)}(y;y^{(-k)}|x)}\right\} \\ &= \pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x;x_*^{(-k)}|y) \\ &\qquad \times \min\left\{\frac{1}{\pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x;x_*^{(-k)}|y)}, \frac{1}{\pi(x)Q^{(k)}(y|x)\bar{w}^{(k)}(y;y^{(-k)}|x)}\right\}.\end{aligned}$$

En procédant aux mêmes étapes, on trouve

$$\begin{aligned}\pi(x)A(y|x) &= \pi(x)\sum_{k=1}^{K}\int_{\mathcal{X}^{K-1}}\int_{\mathcal{X}^{K-1}} Q(y,y^{(-k)}|x)Q^{(-k)}(x_*^{(-k)}|x,y)\bar{w}^{(k)}(y^{(k)};y^{(-k)}|x) \\ &\qquad \times \min\left\{\frac{1}{\pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x|y)}, \frac{1}{\pi(x)Q^{(k)}(y|x)\bar{w}^{(k)}(y|x)}\right\} \\ &\qquad \times \pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x;x_*^{(-k)}|y)\,\mathrm{d}y^{(-k)}\,\mathrm{d}x_*^{(-k)} \\ &= \sum_{k=1}^{K}\int_{\mathcal{X}^{K-1}}\int_{\mathcal{X}^{K-1}} \pi(x)Q(y,y^{(-k)}|x)\bar{w}^{(k)}(y^{(k)}|x)\pi(y)Q(x,x_*^{(-k)}|y)\bar{w}^{(k)}(x|y) \\ &\qquad \times \min\left\{\frac{1}{\pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x;x_*^{(-k)}|y)}, \frac{1}{\pi(x)Q^{(k)}(y|x)\bar{w}^{(k)}(y;y^{(-k)}|x)}\right\}\mathrm{d}y^{(-k)}\,\mathrm{d}x_*^{(-k)}.\end{aligned}$$

Par inspection directe, cette expression est complètement symétrique par rapport à (y,x) et à $(y^{(-k)},x_*^{(-k)})$. C'est donc dire que

$$\pi(x)A(y|x) = \pi(y)A(x|y),$$

ce qui complète la preuve. □

La proposition 4.2 est exprimée d'une façon très générale. En effet, seule la densité conjointe Q et les poids $w^{(k)}$ doivent être spécifiés afin de produire une transition valide. Le reste de cette section sera consacré à l'étude de choix possibles sur Q et sur $w^{(k)}$, puis à certaines généralisations et variantes possibles.

4.1.1 Choix de la densité de propositions

Le choix de la densité de proposition Q peut être fait d'une manière arbitraire si l'on se fie à la conclusion de la proposition 4.1. Cependant, certains choix de Q seront préférables à d'autres. La seule condition sur Q qui a due être faite est qu'il s'agisse d'une densité ; la structure de dépendance entre les candidats $y^{(1)}, \dots, y^{(K)}$ ainsi que leurs distributions marginales ne sont soumises à aucune restriction. Certains cas particuliers de choix de Q sont présents dans la littérature ; en voici une brève revue.

Algorithme 4.1 Algorithme Metropolis à essais multiples général (MTM)

Données La densité cible π, la densité instrumentale conjointe $Q(\cdot|x)$ sur $\mathcal{X}^K$, la densité conditionnelle de $y^{(-k)}$ sachant $y^{(k)}$ $Q^{(-k)}(\cdot|y^{(k)},x)$, et les poids positifs $w^{(k)}(\cdot,\cdot)$, $k = 1,\dots,K$.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0.
2. *Itérations MCMC.* Pour $n = 1,\dots,N$,
   (a) *Génération des candidats.* Échantillonner $(y^{(1)},\dots,y^{(K)}) \sim Q(\cdot|x_n)$;
   (b) *Calcul des poids.* Calculer les poids de sélection $w^{(k)}(y^{(k)}|x_n)$ pour $k = 1,\dots,K$;
   (c) *Sélection de la proposition.* Échantillonner un indice k parmi $\{1,\dots,K\}$ selon les poids calculés en (b), puis poser $y = y^{(k)}$, la proposition ;
   (d) *Points de référence.* Échantillonner les points de référence $x_*^{(-k)} \sim Q^{(-k)}(\cdot|y^{(k)},x_n)$ puis poser $x_*^{(k)} = x_n$;
   (e) *Calcul des poids inverses.* Calculer les poids inverses $w^{(k)}(x_*^{(k)}|y)$ pour $k = 1,\dots,K$;
   (f) *Probabilité d'acceptation.* Calculer la probabilité d'acceptation M.-H. généralisée (4.5) ;
   (g) *Acceptation.* Accepter la proposition $x_{n+1} = y$ selon la probabilité calculée en (f) ; sinon la rejeter et poser $x_{n+1} = x_n$.

Sortie L'échantillon $x_{0:N}$.

4.1.1.1 Candidats indépendants

Initialement, Liu et collab. (2000) proposent l'utilisation de candidats indépendants et identiquement distribués :

$$Q(y^{(1:K)}|x) = \prod_{k=1}^{K} q(y^{(k)}|x),$$

où $q(\cdot|\cdot)$ est la densité marginale commune. Il s'agit du cas le plus simple de l'algorithme MTM. Évidemment, ce choix de Q ne construit pas les K candidats d'une manière particulièrement judicieuse puisque ceux-ci sont générés d'une manière indépendante depuis une même distribution : certains candidats pourraient être près l'un de l'autre ce qui réduit le potentiel d'efficacité.

4.1.1.2 Candidats conditionnellement indépendants

Liu et collab. (2000) considèrent également une variante où une dépendance entre les candidats est produite via une variable aléatoire auxiliaire. La procédure est définie ainsi : une variable aléatoire $z \in \mathbb{R}^{d_*}$, pour une certaine dimension $d_* \in \mathbb{N}$, est générée selon une certaine distribution $h(\cdot|x)$ possiblement dépendante sur l'état actuel de la chaîne x, puis les K candidats sont générés d'une manière i.i.d. conditionnellement à z. On trouve donc

$$Q(y^{(1:K)}|x) = \int_{\mathbb{R}^{d_*}} \prod_{k=1}^{K} q(y^{(k)}|x,z)h(z|x)\,\mathrm{d}z,$$

où $q(\cdot|\cdot)$ est la densité commune. Les poids dans l'étape de sélection doivent être modifiés afin de satisfaire la condition d'équilibre :

$$w^{(k)}(y|z,x) = \pi(x)q(y|x,z)h(z|x)s(y,x|z),$$

pour une certaine fonction symétrique $s(y,x|z)$ qui peut possiblement dépendre de la variable auxiliaire z. Notons que si $h(z|x) = h(z)$ est indépendant de l'état de la chaîne, alors $h(z)$ se simplifiera dans la fraction de l'acceptation M.-H. et l'on peut effectivement omettre $h(z)$ du calcul de w. Ce type de candidats s'applique particulièrement bien à l'algorithme *Hit-and-run* ; voir l'exemple 4.1.

Exemple 4.1 Algorithme *Hit-and-run* à essais multiples (Liu et collab., 2000)

L'algorithme *Hit-and-run* se voit une extension de l'échantillonneur de Gibbs 2.6. L'algorithme de Gibbs échantillonne le nouvel état de la chaîne par une mise à jour d'une composante du vecteur de la chaîne suivant la distribution conditionnelle aux autres composantes d'état. Quant à lui, l'algorithme *Hit-and-run* détermine d'abord une direction aléatoire dans $\mathbb{R}^d$, puis échantillonne le nouvel état selon la distribution conditionnelle du sous-espace engendré par cette direction et passant par l'état actuel x. L'intérêt de cet algorithme vis-à-vis l'algorithme Gibbs est que la direction d'échantillonnage n'est donc pas nécessairement parallèle à un des axes de $\mathbb{R}^d$, ce qui peut être particulièrement pertinent lorsque π comporte de fortes corrélations.

Soit $U \sim \text{uniforme}(B_1(\mathbf{0}))$, la variable aléatoire correspondant à la direction, où $B_r(\mathbf{c})$ dénote la sphère de rayon r centrée en $\mathbf{c}$. Le nouvel état est donc échantillonné parmi tous les points de la forme $x + ru$, $r \in \mathbb{R}$ et la distribution de r est proportionnelle à $\pi(x + ru)$. Alors que l'algorithme Gibbs bénéficie souvent de l'expression exacte de ces distributions conditionnelles, celles-ci ne seront généralement pas disponibles pour une direction arbitraire. Il convient donc de substituer la distribution conditionnelle par une étape Metropolis, comme dans l'algorithme *Metropolis-within-Gibbs* 2.7.

Maintenant, la distribution instrumentale dans l'étape Metropolis doit être mise au point afin d'obtenir une certaine optimalité d'échantillonnage (e.g. une acceptation de 0,441, tableau 2.1.) Lorsque la variance varie fortement selon la direction u, le choix des paramètres de la distribution instrumentale pourraient ne pas être bien ajusté au sous-espace. Dans ce cas, la stratégie des essais multiples peut s'avérer fructueuse. En choisissant une distribution instrumentale sur $\mathbb{R}$ qui est sur-dispersée par rapport à toutes les directions pour générer les candidats, il sera possible de contourner d'une certaine manière le problème des variances différentes.

Pour une densité sur-dispersée T sur $\mathbb{R}$, les candidats sont générés d'une manière i.i.d. conditionnellement à la direction $U = u$: les pas dans la direction de u sont d'abord générés selon

$$r^{(1)}, \ldots, r^{(K)} \overset{\text{i.i.d.}}{\sim} T(\cdot),$$

puis les candidats sont calculés selon

$$y^{(k)} = x + r^{(k)}u, \qquad k = 1, \ldots, K.$$

Avec un choix judicieux de poids, les candidats aux $r^{(k)}$ petits seront favorisés dans l'étape de sélection lorsque la variance dans la direction de u est petite et vice versa dans la situation opposée.

4.1.1.3 Candidats séquentiellement dépendants

Qin et Liu (2001) proposent de généraliser l'algorithme MTM de Liu et collab. (2000) en introduisant une dépendance séquentielle au sein des candidats. Pour ce faire, les candidats $y^{(1)}, \ldots, y^{(K)}$ sont générés suivant l'ordre naturel $1, \ldots, K$ et les densité instrumentales utilisées dépendent des candidats précédents. Ainsi, $y^{(1)}$ est généré à partir de la densité $Q^{(1)}(\cdot|x)$ où x est l'état actuel de la chaîne,

puis les candidats suivants sont générés à partir des densités conditionnelles suivantes :

$$Y^{(k)}|y^{(k-1:1)} \sim Q^{(k)}(\cdot|y^{(k-1:1)},x), \qquad k = 2\,,\ldots,K.$$

On trouve donc la factorisation de la densité conjointe via les densités conditionnelles successives :

$$Q(y^{(1:K)}|x) = Q^{(1)}(y^{(1)}|x)\prod_{k=2}^{K} Q^{(k)}(y^{(k)}|y^{(k-1:1)},x).$$

Ce type de candidats n'est pas sans rappeler l'algorithme M.-H. à rejet retardé 2.8 : les candidats y sont également générés d'une manière séquentiellement dépendante. Il n'y a cependant pas de poids associé à chaque candidat et il n'y a donc pas d'étape de sélection MTM particulièrement explicite étant donné que les candidats subissent successivement le test d'acceptation M.-H. d'une manière systématique.

4.1.1.4 *Candidats extrêmement antithétiques*

Ensuite, Craiu et Lemieux (2007) proposent d'autres structures de dépendance afin de produire un ensemble de candidats qui soient « bien distribués » dans l'espace. Pour ce faire, deux méthodes sont considérées où la distribution marginale de chaque candidat est la même, c.-à-d.,

$$Q^{(k)}(y|x) = q(y|x), \qquad k = 1\,,\ldots,K.$$

La première méthode, appelée **extrêmement antithétique**, cherche à produire un échantillon de candidats qui soit tel que la distance (euclidienne) entre chaque paire de candidats soit maximale. Si la distribution marginale des candidats est la même, alors il est possible de voir que ce critère correspond à minimiser la corrélation paire-à-paire entre les candidats. En effet, si les candidats $y^{(1)}\,,\ldots,y^{(K)}$ sont distribués selon une certaine distribution F telle que, sans perte de généralité, $\mathbb{E}_F\{Y^{(k)}\} = 0$ pour tout $k = 1\,,\ldots,K$, alors on trouve que la distance carrée moyenne entre les candidats est donnée par

$$\begin{aligned}\mathbb{E}_F\left\{(Y^{(k)} - Y^{(j)})^2\right\} &= \mathrm{Var}_F(Y^{(k)}) + \mathrm{Var}_F(Y^{(j)}) - 2\,\mathrm{Cov}_F(Y^{(k)},Y^{(j)}) \\ &= 2\,\mathrm{Var}_F(Y^{(k)})\left(1 - \mathrm{Corr}_F(Y^{(k)},Y^{(j)})\right),\end{aligned}$$

qui sera maximale lorsque la corrélation entre $Y^{(k)}$ et $Y^{(j)}$ est minimale. À l'exemple 4.2, il est question de comment produire un ensemble de candidats à corrélation minimale dans une marche aléatoire Metropolis à propositions gaussiennes.

La figure 4.1 contient une comparaison entre des candidats générés indépendamment et des candidats générés par la méthode antithétique. On peut voir que les candidats indépendants peuvent ne pas bien être bien répartis par rapport à la densité instrumentale alors que les candidats antithétiques couvrent mieux le support de cette distribution. Comme on s'y attend, le distance euclidienne entre les candidats est supérieure dans l'ensemble antithétique.

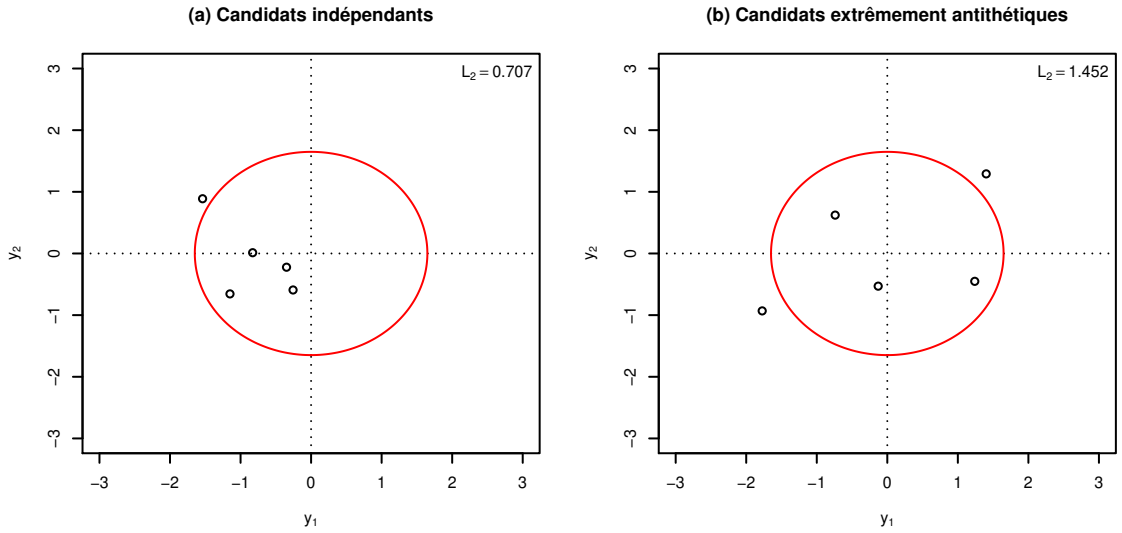


Figure 4.1 *Comparaison entre (a) des candidats générés indépendamment et (b) des candidats générés par la méthode extrêmement antithétique. Dans les deux cas, $K = 5$ candidats sont générés à partir d'une densité marginale normale centrée réduite en $d = 2$ dimensions. L_2 est la distance euclidienne moyenne entre les candidats et l'ellipse représente la région de couverture 95%.*

Exemple 4.2 Algorithme Metropolis à essais multiples à propositions gaussiennes extrêmement antithétiques (Craiu et Lemieux, 2007)

On considère un algorithme Metropolis de type marche aléatoire sur $\mathcal{X} = \mathbb{R}^2$ afin de simplifier la discussion. On suppose donc que chaque candidat aura la même distribution marginale $\mathcal{N}_2((x_1,x_2)^\top,\Sigma)$ où $x = (x_1,x_2)^\top$ est l'état actuel de la chaîne. Cependant, afin de produire un ensemble de K candidats antithétiques, une corrélation sera introduite entre chaque paire de candidats. Pour ce, on considère la distribution conjointe d'une paire de candidats :

$$(Y_1^{(k)},Y_2^{(k)},Y_1^{(j)},Y_2^{(j)})^\top \sim \mathcal{N}_4\left((x_1,x_2,x_1,x_2)^\top, \begin{pmatrix} \Sigma & \Psi \\ \Psi & \Sigma \end{pmatrix}\right),$$

où Ψ correspond à la matrice de covariance entre $Y^{(k)}$ et $Y^{(j)}$. Explicitement, on peut écrire

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \qquad \Psi = \begin{pmatrix} \rho_1\sigma_1^2 & \rho_2\sigma_1\sigma_2 \\ \rho_2\sigma_1\sigma_2 & \rho_1\sigma_2^2 \end{pmatrix}.$$

Les valeurs de σ_1, de σ_2 et de ρ doivent être spécifiées afin de définir la distribution marginale. On est donc intéressé seulement par la valeur de ρ_1 et de ρ_2 afin de produire deux candidats à distance euclidienne maximale. Dans un but de clarté d'exposition, on supposera, sans perte de généralité après transformation orthogonale, que $\rho = 0$.

On considère donc la distribution de la différence entre deux candidats :

$$\begin{aligned} \mathbb{E}\left\{Y^{(j)} - Y^{(k)}\right\} &= \mathbb{E}\left\{Y^{(j)}\right\} - \mathbb{E}\left\{Y^{(k)}\right\} = \mathbf{0}, \\ \mathrm{Var}\left(Y^{(j)} - Y^{(k)}\right) &= 2\,\mathrm{Var}\left(Y^{(j)}\right) - 2\,\mathrm{Cov}\left(Y^{(j)},Y^{(k)}\right) = 2\Sigma - 2\Psi. \end{aligned}$$

Par les propriété des normales multivariées, on trouve donc que

$$Y^{(j)} - Y^{(k)} \sim \mathcal{N}_2\left(\mathbf{0}, 2(\Sigma - \Psi)\right).$$

Maintenant, la distance euclidienne moyenne sera donnée par (Craiu et Lemieux, 2007, propriété reliée à la

distribution du khi, c.-à-d., la racine carrée d'une khi-deux)

$$\mathbb{E}\left\{\left\|Y^{(j)} - Y^{(k)}\right\|_2\right\} = 2\left(\sigma_1^2 + \sigma_2^2\right)(1 - \rho_1).$$

Ainsi, la distance euclidienne sera maximisée lorsque ρ_1 sera minimale (sous contrainte que la matrice globale soit définie semi-positive) et pour n'importe quelle valeur de ρ_2, que l'on peut choisir nulle par simplicité.

Plus généralement, on trouve que la matrice de covariance entre deux candidats en dimension d peut être choisie comme

$$\Psi = \mathrm{diag}(\rho\sigma_1^2, \ldots, \rho\sigma_d^2),$$

où $\rho = -\frac{1}{K-1}$ pour assurer la semi-positivité de la matrice globale

$$\mathrm{Var}\left((Y^{(1)}, \ldots, Y^{(K)})^\top\right) = \begin{pmatrix} \Sigma & \Psi & \cdots & \Psi \\ \Psi & \Sigma & \cdots & \Psi \\ \vdots & \vdots & \ddots & \vdots \\ \Psi & \Psi & \cdots & \Sigma \end{pmatrix}.$$

4.1.1.5 Candidats quasi-Monte Carlo randomisés

La seconde méthode proposée par Craiu et Lemieux (2007) afin de produire un ensemble de candidats ayant la même distribution marginale, mais remplissant bien l'espace se base sur l'idée **quasi-Monte-Carlo**. Plutôt que d'utiliser une suite de points (pseudo-) aléatoires afin d'approcher l'intégrale $\pi(f)$, les méthodes quasi-Monte-Carlo utilisent une suite de points dite **à discrépance faible**. Sans entrer dans les détails mathématiques, il s'agit d'une suite déterministe de points qui remplace un échantillonnage par rapport à une loi uniforme. Ce type de suite est souvent utilisé dans le cadre d'intégration numérique sur des domaines réguliers tels qu'un hypercube.

La transformée intégrale de probabilité (par la fonction des quantiles) permet de faire le lien entre une distribution uniforme sur [0,1] et la fonction de répartition d'une distribution unidimensionnelle. Ainsi, à partir d'une suite à discrépance faible sur [0,1], il est possible d'obtenir une suite hautement uniforme, par rapport à la densité, sur le support de la distribution. Ce procédé peut être étendu à une distribution multidimensionnelle à condition qu'il existe une fonction permettant la transformation de l'hypercube vers le support de la distribution. Par exemple, lorsque les composantes sont indépendantes, il est possible d'appliquer la transformée intégrale de probabilité à chacune des composantes.

Dans le contexte de l'algorithme Metropolis à essais multiples, l'ensemble des candidats doit être généré aléatoirement. Alors, une suite déterministe ne peut pas être utilisée directement afin de produire l'ensemble de candidats. Pour y arriver, une composante aléatoire doit être introduite dans la génération de la suite. On parle alors d'une méthode **quasi-Monte-Carlo aléatoire**. L'exemple 4.3 détaille une méthode particulière de produire ce type de candidats utilisant la règle de Korobov. D'abord une suite à discrépance faible est produite sur l'hypercube, puis elle est déplacée par translation (modulo 1) pour obtenir une suite aléatoire. Ensuite, l'ensemble des points est transformé vers le support de la distribution et est utilisé comme ensemble de candidats à l'étape de sélection de l'algorithme MTM. Une attention particulière doit par la suite être portée au calcul de la probabilité d'acceptation puisque les candidats sont fortement dépendants. En fait, connaître un des points permet de retrouver tous les autres puisqu'une seule variable aléatoire entre dans la génération de l'ensemble.

De plus, plutôt que d'utiliser une densité uniforme sur l'hypercube, Craiu et Lemieux (2007) considèrent également des densités non-uniformes en effectuant une transformation sur l'hypercube.

Par exemple, s'il est d'intérêt d'obtenir des candidats dans les ailes de la distribution instrumentale, il pourra être pertinent de considérer une transformation telle que les extrémités de l'hypercube soient plus denses. Soit $g : [0,1] \to [0,1]$ une bijection qui met un poids plus élevé vers 0 et 1 (e.g. $g(u) = (\sin(u-1/2)\pi+1)/2$). En appliquant g à chaque composante des points hautement uniformes, ces points seront poussés vers les extrémités de $[0,1]^d$ ce qui aura pour effet d'alourdir artificiellement les ailes de la densité instrumentale.

Exemple 4.3 Algorithme Metropolis à essais multiples à propositions de Korobov (Craiu et Lemieux, 2007)

Supposons que l'on cherche à produire des candidats $y^{(1)}, \ldots, y^{(K)}$ tous distribués selon une certaine distribution F à support dans $\mathbb{R}^d \supseteq \mathcal{X}$. Supposons de plus qu'il existe une fonction $G : [0,1)^d \to \mathbb{R}^d$ telle que si $\mathbf{u} \sim \text{uniforme}[0,1)^d$ alors $G(\mathbf{u}) \sim F$.

La première étape consiste à produire un ensemble de points à discrépance faible (hautement uniforme) sur l'hypercube $[0,1)^d$. Pour ce, la **règle de Korobov** sera considérée (à noter que d'autres méthodes sont également possibles, mais que celle-ci est parmi les plus simples.) On choisit d'abord un nombre entier $a \in \{1, \ldots, K-1\}$ puis l'ensemble de points est définit par

$$P_{a,K} = \left\{ \frac{k-1}{K} \left(1, a, \ldots, a^{d-1}\right) \mod 1 \;\middle|\; k = 1, \ldots, K \right\}.$$

Pour chaque choix de a et de K, cet ensemble est entièrement déterminé.

La seconde étape effectue une transformation de $P_{a,K}$ de sorte à rendre l'ensemble aléatoire. Pour ce, un point $\mathbf{u}$ uniformément aléatoire sur $[0,1)^d$ est généré et tous les points de $P_{a,K}$ sont glissés de $\mathbf{u}$ en forçant la somme à rester dans l'hypercube. Si $\mathbf{u}^{(k)}$ dénote le k-ième point de $P_{a,K}$, alors on définit

$$\tilde{\mathbf{u}}^{(k)} = \left(\mathbf{u}^{(k)} + \mathbf{u}\right) \mod 1, \qquad k = 1, \ldots, K,$$

et on dénote l'ensemble de ces points par $\tilde{P}_{a,K}(\mathbf{u}) = \{\tilde{\mathbf{u}}^{(k)} | k = 1, \ldots, K\}$. Désormais, l'ensemble $\tilde{P}_{a,K}(\mathbf{u})$ définit un ensemble hautement uniforme sur l'hypercube, mais qui est alors aléatoire. Notons que la valeur de a peut demeurer constante entre les itérations MCMC.

La troisième étape dans la génération des candidats est de transformer les points de $\tilde{P}_{a,K}(\mathbf{u})$ selon la fonction $G(\cdot)$ de sorte à produire un ensemble de candidats hautement uniforme par rapport à la distribution F. Enfin, les candidats sont donnés par

$$y^{(k)} = G(\tilde{\mathbf{u}}^{(k)}), \qquad k = 1, \ldots, K.$$

En particulier, si on cherche à produire une marche aléatoire gaussienne, on aura $F = \mathcal{N}_d(x, \Sigma)$, où x correspond à l'état actuel de la chaîne. La fonction G prend alors la forme suivante. D'abord, les K points sont transformés vers une distribution $\mathcal{N}(\mathbf{0}, I_d)$ via la transformée intégrale de probabilité indépendante sur chaque composante des points hautement uniformes. Puis, la distribution normale arbitraire est obtenue par la transformation usuelle $y^{(k)} = x + \Sigma^{1/2} z^{(k)}$, où $z^{(k)}$ correspond à la transformée intégrale de probabilité de $\tilde{\mathbf{u}}^{(k)}$ par la normale centrée réduite et $\Sigma^{1/2}$ correspond à la « racine carrée » de Σ dans le sens où $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. La figure 4.2 contient une réalisation de la procédure pour produire des candidats venant d'une normale bivariée centrée réduite ainsi qu'un exemple de transformation alourdissant les ailes de la densité normale.

Maintenant que les candidats sont générés, l'étape de sélection MTM selon les poids (4.4) peut être effectuée sans problème. Par construction, les distributions marginales de $y^{(1)}, \ldots, y^{(K)}$ seront toutes données par F. Une fois la proposition M.-H. sélectionnée, l'étape d'acceptation requiert l'échantillonnage de points de référence afin d'assurer la condition d'équilibre. Ces points de référence doivent être choisis de sorte à provenir du chemin inverse et à ce qu'un des points soit exactement l'état actuel de la chaîne x. Tel que mentionné précédemment, connaître un point de l'ensemble de candidats détermine tous les autres. Ainsi, savoir que x est dans l'ensemble des points de référence détermine tous les autres points. Il faut donc trouver le décalage $\mathbf{w}$ (correspondant au $\mathbf{u}$ plus haut) assurant le passage de la proposition y vers x. Il est à noter qu'il faut également considérer une permutation des indices $1, \ldots, K$ dans l'attribution de x à un des points de référence afin d'assurer la condition d'équilibre, mais Craiu et Lemieux (2007, proposition 3.1) montrent que cette permutation n'est pas requise et que l'on peut poser systématiquement $x_*^{(1)} = x$.

Dans l'exemple d'un algorithme MTM à propositions gaussiennes sans transformation G, retrouver $\mathbf{w}$

peut être fait comme suit. Si $\Phi(\cdot)$ dénote la fonction de répartition d'une distribution normale centrée réduite et y dénote la proposition, alors

$$\mathbf{w}_j = \Phi\left([\Sigma^{-1/2}(y-x)]_j\right), \qquad j = 1, \ldots, d.$$

Une fois $\mathbf{w}$ déterminé, les points de référence sont produits de la même façon que les candidats. Enfin la probabilité d'acceptation M.-H. généralisée (4.2) est calculée et la proposition y est acceptée avec cette probabilité.

4.1.1.6 Candidats indépendants à distributions marginales distinctes

Les différentes manières de générer les candidats dans l'algorithme MTM étudiées jusqu'ici, à l'exception des candidats séquentiellement dépendants, supposaient toutes que la distribution marginale des candidats était la même à travers l'ensemble des candidats. Comme le suggère la proposition 4.1, une telle supposition n'est pas requise. Casarin et collab. (2013, algorithme 2) proposent justement l'utilisation de candidats générés indépendamment, mais ayant des distributions marginales distinctes.

Dans le contexte d'un algorithme Metropolis, on peut réinterpréter ce type de candidats comme une mixture locale. En effet, chaque distribution marginale peut être vue comme une composante de la mixture et les poids (4.4) agissent comme les poids de la mixture. Ainsi, si une région du support de π est mieux représentée par la k-ième distribution, alors le poids associé au candidat venant de cette distribution $Q^{(k)}(\cdot|x)$ risque d'être plus élevé. Ce genre de construction n'est donc pas sans rappeler l'algorithme Metropolis adaptatif à mixture locale de Andrieu et Thoms (2008, algorithme 7). La différence fondamentale est que les poids de la mixture sont alors calculés par rapport à tout le passé de la chaîne plutôt que par rapport à l'état actuel de la chaîne et aux candidats générés.

Évidemment, la supposition d'indépendance peut également être relâchée ; il est donc possible de combiner les méthodes exposées précédemment à des distributions marginales distinctes. La section suivante contient également un exemple supplémentaire de cette situation.

4.1.1.7 Candidats par variable aléatoire commune

Une généralisation de la méthode quasi-Monte-Carlo aléatoire est l'utilisation d'une seule variable aléatoire commune afin de générer l'ensemble des candidats. En effet, au cours de la génération des candidats à l'exemple 4.3, la seule variable aléatoire utilisée est celle effectuant la translation aléatoire des points hautement uniformes.

La définition plus générale de cette stratégie est la suivante (Bédard et collab., 2012). Soit $U \sim \text{uniforme}[0,1)^d$ une variable aléatoire sur l'hypercube de dimension d. Pour chaque $k = 1, \ldots, K$, il existe une fonction mesurable $G^{(k)} : [0,1)^d \times \mathcal{X} \to \mathcal{X}$ de sorte que la variable aléatoire donnée par la transformation $Y^{(k)} = G^{(k)}(U|x)$ admette la distribution marginale $Q^{(k)}(\cdot|x)$. De plus, pour chaque paire $(j,k) \in \{1, \ldots, K\}^2$, il existe une fonction mesurable $G^{(j,k)} : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ qui permet de retrouver $y^{(k)}$ à partir de $y^{(j)}$ et de x, c.-à-d., $y^{(k)} = G^{(j,k)}(y^{(j)}|x)$.

Par exemple, si les différentes densité marginales sont constituées de normales à covariances différentes $\{\Sigma^{(k)}\}_{k=1}^K$, on pourra d'abord générer un vecteur aléatoire venant d'une normale centrée réduite puis utiliser ce même vecteur pour générer tous les candidats via la transformation usuelle :

$$Z \sim \mathcal{N}_d(\mathbf{0}, I_d), \qquad\qquad y^{(k)} = x + (\Sigma^{(k)})^{1/2} z, \qquad k = 1, \ldots, K.$$

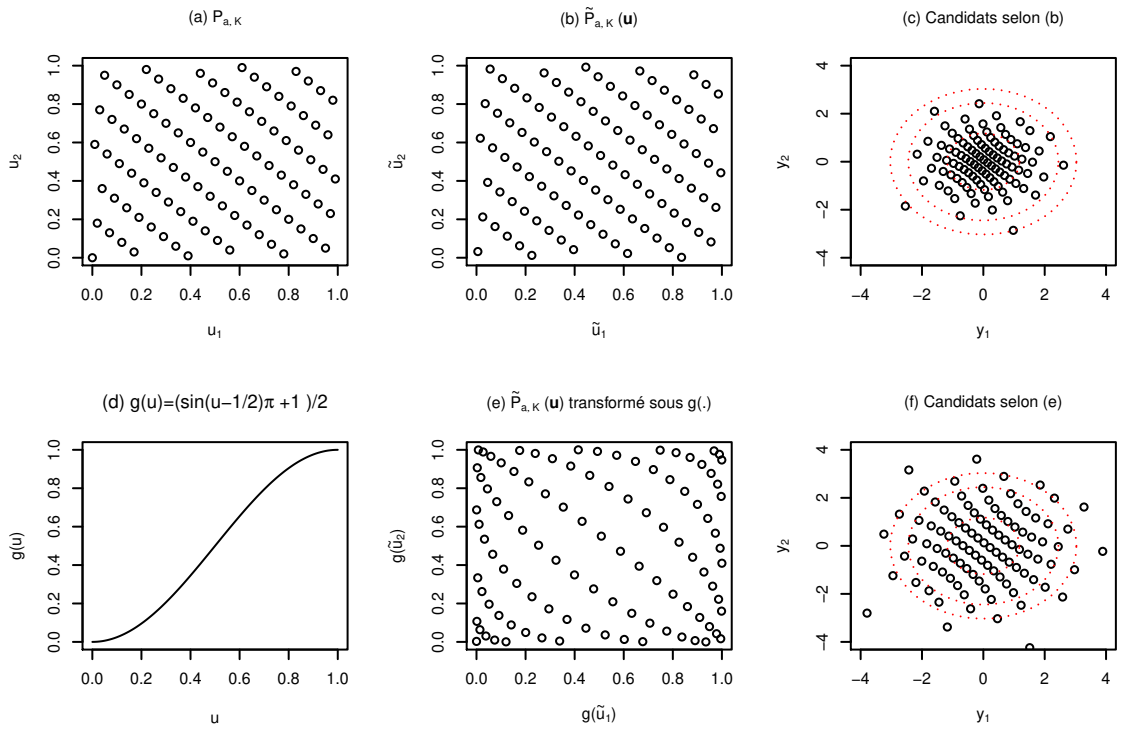


Figure 4.2 *Candidats quasi-Monte Carlo via une règle de Korobov avec $K = 100$ points et $a = 59$ en $d = 2$ dimensions : (a) l'ensemble $P_{a,K}$ de points hautement uniforme sur l'hypercube, (b) l'ensemble des points décalés aléatoirement $\tilde{P}_{a,K}(\mathbf{u})$, (c) l'ensemble des candidats obtenus par transformation vers une normale centrée réduite à partir des points de (b), (d) la transformation $G(u) = (\sin(u-1/2)\pi+1)/2$, (e) l'ensemble des points décalés aléatoirement $\tilde{P}_{a,K}(\mathbf{u})$ transformé sous $G(\cdot)$ et (f) l'ensemble des candidats obtenus à partir de (e). Les tracés rouges pointillés définissent respectivement les régions de 50%, 95% et de 99% de la distribution normale.*

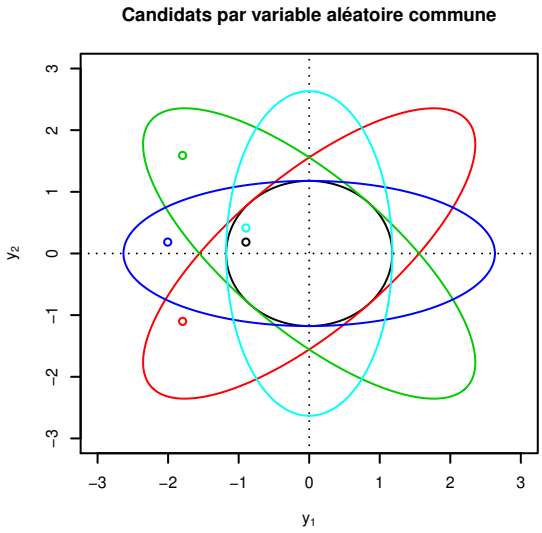


Figure 4.3 *Candidats générés par une variable aléatoire commune vers des densités normales bivariées à covariances différentes. Les points représentent les candidats alors que les ellipses définissent les régions de couverture 50% de chacune des densités gaussiennes.*

La variable aléatoire uniforme U est ici implicite dans la génération de Z par transformée intégrale de probabilité gaussienne sur chacune des composante de U. La figure 4.3 contient une représentation de cette méthode.

4.1.2 Choix des poids

Frenkel et Smit (1996) proposent l'algorithme Monte Carlo à biais orienté (*Orientationally biased Monte Carlo*) dans le contexte de la biologie moléculaire. Il s'agit en fait du cas particulier de l'algorithme MTM à candidats indépendants et identiquement distribués pour le choix de poids $w^{(k)}(y|x) = \pi(y)$. Liu et collab. (2000) généralisent alors cette technique en considérant une forme offrant plus de liberté dans les poids $w^{(k)}(y|x) = \pi(y)q(x|y)s(x,y)$ en supposant une distribution instrumentale marginale identique pour tous les candidats. Les conditions sur q et sur s sont que $q(y|x) > 0$ si et seulement si $q(x|y) > 0$, que s soit symétrique (c.-à-d., $s(x,y) = s(y,x)$) et que $s(x,y) > 0$ dès que $q(y|x) > 0$. Ce type de poids se généralise trivialement à des densités marginales différentes ainsi qu'à des fonctions symétriques différentes parmi l'ensemble des candidats. Enfin, la généralisation due à Pandolfi et collab. (2010) relâchent toute supposition sur l'expression analytique de $w^{(k)}$. En fait, en utilisant la probabilité d'acceptation généralisée (4.5), on ne requiert que la positivité des poids : $w^{(k)}(y|x) > 0$.

Devant si peu de restrictions, une question fondamentale dans l'élaboration d'un algorithme MTM est donc le choix de la fonction $w^{(k)}$. Alors que certains résultats expérimentaux peuvent indiquer la supériorité de certains choix sous certaines conditions, aucun résultat théorique ne permet d'établir qu'une expression de poids soit optimale par rapport à un certain critère. Ainsi, l'utilisateur doit effectuer le choix des poids de sorte à favoriser le comportement recherché de la part de l'algorithme. En rappelant que l'objectif principal des algorithmes à essais multiples est une exploration plus rapide et plus efficace du support de la distribution cible π, l'expression de $w^{(k)}$ peut être choisie de sorte à

favoriser la sélection de propositions qui sont le plus éloignées de l'état actuel, par exemple.

Cette section sera donc consacrée à l'étude des différentes possibilités dans l'expression de $w^{(k)}$ ainsi que des raisons justifiant ces choix.

4.1.2.1 Choix de la fonction symétrique

Bien que $w^{(k)}$ puisse prendre une forme arbitraire positive, les choix sensés de poids seront souvent parmi les expressions de la forme de (4.4). En effet, inclure le facteur $\pi(y)$ dans l'expression des poids est presque primordial afin d'effectuer une sélection d'une proposition qui explore bien le support de π. De plus, la densité de transition $Q^{(k)}(x|y)$ sera souvent symétrique et pourra donc être absorbée par le choix de la fonction symétrique $s^{(k)}$. On considère ici quelques exemple de choix de $s^{(k)}$.

Lorsque $Q^{(k)}(x|y)$ est symétrique, il est possible de choisir $s^{(k)}(x,y) = Q^{(k)}(x|y)^{-1}$ de sorte à obtenir un poids égal à la densité cible :

$$w^{(k)}(y|x) = \pi(y)Q^{(k)}(x|y)Q^{(k)}(x|y)^{-1} = \pi(y). \tag{4.6}$$

Ceci engendre alors un algorithme MTM où les candidats à haute densité seront favorisés au moment de la sélection. Lorsque la densité instrumentale n'est pas symétrique, il est possible d'utiliser la fonction suivante afin d'obtenir un comportement similaire :

$$s^{(k)}(x,y) = \left(\frac{Q^{(k)}(y|x) + Q^{(k)}(x|y)}{2}\right)^{-1} \tag{4.7}$$

Le choix le plus simple pour $s^{(k)}$ est de poser $s^{(k)}(x,y) \equiv 1$, ce qui a pour effet de produire des poids semblables aux termes d'acceptation M.-H. :

$$w^{(k)}(y|x) = \pi(y)Q^{(k)}(x|y). \tag{4.8}$$

Les candidats favorisés seront ceux tels que la paire (y, x) a une densité conjointe élevée, ce qui advient lorsque y a une densité cible élevée et la transition $y \to x$ a une densité élevée.

Considérons la fonction symétrique suivante :

$$s^{(k)}(x,y) = \left(Q^{(k)}(y|x)Q^{(k)}(x|y)\right)^{-\alpha}, \qquad \alpha > 0. \tag{4.9}$$

Lorsque $\alpha = 1$, on trouve des poids semblables aux poids d'importance de l'algorithme Monte Carlo par échantillonnage préférentiel 2.2 :

$$w^{(k)}(y|x) = \pi(y)Q^{(k)}(x|y)\left(Q^{(k)}(y|x)Q^{(k)}(x|y)\right)^{-\alpha} \overset{\alpha=1}{=} \frac{\pi(y)}{Q^{(k)}(y|x)}. \tag{4.10}$$

Les candidats favorisés dans la sélection seront ceux à densité cible élevée ou bien à transition $x \to y$ de faible densité, ce qui favorisera à la fois l'exploration du support de π ainsi que de grands sauts.

Une des mesures d'efficacité d'un algorithme MCMC est le saut quadratique moyen donné par la moyenne des normes euclidienne entre deux états consécutifs de la chaîne, $\frac{1}{N}\sum_{n=0}^{N-1} \|x_{n+1} - x_n\|_2$. Un saut quadratique moyen élevé est associé à un faible temps d'autocorrélation et donc une estimation plus efficace (section 2.4.3.3.) Puisqu'une norme est une fonction symétrique, il est possible d'utiliser ce critère à même les poids de sélection pour favoriser un algorithme tel que le saut quadratique moyen

soit élevé. C'est l'idée que proposent Yang et collab. (2019) dans un algorithme MTM par composante. Dans un algorithme MTM régulier, on peut donc choisir, en supposant une transition symétrique,

$$s^{(k)}(x,y) = Q^{(k)}(y|x)^{-1}\|y-x\|_2^{\alpha}; \tag{4.11}$$

$$w^{(k)}(y|x) = \pi(y)\|y-x\|_2^{\alpha}. \tag{4.12}$$

Empiriquement, Yang et collab. (2019) trouvent que $\alpha \in (2,4)$ procure des résultats satisfaisants dans le contexte par composante.

4.1.2.2 Choix de poids généraux

Tel que mentionné, la seule supposition sur $w^{(k)}$ vraiment nécessaire est la positivité. On considère ici d'autres expressions pour les poids qui ne sont pas de la forme (4.4).

Dans une étude de la flexibilité des algorithmes MTM, Martino et Read (2013) effectuent une expérimentation sur l'efficacité de différents types de poids. En plus des expressions (4.6), (4.8) et (4.10) considérées précédemment, les auteurs explorent les choix suivants.

D'abord, des poids égaux à une puissance de la densité cible sont étudiés, ce qui généralise (4.6). Ces poids sont de la forme

$$w^{(k)}(y|x) = [\pi(y)]^{\alpha}, \qquad \alpha \in \mathbb{R}\,. \tag{4.13}$$

Martino et Read (2013) considèrent $\alpha \in \{0,1/2,1,2,3\}$. Lorsque $\alpha \in (0,1)$, les régions de densité plus faibles ($\pi(x) \ll 1$) seront visitées plus facilement que lorsque $\alpha = 1$, mais les régions de haute densité ($\pi(x) > 1$) risquent d'être sous-représentées. Inversement, lorsque $\alpha > 1$, les régions de faible densité ($\pi(x) \ll 1$) seront davantage difficiles à atteindre.

Ensuite, Martino et Read (2013) considèrent des poids donnés par des puissances de la densité instrumentale :

$$w^{(k)}(y|x) = [Q^{(k)}(y|x)]^{\alpha}, \qquad \alpha \in \mathbb{R}\,. \tag{4.14}$$

Dans leurs expérimentations, ils étudient les cas $\alpha = 1$ et $\alpha = -1$.

Parmi toutes ces options de poids à l'étape de sélection, Martino et Read (2013) observent que les meilleurs résultats sont obtenus en utilisant les expressions (4.6), (4.8) et (4.10), alors que les poids puissance de la densité cible ou instrumentale performent significativement moins bien en terme d'acceptation et d'autocorrélation. C'est donc dire que les choix sensés de la section précédente sont tous préférables et que la généralisation aux poids arbitraires n'apporte pas, empiriquement, d'avantage en terme d'efficacité. De plus, ils notent que les poids d'importance (4.10) affichent des résultats légèrement supérieurs aux autres types de poids. Liu et collab. (2000) arrivent à des conclusions similaires dans leurs expériences : le choix de α dans la famille (4.9) n'influence pas particulièrement les résultats, mais α près de -1 semble être légèrement préférable. Casarin et collab. (2013) observent également une légère supériorité de poids utilisant (4.9) plutôt que (4.7).

Notons que le cas $\alpha = 0$ dans (4.13) ou dans (4.14) correspond à des poids égaux : cette stratégie est en fait équivalente à n'utiliser qu'un seul candidat (c.-à-d., l'algorithme M.-H. régulier) puisqu'aucune information sur les candidats n'est utilisée : une proposition est choisie au hasard uniformément dans l'ensemble et l'information recueillie par les candidats additionnels n'est jamais utilisée. Empiriquement, Martino et Read (2013) observent des résultats identiques à ceux d'un algorithme M.-H. à un seul candidat.

4.1.2.3 Approximation quadratique de la densité cible

Pour tout algorithme MCMC, la partie la plus couteuse computationnellement est souvent le calcul de la densité cible π. Les algorithmes MTM exigent plusieurs – jusqu'à $2K-1$ dans certains cas – calculs de la sorte à chaque itération. Afin de pallier à ce problème, Pandolfi et collab. (2010) proposent d'utiliser une approximation de π dans le calcul des poids plutôt que π elle-même. Si cette approximation est peu coûteuse par rapport à celle de π et que les poids résultant représentent bien les vrais poids, il peut en sortir une grande augmentation dans l'efficacité par rapport au temps de calcul.

Lorsque la chaîne se trouve en x, la densité cible est approximée localement de la façon suivante. On considère

$$\pi^*(y) = \pi(x)A(y|x),$$

où

$$\log A(y|x) = [\nabla \log \pi(x)]^\top (y-x) + \frac{1}{2}(y-x)^\top [\nabla^2 \log \pi(x)](y-x),$$

$\nabla \log \pi(x)$ est le gradient par rapport à x de $\log \pi$ évalué en x et $\nabla^2 \log \pi(x)$ est la matrice Hessienne par rapport à x de $\log \pi$ évaluée en x. Cette approximation permet alors de produire des poids approximatifs en substituant π^* à π. Par exemple, les poids (4.4) deviennent

$$w^{*(k)}(y|x) = \pi^*(y)Q^{(k)}(x|y)s(x,y).$$

L'avantage principal de cette construction est que le terme $\pi(x)$ se simplifie lors du calcul des poids standardisés :

$$\begin{aligned}\bar{w}^{*(k)}(y; y^{(-k)}|x) &= \frac{w^{*(k)}(y|x)}{\sum_{j=1}^K w^{*(j)}(y^{(j)}|x)} \\ &= \frac{A(y|x)Q^{(k)}(x|y)s(x,y)}{\sum_{j=1}^K A(y^{(j)}|x)Q^{(j)}(x|y^{(j)})s(x,y^{(j)})}.\end{aligned}$$

Ainsi, seulement deux calculs de densité cible seront requis au cours d'une itération puisque la probabilité d'acceptation généralisée requiert $\pi(x)$ et $\pi(y)$. Lorsque $\nabla \log \pi(x)$ et $\nabla^2 \log \pi(x)$ sont relativement simples à calculer par rapport à π, cette méthode s'avère donc avantageuse.

Les expérimentations effectuées par Pandolfi et collab. (2010) sur cet algorithme montrent que le coût computationnel réduit compense suffisamment la perte en précision, ce qui produit un algorithme plus efficace globalement, comparativement à des algorithmes MTM réguliers.

4.1.2.4 Candidats séquentiellement dépendants

Dans leur algorithme Metropolis multipoint, Qin et Liu (2001) considèrent des candidats séquentiellement dépendants. Les poids qu'ils utilisent sont les suivants :

$$w^{(k)}(y^{(k)}|y^{(k-1:1)}, x) = \pi(y^{(k)})Q^{(k)}(x, y^{(1:k-1)}|y^{(k)})s^{(k)}(y^{(k:1)},x),$$

où $s^{(k)}$ est une fonction positive, bornée et **séquentiellement symétrique**, c'est-à-dire

$$s^{(k)}(y^{(k:1)},x) = s^{(k)}(x,y^{(1:k)}).$$

La probabilité d'acceptation M.-H.

$$\alpha(y,y^{(-k)}|x,x_*^{(-k)}) = \min\left\{1, \frac{\sum_{k=1}^K w^{(k)}(y^{(k)}|y^{(k-1:1)},x)}{\sum_{k=1}^K w^{(k)}(x_*^{(k)}|x_*^{(k-1:1)},y)}\right\}$$

est utilisée afin d'accepter ou de rejeter la proposition choisie selon les poids $w^{(k)}(y^{(k)}|y^{(k-1:1)},x)$, $k = 1,\ldots,K$. Les choix de fonctions séquentiellement symétriques proposées incluent $s^{(k)}(y^{(k:1)},x) \equiv 1$ et, supposant la symétrie séquentielle de la transition $Q^{(k)}$, nous pouvons choisir $s^{(k)}(y^{(k:1)},x) = [Q^{(k)}(y^{(k:1)}|x)]^{-1}$ dans lequel cas les poids se réduisent à $w^{(k)}(y^{(k)}|y^{(k-1:1)},x) = \pi(y^{(k)})$.

Martino et collab. (2012) appliquent la généralisation des poids de Pandolfi et collab. (2010) à l'algorithme Metropolis multipoint. Ainsi, pour les poids standardisés

$$\bar{w}^{(k)}(y^{(k)}|y^{(k-1:1)},x) = \frac{w^{(k)}(y^{(k)}|y^{(k-1:1)},x)}{\sum_{j=1}^K w^{(j)}(y^{(j)}|y^{(j-1:1)},x)},$$

on trouve la probabilité d'acceptation M.-H.

$$\alpha(y,y^{(-k)}|x,x_*^{(-k)}) = \min\left\{1, \frac{\pi(y)Q^{(k)}(x_*^{(1:k)}|y)\bar{w}^{(k)}(x_*^{(k)}|x_*^{(k-1:1)},y)}{\pi(x)Q^{(k)}(y^{(1:k)}|x)\bar{w}^{(k)}(y^{(k)}|y^{(k-1:1)},x)}\right\}.$$

Ce choix différent de probabilité d'acception satisfait tout de même la condition d'équilibre (Martino et collab., 2012, section 4), produisant ainsi un algorithme valide du point de vue théorique.

Évidemment, le choix de $w^{(k)}$ peut être effectué de n'importe quelle manière, mais des expressions similaires à celles des sections précédentes seront plus pertinentes. Les auteurs en proposent quelques unes :

$$\begin{aligned}
&w^{(k)}(y^{(k)}|y^{(k-1:1)},x) = \pi(y^{(k)});\\
&w^{(k)}(y^{(k)}|y^{(k-1:1)},x) = \pi(y^{(k)})\pi(y^{(k-1)})\cdots\pi(y^{(1)})\pi(x);\\
&w^{(k)}(y^{(k)}|y^{(k-1:1)},x) = \left[\frac{\pi(y^{(k)})}{Q^{(k)}(y^{(k:1)}|x)}\right]^{\alpha}, \qquad \alpha > 0;\\
&w^{(k)}(y^{(k)}|y^{(k-1:1)},x) = \frac{\pi(y^{(k)})}{Q^{(k)}(y^{(k:1)}|x)}\frac{\pi(y^{(k-1)})}{Q^{(k-1)}(y^{(k-1:1)}|x)}\cdots\frac{\pi(y^{(1)})}{Q^{(1)}(y^{(1)}|x)};\\
&w^{(k)}(y^{(k)}|y^{(k-1:1)},x) = \frac{\pi(y^{(k)})}{Q^{(k)}(y^{(k)}|y^{(k-1:1)},x)}.
\end{aligned} \tag{4.15}$$

Dans leur expérimentation, les auteurs trouvent que les poids (4.15), similaires aux poids d'importance, performent le mieux en terme d'autocorrélation de la chaîne.

4.1.3 Généralisations et variantes

Comme tout algorithme MCMC, il est possible de combiner plusieurs approches ensemble. Par exemple, Pandolfi et collab. (2010) considèrent un algorithme MTM à sauts réversibles. Les algorithmes à sauts réversibles, sans entrer dans les détails, sont utilisés principalement dans le contexte de la sélection de modèles bayésiens et l'introduction d'essais multiples peut améliorer la performance de ce type d'algorithme. Dans cette section, d'autres généralisations et variantes de l'algorithme MTM seront étudiées.

4.1.3.1 Algorithmes à chaînes parallèles

Dans l'esprit des méthodes Monte Carlo par population, il est possible d'augmenter un algorithme MTM de chaînes paralèlles. C'est ce que proposent Casarin et collab. (2013) par leur algorithme IMTM (*Interacting Multiple-Try Metropolis.*)

On considère une population de C chaînes $\{x_{n,c}\}_{n\in\mathbb{N}}$, $c = 1,\dots,C$. Pour chaque chaîne c, on procède à un algorithme MTM à K_c candidats indépendants et de densités marginales $Q_c^{(k)}$, $k = 1,\dots,K_c$, possiblement différentes. De plus, afin de faire en sorte qu'une chaîne puisse profiter de l'information additionnelle contenue dans les autres chaînes de la population, les distributions instrumentales peuvent dépendre de l'état actuel de toutes les chaînes plutôt que de l'état actuel de sa propre chaîne seulement.

Notons que l'on pourrait interpréter cette construction comme un algorithme à adaptation externe (cf. sous-section 3.1.2), mais le fait de ne dépendre que de l'état actuel de toutes les chaînes préserve la propriété markovienne de l'ensemble des chaînes vue comme une chaîne unique dans $\mathcal{X}^C$. Ainsi, le contexte adaptatif n'est pas nécessaire à l'analyse de cet algorithme.

En définissant $\Xi_n = \{x_{n,c}\}_{c=1}^C$ comme l'ensemble des états des chaînes au temps n, on écrit alors la densité de proposition dans la chaîne c par $Q_c^{(k)}(y|\Xi_n)$. Au temps n, on génère donc indépendamment $Y_c^{(k)} \sim Q_c^{(k)}(\cdot|\Xi_n)$ pour chaque $k = 1,\dots,K_c$ et $c = 1,\dots,C$. On écrit ensuite $\Xi_{n,c}(x) = (x_{n,1:c-1}, x, x_{n,c+1:C})$, c.-à-d., Ξ_n où l'on remplace à c-ième entrée par x. Ceci permet d'écrire les poids de sélection :

$$w_c^{(k)}(y|x) = \pi(y)Q_c^{(k)}(x|\Xi_{n,c}(y))s_c^{(k)}(y,x), \qquad k = 1,\dots,K_c,\ c = 1,\dots,C,$$

où $s_c^{(k)}$ est une fonction positive et symétrique. Une proposition est ensuite choisie pour chaque chaîne c selon les poids $w_c^{(k)}$, $k = 1,\dots,k_c$. Soit k_c l'indice du candidat choisi dans la chaîne c de sorte que $y_c = y_c^{(k_c)}$ est la proposition choisie dans cette même chaîne. Afin d'assurer la réversibilité de la chaîne, des points de référence doivent alors être échantillonnés selon

$$\begin{aligned} x_{*,c}^{(k)} &\sim Q_c^{(k)}(\cdot|\Xi_{n,c}(y_c)), \qquad k \neq k_c,\ c = 1,\dots,C, \\ x_{*,c}^{(k_c)} &= x_{n,c}. \end{aligned}$$

Enfin, la probabilité d'acceptation M.-H. généralisée (4.2) est utilisée dans chaque chaîne afin de décider du nouvel état de la chaîne. Suivant cette construction, Casarin et collab. (2013, théorème 1) montrent que la transition induite par l'algorithme satisfait la condition d'équilibre (4.1) pour chacune des C chaînes. Puisque chaque chaîne satisfait la condition d'équilibre, alors la chaîne unique sur $\mathcal{X}^C$ satisfait à cette condition également, ce qui assure la validité de l'algorithme.

Maintenant, il reste à déterminer les densités instrumentales ainsi que les fonctions symétriques d'une manière à utiliser efficacement l'ensemble des chaînes à chaque itération. En effet, l'augmentation considérable du temps de calcul due à l'ajout de chaînes parallèles doit être compensée par une augmentation de l'efficacité. Ceci peut être atteint en produisant des candidats de meilleure qualité ainsi qu'en sélectionnant les propositions judicieusement.

Le choix des densités instrumentales dépend fortement du cas spécifique d'application, mais l'utilisation de l'état des autres chaînes devrait être priorisée afin de tirer au maximum profit des chaînes parallèles. Par exemple, on peut considérer une marche aléatoire gaussienne où la moyenne de la densité normale est choisie aléatoirement parmi les différents états de la chaîne. Ceci pourrait permettre aux chaînes de sauter de modes en modes et donc de visiter la totalité du support de π

plutôt que de rester prises dans un seul mode.

Quant au choix de la fonction symétrique, l'information contenue dans les autres chaînes peut être utilisée afin d'ajuster les poids. Par exemple, Casarin et collab. (2013) proposent d'utiliser les poids suivants :

$$s_c^{(k)}(x,y) = v^{(k)}[Q_c^{(k)}(x|y)Q_c^{(k)}(y|x)]^{-\alpha}, \qquad \alpha > 0,$$

où $v^{(k)}$ est un facteur accordant plus de poids si les candidats k ont été sélectionnés au temps précédent par les chaînes parallèles :

$$v^{(k)} = \frac{1}{C}\left[1 + \sum_{c=1}^{C} \mathbb{1}(k = k_c)\right].$$

Empiriquement, les auteurs observent cependant que cet ajustement n'augmente pas particulièrement la performance par rapport à choisir $v^{(k)} \equiv 1$.

De plus, le contexte des chaînes parallèles se prête bien au tempérage (section 3.1.3.3). En effet, lorsque la densité cible est une version tempérée de π pour chaque chaîne (e.g. $\pi^{1/c}$, $c = 1, \ldots, C$), l'échange d'information entre les chaînes peut aider la chaîne principale (celle avec π) à explorer plus facilement son support.

4.1.3.2 Généralisations de la probabilité d'acceptation

Martino et Read (2013) étudient la possibilité d'utiliser d'autres expressions que la probabilité d'acceptation M.-H. généralisée (4.5) à l'étape d'acceptation/rejet de l'algorithme MTM.

La condition principale que l'on doit imposer à la probabilité d'acceptation est la condition d'équilibre (4.1). Martino et Read (2013, section 3.2) montrent que la probabilité d'acceptation $\alpha(y|x)$ doit satisfaire la condition suivante :

$$\pi(x)Q^{(k)}(y|x)\bar{w}^{(k)}(y|x)\alpha(y|x) = \pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x|y)\alpha(x|y), \tag{4.16}$$

où la dépendance sur les autres candidats et points de référence est implicite par simplicité d'écriture. Les auteurs montrent également qu'une certaine forme de probabilité d'acceptation satisfait à cette condition.

Proposition 4.3 (Martino et Read, 2013, section 3) *Supposons que la probabilité d'acceptation soit de la forme*

$$\alpha(y|x) = \beta(y|x)\gamma(y,y^{(-k)}|x,x^{(-k)}) \in [0,1], \tag{4.17}$$

où β et γ satisfont aux conditions suivantes :

$$\pi(x)Q^{(k)}(y|x)\beta(y|x) = \pi(y)Q^{(k)}(x|y)\beta(x|y), \qquad k = 1, \ldots, K; \tag{4.18}$$

$$\bar{w}^{(k)}(y|x)\gamma(x,x^{(-k)}|y,y^{(-k)}) = \bar{w}^{(k)}(x|y)\gamma(y,y^{(-k)}|x,x^{(-k)}); \tag{4.19}$$

Alors, $\alpha(y|x)$ satisfait à la condition d'équilibre (4.16).

La proposition 4.3 énonce des conditions suffisantes sur la forme possible de l'expression de la probabilité d'acceptation. Intuitivement, (4.18) exige que la fonction β satisfasse la condition d'équilibre pour toutes les densités instrumentales et (4.19) exige quant à elle une condition d'équilibre dans l'étape de sélection. Parmi plusieurs choix considérés, la paire de fonctions qui suit est celle

affichant les meilleurs résultats empiriques :

$$\beta(y|x) = \min\left\{1, \frac{\pi(y)Q^{(k)}(x|y)}{\pi(x)Q^{(k)}(y|x)}\right\};$$
$$\gamma(y,y^{(-k)}|x,x^{(-k)}) = \min\left\{1, \frac{\bar{w}^{(k)}(x|y)}{\bar{w}^{(k)}(y|x)}\right\}.$$

Il s'agit en fait que la probabilité d'acceptation M.-H. régulière multipliée par le rapport des poids restreint à [0,1]. Notons qu'il ne s'agit pas exactement de la probabilité d'acceptation généralisée (4.5) étant donné que le minimum avec 1 est pris séparément dans chacune des fonctions plutôt qu'après le produit. Cependant, l'expression (4.5) donne de meilleurs résultats en terme d'acceptation et d'autocorrélation que toutes les combinaisons de β et de γ considérées dans l'analyse. Cette observation semble indiquer que la probabilité d'acceptation M.-H. généralisée (4.5) est le choix optimal pour définir l'étape d'acceptation/rejet d'un algorithme MTM.

4.1.3.3 Algorithmes par composantes et par blocs

À l'image de l'algorithme Metropolis-Hastings qui peut être décomposé en mises-à-jour successives des composantes de la chaîne (algorithmes de Gibbs ou *Metropolis-within-Gibbs*), les algorithmes à essais multiples peuvent également être construits à cette fin.

Un exemple de ce type de mise-à-jour est l'algorithme CMTM (*Component-wise Multiple-Try Metropolis*) de Yang et collab. (2019). Spécifiquement, à partir de l'état actuel de la chaîne $x \in \mathcal{X}$, les K candidats $y_j^{(1)}, \ldots, y_j^{(K)}$ pour la mise-à-jour de la composante $j \in \{1, \ldots, d\}$ sont générés indépendamment à partir de densités unidimensionnelles distinctes $Q_j^{(k)}(\cdot|x_j)$, $k = 1, \ldots, K$. Les poids utilisés dans la sélection sont donnés par

$$w_j^{(k)}(y_j^{(k)}|x) = \pi(y_j^{(k)}|x_{-j})Q_j^{(k)}(x_j|y_j^{(k)})s_j^{(k)}(y_j^{(k)}, x_j),$$

où $\pi(\cdot|x_{-j})$ correspond à la densité conditionnelle de la j-ième composante sachant les autres composante x_{-j} et $s_j^{(k)}(\cdot,\cdot)$ est une fonction symétrique satisfaisant $s_j^{(k)}(x_k,y_k) > 0$ si et seulement si $Q_j^{(k)}(y_j|x_j) > 0$. Soient y_j la proposition obtenue, k l'indice du candidat sélectionné, i.e $y_j = y_j^{(k)}$, et $y = (x_{1:j-1},y_j,x_{j+1:d})$ le nouvel état global proposé. Pour calculer la probabilité d'acceptation, des points de référence sont également générés selon la méthode habituelle : $x_{*j}^{(l)} \sim Q_j^{(l)}(\cdot|y_j)$ pour tout $l \neq k$ et $x_{*j}^{(k)} = x_j$. Enfin, la proposition y_j est acceptée avec probabilité égale à

$$\alpha(y,y_j^{(-k)}|x,x_{*j}^{(-k)}) = \min\left\{1, \frac{\sum_{k=1}^K w_j^{(k)}(y_j^{(k)}|x)}{\sum_{k=1}^K w_j^{(k)}(x_{*j}^{(k)}|y)}\right\}.$$

Tel que les auteurs le mentionnent, cette procédure peut être généralisée de plusieurs manières : les densités instrumentales $Q_j^{(k)}$ pourraient dépendre de x et non seulement de x_j, de la dépendance entre les candidats pourrait être introduite, etc. Pour ce qui est de la fonction symétrique $s_j^{(k)}$, ils suggèrent d'utiliser l'expression (4.11) où le saut quadratique moyen est directement le critère optimisé. On trouve alors les poids suivants

$$w_j^{(k)}(y_j^{(k)}|x) = \pi(y_j^{(k)}|x_{-j})\left|y_j^{(k)} - x_j\right|^\alpha,$$

où $\alpha \in \mathbb{R}$ est un paramètre ; les expérimentations des auteurs montrent que $\alpha \in (2,4)$ est généralement un choix donnant de bons résultats. De plus, les auteurs considèrent d'intégrer de l'adaptation au sein

de leur algorithme : il en sera question à la section 5.2.

D'une manière plus générale, les composantes peuvent être mises à jour en blocs : un bloc actualise plusieurs composantes simultanément et l'ensemble des blocs forme une partition des composantes de la chaîne. So (2006) propose un tel algorithme dans le contexte de modèles à espace d'états non-gaussiens. Les poids utilisés dans cette étude sont égaux à la densité conditionnelle aux autres blocs laissés fixes, mais d'autre options sont également possibles.

4.2 Propriétés des algorithmes à essais multiples

Dans cette section, les résultats théoriques associés aux algorithmes à essais multiples seront abordés. D'abord, il sera question de la validité de tels algorithmes dans le contexte des méthodes MCMC. En appliquant les résultats des sections 2.2 et 2.3 aux définitions de la section précédente, il sera possible de vérifier l'ergodicité des algorithmes à essais multiples et de dégager des conditions suffisantes aux propriétés de l'estimation Monte Carlo. Ensuite, l'optimisation du choix de la densité instrumentale sera étudié dans une situation particulière suivant le travail de Bédard et collab. (2012).

Les algorithmes à essais multiples constituent une généralisation de l'algorithme Metropolis-Hastings (définition 2.24), où la génération de la proposition et la probabilité d'acceptation sont plus complexes. Ainsi, il n'est pas possible d'appliquer directement les conclusions des différents résultats sur l'algorithme M.-H., présentés à la section 2.3, à l'algorithme MTM. Cependant, étant donné la grande similitude entre les deux méthodes, il est possible de s'inspirer du traitement de l'algorithme M.-H. afin d'étudier les propriétés de l'algorithme MTM.

4.2.1 Ergodicité

D'abord, l'algorithme MTM satisfait la condition d'équilibre (définition 2.13), via la proposition 2.6, par construction pour des poids de sélection particuliers (proposition 4.1) ainsi que pour des poids généraux (proposition 4.2). Ainsi, la proposition 2.7 indique que le noyau de l'algorithme MTM est réversible et admet la densité cible π comme distribution invariante.

Dans la littérature, seule la condition d'équilibre du noyau MTM est démontrée. Ainsi, l'ergodicité de la chaîne n'est jamais démontrée et on se fie sur l'intuition que les résultats valides pour les noyaux M.-H. sont également valides pour les noyaux MTM. On observe, par exemple, la justification suivante pour l'ergodicité de l'algorithme IMTM (Casarin et collab., 2013) sans plus de détails : « *Fixed the ith chain, the conditional detailed balance is proved. This ensures the ergodicity of the chain.* » Aucune référence ne traite explicitement ni de la loi des grands nombres ni d'un théorème limite central pour les algorithmes à essais multiples.

Pour assurer l'ergodicité, la π-irréductibilité et l'apériodicité sont deux conditions suffisantes qui peuvent être vérifiées facilement dans la plupart des cas. On considère ici des conditions suffisantes sur la densité instrumentale Q et sur la densité cible π qui permettent d'assurer ces deux conditions.

La π-irréductibilité et l'apériodicité du noyau ne peuvent être montrées en utilisant la proposition 2.25 directement puisque ce résultat est basé sur les noyaux Metropolis-Hastings seulement. Par contre, une inspection de la preuve de Robert et Casella (2004, lemme 6.2.7) permet d'identifier que la forme explicite de la probabilité d'acception est peu utilisée et qu'il est possible d'étendre ce résultat à l'algorithme MTM. Pour ce, on réécrit la densité MTM d'une manière similaire à celle d'un noyau M.-H. La densité de transition est alors donnée par

$$p(y|x) = \sum_{k=1}^{K} A^{(k)}(y|x)Q^{(k)}(y|x) + R(x)\delta_x(y),$$

où $A^{(k)}(y|x)$ correspond à la probabilité d'acceptation intégrée de x vers y via le k-ième candidat,

$$A^{(k)}(y|x) = \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} Q^{(-k)}(y^{(-k)}|x,y)\bar{w}^{(k)}(y; y^{(-k)}|x)Q^{(-k)}(x_*^{(-k)}|x,y) \\ \times \alpha(y,y^{(-k)}|x,x_*^{(-k)})\,\mathrm{d}x_*^{(-k)}\,\mathrm{d}y^{(-k)},$$

et $R(x)$ correspond à la probabilité de rejet :

$$R(x) = 1 - \sum_{k=1}^{K} \bar{A}^{(k)}(x),$$

où

$$\bar{A}^{(k)}(x) = \int_{\mathcal{X}} A^{(k)}(y|x)Q^{(k)}(y|x)\,\mathrm{d}y.$$

Proposition 4.4 *Soit P un noyau MTM pour une densité π à support connexe $\mathcal{X}$. Supposons que π est borné par le haut et par le bas sur tout sous-ensemble compact de $\mathcal{X}$, qu'il existe $\delta, \varepsilon > 0$ tels que les densités instrumentales marginales symétriques satisfont*

$$Q^{(k)}(y|x) > \varepsilon, \qquad \forall\, \|x-y\|_2 < \delta, \qquad k = 1, \ldots, K, \tag{4.20}$$

et que les poids $w^{(k)}(\cdot|x)$ soient bornés par le haut uniformément et par le bas sur tout compact pour x fixé, alors le noyau P est π-irréductible et apériodique.

Démonstration. Afin de montrer la π-irréductibilité, on doit montrer que pour tout $B \in \mathcal{B}(\mathcal{X})$ tel que $\pi(B) > 0$ et pour tout $x_0 \in \mathcal{X}$, on a $P^m(B|x_0) > 0$ pour un certain $m \in \mathbb{N}$, qui peut possiblement dépendre de x et de B. Par la connectivité de $\mathcal{X}$, on peut trouver $m \in \mathbb{N}$ ainsi qu'une suite de m points $(x_1, \ldots, x_m) \in \mathcal{X}$ tels que deux points consécutifs soient à une distance inférieure à δ (c.-à-d., $\|x_i - x_{i+1}\|_2 < \delta$) et que $x_m \in B$. Ceci constitue un chemin entre x_0 et B de pas inférieurs à δ et dont tous les points ont une densité positive sous π. Autour de chaque x_i, on considère une boule de rayon δ que l'on dénotera $B_\delta(x_i)$. Chacune de ces boules constitue un ensemble compact où les densités marginales sont bornées par le bas par ε par hypothèse. Maintenant, on montre que le passage d'une boule à la suivante s'effectue avec probabilité positive. Soit $x \in B_\delta(x_i)$ où $x = x_0$ pour $i = 0$; on cherche à montrer $P(B_\delta(x_{i+1})|x) > 0$. On a

$$P(B_\delta(x_{i+1})|x) \geqslant \sum_{k=1}^{K} \int_{B_\delta(x_{i+1})} A^{(k)}(y|x)Q^{(k)}(y|x)\,\mathrm{d}y.$$

Chacun des termes de la somme prend la forme suivante

$$\int_{B_\delta(x_{i+1})} \Bigg[\int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \bar{w}^{(k)}(y; y^{(-k)}|x)\alpha(y,y^{(-k)}|x,x_*^{(-k)}) \\ \times Q^{(-k)}(y^{(-k)}|x,y)Q^{(-k)}(x_*^{(-k)}|y,x)\,\mathrm{d}x_*^{(-k)}\,\mathrm{d}y^{(-k)} \Bigg] Q^{(k)}(y|x)\,\mathrm{d}y.$$

Il s'agit de l'espérance d'une fonction positive (entre crochets), sur un domaine de mesure positive par rapport à la densité. La fonction entre crochets est positive puisqu'il s'agit de l'espérance d'une fonction positive sur tout le support : les poids $\bar{w}^{(k)}$ sont positifs par définition et la probabilité d'acceptation est positive. En effet, le numérateur de la probabilité d'acceptation,

$$\pi(y)Q^{(k)}(x|y)\bar{w}^{(k)}(x; x_*^{(-k)}|y),$$

est positif puisque la densité cible est bornée par le bas sur tout compact, la densité instrumentale est bornée par le bas à l'intérieur de la boule et les poids sont positifs. Ainsi,

$$P(B_\delta(x_{i+1})|x) > 0, \qquad \forall\, x \in B_\delta(x_i), i = 0, \ldots, m-1.$$

Puis, par induction, on peut montrer que les transitions itérées sont également positives

$$\begin{aligned} P^i(B_\delta(x_i)|x_0) &= \int_{\mathcal{X}} P(B_\delta(x_i)|y) P^{i-1}(\,\mathrm{d}y|x_0) \\ &\geqslant \int_{B_\delta(x_{i-1})} P(B_\delta(x_i)|y) P^{i-1}(\,\mathrm{d}y|x_0) > 0 \end{aligned}$$

puisqu'on intègre une fonction positive sur une domaine de mesure positive. Ainsi, on trouve

$$P^{m-1}(B_\delta(x_{m-1})|x_0) > 0.$$

Enfin, pour la dernière étape de $B_\delta(x_{m-1})$ vers x_m, on note qu'un argument similaire permet de montrer que $p(x_m|y) > 0$ pour tout $y \in B_\delta(x_{m-1})$. Ceci permet de conclure

$$p^m(x_m|x_0) = \int_{\mathcal{X}} p(x_m|y) P^{m-1}(\,\mathrm{d}y|x_0) \geqslant \int_{B_\delta(x_{m-1})} p(x_m|y) P^{m-1}(\,\mathrm{d}y|x_0) > 0.$$

En résumé, on vient de montrer que pour tout $x \in \mathcal{X}$ et $y \in B$ il existe $m \in \mathbb{N}$ tel que $p^m(y|x) > 0$. On cherche cependant à montrer qu'il existe $M \in \mathbb{N}$ tel que $P^M(B|x) > 0$. Afin de passer de la première à la seconde conclusion, on considère la partition suivante de B :

$$\begin{aligned} B &= \bigcup_{m \in \mathbb{N}} B_m, \\ B_m &= \left\{ y \in B \;\middle|\; p^m(y|x) > 0, p^{m-1}(y|x) = 0 \right\}, \end{aligned}$$

c.-à-d., B_m correspond au sous-ensemble de points de B tel que le passage de x à B en m pas a une densité positive et tel que m est minimal. Puisque $\pi(B) > 0$, alors il existe $M \in \mathbb{N}$, qui dépend de x et de B, tel que $\pi(B_M) > 0$ et qui a donc une mesure de Lebesgue positive. On trouve alors la conclusion désirée :

$$P^M(B|x) \geqslant P^M(B_M|x) = \int_{B_M} p^M(y|x)\,\mathrm{d}y > 0.$$

La proposition 2.2 confirme finalement la π-irréductibilité de la chaîne.

Maintenant, on montre l'apériodicité forte de la chaîne. Soit $x \in \mathcal{X}$ tel que $\pi(x) > 0$, $B \in \mathcal{B}(\mathcal{X})$ et $y \in B_{\delta/2}(x)$. On cherche à montrer que $B_{\delta/2}(x)$ est ν_1-petit. À nouveau, on trouve

$$\begin{aligned} P(B|x) &= \sum_{k=1}^{K} \int_B A^{(k)}(y|x) Q^{(k)}(y|x)\,\mathrm{d}y \\ &\geqslant \sum_{k=1}^{K} \int_{B \cap B_{\delta/2}(x)} A^{(k)}(y|x) Q^{(k)}(y|x)\,\mathrm{d}y. \end{aligned} \tag{4.21}$$

Par la symétrie des densités instrumentales, on a

$$\alpha(z,z^{(-k)}|y,y_*^{(-k)}) = \min\left\{1, \frac{\pi(z)Q^{(k)}(y|z)\bar{w}^{(k)}(y;y^{(-k)}|z)}{\pi(y)Q^{(k)}(z|y)\bar{w}^{(k)}(z;z^{(-k)}|y)}\right\} = \min\left\{1, \frac{\pi(z)\bar{w}^{(k)}(y;y^{(-k)}|z)}{\pi(y)\bar{w}^{(k)}(z;z^{(-k)}|y)}\right\}.$$

Pour chaque candidat, on considère la partition de l'espace $\mathcal{X}$ selon le fait que la proposition est automatiquement acceptée ou non :

$$D^{(k)}(y|y_*^{(-k)},z^{(-k)}) = \left\{ z \in \mathcal{X} \;\middle|\; \pi(z)\bar{w}^{(k)}(y;y^{(-k)}|z) \leqslant \pi(y)\bar{w}^{(k)}(z;z^{(-k)}|y) \right\}.$$

On écrira parfois $D^{(k)}(y)$ pour alléger la notation, mais il faudra garder en tête que cette région dépend des autres candidats et des points de référence. Pour $z \in D^{(k)}(y)$, la probabilité d'acceptation s'écrit

$$\alpha(z,z^{(-k)}|y,y_*^{(-k)}) = \frac{\pi(z)\bar{w}^{(k)}(y|z)}{\pi(y)\bar{w}^{(k)}(z|y)},$$

alors que pour $z \notin D^{(k)}(y)$, on a plutôt

$$\alpha(z,z^{(-k)}|y,y_*^{(-k)}) = 1.$$

Soit $C = B \cap B_{\delta/2}(x)$. Le k-ième terme de (4.21) s'écrit

$$\begin{aligned}
\int_C A^{(k)}(z|y)Q^{(k)}(z|y)\,\mathrm{d}z &= \int_C \Bigg[\int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \bar{w}^{(k)}(z; z^{(-k)}|y)\alpha(z,z^{(-k)}|y,y_*^{(-k)}) \\
&\qquad \times Q^{(-k)}(z^{(-k)}|y,z)Q^{(-k)}(y_*^{(-k)}|z,y)\,\mathrm{d}y_*^{(-k)}\,\mathrm{d}z^{(-k)} \Bigg] Q^{(k)}(z|y)\,\mathrm{d}z \\
&= \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \int_C \bar{w}^{(k)}(z; z^{(-k)}|y)\alpha(z,z^{(-k)}|y,y_*^{(-k)})Q^{(k)}(z|y) \\
&\qquad \times Q^{(-k)}(z^{(-k)}|y,z)Q^{(-k)}(y_*^{(-k)}|z,y)\,\mathrm{d}z\mathrm{d}y_*^{(-k)}\,\mathrm{d}z^{(-k)} \\
&= \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \int_{C\cap D^{(k)}(y)} \bar{w}^{(k)}(z; z^{(-k)}|y)\frac{\pi(z)\bar{w}^{(k)}(y; y^{(-k)}|z)}{\pi(y)\bar{w}^{(k)}(z; z^{(-k)}|y)}Q^{(k)}(z|y) \\
&\qquad \times Q^{(-k)}(z^{(-k)}|y,z)Q^{(-k)}(y_*^{(-k)}|z,y)\,\mathrm{d}z\mathrm{d}y_*^{(-k)}\,\mathrm{d}z^{(-k)} \\
&\quad + \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \int_{C\cap[D^{(k)}(y)]^C} \bar{w}^{(k)}(z; z^{(-k)}|y)Q^{(k)}(z|y) \\
&\qquad \times Q^{(-k)}(z^{(-k)}|y,z)Q^{(-k)}(y_*^{(-k)}|z,y)\,\mathrm{d}z\mathrm{d}y_*^{(-k)}\,\mathrm{d}z^{(-k)}
\end{aligned}$$

Maintenant, pour $y \in B_{\delta/2}(x)$ et $z \in C \subseteq B_{\delta/2}(x)$ on a $\|z-y\|_2 \leqslant \delta$ et donc $Q^{(k)} \geqslant \varepsilon$. Par les conditions sur les poids, on sait qu'il existe $a > 0$ et $A < \infty$ tels que $w^{(k)}(z|y) > a$ pour tout $\|z-y\|_2 \leqslant \delta$ et $w^{(j)}(z^{(j)}|y) < A$ pour tout $z^{(j)}$, $j = 1, \ldots, K$. Ainsi, on trouve que $\bar{w}^{(k)}(z; z^{(-k)}|y) \geqslant a/KA$ pour tout $\|z-y\|_2 \leqslant \delta$. Puis,

$$\bar{w}^{(k)}(z; z^{(-k)}|y)\frac{\pi(z)\bar{w}^{(k)}(y; y^{(-k)}|z)}{\pi(y)\bar{w}^{(k)}(z; z^{(-k)}|y)}Q^{(k)}(z|y) = \bar{w}^{(k)}(y; y^{(-k)}|z)\frac{\pi(z)}{\pi(y)}Q^{(k)}(z|y) \geqslant \frac{a}{KA}\frac{\inf_{B_{\delta/2}(x)}\pi}{\sup_{B_{\delta/2}(x)}\pi}\varepsilon.$$

Similairement,

$$\bar{w}^{(k)}(z; z^{(-k)}|y)Q^{(k)}(z|y) \geqslant \frac{a}{KA}\varepsilon \geqslant \frac{a}{KA}\frac{\inf_{B_{\delta/2}(x)}\pi}{\sup_{B_{\delta/2}(x)}\pi}\varepsilon,$$

puisque $\inf_{B_{\delta/2}(x)}\pi / \sup_{B_{\delta/2}(x)}\pi \leqslant 1$. Ainsi, on trouve

$$\begin{aligned}
\sum_{k=1}^K \int_C A^{(k)}(z|y)Q^{(k)}(z|y)\,\mathrm{d}z &\geqslant \sum_{k=1}^K \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \int_C \frac{a\varepsilon}{KA}\frac{\inf_{B_{\delta/2}(x)}\pi}{\sup_{B_{\delta/2}(x)}\pi} \\
&\qquad \times Q^{(-k)}(z^{(-k)}|y,z)Q^{(-k)}(y_*^{(-k)}|z,y)\,\mathrm{d}z\mathrm{d}y_*^{(-k)}\,\mathrm{d}z^{(-k)} \\
&\geqslant \sum_{k=1}^K \int_C \frac{a\varepsilon}{KA}\frac{\inf_{B_{\delta/2}(x)}\pi}{\sup_{B_{\delta/2}(x)}\pi} \Bigg[\int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \\
&\qquad \times Q^{(-k)}(z^{(-k)}|y,z)Q^{(-k)}(y_*^{(-k)}|z,y)\,\mathrm{d}y_*^{(-k)}\,\mathrm{d}z^{(-k)} \Bigg] \mathrm{d}z \\
&= K\frac{a\varepsilon}{KA}\frac{\inf_{B_{\delta/2}(x)}\pi}{\sup_{B_{\delta/2}(x)}\pi}\int_C 1\,\mathrm{d}z \\
&= \frac{a\varepsilon}{A}\frac{\inf_{B_{\delta/2}(x)}\pi}{\sup_{B_{\delta/2}(x)}\pi}\lambda^{\mathrm{Leb}}(B \cap B_{\delta/2}(x)).
\end{aligned}$$

Ceci montre en fait que $B_{\delta/2}(x)$ est un ν_1-petit ensemble où

$$\nu_1(B) = \frac{a\varepsilon}{A}\frac{\inf_{B_{\delta/2}(x)}\pi}{\sup_{B_{\delta/2}(x)}\pi}\lambda^{\mathrm{Leb}}\left(B \cap B_{\delta/2}(x)\right)$$

est bel et bien une mesure concentrée sur $B_{\delta/2}(x)$ qui est non-triviale puisque $\inf_{B_{\delta/2}(x)}\pi > 0$ étant donnée la compacité de $B_{\delta/2}(x)$. Finalement, par l'existence d'un ensemble ν_1-petit, la chaîne est (fortement) apériodique par définition. □

La condition de positivité locale des densités marginales (4.20) se vérifie facilement lorsque les

candidats sont générés par une marche aléatoire, c.-à-d.,

$$Q^{(k)}(y|x) = Q^{(k)}(y-x), \qquad k = 1, \ldots, K.$$

Dans ce cas, cette condition correspond à ce que les densités marginales soient positive près de l'origine, ce qui devrait être le cas généralement. Ensuite, les conditions sur les poids peuvent sembler restrictives dans le choix de l'expression de $w^{(k)}(\cdot|\cdot)$, mais elles ne seront généralement pas problématiques en pratique. La borne supérieure uniforme pourrait ne pas être satisfaite pour des poids du type $w^{(k)}(y|x) = \pi(y)/Q^{(k)}(y|x)$ où la densité marginale pourrait être arbitrairement petite, mais le ratio sera borné lorsque les densités marginales ont des ailes au moins aussi lourdes que celles de π. De plus, la borne inférieure locale ne posera généralement pas de problème étant donné les suppositions sur π et sur les densités marginales. Évidemment, il sera toujours possible de forcer artificiellement $w^{(k)} \in [a,A]$ d'une manière uniforme pour assurer ces conditions pour des choix plus généraux de poids ; ceci pourrait également éviter certains problèmes numériques en pratique.

Il n'est cependant pas clair que la condition sur les poids soit nécessaire. En effet, par la normalisation, il y aura toujours un des poids qui soit supérieur à $1/K$ et cette borne pourrait être utilisée plutôt que a/KA, ce qui rendrait la condition obsolète. Le problème provient du fait que la borne $1/K$ n'est valide que pour un ensemble fixe $z^{(1:K)}$ et que les candidats ne sont pas interchangeables ; il n'est donc pas possible d'inverser l'ordre d'intégration d'une manière triviale.

4.2.1.1 Ergodicité V-géométrique

La proposition 4.4 permet de vérifier l'ergodicité d'algorithmes MTM sous des conditions relativement faibles. L'ergodicité seule permet de vérifier également la loi des grands nombres en se basant sur le théorème 2.18. Cependant, afin de vérifier un théorème limite central, des conditions plus fortes sont nécessaires : le théorème 2.22 exige une ergodicité V-géométrique. Cette section sera donc portée sur la vérification de l'ergodicité V-géométrique des algorithmes MTM.

Tout comme il a été possible d'étendre la proposition 2.25 aux algorithmes MTM afin de produire la proposition 4.4, on tente ici d'étendre le théorème 2.32 généralisant l'ergodicité V-géométrique des algortihmes Metropolis, aux algorithmes MTM où les candidats sont générés selon une marche aléatoire sphérique. Pour ce, on considère une généralisation des résultats menant à ce théorème : les théorèmes 2.30 et 2.31.

Lemme 4.5 *Soit P un noyau MTM de densité instrumentale Q tel qu'il existe $\delta, \varepsilon > 0$ avec*

$$Q^{(k)}(y|x) = Q^{(k)}(\|y-x\|_2), \tag{4.22}$$

$$Q^{(k)}(z) \geqslant \varepsilon, \qquad \forall\, |z| \leqslant \delta, \tag{4.23}$$

de poids $w^{(k)} > 0$ satisfaisant les mêmes conditions qu'à la proposition 4.4 et de densité cible π positive et continue. Supposons qu'il existe une petit ensemble C pour lequel P satisfait la dérive géométrique pour une certaine fonction $V \geqslant 1$ continue et que les conditions suivantes soient satisfaites ,

$$\limsup_{\|x\|_2 \to \infty} \frac{PV(x)}{V(x)} < 1, \qquad \sup_{x \in \mathcal{X}} \frac{PV(x)}{V(x)} < \infty. \tag{4.24}$$

Alors, la chaîne est V-géométriquement ergodique à π.

Démonstration. Soit $V \geqslant 1$ satisfaisant (4.24). On cherche à vérifier les conditions du théorème 2.17. Par la preuve de la proposition 4.4, on sait que P est π-irréductible et apériodique. Il reste donc à trouver un petit ensemble C pour lequel P satisfait la condition de dérive géométrique vers C avec la fonction V donnée.

Par la première partie de l'hypothèse (4.24), il existe $R > 0$ tel que $\|x\|_2 \geqslant R$ implique

$$\frac{PV(x)}{V(x)} \leqslant 1 - \varepsilon =: \lambda,$$

pour un certain $\varepsilon > 0$ et donc $\lambda < 1$. Ainsi, en choisissant $C = B_R(\mathbf{0})$ on a $x \notin C \Rightarrow PV(x) \leqslant \lambda V(x) + b\,\mathbb{1}_C(x)$ pour un certain $\lambda < 1$ et tout b. Ensuite, on considère

$$b := \left(\sup_{x \in C} V(x)\right)\left(\sup_{x \in \mathcal{X}} \frac{PV(x)}{V(x)}\right).$$

Par la seconde partie de l'hypothèse, on a $\sup_{x \in \mathcal{X}} \frac{PV(x)}{V(x)} < \infty$. Puis, étant donné que V est une fonction continue et que C est un ensemble compact, on a $\sup_{x \in C} V(x) < \infty$, ce qui montre que $b < \infty$. Enfin, pour tout $x \in C$, on trouve

$$PV(x) = V(x)\frac{PV(x)}{V(x)} \leqslant \left(\sup_{x \in C} V(x)\right)\left(\sup_{x \in \mathcal{X}} \frac{PV(x)}{V(x)}\right) = b \leqslant \lambda V(x) + b\,\mathbb{1}_C(x).$$

Finalement, la preuve de la proposition 4.4 montre que tout ensemble compact est petit. Pour voir ceci, on requiert la notion d'ensemble *petite* qui généralise la notion d'ensembles petits. On réfère à Meyn et Tweedie (2009, section 5.5.2) pour une définition ainsi que pour les résultats nécessaires (proposition 5.5.5 et théorème 5.5.7), mais les détails seront omis de cette preuve. Ceci est suffisant pour montrer que C est un ensemble petit et donc que la condition de dérive géométrique est satisfaite. □

4.2.2 Théorèmes limites

4.2.2.1 *Loi des grands nombres*

Le théorème 2.18 indique que toute chaîne de Markov Harris-positive (pour une distribution stationnaire π) satisfait la loi forte des grands nombres.

Les algorithmes MTM sont réversibles par construction (proposition 4.2) et des résultats tels que la proposition 4.4 permettent de vérifier, entre autres, la π-irréductibilité de la chaîne. Puis, le théorème 2.10 implique que la chaîne est récurrente.

En introduisant d'abord la notion de fonction harmonique, il nous est possible d'étendre la proposition 2.27, dûe à Tierney (1994, corollaire 2), aux algorithmes MTM, ce qui permet de vérifier l'Harris-récurrence à partir de la récurrence seulement.

Définition 4.1 (Fonction harmonique) *Une fonction $h : \mathbb{R}^d \to [0,\infty]$ est dite harmonique pour un noyau P si $h = Ph$, au sens où*

$$h(x) = \int_{\mathcal{X}} h(y)P(\,\mathrm{d}y|x), \qquad \forall\, x \in \mathbb{R}^d\,.$$

Théorème 4.6 (Tierney, 1994, théorème 2) *Soit P un noyau de transition de Markov qui soit récurrent. Alors, P est Harris-récurrent si et seulement si toute fonction harmonique bornée est constante.*

Proposition 4.7 (Nummelin, 1984, proposition 3.13(i)) *Soit P un noyau de transition de Markov qui soit récurrent (avec ϕ-irréductibilité). Alors, toute fonction harmonique est constante ϕ-presque partout.*

En utilisant ces résultats, il nous est alors possible de démontrer le résultat suivant. Enfin, cette propriété permet de vérifier la loi des grands nombres sous certaines conditions supplémentaires.

Proposition 4.8 *Soit P un noyau MTM pour une certaine distribution cible π. Si P est π-irréductible, alors P est Harris-récurrent.*

Démonstration. Par le théorème 2.10, on sait que P est récurrent. Soit h une fonction harmonique pour P qui soit bornée. Par la proposition 4.7, on sait déjà que h est constante π-presque partout et donc que $h(x) = \pi h$ pour π-presque tout x. Soit $H = \{x \in \mathbb{R}^d \mid h(x) \neq \pi h\}$, l'ensemble des points où h n'est pas égale à son espérance. Alors, $\pi(H) = 0$ et la probabilité de passer de $x \in \mathcal{X}$ à H via une acceptation est nulle,

$$\int_H \sum_{k=1}^{K} A^{(k)}(y|x) Q^{(k)}(\,\mathrm{d}y|x) = 0,$$

étant donné que π et $Q^{(k)}(\cdot|x)$ sont toutes deux absolument continues par rapport à la mesure de Lebesgue.
Puisque h est harmonique, on sait que

$$\begin{aligned} h(x) &= \int_{\mathbb{R}^d} h(y) P(\,\mathrm{d}y|x) \\ &= \int_H h(y) P(\,\mathrm{d}y|x) + \int_{H^C} h(y) P(\,\mathrm{d}y|x). \end{aligned}$$

D'une part, on a

$$\begin{aligned} \int_H h(y) P(\,\mathrm{d}y|x) &= \int_H h(y) \left[\sum_{k=1}^{K} A^{(k)}(y|x) Q^{(k)}(y|x) \right] + R(x)\delta_x(y)\,\mathrm{d}y \\ &= 0 + h(x)R(x)\,\mathbb{1}(h(x) \neq \pi h). \end{aligned}$$

D'autre part, on a

$$\begin{aligned} \int_{H^C} h(y) P(\,\mathrm{d}y|x) &= \int_H \pi h \left[\sum_{k=1}^{K} A^{(k)}(y|x) Q^{(k)}(y|x) \right] + R(x)\delta_x(y)\,\mathrm{d}y \\ &= \pi h(1 - R(x)) + \pi h R(x)\,\mathbb{1}(h(x) = \pi h). \end{aligned}$$

En combinant, on trouve

$$\begin{aligned} h(x) &= h(x)R(x)\,\mathbb{1}(h(x) \neq \pi h) + \pi h(1 - R(x)) + \pi h R(x)\,\mathbb{1}(h(x) = \pi h) \\ &= h(x)R(x)\,\mathbb{1}(h(x) \neq \pi h) + \pi h(1 - R(x)) + \pi h R(x)(1 - \mathbb{1}(h(x) \neq \pi h)) \\ &= \pi h + R(x)(h(x) - \pi h)\,\mathbb{1}(h(x) \neq \pi h). \end{aligned}$$

Après factorisation, on obtient enfin

$$0 = (h(x) - \pi h)\left[R(x)\,\mathbb{1}(h(x) \neq \pi h) - 1\right].$$

Maintenant, si $h(x) \neq \pi h$, alors on doit avoir $R(x) = 1$. Enfin, la π-irréductibilité implique que la probabilité de rester sur place ne peut être totale de sorte que $R(x) = 1$ est une contradiction lorsque $x \in \mathcal{X}$. Si $x \notin \mathcal{X}$, alors $R(x) = 0$ et on obtient directement $h(x) = \pi h$, ce qui montre finalement que $h \equiv \pi h$. □

Corollaire 4.9 *Soit P un noyau MTM pour une certaine distribution cible π. Si P est π-irréductible, alors P satisfait la loi forte des grands nombres.*

Démonstration. Conséquence directe des propositions 4.2 et 4.8 et du théorème 2.18. □

Corollaire 4.10 *Soit P un noyau MTM pour une certaine distribution cible π satisfaisant les conditions de la proposition 4.4. Alors, P satisfait la loi forte des grands nombres.*

Démonstration. Conséquence directe de la proposition 4.4 et du corollaire 4.9. □

4.2.2.2 Théorème limite central

Pour ce qui est de la distribution asymptotique de l'estimateur Monte Carlo, les théorèmes limites centraux peuvent être utilisés dans le cadre des algorithmes à essais multiples. Par exemple, le théorème 2.22 exige que la chaîne soit V-géométriquement ergodique afin d'obtenir un théorème

limite central. Le lemme 4.5 énonce certaines conditions suffisantes pour établir ce type d'ergodicité. Plus généralement, le théorème 2.23 indique que tout algorithme MTM ergodique (e.g. en vérifiant la proposition 4.4) vérifie le théorème limite central si la variance asymptotique est finie.

4.2.3 Échelle optimale

Bédard et collab. (2012) proposent une étude du problème de l'échelle optimale pour quelques variations de l'algorithme MTM. Les résultats théoriques obtenus sont asymptotiques ($d \to \infty$) comme la plupart des résultats présentés à la section 2.5. De plus, ils supposent des poids proportionnels à la densité cible ($w^{(k)}(y|x) = \pi(y)$), des candidats de distribution gaussienne (pour certaines classes de variances) et une densité cible à composantes i.i.d.

La première version de l'algorithme considérée, appelée MCTM (*Multiple Correlated-try Metropolis*), utilise des candidats générés de la manière suivante. Puisque la densité cible se factorise, on considère chacune des composantes indépendantes lors de la production des candidats. Une matrice de covariance $\Sigma \in \mathcal{C}_K^+$ détermine la corrélation entre les différents candidats pour une composante donnée, c.-à-d.,

$$Y_j^{(k)} = x_j + d^{-1/2} Z_j^{(k)}, \qquad k = 1, \dots, K,\ j = 1, \dots, d,$$

où $Z_j \sim \mathcal{N}_d(\mathbf{0}, \Sigma)$ sont i.i.d. pour $j = 1, \dots, d$. Marginalement, chacun des candidats $Y^{(k)}$ sera distribué selon une loi gaussienne de covariance proportionnelle à la matrice identité, mais une covariance entre les candidats existera selon l'expression de Σ. En fait, la distribution conjointe des K candidats est donnée par

$$Y^{(1:K)} | X = x \sim \mathcal{N}_{Kd}\left(\mathbf{1} \otimes x, d^{-1/2}\Sigma \otimes I_d\right).$$

Les auteurs considèrent quatre sous-classes de matrices et obtiennent des résultats d'échelle optimale pour chacune de ces classes :

— La classe des matrices multiples de l'identité où les candidats sont alors indépendants et identiquement distribués :

$$\mathcal{C}_{K,Id}^+ = \left\{\Sigma = \ell^2 I_d \mid \ell \in \mathbb{R}\right\};$$

— La classe des matrices diagonales où les candidats sont alors indépendants, mais d'échelles différentes :

$$\mathcal{C}_{K,Ind}^+ = \left\{\Sigma = \mathrm{diag}(\ell_1^2, \dots, \ell_K^2) \mid (\ell_1, \dots, \ell_K) \in \mathbb{R}^K\right\};$$

— La classe des matrices extrêmement antithétiques :

$$\mathcal{C}_{K,EA}^+ = \left\{\Sigma = \ell^2 \Sigma_{EA} \mid \ell \in \mathbb{R}\right\},$$

où

$$\Sigma_{EA} = \frac{K}{K-1} I_K - \frac{1}{K-1} \mathbf{1}\mathbf{1}^\top;$$

— La classe des matrices symétriques définies positives de dimension $K \times K$, $\mathcal{C}_K^+$.

$\mathcal{C}^+_{K,Id}$, $\mathcal{C}^+_{K,Ind}$: **Covariance multiple de l'identité**					
	1	2	3	4	5
Échelle (ℓ)	2,38	2,64	2,82	2,99	3,12
Vitesse	1,32	2,24	2,94	3,51	4,00
Pr. d'acceptation (α)	0,23	0,32	0,37	0,39	0,41
$\mathcal{C}^+_{K,EA}$: **Covariance extrêmement antithétique**					
	1	2	3	4	5
Échelle (ℓ)	2,38	2,37	2,64	2,83	2,99
Vitesse	1,32	2,64	3,66	4,37	4,91
Pr. d'acceptation (α)	0,23	0,46	0,52	0,54	0,55
$\mathcal{C}^+_K$: **Covariance générale**					
	1	2	3	4	5
Échelle (ℓ)	2,38	2,37	2,66	2,83	2,98
Vitesse	1,32	2,64	3,70	4,40	4,93
Pr. d'acceptation (α)	0,23	0,46	0,52	0,55	0,56

Tableau 4.1 *Échelle optimale de l'algorithme MCTM pour une densité cible à composantes i.i.d. en fonction du nombre de candidats K (Bédard et collab., 2012). L'optimisation est effectuée sur les quatre classes de covariance décrites dans le texte.*

Bédard et collab. (2012, proposition 4) montrent que la matrice optimale de $\mathcal{C}^+_{K,Ind}$ est en fait multiple de l'identité, de sorte que cette sous-classe a la même échelle optimale que $\mathcal{C}^+_{K,Id}$. Le tableau 4.1 contient l'échelle optimale, la probabilité d'acceptation optimale ainsi que la mesure d'efficacité (la **vitesse**) au maximum pour les trois autres classes et pour $K \in \{1,2,3,4,5\}$. Notons que les résultat pour la classe générale $\mathcal{C}^+_K$ sont obtenus numériquement puisque des résultats analytiques sont impossibles à obtenir. Sans définir directement la vitesse de l'algorithme, on retiendra surtout qu'il s'agit d'une quantité directement proportionnelle à l'efficacité de l'algorithme. L'analyse de ces résultat permet de dégager certaines remarques quant à l'efficacité de l'algorithme en fonction du nombre de candidats.

Premièrement, sans surprise, l'efficacité augmente avec le nombre de candidats : intuitivement, un plus grand ensemble de candidats permet de produire une proposition de meilleure qualité qui sera alors acceptée plus souvent et qui réduira alors l'autocorrélation. Ainsi, lorsque $K \to \infty$, l'échantillon Monte Carlo se rapprochera d'un échantillon i.i.d. L'augmentation du temps de calcul doit cependant être prise en compte et un compromis entre temps de calcul et efficacité doit être recherchée.

Deuxièmement, l'utilisation de candidats antithétiques améliore l'efficacité par rapport à des candidats indépendants telle que le montre la vitesse. On obtient également une probabilité d'acceptation supérieure, ce qui indique une meilleure qualité de candidats. Par contre, l'échelle optimale est légèrement inférieure à celle des covariances diagonales : il semblerait donc que des propositions plus agressives soient nécessaires lorsque les candidats sont indépendants. L'augmentation de la vitesse est particulièrement forte lors du passage de $K = 1$ à $K = 2$ candidats et tend à ralentir par la suite ; il ne semble donc pas nécessaire d'opter pour un grand nombre de candidats pour améliorer significativement l'efficacité : un compromis entre efficacité et temps de calcul est donc envisageable.

Troisièmement, l'utilisation de covariance générale ne semble pas améliorer l'efficacité par rapport à l'utilisation de candidats antithétiques. Ceci indique donc que cette classe de candidats est pratiquement optimale parmi tout choix de covariance. Intuitivement, cette observation n'est pas surprenante puisque l'utilisation de candidats antithétiques était justifiée par une volonté d'exploration rapide de l'espace, ce qui est directement lié à l'efficacité de l'algorithme.

Algorithme Hit-and-run à étendue $[-\ell,\ell]$					
	1	2	3	4	5
Échelle (ℓ)	2,38	2,37	7,11	11,85	16,75
Vitesse	1,32	2,64	2,65	2,65	2,65
Pr. d'acceptation (α)	0,23	0,46	0,46	0,46	0,46

Tableau 4.2 *Échelle optimale de l'algorithme MTM à variable aléatoire commune pour une densité cible à composantes i.i.d. en fonction du nombre de candidats K (Bédard et collab., 2012). L'optimisation est effectuée sur l'étendue des pas $[-\ell,\ell]$.*

La seconde version de l'algorithme MTM considérée utilise une variable aléatoire commune et fonctionne à la manière de l'algorithme *Hit-and-run*. D'abord, une direction aléatoire est générée à partir d'une distribution normale centrée réduite. Puis, les candidats sont produits en considérant divers pas déterministes le long de cette direction. En particulier, pour une direction $Z \sim \mathcal{N}_d(\mathbf{0}, I_d)$, les candidats sont donnés par

$$Y_j^{(k)} = x_j + d^{-1/2}\gamma^{(k)}z_j, \qquad j = 1, \ldots, d, \qquad k = 1, \ldots, K,$$

pour une certaine suite régulière de pas $\gamma^{(1:K)} \in [-\ell,\ell]$. Les résultats sont contenus dans le tableau 4.2. Pour cet algorithme, on note une augmentation de l'efficacité dès $K = 2$, mais aucune augmentation au-delà de ce nombre de candidats. Ainsi, l'utilisation de deux candidats opposés en direction semble grandement améliorer l'algorithme, mais des candidats intermédiaires (entre les deux extrêmes) ne semblent pas pertinents. Intuitivement, ce principe s'explique par l'observation suivante : si la direction pointe vers une région de plus faible densité (gradient négatif), la direction opposée (gradient positif) risque fortement de pointer vers une région de forte densité.

Enfin, Bédard et collab. (2012) considèrent l'utilisation de poids d'importance $w^{(k)}(y|x) = \pi(y)/q^{(k)}(y|x)$. Leurs simulations indiquent que ces poids ne performent pas aussi bien que les poids proportionnels à la densité cible. Malgré l'augmentation de la vitesse attendue avec l'augmentation de K, ces poids ne permettent pas d'atteindre la même efficacité que l'autre choix considéré. De plus, ils trouvent que la probabilité d'acceptation diminue avec K plutôt qu'augmenter de sorte que les candidats doivent être de plus en plus agressifs lorsque le nombre de candidats augmente.

Ces résultats n'étant seulement qu'asymptotiques ($d \to \infty$), les auteurs étudient également la situation en dimension finie à l'aide de simulations. Ils trouvent que les résultats obtenus sont relativement robustes lorsque le nombre de candidats K est petit, mais que de plus grandes différences entre le cas asymptotique et le cas fini peuvent être observées pour de grands K. Ils préviennent donc contre l'application aveugle de ces résultats pour de grands K et de petits d.

4.3 Efficacité empirique des algorithmes MTM

Les résultats théoriques d'échelle optimale présentés à la sous-section 4.2.3 montrent bien que l'utilisation d'essais multiples peut théoriquement améliorer l'efficacité par rapport à l'utilisation d'un seul candidat. Cependant, tel qu'annoncé en introduction, la complexité additionnelle de l'algorithme est accompagnée d'une augmentation du temps de calcul. En effet, chaque itération de l'algorithme MTM exige jusqu'à $2K-1$ évaluations de la densité cible comparativement à une seule pour l'algorithme Metropolis-Hastings. Ces évaluations constituent bien souvent la majeure partie du calcul dans une itération MCMC et la complexité de calcul sera alors directement liée au nombre d'évaluations à chaque itération.

4.3.1 Études d'efficacité non corrigée

Plusieurs travaux portant sur les algorithmes MTM ne prennent pas en compte cette problématique lorsque vient le temps d'évaluer l'efficacité de l'algorithme. D'une part, le temps de calcul ou bien le nombre d'évaluations n'est souvent pas pris en compte et, d'autre part, une comparaison avec un algorithme plus simple et plus établi, tel que l'algorithme M.-H., n'est pas produite. Ainsi, le gain en efficacité garanti théoriquement n'est pas clairement démontré empiriquement. On considère ici quelques-unes de ces études.

Martino et Read (2013, tableaux 2, 3, 9 et 10) effectuent des simulations comparant l'algorithme MTM avec $K = 1,2,5,100,1\,000$ candidats pour une densité instrumentale gaussienne i.i.d. et deux densités cibles multimodales. Tel qu'attendu, ils trouvent que la probabilité d'acceptation augmente avec K et que l'autocorrélation d'ordre un diminue avec K. Le cas $K = 1$ correspond à l'algorithme M.-H. ; l'étude confirme donc que les candidats additionnels augmentent effectivement l'efficacité de l'algorithme (par rapport au nombre d'itérations), mais elle ne prend pas en compte le temps de calcul additionnel.

Pandolfi et collab. (2010, tableau 1) considèrent un algorithme MTM à candidats gaussiens de type marche aléatoire pour une densité cible correspondant à une régression logistique : le nombre de candidats varie dans $K \in \{1,10,50,100\}$. À nouveau, la probabilité d'acceptation augmente avec K. Au tableau 2 de leur article, les auteurs comparent l'algorithme MTM à $K = 50$ candidats à l'algorithme M.-H. Lorsque le temps de calcul n'est pas pris en compte, l'algorithme MTM est clairement supérieur pour une certaine mesure comparant les variances asymptotiques. Par contre, lorsque la mesure est ajustée pour le temps de calcul, la supériorité de l'utilisation d'essais multiples n'est plus aussi claire.

Stormark (2006, section 7.2) considère une densité cible formée d'un mélange de deux lois normales et un algorithme MTM à densité instrumentale gaussienne à échelle approximativement optimale. Pour $K \in \{1,2,3,5,7,10,15,20\}$, il observe un ESS croissant avec K pour plusieurs variantes de l'algorithme. Des résultats semblables sont également obtenus pour une densité cible gaussienne (Stormark, 2006, figure 15). L'augmentation de l'efficacité est plus notable lorsque la corrélation dans la densité cible est forte ; peu d'amélioration est obtenue en augmentant K lorsque la cible a une covariance multiple de l'identité. Le temps de calcul n'est cependant jamais pris en compte.

Yang et collab. (2019, section 4) comparent un algorithme MTM par composante avec $K = 1,5,30$ composantes pour un modèle bayésien hiérarchique, une densité cible courbée et un mélange de 20 densités normales. Par inspection de l'ESS, ils trouvent que cette mesure augmente systématiquement avec K pour ces trois choix de densité cible.

Globalement, ces différentes études de simulations confirment tous la théorie sur les algorithmes MTM. On trouve que l'efficacité de l'algorithme augmente toujours avec une augmentation du nombre de candidats K. La plupart considèrent le cas $K = 1$, correspondant à l'algorithme M.-H., et on comprend donc qu'il est possible d'obtenir une efficacité supérieure à cet algorithme en utilisant des essais multiples afin de générer la proposition. Il n'est évidemment pas clair que ces résultats s'étendent à l'efficacité empirique où le temps de calcul est considéré.

4.3.2 Mesures d'efficacité empirique

Afin d'évaluer l'efficacité des algorithmes MTM honnêtement, il est donc primordial de moduler l'efficacité, calculée pour un nombre d'itérations fixé, par le temps de calcul. Plusieurs corrections sont possibles à cette fin. On rappelle que l'efficacité des algorithmes MCMC est mesurée par la variance asymptotique de l'estimateur Monte Carlo, mais d'autres mesures semblables ou équivalentes sont souvent préférées (sous-section 2.4.3.)

L'ESS estime la taille d'un échantillon i.i.d. requise pour obtenir la même variance asymptotique : en divisant l'ESS par le temps de calcul, on obtient alors une estimation de la taille échantillonale produite par unité de temps de calcul. Yang et collab. (2019) utilisent ce critère, dénoté ESS/CPU, afin de comparer un algorithme MTM par composante à un algorithme M.-H. par composante.

Ensuite, l'ESJD estime le saut quadratique moyen effectué par une itération de l'algorithme. Il s'agit en fait d'un proxy de la variance asymptotique puisque l'ordre induit par chacune de ces mesures est généralement semblable. Bédard et Mireuta (2013) proposent d'utiliser l'ESJD modulé par le temps de calcul comme mesure d'efficacité empirique. Cette mesure, dénotée par ESJD/CPU, est standardisée de sorte à ce qu'elle soit égale à l'ESJD pour un algorithme M.-H. On obtient donc une mesure de l'efficacité de l'exploration de l'espace corrigée pour le temps de calcul additionnel. Puisque l'ESJD est lié à l'ACF (section 2.4.3.3) et donc à la variance asymptotique, cette mesure constitue donc une mesure de l'efficacité empirique de l'algorithme.

Similairement, Casarin et collab. (2013, section 4.1.5) proposent d'utiliser la réduction relative du RMSE (*root mean squared error*) entre un algorithme et l'algorithme M.-H. par unité de temps de calcul additionnel comme mesure d'efficacité empirique. Le RMSE d'un estimateur $\hat{\theta}$ pour un paramètre d'intérêt θ est donné par

$$\mathrm{RMSE}_\theta(\hat{\theta}) = \sqrt{\mathbb{E}\{(\hat{\theta} - \theta)^2\}}.$$

Dans le contexte d'estimateurs sans biais, le RMSE correspond à l'écart-type de l'estimateur. Les estimateurs MCMC sont généralement asymptotiquement sans biais : lorsque plusieurs chaînes parallèles sont utilisées, la variance empirique des réalisations de l'estimateur estime donc la variance asymptotique de l'estimateur. La mesure d'efficacité empirique (corrigée pour le temps de calcul additionnel) est donnée par :

$$\mathrm{RF} = 100 \times \frac{\Delta_{\mathrm{RMSE}}/\mathrm{RMSE}_{\mathrm{MH}}}{T_{\mathrm{MTM}} - T_{\mathrm{MH}}}, \tag{4.25}$$

où T_A est le temps de calcul de l'algorithme A et Δ_{RMSE} est la réduction de RMSE :

$$\Delta_{\text{RMSE}} = \text{RMSE}_{\text{MTM}} - \text{RMSE}_{\text{MH}}.$$

D'une manière alternative, la conception de l'étude peut également compenser pour le nombre additionnel d'évaluations. Par exemple, Casarin et collab. (2013, section 4.1) comparent des algorithmes MTM à $K = 10$ candidats produisant des chaînes de longueur $N = 10\,000$ à des algorithmes M.-H. produisant des chaînes de longueur $N = 100\,000$ de sorte que le temps de calcul est relativement comparable.

4.3.3 Études d'efficacité empirique

À lu lumière des études d'efficacité empirique produites par Casarin et collab. (2013), par Yang et collab. (2019) et par Bédard et Mireuta (2013), il est possible de dégager certaines observations générales quant à l'efficacité des algorithmes à essais multiples par rapport à l'algorithme M.-H. Un résumé plus détaillé de ces études est inclut à l'annexe 4.4.1.

— Pour un nombre d'itérations fixé et sans prendre le temps de calcul en compte, un algorithme à essais multiples est systématiquement plus efficace que son algorithme équivalent à un seul candidat. De plus, cette efficacité augmente avec K ;

— Les différentes méthodes d'évaluation de l'efficacité produisent généralement le même ordre, ce qui est en accord avec la théorie d'échelle optimale où toutes les mesures d'efficacité sont asymptotiquement équivalentes ;

— L'augmentation de l'efficacité est liée à la densité cible : les densités à géométrie complexes ou multimodales profitent le plus de propositions de meilleure qualité ;

— Lorsque le temps de calcul est pris en compte, les algorithmes MTM ne sont pas systématiquement plus efficaces qu'un équivalent à un seul candidat. On observe certains choix de K, d'algorithmes et de densités cible où l'un ou l'autre est plus performant ;

— Pour une situtation donnée, il existe un $K \in \mathbb{N}$ tel que l'algorithme MTM est le plus efficace lorsque le temps est pris en compte. De plus, ce K semble généralement petit (< 5), mais de grands K sont parfois aussi optimaux (e.g. $K = 30$.) En particulier, le cas $K = 1$ est occasionnellement le meilleur choix ;

— La méthode utilisée pour générer les K candidats peut affecter ces conclusions. Pour deux exemples, des candidats antithétiques (MTM-HR) produisent un algorithme plus efficace qu'un algorithme Metropolis alors que des candidats indépendants n'arrivent pas à battre l'algorithme Metropolis ;

— Le coût computationnel associé à l'évaluation de la densité cible peut affecter ces conclusions. Le temps de calcul requis par l'algorithme en soit peut être particulièrement faible ou élevé en comparaison de sorte que les essais multiples soient profitables ou non.

4.4 Annexes au chapitre 4

4.4.1 Études d'efficacité empirique

Casarin et collab. (2013, section 4) comparent divers algorithmes MTM à $K = 10$ essais à l'algorithme M.-H. où les tailles échantillonnales sont ajustées afin de produire des algorithmes à coût computationnel relativement semblable. En comparant le MSE (*mean squared error*) des estimés, ils trouvent une réduction importante dans les algorithmes à essais multiples. Des résultats semblables sont obtenus en comparant le biais absolu. Le MSE et le biais absolu sont deux mesures qui se comportent similairement à la variance asymptotique de sorte que ces observations indiquent que l'utilisation d'essais multiples augmente l'efficacité de l'algorithme, et ce, même lorsque le design de l'expérience compense l'augmentation du temps de calcul. L'inspection graphique des fonctions d'autocorrélations révèle aussi une supériorité des algorithmes à essais multiples.

De plus, Casarin et collab. (2013, section 4.1.2) effectuent une étude de simulations pour comparer leur algorithme MTM à l'algorithme M.-H. sur une densité cible formée d'un mélange de deux lois normales. Pour $K = 3{,}6{,}10$ candidats, ils calculent la mesure d'efficacité corrigée RF (4.25) et observent que cette mesure diminue avec K, et ce, indépendamment de la dimension d du problème. Il semble donc que la réduction du RMSE de l'estimation par rapport au temps de calcul additionnel soit plus grande pour un petit nombre de candidats, ce qui suggère que leur algorithme a un nombre optimal de candidats pour balancer amélioration de l'efficacité et temps de calcul. Dans tous les cas, la mesure RF est négative indiquant une amélioration par rapport à l'algorithme M.-H. ; l'étude graphique des fonctions d'autocorrélation montre aussi que leur algorithme à essais multiples est plus efficace que l'algorithme M.-H.

En plus d'étudier l'ESS d'algorithmes MTM par composante pour différents nombres de candidats, Yang et collab. (2019) considèrent également la mesure corrigée pour le temps ESS/CPU. Lorsque le temps de calcul est ainsi pris en compte, leurs conclusions diffèrent entre les trois densités cibles étudiées. En effet, pour une densité à géométrie courbe (en forme de banane), l'algorithme M.-H. est celui avec des valeurs de ESS/CPU les plus élevée pour la majorité des composantes du problème. De plus, pour une densité venant d'un modèle bayésien hiérarchique, l'algorithme M.-H. affiche des performances similaires à leur algorithme avec $K = 30$ candidats pour une majorité de composantes. Pour un mélange de 20 densités gaussiennes, les deux algorithmes MTM sont les plus efficaces, par rapport à ce critère corrigé, uniformément sur les composantes de modèle.

Dans une série d'exemples, Bédard et Mireuta (2013) effectuent des simulations pour comparer deux algorithmes MTM à l'algorithme Metropolis à propositions gaussiennes (dénoté RWM). Le premier algorithme MTM utilise $K = 2$ candidats indépendants gaussiens alors que le second est l'algorithme MTM *Hit-and-run* (dénoté MTM-HR) où $K = 2$ candidats opposés sont produit à partir de la même direction. L'utilisation de seulement deux candidats est justifiée par les résultats théoriques d'échelle optimale (tableaux 4.2 et 4.1) où la plus grande augmentation de l'efficacité est lorsque $K = 1$ passe à $K = 2$: on présume donc que le meilleur équilibre entre efficacité et temps de calcul sera atteint pour $K = 2$. Les comparaisons sont effectuées sur la base de l'ESJD et de sa version corrigée par le temps de calcul, ESJD/CPU.

Dans un modèle de régression bayésien à $d = 4$ dimensions, Bédard et Mireuta (2013, tableau 4) obtiennent les résultats suivants. Les deux algorithmes MTM afichent une valeur de ESJD supérieure à celle de l'algorithme Metropolis (RWM : 0,1976 ; MTM : 0,3297 et MTM-HR : 0,3875.) Cependant, ces algorithmes sont deux (MTM-HR) ou trois (MTM) fois plus longs à exécuter que l'algorithme Metropolis et l'augmentation de l'efficacité n'est pas particulièrement grande en comparaison. Lorsque l'ESJD est ajusté pour le temps de calcul, on trouve que l'algorithme Metropolis est le plus efficace (RWM : 0,1976 ; MTM : 0,1194 et MTM-HR : 0,1813.) De plus, pour une densité cible formée d'un mélange de deux densités gaussiennes, Bédard et Mireuta (2013, tableau 8) obtiennent des résultat semblables.

Dans un contexte d'inférence pour un modèle de régression linéaire ($d = 8$), les auteurs effectuent une comparaison similaire entre les algorithmes. Notons cependant que la covariance de la densité cible doit être estimée préalablement afin d'être utilisée par les algorithmes : sans ajustement des candidats par la covariance, tous les algorithmes affichent une performance faible. En comparant les algorithmes RWM, MTM et MTM-HR utilisant tous la covariance estimée, Bédard et Mireuta (2013, tableau 6) arrivent aux conclusions suivantes. Sans ajustement pour le temps de calcul, les algorithmes à essais multiples sont plus efficaces en terme d'ESJD (RWM : 138,7 ; MTM : 226,0 et MTM-HR : 274,1.) Par contre, lorsque le temps de calcul est pris en compte (ESJD/CPU), l'algorithme MTM-HR obtient la plus forte efficacité empirique (RWM : 36,72 ; MTM : 28,21 et MTM-HR : 55,82.) Il faut cependant mentionner que l'utilisation de la covariance augmente sensiblement le temps de calcul de tous les algorithmes et que la densité cible est en comparaison peu coûteuse à évaluer ; l'augmentation du temps de calcul par l'utilisation d'essais multiples n'est donc plus causée par le même phénomène. L'algorithme MTM-HR profite donc d'un rabais considérable en utilisant une variable aléatoire commune pour les deux candidats.

Dans un modèle bayésien hiérarchique à haute dimension ($d = 213$), Bédard et Mireuta (2013, tableau 7) trouvent à nouveau que les algorithmes à essais multiples sont plus efficaces en terme d'ESJD (RWM : 0,113 ; MTM : 0,191 et MTM-HR : 0,227), mais que l'algorithme MTM-HR performe légèrement mieux lorsque le temps de calcul est pris en compte (RWM : 0,113 ; MTM : 0,062 et MTM-HR : 0,118.)

Dans tous ces exemples, il est important de noter que l'algorithme MTM-HR s'exécute plus rapidement que l'algorithme MTM et que ce même algorithme est systématiquement plus efficace que l'algorithme MTM à candidats indépendants. Ces deux propriétés sont donc parfois suffisantes pour produire une efficacité empirique supérieure à celle de l'algorithme Metropolis. La réduction du temps de calcul ainsi que la qualité améliorée des candidats, résultant tous deux d'un ensemble de candidats corrélés, semble donc une stratégie permettant de battre (ou du moins égaler) l'efficacité de l'algorithme Metropolis en pratique.

Chapitre 5

Algorithme MTM adaptatif

L'algorithme Metropolis est une des méthodes MCMC les plus communément utilisées afin d'effectuer une estimation Monte Carlo d'une espérance $\pi(f)$. L'efficacité de cette estimation peut s'avérer optimale sous certaines conditions que l'on peut classifier en deux types. D'abord, la distribution cible π doit satisfaire à certaines conditions de régularité (section 2.5.1). Ensuite, pour ce choix de distribution cible, les paramètres de la distribution instrumentale doivent être choisis adéquatement.

L'algorithme AM, introduit par Haario et collab. (2001), s'attaque à la deuxième condition d'optimalité. En effet, pour une densité cible suffisamment régulière, cet algorithme permet de trouver automatiquement les paramètres de la densité instrumentale qui produisent une chaîne optimale en terme d'efficacité d'estimation. Pour ce faire, la covariance d'une densité gaussienne est adaptée au fur et à mesure que la chaîne progresse et certaines garanties théoriques permettent d'obtenir la convergence de la covariance vers la covariance optimale. Par l'utilisation d'un noyau de transition Metropolis, les densités cibles qui produiront un comportement favorable de l'algorithme AM doivent être suffisamment simples. En effet, l'algorithme AM peut éprouver des difficulté à bien échantillonner de la densité cible lorsque celle-ci est multimodale ou comporte des régions aux covariances locales distinctes.

Dans ce chapitre, nous proposons une nouvelle méthode cherchant à imiter l'application de l'algorithme AM à des densités cibles plus générales. Pour y arriver, nous considérons un noyau de transition de type Metropolis à essais multiples (MTM) qui présente une plus grande flexibilité qu'un noyau Metropolis régulier : les différentes composantes du noyau peuvent mieux modéliser les différentes particularités de la densité cible. Comme c'est le cas pour l'algorithme Metropolis, le choix des paramètres du noyau MTM est crucial pour assurer l'efficacité de l'algorithme. Ainsi, l'introduction d'adaptation au sein des méthodes MTM s'avère le chemin logique dans la mise au point automatique des paramètres du noyau MTM.

À la section 5.1, nous considérons un exemple numérique montrant les limitations des algorithmes AM et MTM et justifiant du même fait la proposition d'un algorithme MTM adaptatif. La section 5.2 contient une revue des différentes tentatives d'algorithmes MTM adaptatifs que l'on retrouve dans la littérature. Ensuite, l'algorithme MTM adaptatif (aMTM) est proposé à la section 5.3. Cette section contient également une description de plusieurs déclinaisons possibles pour cetalgorithme. Enfin, la justification théorique des propriétés de l'algorithme est effectuée à la section 5.4.

5.1 Motivation

Afin de motiver l'introduction de l'algorithme aMTM, nous considérons d'abord un exemple qui mettra en lumière certaines des limitations des algorithmes AM 3.1 et MTM 4.1. L'algorithme AM s'ajuste automatiquement à la covariance de la densité cible, mais cette méthode éprouve certaines difficultés lorsque la densité cible est multimodale ou à géométrie complexe. L'algorithme MTM, quant à lui, nécessite une mise au point des différents paramètres de ses densités instrumentales afin d'échantillonner efficacement.

Soit π une densité cible bimodale formée d'un mélange de deux densités gaussiennes bivariées :

$$\pi = w_1 \mathcal{N}_2(\mu_1,\Sigma_1) + w_2 \mathcal{N}_2(\mu_2,\Sigma_2),$$

où les paramètres sont donnés par

$$\begin{aligned} w_1 &= 0.3, & \mu_1 &= (0,20)^\top, & \Sigma_1 &= \mathrm{diag}(9,1); \\ w_2 &= 0.7, & \mu_2 &= (8,0)^\top, & \Sigma_2 &= \mathrm{diag}(1,9). \end{aligned}$$

La figure 5.1(a) contient une représentation de cette densité cible. Il est possible d'échantillonner directement à partir de π d'une manière i.i.d. : la figure 5.1(b) contient un tel échantillon de taille $N = 10{,}000$.

Les conditions initiales influenceront grandement le comportement de l'algorithme AM pour cette densité cible. Lorsque l'algorithme est initialisé dans un des modes, il peut être difficile d'atteindre l'autre mode. En restant longtemps dans un des modes, la covariance adaptative tend à se stabiliser vers la covariance de ce mode et un saut entre les modes devient de moins en moins probable. C'est ce qu'on observe à a figure 5.1(c) où la valeur initiale de la chaîne est $x_0 = (0,10)^\top$ et où la covariance initiale est petite en comparaison à la covariance globale. On trouve bien un taux d'acceptation satisfaisant de 0,38 (pour une densité cible à deux composantes i.i.d., cf. tableau 2.1, le taux optimal est de 0,352), mais la distribution empirique ne représente pas bien la densité cible dans un des modes. Lorsque la covariance initiale est choisie grande par rapport à la covariance cible, la covariance instrumentale tend à se stabiliser vers la covariance globale de la densité cible comme on peut le voir à la figure 5.1(d). Bien que ceci permette d'explorer les deux modes de π, il en résulte néanmoins un faible taux d'acceptation et une exploration incomplète des ailes de chacun des modes. Ainsi, pour toute initialisation, l'algorithme AM ne semble pas pouvoir arriver à un échantillonnage particulièrement représentatif de π.

Une algorithme MTM à $K = 3$ ou $K = 2$ composantes peut échantillonner efficacement de cette densité cible comme le montrent les figures 5.1(e) et (f). Si les différentes densités instrumentales sont choisies convenablement, l'échantillon résultant représentera bien la densité cible. Par exemple, pour $K = 3$, si deux composantes ressemblent respectivement à chacune des deux composantes du mélange et si la troisième densité est sur-dispersée afin de permettre les sauts entre les modes, alors l'algorithme MTM donnera un échantillon dont la distribution s'approche de π. Pour $K = 2$, il est possible de choisir une composante telle qu'elle permette principalement le saut entre les modes et une composante permettant l'exploration locale de π. Cependant, dans les deux cas, le choix particulier de ces distributions instrumentales a été fait en connaissant les propriétés de la densité cible. Ceci ne sera

généralement pas possible et une période de mise-au-point devra être effectuée avant de trouver une collection de densités instrumentales appropriés à l'échantillonnage. Il s'agit de la même problématique qui a forcé le développement des algorithmes adaptatifs ; il devient donc évident que l'automatisation de la mise-au-point via l'adaptation peut s'avérer pertinente.

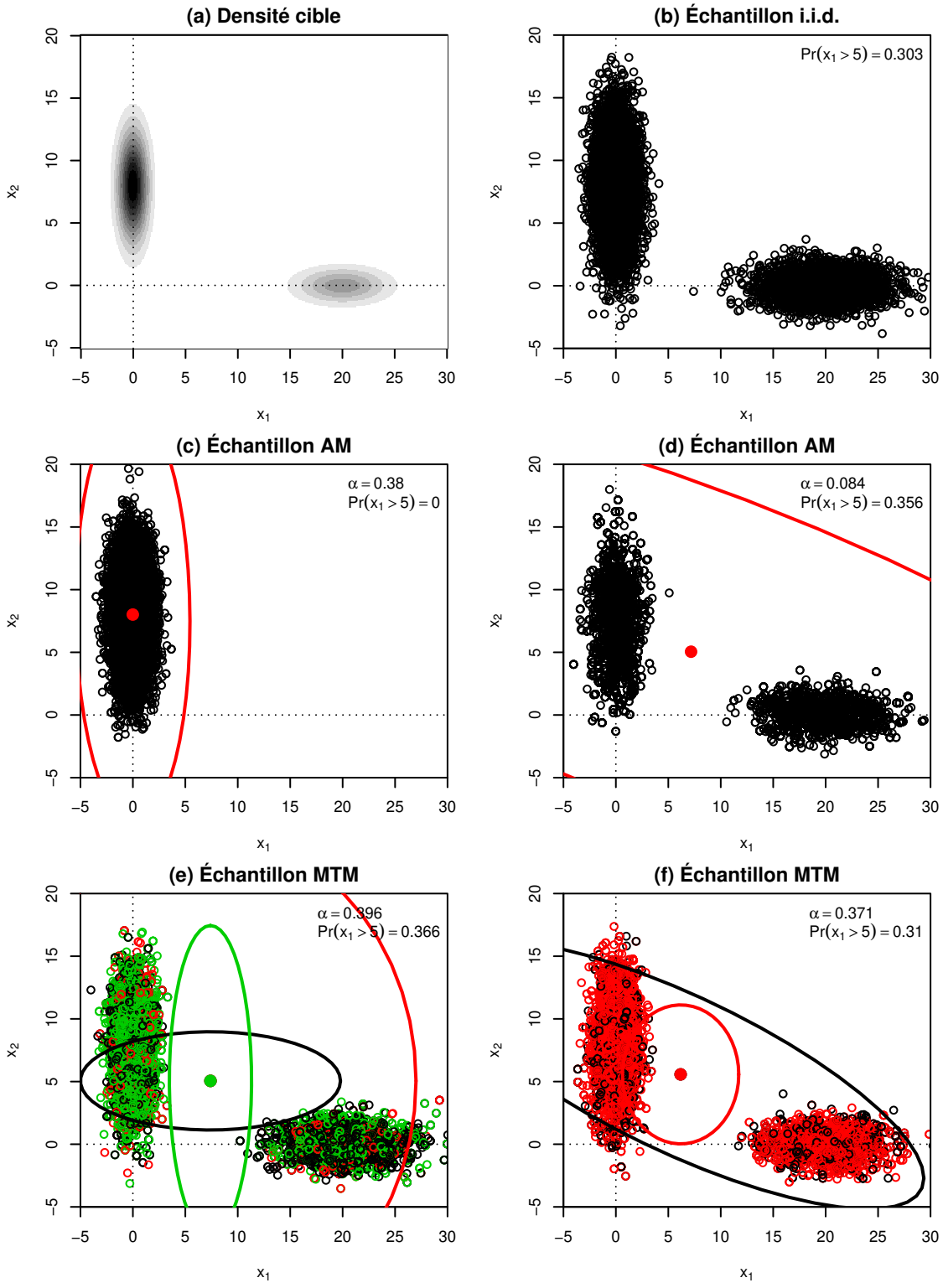


Figure 5.1 *Échantillonnage d'une densité bimodale ($N = 10{,}000$). (a) Représentation de la densité cible. (b) Échantillonnage i.i.d. (c) Échantillonnage par l'algorithme AM 3.1 où l'état initial est $(0,10)^\top$ et la covariance initiale est petite. (d) Échantillonnage par l'algorithme AM 3.1 où l'état initial est $(20,0)^\top$ et la covariance initiale est grande. (e) et (f) Échantillonnage par l'algorithme MTM 4.1 à $K = 3$ et $K = 2$ candidats indépendants de covariances distinctes ; la couleur des points représente quelle densité a été choisie à l'étape de sélection MTM. Pour (b) à (f), α est le taux d'acceptation de l'algorithme, $Pr(x_1 > 5)$ est la proportion de points avec $x_1 > 5$ et les ellipses représentent les covariances des différentes densités instrumentales.*

5.2 Revue de littérature

On retrouve très peu d'instances d'incorporation d'adaptation au sein d'un algorithme MCMC à essais multiples ou d'utilisation d'essais multiples comme noyau de transition d'un algorithme adaptatif. Cette section considère quelques-unes des tentatives présentes dans la littérature. Notons que des auteurs mentionnent que de telles stratégies pourraient résoudre certains problèmes auxquels leurs algorithmes font face sans tester directement l'idée ; par exemple, Tran et collab. (2016, section 8) proposent d'utiliser des essais multiples dans le noyau de transition d'un algorithme MH indépendant (définition 2.25) adaptatif afin d'améliorer l'exploration de l'espace d'états.

5.2.1 Adaptation par estimation non-paramétrique

Martino et collab. (2018) considèrent un algorithme MCMC adaptatif constitué d'un noyau IMH (définition 2.25) où la densité instrumentale est une estimation non-paramétrique q_n de la densité cible construite à partir d'états précédents de la chaîne. Une des extensions considérée par les auteurs est l'utilisation d'essais multiples à la phase d'échantillonnage. L'algorithme AISMTM (*Adaptive Independant Sticky Multiple-try Metropolis*) est défini pour une densité cible univariée π à support univarié $\mathcal{X} \subseteq \mathbb{R}$. Un ensemble de points de support $\mathcal{S}_n$ est maintenu suivant la progression de la chaîne : l'ensemble $\mathcal{S}_n$ est constitué de certains des points de référence utilisés lors des étapes d'acceptation de l'algorithme MTM. Ces points de référence sont alors utilisés pour définir l'estimation non-paramétrique de π qui est utilisée pour la génération des candidats. L'algorithme 5.1 détaille la procédure.

L'estimation $q_n(\cdot|\,\mathcal{S}_n)$ peut être effectuée de diverses manières : Martino et collab. (2018, section 3) proposent une approximation constante par morceaux ou linéaire par morceaux augmentée d'ailes exponentielles afin de couvrir entièrement le support de π. Quant aux poids de sélection MTM, ils suggèrent l'utilisation de poids d'importance $w_n(x|\,\mathcal{S}_n) = \pi(x)/q_n(x|\,\mathcal{S}_n)$. Les poids de sélection à l'étape d'adaptation sont définis selon

$$\varphi_n(z) = \max\left\{w_n(z|\,\mathcal{S}_n), \frac{1}{w_n(z|\,\mathcal{S}_n)}\right\} = \frac{d_n(z|\,\mathcal{S}_n)}{\min\{\pi(z), q_n(z|\,\mathcal{S}_n)\}} + 1,$$

où $d_n(z|\,\mathcal{S}_n) = |\pi(z) - q_n(z|\,\mathcal{S}_n)|$ dénote la distance entre la densité cible et l'estimation non-paramétrique évaluée en z. La probabilité d'inclure le nouveau poids sélectionné z est donnée par

$$P_a(z|\,\mathcal{S}_n) = \eta_n\left(z, d_n(z|\,\mathcal{S}_n)\right),$$

pour un certain choix de fonction $\eta_n(\cdot)$. Martino et collab. (2018, section 4.1) considèrent quelques exemples de $\eta_n(z,d)$:

$$\begin{aligned}
&\eta_n(z,d) = 1 - e^{-\beta d}, && \beta > 0\,;\\
&\eta_n(z,d) = \mathbb{1}(d > \varepsilon), && \varepsilon > 0\,;\\
&\eta_n(z,d) = \frac{d}{\max\{\pi(z), q_n(z, \mathcal{S}_n)\}}.
\end{aligned}$$

Par construction, cet algorithme satisfait à la condition d'adaptation indépendante (définition 3.1.) En effet, l'étape d'ajout à l'ensemble $\mathcal{S}_n$ est telle que le nouvel état de la chaîne n'est jamais inclus dans $\mathcal{S}_{n+1}$ de sorte que la densité instrumentale $q_{n+1}(\cdot|\,\mathcal{S}_{n+1})$ ne dépend jamais de x_{n+1}. Ainsi, les résultats théoriques développés par Holden et collab. (2009) sur les algorithmes à adaptation indépendante peuvent être appliqués à cet algorithme. Afin de vérifier l'ergodicité de l'algorithme, il est alors possible de supposer la condition de Doeblin forte. Cette condition suppose principalement que la densité instrumentale a des ailes uniformément plus lourdes que celles de la densité cible π.

Condition 5.1 (Condition de Doeblin forte, Martino et collab., 2018, annexe A) *Soit une densité instrumentale $q_n(\cdot|\,\mathcal{S}_n)$ univariée donnée par une estimation non-paramétrique d'une densité cible π à support $\mathcal{X} \subseteq \mathbb{R}$ à partir de l'ensemble de points de support $\mathcal{S}_n \subset \mathcal{X}$. Il existe une constante $a_n \in (0,1]$ telle que*

$$\frac{1}{a_n} q_n(x|\,\mathcal{S}_n) \geqslant \pi(x), \qquad \forall\, x \in \mathcal{X}\,.$$

Si les estimations non-paramétriques satisfont la condition de Doeblin forte pour chaque $n \geqslant 1$, alors l'algorithme sera géométriquement ergodique sous une certaine condition sur la suite des constantes a_n, $n \geqslant 1$.

Théorème 5.1 (Holden et collab., 2009, théorème 2) *Soit un algorithme MH à adaptation indépendante satisfaisant la condition de Doeblin forte (condition 5.1) pour une suite de constantes $\{a_n\}_{n\geqslant 1}$. Alors, l'algorithme est géométriquement ergodique à π si*

$$\prod_{n=1}^{N} (1 - a_n) \overset{n\to\infty}{\longrightarrow} 0. \tag{5.1}$$

Puisque $q_n(\cdot|\,\mathcal{S}_n)$ est une estimation non-paramétrique de π, alors on aura $q_n(x|\,\mathcal{S}_n) \to \pi(x)$ avec $n \to \infty$. Ainsi, il sera possible de choisir $a_n \to 1$ dans la condition de Doeblin forte et la condition (5.1) pourra être vérifiée. Enfin, l'ergodicité de la chaîne est donc garantie par construction dès que l'estimation non-paramétrique est convergente uniformément sur $\mathcal{X}$.

Martino et collab. (2018, section 9) effectuent des simulations numériques afin d'évaluer la performance de leur algorithme. Pour une densité cible univariée bimodale, ils considèrent des nombres de candidats parmi $K = 1{,}10{,}50$ ainsi que quatre méthodes de construction de l'estimation non-paramétrique. De plus, ils comparent leur algorithme à l'algorithme ARMS de Gilks et collab. (1995). L'utilisation d'essais multiples améliore systématiquement l'efficacité par rapport à un seul candidat. Cependant, le temps de calcul n'est pas pris en compte dans les comparaisons, alors il n'est pas clair que l'augmentation de l'efficacité soit réelle. Dans un second exemple bimodal en $d = 2$ dimensions, des résultats similaires sont obtenus.

5.2.2 Adaptation à l'aide de chaînes parallèles

Casarin et collab. (2013) proposent un algorithme à essais multiples où des chaînes parallèles sont utilisées (section 4.1.3.1). Lorsque l'information contenue dans les chaînes est partagée aux autres chaînes pour modifier le noyau de transition, alors l'algorithme comporte de l'adaptation externe (section 3.1.2).

Une des variantes de l'algorithme IMTM suggérées par Casarin et collab. (2013) est de modifier les poids de sélection MTM à partir de l'ensemble des chaînes parallèles. En particulier, les poids de

sélection contiennent un facteur

$$v^{(k)} = \frac{1}{M}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left(k = k_{n-1,m}\right)\right), \quad k = 1, \ldots, K,$$

où M est le nombre de chaînes, K le nombre de candidats et $\mathbb{1}\left(k = k_{n-1,m}\right)$ prend la valeur 1 si et seulement si le k-ième candidat fut sélectionné à l'étape précédente dans la m-ième chaîne ($k_{n-1,m}$ correspond à l'index du candidat choisi à l'itération $n-1$ dans la chaîne m). Ainsi, les poids de sélection sont artificiellement gonflés lorsque la même densité densité a été choisie plus souvent à l'étape précédente.

Une seconde variante proposée par Casarin et collab. (2013) est le transfert de localisation entre les chaînes ; ce transfert est effectué par l'utilisation de tempérage tel que décrit à la section 3.1.3.3. L'algorithme AIMTM (*Annealed IMTM*) échantillonne différemment dans la chaîne principale (avec π comme densité cible) que dans les autres chaînes tempérées. Les chaînes tempérées explorent l'espace d'états à l'aide d'un algorithme MH régulier alors que la chaîne principale progresse à l'aide d'un noyau MTM où la marche aléatoire est centrée à l'un des états actuels des autres chaînes. La procédure est décrite à l'algorithme 5.2.

D'un point de vue théorique, Casarin et collab. (2013) montrent seulement que l'algorithme IMTM (c.-à-d., sans adaptation) satisfait la condition d'équilibre 2.13. Les différents résultats de la section 4.2 peuvent être étendus aux chaînes parallèles MTM sans interaction de sorte qu'il est possible de montrer que l'algorithme IMTM est ergodique et satisfait une loi des grands nombres en supposant des conditions similaires à celles de la proposition 4.4. Cependant, en introduisant l'un des deux types d'adaptation précédents, des conditions supplémentaires devront être considérées ; les auteurs ne s'attardent pas à cette problématique.

Casarin et collab. (2013, section 4) effectuent une série d'expériences numériques afin d'évaluer les différentes versions de leur algorithme. Ils trouvent que l'utilisation des poids adaptatifs n'améliore pas significativement l'efficacité de l'algorithme par rapport à un algorithme identique utilisant des poids MTM réguliers. De plus, les auteurs observent que l'utilisation d'essais multiples est systématiquement supérieure à un algorithme MH équivalent, que ce soit avec ou sans tempérage. Le temps de calcul n'est cependant pas pris en compte pour les comparaisons entre les versions adaptatives et non-adaptatives, mais le coût computationnel introduit par l'adaptation ne devrait pas être substantiel par construction.

5.2.3 Adaptation d'un noyau MTM par composante

Yang et collab. (2019) considèrent un algorithme MTM effectuant une mise à jour de l'état de la chaîne composante par composante. Au moment de mettre à jour la j-ième composante, l'algorithme propose K candidats $y_j^{(1)}, \ldots, y_j^{(K)}$ à partir de l'état actuel via des densités instrumentales univariées $T_j^{(1)}, \ldots, T_j^{(K)}$. L'algorithme CMTM (*Component-wise MTM*) est donc une extension à essais multiples de l'algorithme MwG 2.7. Maintenant, les densités instrumentales doivent être fixée d'avance et demeurent fixes tout au long de l'échantillonnage : Yang et collab. (2019) proposent plutôt d'adapter ces densités. Pour ce faire, ils choisissent que chacune des densités $T_j^{(k)}$ proviennent d'une même famille paramétrique (e.g. gaussienne), mais que les paramètres d'échelle $\sigma_j^{(1)}, \ldots, \sigma_j^{(K)}$ diffèrent avec k. En particulier, les paramètres d'échelle sont ordonnés suivant l'ordre naturel $\sigma_j^{(1)} < \sigma_j^{(2)} < \cdots < \sigma_j^{(K)}$.

L'adaptation consiste alors en une transformation de l'étendue des paramètres d'échelle $[\sigma_j^{(1)}, \sigma_j^{(K)}]$

de façon à produire des taux de sélection relativement uniformes sur l'étendue. Lorsque le taux de sélection de la K-ième densité est trop élevé alors $\sigma_j^{(K)}$ est augmenté ; si ce taux est trop faible, alors $\sigma_j^{(K)}$ est diminué. Similairement, si le taux de sélection de la première densité est trop élevé alors $\sigma_j^{(1)}$ est réduit ; si le taux de sélection est trop faible, alors $\sigma_j^{(1)}$ est augmenté. L'intuition derrière ces choix repose sur l'utilisation de poids proportionnels au saut entre l'état actuel et le prochain état : les petits sauts sont donc associés à de faibles probabilités de sélection. L'algorithme ACMTM (*Adaptive Component-wise MTM*) ainsi défini est détaillé à l'algorithme 5.3.

L'ergodicité de l'algorithme ACMTM est démontrée par Yang et collab. (2019, section 3.5) via les conditions d'adaptation diminuante 3.1 et de convergence bornée 3.2. L'adaptation diminuante est assurée directement en adaptant selon une probabilité tendant vers 0 avec $n \to \infty$, et ce, malgré le fait que la quantité même d'adaptation, lorsqu'elle se produit, ne tend pas vers 0. La convergence bornée est quant à elle assurée en forçant la chaîne à l'intérieur d'un ensemble compact $\mathcal{K} \subseteq \mathcal{X}$ et en forçant les paramètres d'échelle dans un intervalle $[\varepsilon, L]$ lui aussi compact.

Dans une série d'applications à diverses densités cibles, Yang et collab. (2019, section 4) évaluent la performance de leurs algorithmes CMTM et ACMTM et la comparent à un algorithme MwG ainsi qu'une version adaptative de l'algorithme MwG. Pour une densité cible formée d'un mélange de lois gaussiennes à $d = 20$ dimensions et dans un modèle de composantes de variance, l'algorithme ACMTM à $K = 20$ ou $K = 30$ candidats affiche les meilleurs valeurs d'ESS par unité de temps de calcul. Pour une densité en forme de « banane », c'est l'algorithme MwG adaptatif qui est le plus efficace empiriquement.

Algorithme 5.1 *Adaptive Independant Sticky MTM* (AISMTM, Martino et collab., 2018)

Données Densité cible π à support $\mathcal{X} \subseteq \mathbb{R}$, nombre de candidats K, estimation non-paramétrique $q_n(\cdot|\,\mathcal{S}_n)$, fonction de poids MTM $w_n(\cdot|\,\mathcal{S}_n)$, fonction de poids d'ajout $\varphi_n(\cdot)$ et probabilité d'ajout $P_a(\cdot|\,\mathcal{S}_n)$.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0 et ensemble de points de support $\mathcal{S}_0$.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonnage MTM.*
      i. *Candidats.* Générer
      $$\left(y^{(1)}, \ldots, y^{(K)}\right) \overset{\text{i.i.d.}}{\sim} q_n(\cdot|\,\mathcal{S}_n);$$
      ii. *Poids.* Calculer $w_n(y^{(k)}|\,\mathcal{S}_n)$, $k = 1, \ldots, K$;
      iii. *Sélection.* Choisir $y = y^{(j)}$ avec probabilité proportionnelle à $w_n(y^{(j)}|\,\mathcal{S}_n)$, $j = 1, \ldots, K$;
      iv. *Points de référence.* Poser $x_*^{(k)} = y^{(k)}$ et $z^{(k)} = y^{(k)}$ pour $k \neq j$ ainsi que $x_*^{(j)} = x_n$ (notons qu'une densité instrumentale indépendante ne requiert pas de points de référence aléatoires ; les points $z^{(k)}$, $k = 1, \ldots, K$, formeront un ensemble parmi lequel un nouveau point sera choisi pour être ajouté à $\mathcal{S}_n$)) ;
      v. *Poids inverses.* Calculer $w_n(x_*^{(k)}|\,\mathcal{S}_n)$, $k = 1, \ldots, K$;
      vi. *Probabilité d'acceptation.* Calculer
      $$\alpha = \min\left\{1, \frac{\sum_{k=1}^{K} w_n(y^{(k)}|\,\mathcal{S}_n)}{\sum_{k=1}^{K} w_n(x_*^{(k)}|\,\mathcal{S}_n)}\right\};$$
      vii. *Acceptation.* Avec probabilité α, accepter la proposition ($x_{n+1} = y$ et $z^{(j)} = x_n$) sinon rejeter la proposition ($x_{n+1} = x_n$ et $z^{(j)} = y$.)
   (b) *Adaptation.* Mise à jour de l'ensemble des points de support :
      i. *Sélection* Avec probabilité proportionnelle à $\varphi_n(z^{(k)})$, $k = 1, \ldots, K$, choisir $z = z^{(k)}$;
      ii. *Acceptation* Avec probabilité $P_a(z|\,\mathcal{S}_n)$, ajouter z à l'ensemble (c.-à-d., $\mathcal{S}_{n+1} = \mathcal{S}_n \cup \{z\}$) sinon conserver l'ensemble (c.-à-d., $\mathcal{S}_{n+1} = \mathcal{S}_n$.)

Sortie L'échantillon $x_{1:N}$.

Algorithme 5.2 *Annealed Interactive MTM* (AIMTM, Casarin et collab., 2013)

Données Densité cible π à support $\mathcal{X} \subseteq \mathbb{R}^d$, nombre de chaînes M, nombre de candidats K, densités instrumentales $q_1^{(k)}$, $k = 1, \ldots, K$, et q_m, fonctions de poids MTM $w_1^{(k)}$ pour $k = 1, \ldots, K$ et $m = 1, \ldots, M$ et échelle de températures $1 > \xi_2 > \cdots > \xi_M$.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne $x_{0,m}$, $m = 1, \ldots, M$.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonnage MTM.* Pour la chaîne principale ($m = 1$),
      i. *Choix des centres.* Échantillonner $(I^{(1)}, \ldots, I^{(k)}) \overset{\text{i.i.d.}}{\sim} \text{uniforme}\{1, \ldots, M\}$;
      ii. *Candidats.* Générer
      $$y^{(k)} \sim q_1^{(k)}(\cdot | x_{n,I^{(k)}}), \quad k = 1, \ldots, K;$$
      iii. *Poids.* Calculer $w_1^{(k)}(y^{(k)} | x_{n,1})$, $k = 1, \ldots, K$;
      iv. *Sélection.* Choisir $y = y^{(j)}$ avec probabilité proportionnelle à $w_n(y^{(j)} | x_{n,1})$;
      v. *Points de référence.* Poser $x_*^{(j)} = x_{n,1}$ et, pour $k \neq j$, échantillonner
      $$x_*^{(k)} \sim \begin{cases} q_1^{(k)}(\cdot | x_{n,I^{(k)}}), & I^{(k)} \neq 1; \\ q_1^{(k)}(\cdot | y), & I^{(k)} = 1; \end{cases}$$
      vi. *Poids inverses.* Calculer $w_1^{(k)}(x_*^{(k)} | y)$, $k = 1, \ldots, K$;
      vii. *Probabilité d'acceptation.* Calculer
      $$\alpha = \min \left\{ 1, \frac{\sum_{k=1}^K w_1^{(k)}(y^{(k)} | x_{n,1})}{\sum_{k=1}^K w_1^{(k)}(x_*^{(k)} | y)} \right\};$$
      viii. *Acceptation.* Avec probabilité α, accepter la proposition ($x_{n+1,1} = y$) sinon rejeter la proposition ($x_{n+1,1} = x_{n,1}$.)
   (b) *Échantillonnage MH.* Pour $m = 2, \ldots, M$,
      i. *Proposition.* Générer $y \sim q_m(\cdot | x_{n,m})$;
      ii. *Probabilité d'acceptation.* Calculer
      $$\alpha = \min \left\{ 1, \frac{\pi^{\xi_m}(y) q_m(x_{n,m} | y)}{\pi^{\xi_m}(x_{n,m}) q_m(y | x_{n,m})} \right\};$$
      iii. *Acceptation.* Avec probabilité α, accepter la proposition ($x_{n+1,m} = y$) sinon rejeter la proposition ($x_{n+1,m} = x_{n,m}$).

Sortie L'échantillon $x_{1:N,m}$.

Algorithme 5.3 *Adaptive Component-wise MTM* (ACMTM, Yang et collab., 2019)

Données Densité cible π à support $\mathcal{X} \subseteq \mathbb{R}^d$, nombre de candidats K, densités instrumentales $q_j^{(k)}(\cdot|x) = \mathcal{N}(\cdot; x, \sigma_j^{(k)2})$ et fonctions de poids $w_j^{(k)}$ pour $j = 1, \ldots, p$ et pour $k = 1, \ldots, K$, β le nombre d'itérations entre les adaptations.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0, paramètres d'échelle initiaux $\sigma_j^{(k)}$ et taux de sélection initiaux $S_j^{(k)} = 1/K$, $j = 1, \ldots, p$, $k = 1, \ldots, K$.
2. Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonnage CMTM.* Pour $j = 1, \ldots, d$, poser $x = x_{n,j}$ puis :
      i. *Candidats.* Générer $y^{(k)} \sim q_j^{(k)}(\cdot|x)$, $k = 1, \ldots, K$;
      ii. *Poids.* Calculer $w_j^{(k)}(y^{(k)}|x)$, $k = 1, \ldots, K$;
      iii. *Sélection.* Choisir $y = y^{(s)}$ avec probabilité proportionnelle à $w_j^{(k)}(y^{(k)}|x)$. Mettre à jour les taux de sélection $S_j^{(k)}$;
      iv. *Points de référence.* Poser $x_*^{(s)} = x$ et échantillonner $x_*^{(k)} \sim q_j^{(k)}(\cdot|y)$, $k \neq s$;
      v. *Poids inverses.* Calculer $w_j^{(k)}(x_*^{(k)}|y)$, $k = 1, \ldots, K$;
      vi. *Probabilité d'acceptation.* Calculer
      $$\alpha = \min\left\{1, \frac{\sum_{k=1}^K w_j^{(k)}(y^{(k)}|x)}{\sum_{k=1}^K w_j^{(k)}(x_*^{(k)}|y)}\right\};$$
      vii. *Acceptation.* Avec probabilité α, accepter la proposition $(x_{n+1,j} = y)$ sinon rejeter la proposition $(x_{n+1,j} = x_{n,j}.)$
   (b) *Adaptation.* Si $n \equiv 0 \mod \beta$ et avec probabilité $P_a = \max\{0.99^{a-1}, a^{-1/2}\}$ où $a = n/\beta$: pour $j = 1, \ldots, d$,
      i. Si $S_j^{(K)} > 2/K$, $\sigma_j^{(K)} \leftarrow 2\sigma_j^{(K)}$;
      ii. Si $S_j^{(K)} < 1/2K$ et $\sigma_j^{(1)} < \sigma_j^{(K)}/2$, $\sigma_j^{(K)} \leftarrow \sigma_j^{(K)}/2$;
      iii. Si $S_j^{(1)} > 2/K$, $\sigma_j^{(1)} \leftarrow \sigma_j^{(1)}/2$;
      iv. Si $S_j^{(1)} < 1/2K$ et $2\sigma_j^{(1)} < \sigma_j^{(K)}$, $\sigma_j^{(1)} \leftarrow 2\sigma_j^{(1)}$;
      v. Mettre à jour $\{\sigma_j^{(k)}\}_{k=1}^K$ pour former une échelle logarithmique sur $[\sigma_j^{(1)}, \sigma_j^{(K)}]$;
      vi. Réinitialiser les taux de sélection $S_j^{(k)}$, $k = 1, \ldots, K$.

Sortie L'échantillon $x_{1:N}$.

5.3 Description de l'algorithme

À la section 4.1, il a été question des différents choix qui s'offrent à l'utilisateur d'algorithmes MCMC à essais multiples. D'abord, une partie majeure de la définition d'un algorithme à essais multiples est le processus de génération des candidats : la densité marginale de chaque candidat et la relation entre ces candidats doivent être spécifiées. Il est possible de choisir des densités marginales identiques, d'une même famille paramétrique ou tout simplement différentes ; il est possible de générer un ensemble de candidats indépendants, corrélés ou même complètement déterministes. D'autre part, la méthode de sélection de la proposition au sein de l'ensemble des candidats doit être spécifiée. En particulier, la fonction des poids de sélection peut être choisie pour favoriser certains comportements. En se basant sur divers résultats théoriques ou expérimentaux, il est possible d'identifier heuristiquement certaines combinaisons de densités, de corrélation et de sélection plus efficaces que d'autres.

Algorithme 5.4 Algorithme Metropolis à essais multiples adaptatif (aMTM)

Données La densité cible π, la taille de l'échantillon Monte Carlo N et les fonctions d'adaptation $Q_n \mapsto Q_{n+1}$ et $w_n \mapsto w_{n+1}$.

Procédure

1. *Initialisation.* Valeur initiale de la chaîne x_0, densité instrumentale initiale Q_0 et fonction de poids initiale w_0.
2. *Itérations MCMC.* Pour $n = 0, \ldots, N-1$,
   (a) *Échantillonnage MTM.* Générer le nouvel état x_{n+1} suivant l'étape 2 de l'algorithme MTM 4.1 avec la densité conjointe Q_n et la fonction de poids de sélection w_n ;
   (b) *Adaptation.* Mettre à jour Q_{n+1} et w_{n+1} à l'aide de x_{n+1} et de $\{x_j, Q_j, w_j\}_{j=0}^{n}$;

Sortie L'échantillon $x_{1:N}$.

L'algorithme le plus général est donc identique à l'algorithme MTM 4.1 à l'exception d'une étape supplémentaire d'adaptation à chaque itération MCMC : l'algorithme 5.4 montre les grandes lignes de la procédure générale. Dans la conception de cette étape d'adaptation, il est important de garder en tête que l'étude théorique des propriétés de l'algorithme peut devenir rapidement complexe en raison de la perte de la propriété markovienne. Ainsi, les différents résultats des sections 3.2 et 3.3 peuvent nous guider dans la manière de définir l'adaptation de sorte à assurer certaines des propriétés essentielles à tout algorithme MCMC.

Parmi tous les détails de l'algorithme qui peuvent être adaptés, on concentrera nos efforts sur l'adaptation des densités marginales de la densité instrumentale conjointe. En effet, les structures de corrélations exposées à la section 4.1.1 doivent être choisies et ne permettent pas vraiment d'adaptation. De plus, la fonction de poids w_n pourrait être adaptée au long des itérations afin de modifier progressivement le comportement souhaité par l'algorithme, mais des modifications sur les densités marginales peuvent produire des résultats similaires. Ainsi, on supposera que la fonction de poids ainsi que la relation entre les candidats sera fixe tout au long de l'algorithme.

La densité conjointe Q est entièrement définie à partir des densités marginales $Q^{(k)}$, $k = 1, \ldots, K$, ainsi que par la structure de corrélation entre les candidats. Toutes les méthodes de génération de l'ensemble de candidats présentées à la section 4.1.1 sont telles que les densités marginales peuvent

différer entre les candidats. Ainsi, l'adaptation des densités marginales peut produire des densités qui diffèrent selon k sans altérer la définition de l'algorithme. On considère ici le cas où toutes les densités marginales proviennent d'une même famille paramétrique de sorte que l'adaptation sera effectuée sur les paramètres des densités. Il sera alors possible d'écrire Q_θ pour représenter la densité conjointe avec $\theta = (\theta^{(1)}, \ldots, \theta^{(K)})$ comme ensemble de paramètres ; les densités marginales seront alors dénotées par $Q_\theta^{(k)}$, $k = 1, \ldots, K$. La suite des densités conjointes adaptées $\{Q_n\}_{n \geqslant 0}$ correspond alors exactement à la suite des paramètres adaptés $\{\theta_n\}_{n \geqslant 0}$.

Au cours de l'exposition, on considérera des densités marginales gaussiennes à moyenne nulle pour générer les pas d'une marche aléatoire. En particulier, le paramètre $\theta^{(k)}$, déterminant entièrement la k-ième densité marginale, correspond à la covariance de la densité gaussienne. On dénote alors $\theta^{(k)} = \Sigma^{(k)} \in \mathcal{C}_d^+$, où $\mathcal{C}_d^+$ dénote le cône des matrices $d \times d$ symétriques définies positives. L'adaptation des densités marginales est donc définie par la méthode selon laquelle la covariance $\Sigma_{n+1}^{(k)}$ est calculée à partir du passé de la chaîne.

Notons que le choix de densités gaussiennes peut être remplacé par d'autres familles paramétriques. Par exemple, Andrieu et Thoms (2008, section 5.2.1) considèrent la famille des densités t de Student qui est sur-dispersée par rapport à la famille gaussienne et qui peut donc s'avérer plus efficace lorsque la densité cible a des ailes particulièrement lourdes. Les différentes méthodes présentées ici peuvent être facilement modifiées afin d'utiliser cette famille, mais les détails seront omis.

5.3.1 Sur la fréquence d'adaptation

À chaque itération MCMC, il est possible de procéder à l'adaptation de chacune des covariances $\Sigma^{(k)}$, $k = 1, \ldots, K$. Cependant, ceci peut s'avérer coûteux computationnellement et des stratégies plus parcimonieuses doivent être considérées.

La première méthode qui sera étudiée est l'adaptation de la covariance de la densité à partir de laquelle le candidat sélectionné a été produit. Le candidat choisi aura généralement un poids de sélection plus élevé que les autres, ce qui indique une meilleure qualité du candidat selon la fonction de poids w. Ainsi, si le k-ième candidat a été sélectionné, alors seulement $\Sigma^{(k)}$ sera adapté dans cette itération. Un poids de sélection plus élevé est souvent synonyme d'un meilleur ajustement local de la densité marginale à la densité cible de sorte que l'adaptation améliorera cet ajustement local. À la longue, cette même densité sera de mieux en mieux ajustée à π dans une certaine région et les candidats produits par la k-ième densité seront choisis plus souvent dans cette région. Les différentes densités marginales devraient donc implicitement définir des régions où elles sont mieux ajustées localement à la densité cible. L'avantage de cette construction est que la définition des régions s'effectue automatiquement via les poids de sélection. L'algorithme 5.5 résume la procédure.

Une seconde méthode considérée est construite de sorte à maintenir une densité marginale globale qui facilitera le passage entre les régions définies implicitement par les poids de sélection. En plus d'adapter la covariance de la densité choisie, on adapte également la covariance globale. Celle-ci tendra à la longue à s'ajuster vers la covariance de la densité cible, ce qui permettra le saut entre les régions. On pose par exemple que la première densité est celle adaptée à toutes les itérations ; les $K - 1$ autres densités sont adaptées seulement lorsque choisies. Pour s'assurer que la densité globale englobe bien l'ensemble du support de π, il sera de mise d'initialiser sa covariance à une grande valeur. L'algorithme 5.6 détaille cette méthode.

Algorithme 5.5 Adaptation de la densité sélectionnée uniquement

Données Paramètres des densités marginales $(\theta_n^{(1)}, \ldots, \theta_n^{(K)})$, indice du candidat sélectionné $k_n \in \{1, \ldots, K\}$, état actuel de la chaîne x_n, fonction de mise à jour $\theta(\theta^{(k)},x)$.

Procédure Mettre à jour le k_n-ième paramètre

$$\theta_{n+1}^{(k_n)} = \theta(\theta_n^{(k_n)}, x_n),$$

et poser $\theta_{n+1}^{(k)} = \theta_n^{(k)}$ pour $k \neq k_n$.

Sortie L'ensemble des paramètres $\theta_{n+1} = (\theta_{n+1}^{(1)}, \ldots, \theta_{n+1}^{(K)})$.

Algorithme 5.6 Adaptation de la densité sélectionnée et de la densité globale

Données Paramètres des densités marginales $(\theta_n^{(1)}, \ldots, \theta_n^{(K)})$, indice du candidat sélectionné $k_n \in \{1, \ldots, K\}$, état actuel de la chaîne x_n, fonction de mise à jour $\theta(\theta^{(k)},x)$ et $\theta_1(\theta^{(1)},x)$.

Procédure

1. Si $k_n \neq 1$, mettre à jour le k_n-ième paramètre

$$\theta_{n+1}^{(k_n)} = \theta(\theta_n^{(k_n)}, x_n),$$

et poser $\theta_{n+1}^{(k)} = \theta_n^{(k)}$ pour $k \neq 1,k_n$;

2. Mettre à jour le premier paramètre

$$\theta_{n+1}^{(1)} = \theta_1(\theta_n^{(1)}, x_n).$$

Sortie L'ensemble des paramètres $\theta_{n+1} = (\theta_{n+1}^{(1)}, \ldots, \theta_{n+1}^{(K)})$.

Dans les deux méthodes précédentes, il est possible d'observer une faille potentielle. En effet, si une des densités n'est pratiquement jamais choisie, alors elle ne sera jamais adaptée. À la longue, cette densité demeurera la même de sorte qu'elle restera rarement choisie. Afin d'éviter ce problème, un paramètre d'échelle $\lambda^{(k)} \in (0,\infty)$ de la covariance est considéré et adapté. La covariance de la densité générant le candidat est alors donnée par $\lambda^{(k)}\Sigma^{(k)}$. Au fil des itérations, les poids de sélection de chacun des candidats sont conservés. Ensuite, ces poids sont utilisés pour modifier le paramètre d'échelle de sorte à ce que toutes les densités soient choisies relativement souvent. Dans le contexte de marche aléatoire, ceci peut être réalisé en diminuant le paramètre d'échelle. Similairement, si une densité est choisie trop souvent, alors son facteur d'échelle est augmenté. L'algorithme 5.7 détaille une telle adaptation.

Algorithme 5.7 Adaptation des échelles par le taux de sélection

Données Paramètres d'échelle des densités marginales $(\lambda_n^{(1)}, \ldots, \lambda_n^{(K)})$, taux de sélection $s_n^{(1)}, \ldots, s_n^{(K)}$, fonction de mise à jour $\lambda(\lambda^{(k)}, s^{(k)})$.

Procédure Pour $k = 1, \ldots, K$, mettre à jour

$$\lambda_{n+1}^{(k)} = \lambda(\lambda_n^{(k)},s_n^{(k)}).$$

Sortie L'ensemble des paramètres d'échelle $\lambda_{n+1} = (\lambda_{n+1}^{(1)}, \ldots, \lambda_{n+1}^{(K)})$.

5.3.2 Sur les fonctions de mise à jour

Chacun des algorithmes 5.5, 5.6 et 5.7 est décrit d'une manière générale. La fonction de mise à jour des différents paramètres doit être spécifiée : on considère ici quelques manières de définir ces fonctions.

5.3.2.1 Mise à jour de la covariance

Pour ce qui est de la mise à jour des covariances $\Sigma^{(k)}$, $k = 1, \dots, K$, on retrouve principalement trois méthodes couramment utilisées. D'abord, il est possible d'utiliser une généralisation de la mise à jour utilisée par Haario et collab. (2001) dans leur algorithme AM 3.1. La covariance de la k-ième densité est donnée par $(2{,}38)^2\Sigma_n/d$, où la covariance est calculée par les récursions suivantes :

$$\mu_{n+1}^{(k)} = \mu_n^{(k)} + \gamma_{n+1}\left[x_{n+1} - \mu_n^{(k)}\right], \tag{5.2}$$

$$\Sigma_{n+1}^{(k)} = \Sigma_n^{(k)} + \gamma_{n+1}\left[(x_{n+1} - \mu_n^{(k)})(x_{n+1} - \mu_n^{(k)})^\top - \Sigma_n^{(k)}\right], \tag{5.3}$$

où $(\gamma_n)_{n\geqslant 1}$ est une suite de pas d'adaptation contrôlant le rythme de l'adaptation. On appelera les récursions (5.2) et (5.3) la **mise à jour AM**.

Ensuite, il est possible d'introduire un paramètre d'échelle $\lambda^{(k)}$ lui-même adaptatif. La covariance de la k-ième densité est donnée par $\lambda^{(k)}\Sigma^{(k)}$. Les récursions AM sont basées sur les résultats d'optimalité de l'algorithme Metropolis disant que, sous certaines conditions, le choix optimal de covariance instrumentale est $(2{,}38)^2\Sigma_\pi/d$, où Σ_π est la covariance de la densité cible. Le facteur $(2{,}38)^2/d$ n'est valide que dans certains cas particuliers alors que la probabilité d'acceptation associée de 0,234 est plus robuste au choix de π. Le facteur d'échelle adaptatif permet donc de trouver la valeur produisant un certain taux d'acceptation cible α_*. Ceci correspond à la **mise à jour ASWAM** (de l'algorithme ASWAM 3.10) donnée par les récursions suivantes

$$\mu_{n+1}^{(k)} = \mu_n^{(k)} + \gamma_{n+1}\left[x_{n+1} - \mu_n^{(k)}\right], \tag{5.4}$$

$$\Sigma_{n+1}^{(k)} = \Sigma_n^{(k)} + \gamma_{n+1}\left[(x_{n+1} - \mu_n^{(k)})(x_{n+1} - \mu_n^{(k)})^\top - \Sigma_n^{(k)}\right], \tag{5.5}$$

$$\log\left(\lambda_{n+1}^{(k)}\right) = \log\left(\lambda_n^{(k)}\right) + \gamma_{n+1}\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right], \tag{5.6}$$

où $\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)})$ est le taux d'acceptation MTM (4.2). Enfin, la **mise à jour RAM** (basée sur l'algorithme RAM 3.11) incorpore la probabilité d'acceptation à même la mise à jour de la covariance sans utiliser de facteur d'échelle. En écrivant $\Sigma_n^{(k)} = S_n^{(k)} S_n^{(k)\top}$, on peut mettre à jour la covariance en trouvant $S_{n+1}^{(k)}$ tel que

$$S_{n+1}^{(k)} S_{n+1}^{(k)\top} = S_n^{(k)}\left(I_d + \gamma_{n+1}\left[\alpha(y|x_n) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right) S_n^{(k)\top}, \tag{5.7}$$

où $u_{n+1} = (S_n^{(k)})^{-1}(y - x_n)$. La solution à la récursion (5.7) peut être trouvée efficacement à l'aide d'une mise à jour de Cholesky de rang un (Vihola, 2012).

Un élément important différenciant les deux premières méthodes (les mises à jour AM et ASWAM) de la troisième (la mise à jour RAM) est que le vecteur utilisé pour le calcul de la covariance fait

intervenir la moyenne $\mu_n^{(k)}$ plutôt que l'état actuel x_n. Bien que l'état actuel x_n soit tout de même utilisé dans la génération des candidats, l'adaptation se comportera différemment dû à cette particularité. En effet, en utilisant une moyenne $\mu_n^{(k)}$ lors de l'adaptation, les densités se trouvent donc attachées à une région de l'espace ; si cette densité est utilisée ailleurs dans le support de π, alors le vecteur $(x_{n+1} - \mu_n^{(k)})$ pourrait être grand par rapport au pas utilisé $(x_{n+1} - x_n)$ et la quantité d'adaptation serait alors grande. À la longue, les différentes densités marginales risquent de toutes converger vers la même densité globale avec Σ_π comme covariance. Pour éviter ce potentiel problème, on considère aussi la **mise à jour AM locale** où l'on remplace $\mu_n^{(k)}$ par x_n dans la récursion (5.3) ainsi que la **mise à jour ASWAM locale** où l'on effectue la même substitution dans la récursion (5.5). Notons que le maintien des moyennes $\mu_n^{(k)}$, $k = 1, \ldots, K$, n'est plus requis pour ce type de mise à jour.

5.3.2.2 Mise à jour du paramètre d'échelle

Lorsque le taux de sélection est utilisé pour ajuster le paramètre d'échelle (algorithme 5.7), la fonction qui calcule le nouveau facteur d'échelle doit être définie. Il est plus difficile de définir cette fonction d'une manière générale puisque le choix de la fonction de poids $w^{(k)}$ détermine le taux de sélection. Des choix de fonctions différents mèneront alors à des choix différents de mise à jour du facteur d'échelle.

Pour des poids proportionnels à la densité cible, $w^{(k)}(y|x) = \pi(y)$, des poids faibles signifient que les candidats générés par cette densité se trouvent souvent dans des régions de faible densité de π et ne sont donc pas souvent choisis. Puisque π est généralement supposée continue, de faibles valeurs de π correspondent généralement à des régions plus éloignées. La solution est donc de réduire le facteur d'échelle de la densité en question : de faibles poids sont associés à une valeur de $\pi(y)$ faible et donc un y possiblement trop éloigné. Une possibilité de récursion est donc

$$\log\left(\lambda_{n+1}^{(k)}\right) = \log\left(\lambda_n^{(k)}\right) + \gamma_{n+1}\left[s_n^{(k)} - s_*\right], \tag{5.8}$$

où s_* est un taux de sélection cible (e.g. $1/K$).

Pour des poids proportionnels à la densité instrumentale marginale $w^{(k)}(y|x) = \pi(y)Q_\theta^{(k)}(y|x)$, un poids de sélection faible signifie que les candidats générés sont soit dans des régions de faible densité ou que les candidats dans les régions de haute densités sont éloignés de l'état actuel (petite valeur de densité instrumentale.) Ainsi, il n'est pas clair qu'il faille réduire ou augmenter le facteur d'échelle.

Pour des poids inversement proportionnels à la densité marginale, c.-à-d., les poids d'importance $w^{(k)}(y|x) = \pi(y)/Q_\theta^{(k)}(x|y)$, un poids de sélection faible signifie que les candidats générés sont soit dans des régions de faible densité ou que les candidats dans les régions de haute densité sont près de l'état actuel (grande valeur de densité instrumentale.) Dans les deux cas, une réduction du facteur d'échelle est de mise et une récursion telle que (5.8) peut être utilisée.

5.3.3 Sur les poids de sélection

À la section 4.1.2, il a été question de différentes options pour définir les fonctions de poids de sélection $w^{(k)}$, $k = 1, \ldots, K$, dans les algorithmes à essais multiples. Bien que la seule condition requise sur ces poids soit la positivité (proposition 4.2), certains choix sont plus intéressants que

Type de candidats	Candidats (étape (a))	Points de référence (étape (d))
Indépendants	Échantillonner $Y^{(j)} \sim Q_\theta^{(j)}(\cdot\|x_n)$ indépendamment sur $j = 1, \dots, K$.	Échantillonner $Y^{(j)} \sim Q_\theta^{(j)}(\cdot\|y)$ indépendamment sur $j \neq k$.
Extrêmement antithétiques	Décomposer préalablement $\Psi_K = (X\Lambda^{1/2})(X\Lambda^{1/2})^\top$; Pour $k = 1, \dots, K$: – Échantillonner $Z^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mathbf{0}_d, I_d)$, – Calculer $u^{(k)} = X\Lambda^{1/2} z^{(k)}$, – Calculer $y^{(k)} = x_n + S^{(k)} u^{(k)}$.	Décomposer préalablement $\Phi_{K-1} = (X'\Lambda'^{1/2})(X'\Lambda'^{1/2})^\top$; Pour $j \neq k$: – Échantillonner $Z^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mathbf{0}_d, I_d)$, – Calculer $u_*^{(j)} = X'\Lambda'^{1/2} z^{(j)}$, – Calculer $x_*^{(j)} = y + S^{(j)} \left(u_*^{(j)} - \rho u^{(k)}\right)$.
Quasi-Monte Carlo aléatoires Règle de Korobov : $a \in \{1, \dots, K-1\}$.	Échantillonner $U \sim \text{uniforme}[0,1)^d$; Pour $k = 1, \dots, K$: – Calculer $u^{(k)} \equiv_1 \frac{k-1}{K}\left(1, a, \dots, a^{d-1}\right) + u$, – Calculer $z^{(k)} = F^{-1}(u^{(k)})$, – Calculer $y^{(k)} = x_n + S^{(k)} z^{(k)}$.	Calculer $u_* = F\left((S^{(k)})^{-1}(y - x_n)\right)$; Pour $j \neq k$: – Calculer $u_*^{(j)} \equiv_1 \frac{j-1}{K}\left(1, a, \dots, a^{d-1}\right) + u_*$, – Calculer $z_*^{(j)} = F^{-1}(u_*^{(j)})$, – Calculer $x_*^{(j)} = y + S^{(j)} z_*^{(j)}$.
Variable aléatoire commune	Échantillonner $Z \sim \mathcal{N}_d(\mathbf{0}_d, I_d)$; Pour $k = 1, \dots, K$: – Calculer $y^{(k)} = x_n + S^{(k)} z$.	Calculer $z_* = (S^{(k)})^{-1}(y - x_n)$; Pour $j \neq k$: – Calculer $x_*^{(j)} = y + S^{(j)} z_*$.

Tableau 5.1 *Résumé des différents types de candidats qui sont utilisés à l'étape d'échantillonnage MTM dans l'algorithme aMTM 5.4. La génération des candidats ainsi que des points de référence est décrite, où x_n est l'état actuel de la chaîne, k est l'index du candidat choisi, F est la fonction de répartition normale centrée réduite, $y = y^{(k)}$ et $\Sigma^{(k)} = S^{(k)} S^{(k)\top}$; le reste de l'échantillonnage MTM est le même pour tout type de candidat et se trouve à l'algorithme MTM 4.1.*

d'autres. À travers diverses expériences de simulations discutées à la section 4.1.2, les poids offrant la meilleure efficacité empirique sont les poids proportionnels à la densité cible $w^{(k)}(y|x) = \pi(y)$ et les poids d'importance $w^{(k)}(y|x) = \pi(y)/Q_\theta^{(k)}(y|x)$. Notons que les densités marginales gaussiennes sont symétriques et que la version non-symétrique des poids d'importance se réduit aux poids d'importance réguliers. L'algorithme aMTM 5.4 présenté ici n'utilisera que ces deux déclinaisons puisque l'évidence expérimentale suggère fortement que ces choix sont les plus pertinents.

5.3.4 Sur la structure de corrélation

La section 4.1.1 contient une exposition des principales manières de définir la relation entre les candidats. Parmi ces méthodes, on en identifie quelques-unes qui sont pertinentes à notre algorithme ; celles-ci seront les seules considérées. Pour chacune des méthodes, il est nécessaire de définir la manière selon laquelle les points de référence sont produits pour notre choix de densités marginales gaussiennes. Une description de la procédure requise pour produire ces candidats est comprise dans le tableau 5.1.

5.3.4.1 Candidats indépendants

Les candidats indépendants, comme leur nom l'indique, sont distribués d'une manière indépendante et selon leur distribution marginale :

$$Y^{(k)} \overset{\text{ind.}}{\sim} Q_\theta^{(k)}(\cdot|x_n), \qquad k = 1, \dots, K.$$

La densité conjointe est alors donnée par le produit des densités marginales ; la densité conditionnelle utilisée pour produire les points de référence est alors donnée par

$$Q_\theta^{(-k)}(x_*^{(-k)}|y,x_n) = \frac{Q_\theta(x_*^{(1:K)}|y)}{Q_\theta^{(k)}(x_n|y)} = \prod_{j \neq k} Q_\theta^{(k)}(x_*^{(j)}|y) \tag{5.9}$$

où $x_*^{(k)} = x_n$ et $y = y^{(k)}$ lorsque le k-ième candidat est sélectionné. Ainsi, les points de référence sont tout simplement générés indépendamment selon les densités marginales $Q_\theta^{(j)}(\cdot|y)$, $j \neq k$.

5.3.4.2 Candidats extrêmement antithétiques

Les candidats extrêmement antithétiques sont générés de la manière suivante. On produit d'abord des vecteurs aléatoires de distribution marginale normale centrée réduite, mais avec corrélation extrêmement antithétique :

$$(U^{(1)}, \dots, U^{(K)})^\top \sim \mathcal{N}_{dK}(\mathbf{0}, \Psi_K),$$

où d est la dimension de l'espace d'états $\mathcal{X} \subseteq \mathbb{R}^d$ et où

$$\Psi_K = \begin{pmatrix} I_d & \cdots & \rho I_d \\ \vdots & \ddots & \vdots \\ \rho I_d & \cdots & I_d \end{pmatrix},$$

avec $\rho = -1/(K-1)$. Ensuite, les candidats sont calculés selon

$$y^{(k)} = x_n + S_n^{(k)} u^{(k)}, \qquad k = 1, \dots, K,$$

où $\Sigma_n^{(k)} = S_n^{(k)} S_n^{(k)\top}$. Ceci définit donc la densité conjointe Q_θ qui correspond à

$$(Y^{(1)}, \dots, Y^{(K)})^\top \sim \mathcal{N}_{dK}(\mathbf{1}_K \otimes x_n, \Sigma_n),$$

où

$$\begin{aligned} \Sigma_n &= \operatorname{diag}\left(S^{(1)}, \dots, S^{(K)}\right) \Psi_K \operatorname{diag}\left(S^{(1)\top}, \dots, S^{(K)\top}\right) \\ &= \begin{pmatrix} \Sigma_n^{(1)} & \cdots & \rho S_n^{(1)} S_n^{(K)\top} \\ \vdots & \ddots & \vdots \\ \rho S_n^{(K)} S_n^{(1)\top} & \cdots & \Sigma_n^{(K)} \end{pmatrix}. \end{aligned}$$

Afin de générer les points de référence, on doit trouver la densité conditionnelle $Q_\theta^{(-k)}$ de $x_*^{(-k)}$ sachant $x_*^{(k)} = x_n$. Par les propriétés de la distribution normale, on sait que cette distribution est également normale ; la moyenne et la covariance peuvent aussi être calculées. Pour simplifier les calculs, on

supposera $k = K$, mais des expressions similaires peuvent être trouvées pour tout k. On écrit

$$\mathbf{1}_K \otimes y = \begin{pmatrix} \mathbf{1}_{K-1} \otimes y \\ y \end{pmatrix} =: \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix};$$

$$\Sigma = \left(\begin{array}{ccc|c} \Sigma^{(1)} & \cdots & \rho S^{(1)} S^{(K-1)\top} & \rho S^{(1)} S^{(K)\top} \\ \vdots & \ddots & \vdots & \vdots \\ \rho S^{(K-1)} S^{(1)\top} & \cdots & \Sigma^{(K-1)} & \rho S^{(K-1)} S^{(K)\top} \\ \hline \rho S^{(K)} S^{(1)\top} & \cdots & \rho S^{(K)} S^{(K-1)\top} & \Sigma^{(K)} \end{array}\right) =: \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Puis, la moyenne est donnée par

$$\begin{aligned} \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_n - y) &= \mathbf{1}_{K-1} \otimes y + \begin{pmatrix} \rho S^{(1)} S^{(K)\top} \\ \vdots \\ \rho S^{(K-1)} S^{(K)\top} \end{pmatrix} (S^{(K)\top})^{-1} (S^{(K)})^{-1} (x_n - y) \\ &= \mathbf{1}_{K-1} \otimes y + \rho \begin{pmatrix} S^{(1)} \\ \vdots \\ S^{(K-1)} \end{pmatrix} (S^{(K)})^{-1} (x_n - y) \\ &= \mathbf{1}_{K-1} \otimes y - \rho \begin{pmatrix} S^{(1)} \\ \vdots \\ S^{(K-1)} \end{pmatrix} u^{(K)} \end{aligned}$$

en notant que $y = y^{(K)} = x_n + S^{(K)} u^{(K)}$ implique $(S^{(K)})^{-1}(x_n - y) = -u^{(K)}$. La covariance est quant à elle donnée par

$$\begin{aligned} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} &= \Sigma_{11} - \begin{pmatrix} \rho S^{(1)} S^{(K)\top} \\ \vdots \\ \rho S^{(K-1)} S^{(K)\top} \end{pmatrix} (S^{(K)\top})^{-1} (S^{(K)})^{-1} \begin{pmatrix} \rho S^{(K)} S^{(1)\top} & \cdots & \rho S^{(K)} S^{(K-1)\top} \end{pmatrix} \\ &= \Sigma_{11} - \rho^2 \begin{pmatrix} S^{(1)} \\ \vdots \\ S^{(K-1)} \end{pmatrix} \begin{pmatrix} S^{(1)\top} & \cdots & S^{(K-1)\top} \end{pmatrix} \\ &= \begin{pmatrix} (1-\rho^2)\Sigma^{(1)} & \cdots & (\rho - \rho^2) S^{(1)} S^{(K-1)\top} \\ \vdots & \ddots & \vdots \\ (\rho - \rho^2) S^{(K-1)} S^{(1)\top} & \cdots & (1-\rho^2)\Sigma^{(K-1)} \end{pmatrix} \\ &= (1-\rho) \begin{pmatrix} (1+\rho)\Sigma^{(1)} & \cdots & \rho S^{(1)} S^{(K-1)\top} \\ \vdots & \ddots & \vdots \\ \rho S^{(K-1)} S^{(1)\top} & \cdots & (1+\rho)\Sigma^{(K-1)} \end{pmatrix} \\ &= (1-\rho) \operatorname{diag}\left(S^{(1)}, \ldots, S^{(K-1)}\right) \begin{pmatrix} (1+\rho) I_d & \cdots & \rho I_d \\ \vdots & \ddots & \vdots \\ \rho I_d & \cdots & (1+\rho) I_d \end{pmatrix} \operatorname{diag}\left(S^{(1)}, \ldots, S^{(K-1)}\right). \end{aligned}$$

Ainsi, pour produire les points de référence lorsque le K-ième candidat est sélectionné, on doit générer

$$U_*^{(-K)} \sim \mathcal{N}_{d(K-1)}(\mathbf{0}, \Phi_{K-1}),$$

où

$$\Phi_{K-1} = (1-\rho)\begin{pmatrix} (1+\rho)I_d & \cdots & \rho I_d \\ \vdots & \ddots & \vdots \\ \rho I_d & \cdots & (1+\rho)I_d \end{pmatrix} = (1-\rho)\left[\Psi_{K-1} + \rho I_{d(K-1)}\right];$$

pour ensuite calculer

$$x_*^{(-j)} = \left(y - \rho S^{(j)} u^{(K)}\right) + S^{(j)} u_*^{(j)} = y + S^{(j)}\left(u_*^{(j)} - \rho u^{(K)}\right), \qquad j \neq K.$$

Notons que les covariances Ψ_K et Φ_{K-1} sont constantes tout au long de l'algorithme MCMC de sorte que l'échantillonnage utilisant ces covariances peut être fait efficacement. Par contre, ces matrices ne sont que définies semi-positives de sorte que la décomposition en valeurs singulières est requise plutôt que la décomposition de Choleski. Pour une matrice symétrique, ceci correspond à écrire $\Psi_K = X\Lambda X^\top$ où X est orthogonale (et contient les vecteurs propres) et où $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ contient les valeurs propres. Alors, on peut écrire $\Psi_K = (X\Lambda^{1/2})(X\Lambda^{1/2})^\top$ et ainsi produire les $u^{(k)}$ (et d'une manière similaire $u_*^{(-K)}$) à partir de variables aléatoires normales centrées réduites.

5.3.4.3 Candidats quasi-Monte Carlo randomisés

Une version des candidats quasi-Monte Carlo aléatoires utilisant des densités marginales gaussiennes est complètement décrite à l'exemple 4.3. Notons que les points de référence sont déterministes dans ce cas.

5.3.4.4 Candidats par variable aléatoire commune

Des candidats générés par une variable aléatoire commune peuvent être facilement produits lorsque les densités marginales sont gaussiennes. Une variable aléatoire normale centrée réduite $Z \sim \mathcal{N}_d(\mathbf{0}_d, I_d)$ est générée, puis les candidats sont calculés par la transformation linéaire naturelle

$$y^{(k)} = x_n + S^{(k)} z, \qquad k = 1, \ldots, K.$$

Soit k l'indice du candidat choisi. Les points de référence sont donnés par la même transformation

$$x_*^{(j)} = \begin{cases} y + S^{(j)} z_*, & j \neq k; \\ x_n, & j = k; \end{cases}$$

où le pas normal centré réduit z_* est donné par la standardisation de $y - x_n$ par la covariance du candidat choisi, c.-à-d., $z_* = (S^{(k)})^{-1}(y - x_n)$. Ces points de références sont donc déterministes.

5.4 Propriétés de l'algorithme aMTM

L'algorithme aMTM 5.4 est avant tout un algorithme MCMC adaptatif. Ainsi, l'étude de ses propriétés théoriques s'inscrit dans l'exposition présentée aux sections 3.2 et 3.3 respectivement sur l'ergodicité et sur les théorèmes limites des algorithmes MCMC adaptatifs. L'algorithme aMTM est en fait similaire aux algorithmes AM 3.1, ASWAM 3.10 ou RAM 3.11 en substituant un noyau Metropolis adapté par un noyau MTM adapté. Il en résulte donc que les différentes conditions qui seront utilisées pour montrer les principales propriétés de l'algorithme aMTM afficheront une grande similitude avec celles utilisées pour les l'étude des propriétés de l'algorithme AM.

La section 5.3 décrit plusieurs variations de l'algorithme aMTM. Toutes les déclinaisons introduites ne peuvent être facilement traitées simultanément par une approche unifiée générale. La présentation sera effectuée dans le cas le plus général lorsque possible, mais certaines des propriétés devront être traitées différemment. Lorsque les déclinaisons de l'algorithme influencent grandement la démonstration, nous opterons pour la présentation suivante. L'étude principale considérera le cas où l'adaptation est effectuée suivant l'algorithme 5.5 et utilisant les mises à jour AM, où les candidats sont indépendants et où les poids de sélection sont proportionnels à la densité cible seulement. Ensuite, les différentes déclinaisons seront abordées au fil de l'exposition via des remarques sur l'application et la vérification des conditions aux autres cas.

Dans ce contexte, il nous est possible de fixer davantage de notation. La distribution cible admet une densité π par rapport à la mesure de Lebesgue ; le support de π est noté $\mathcal{X} \subseteq \mathbb{R}^d$ et correspond à

$$\mathcal{X} = \operatorname{supp}(\pi) := \overline{\left\{ x \in \mathbb{R}^d \,\middle|\, \pi(x) > 0 \right\}}.$$

où $\overline{A}$ dénote l'adhérence de l'ensemble A. La fonction dont l'espérance sous π est recherchée sera dénotée par $f : \mathbb{R}^d \to \mathbb{R}^p$ et satisfait $\pi(\|f\|_1) < \infty$. À l'itération $n \in \mathbb{N}$, les paramètres du noyau de transition MTM sont donnés par

$$\Sigma = \left(\Sigma^{(1)}, \ldots, \Sigma^{(K)} \right),$$

où $\Sigma^{(k)} \in \mathcal{C}_d^+$ est une matrice $d \times d$ symétrique et définie positive pour $k = 1, \ldots, K$. L'espace des paramètres est alors donné par

$$\Theta = \underbrace{\mathcal{C}_d^+ \times \cdots \times \mathcal{C}_d^+}_{K \text{ fois}} = \left(\mathcal{C}_d^+ \right)^K.$$

Les densités instrumentales marginales sont alors données par

$$\begin{aligned} Q_\theta^{(k)}(y|x) &= \varphi(y|x, \Sigma^{(k)}) \\ &= (2\pi)^{-d/2} \det(\Sigma^{(k)})^{-1/2} \exp\left\{ -\frac{1}{2}(y-x)^\top (\Sigma^{(k)})^{-1}(y-x) \right\}, \qquad k = 1, \ldots, K. \end{aligned}$$

L'analyse des propriétés de l'algorithme sera effectuée en supposant que le support de la densité cible $\mathcal{X}$ et que l'espace des paramètres adaptatifs Θ sont tous deux compacts. Ces suppositions, similaires à celles faites aux exemples 3.7 et 3.8, facilitent grandement l'étude qui est contenue à la section 5.4.1. La généralisation aux cas non-bornés sera discutée à la section 5.4.1.5. Enfin, les théorèmes limites seront abordés à la section 5.4.2.

5.4.1 Ergodicité de l'algorithme

Afin de vérifier l'ergodicité de l'algorithme aMTM, le théorème 3.1 (Roberts et Rosenthal, 2007, théorème 2) sera utilisé. Pour ce faire, les conditions d'adaptation diminuante 3.1 et de convergence bornée 3.2 doivent être étudiées. Notons que l'invariance des transitions MTM pour une densité cible π est démontrée par la condition d'équilibre qui est toujours satisfaite dans notre contexte.

Théorème 5.2 *Soit $\{P_\theta\}_{\theta\in\Theta}$ une famille de transitions MTM telles que les conditions de la proposition 4.2 sont satisfaites pour chaque $\theta \in \Theta$. Si l'algorithme aMTM satisfait aux conditions d'adaptation diminuante 3.1 et de convergence bornée 3.2, alors l'algorithme est ergodique (en variation totale) pour tout choix de valeur initiale.*

Démonstration. Les conditions de la proposition 4.2 suffisent à montrer la condition d'équilibre de chaque transition P_θ. Le théorème 2.7 implique alors que P_θ admet la distribution cible comme distribution invariante. Enfin l'ergodicité découle du théorème 3.1. □

Lorsque les distributions instrumentales marginales $Q_\theta^{(k)}$ sont choisies gaussiennes, alors elles sont symétriques, ce qui vérifie la première condition de la proposition 4.2. La positivité des poids est une condition qui peut être facilement assurée par construction dans le choix de la fonction de poids.

5.4.1.1 Convergence bornée

Afin de vérifier la convergence bornée (condition 3.2), l'approche considérée utilisera la proposition 3.17. qui demande la vérification de plusieurs conditions.

D'abord, le processus $\{X_n\}_{n\geqslant 1}$ doit être borné en probabilité. Ceci peut être vérifié simplement en supposant $\mathcal{X}$ compact. Cependant, il se peut que l'espace d'états ne soit pas borné : dans ce cas, l'algorithme doit être légèrement altéré et certaines suppositions supplémentaires doivent être faites. En effet, on doit considérer un sous-ensemble borné $K \subseteq \mathcal{X}$ dans lequel l'adaptation se produit. Lorsque $x_n \notin K$ aucune adaptation des paramètres du noyau de transition n'est effectuée et la chaîne se déplace à l'aide d'un noyau de transition fixe P. Cette dernière doit donc être choisie d'avance et doit respecter certaines conditions faciles à vérifier pour des choix communs de transition. Notons cependant que l'ensemble K peut également être choisi arbitrairement grand de façon à que le processus ne sorte jamais de K en pratique et donc que le noyau fixe P ne soit jamais utilisé. Cette construction n'est donc principalement utile que pour des fins théoriques. Notons cependant que cette modification n'est requise que si $\mathcal{X}$ n'est pas borné. On supposera donc plus généralement que $\{X_n\}_{n\geqslant 1}$ est borné en probabilité et on discutera à la section 5.4.1.5 de généralisations au cas $\mathcal{X}$ non-borné.

Ensuite, l'espace des paramètres Θ doit être supposé compact. Puisque $\mathcal{C}_d^+$ est l'espace des matrices $d \times d$ symétriques et définies positives, on trouve que Θ n'est donc pas compact puisque non-borné pour les grandes covariances et ouvert vers les petites covariances. Cependant, il est possible de restreindre l'algorithme à un sous-ensemble compact $\widetilde{\Theta} \subset \Theta$. Ce sous-ensemble peut être choisi arbitrairement grand de sorte que cette restriction ne pose pas de problème en pratique.

De plus, la famille de noyaux de transitions $\{P_\theta\}_{\theta\in\Theta}$ doit être telle que chaque P_θ admette π comme distribution invariante, c'est-à-dire Harris-ergodique avec des sauts bornés par $D < \infty$ en probabilité (3.23). Puisque chaque P_θ est un noyau MTM, toutes ces conditions peuvent être aisément vérifiées en imposant seulement que les densités instrumentales soient nulles au-delà d'une distance D et en imposant certaines conditions sur la densité cible π. En particulier, si $\mathcal{X}$ est compact, alors on peut choisir $D = \operatorname{diam}(\mathcal{X})$.

Finalement, la proposition 3.17 requiert la continuité de l'application $(x,\theta) \mapsto \|P_\theta^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}}$. Cette condition est la seule qui requiert une attention particulière : on rapporte ici certains résultats techniques nécessaires à la démonstration de la convergence bornée. La section 5.5.1.1 contient la démonstration de ces résultats ainsi qu'une discussion des conditions s'y rattachant.

Lemme 5.3 *Pour un noyau MTM à paramètre θ, la probabilité d'acceptation intégrée à partir de x via le k-ième candidat $\overline{A}_\theta^{(k)}(x)$ est une fonction continue de (x,θ) à condition que $A_\theta^{(k)}(y|x)$ soit une fonction continue de (x,y,θ) et que $Q_\theta^{(k)}$ soit une densité par rapport à la mesure de Lebesgue telle qu'il existe $Q^+ : \mathcal{X} \to \mathbb{R}_{\geqslant 0}$ intégrable avec $Q_\theta^{(k)}(y|x) \leqslant Q^+(y)$ uniformément sur (x,θ). De plus, sous ces mêmes conditions, la probabilité de rejet $R_\theta(x)$ est également une fonction continue de (x,θ).*

Démonstration. Voir section 5.5.1.1. □

Lemme 5.4 *Soit un noyau de transition de la forme Metropolis-Hastings,*

$$P(\,\mathrm{d}y|x) = R(x)\delta_x(\,\mathrm{d}y) + p(y|x)\lambda(\,\mathrm{d}y).$$

Alors, le noyau de transition itéré de n pas peut être écrit sous la forme

$$P^n(\,\mathrm{d}y|x) = R^n(x)\delta_x(\,\mathrm{d}y) + p^n(y|x)\lambda(\,\mathrm{d}y), \tag{5.10}$$

où

$$p^n(y|x) = \int p^{n-1}(y|z)p(z|x)\lambda(\,\mathrm{d}z), \qquad p^0(y|z) = \delta_z(y).$$

Démonstration. Voir section 5.5.1.1. □

Lemme 5.5 *Sous les conditions du lemme 5.3 et en supposant que Q^+ est intégrable, $p_\theta^n(y|x)$ est une fonction continue de (x,y,θ) pour tout $n \geqslant 1$.*

Démonstration. Voir section 5.5.1.1. □

À l'aide de ces résultats, il est désormais possible de prouver la convergence bornée de l'algorithme aMTM.

Théorème 5.6 *Soit un algorithme aMTM pour une densité cible π uniformément bornée par le haut et par le bas sur son support $\mathcal{X}$. Supposons que l'espace des paramètres Θ est compact et supposons que $\{X_n\}_{n\geqslant 1}$ est borné en probabilité. Supposons de plus que l'algorithme satisfait les différentes conditions des lemmes 5.3 et 5.5. Alors, l'algorithme satisfait la convergence bornée (condition 3.2).*

Démonstration. Les conditions de la proposition 3.17 sont toutes vérifiées à l'exception de la continuité de $(x,\theta) \mapsto \Delta(x,\theta,n)$. En effet, les noyaux MTM admettent tous π comme distribution invariante étant donné la condition d'équilibre (proposition 4.2) et sont tous Harris-ergodiques à π. L'ergodicité des algorithmes MTM est montrée à la proposition 4.4 où les densités marginales gaussiennes satisfont aux conditions requises en supposant la compacité de Θ et l'Harris-ergodicité est ensuite assurée par la proposition 4.8.

La vérification de la continuité de $\Delta(x,\theta,n)$ s'inspire de la preuve du corollaire 11 de Roberts et Rosenthal (2007) montrant la même continuité dans le cas d'une transition Metropolis-Hastings. Le noyau itéré P_θ^n peut être écrit sous la forme (5.10),

$$P_\theta^n(\,\mathrm{d}y|x) = R_\theta^n(x)\delta_x(\,\mathrm{d}y) + p_\theta^n(y|x)\lambda(\,\mathrm{d}y).$$

On développe alors $\Delta(x,\theta,n)$ afin de l'exprimer en fonction d'expressions que l'on sait continues par

rapport à (x,θ).

$$\begin{aligned}
\Delta(x,\theta,n) &= \|P_\theta^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}} \\
&= \sup_{B\in\mathcal{B}(\mathcal{X})} |P_\theta^n(B|x) - \pi(B)| \\
&= \sup_{B\in\mathcal{B}(\mathcal{X})} \left| \int_B P_\theta^n(\,\mathrm{d}y|x) - \int_B \pi(\,\mathrm{d}y) \right| \\
&= \sup_{B\in\mathcal{B}(\mathcal{X})} \left| R_\theta^n(x)\delta_x(B) + \int_B p_\theta^n(\,\mathrm{d}y|x) - \int_B \pi(\,\mathrm{d}y) \right| \\
&= R_\theta^n(x) + \frac{1}{2}\int_{\mathcal{X}} |p_\theta^n(y|x) - \pi(y)|\lambda(\,\mathrm{d}y), \qquad (5.11)
\end{aligned}$$

où la dernière égalité provient du choix $B = \{y : p_\theta^n(y|x) > \pi(y)\}$ ou $B^c = \{y : p_\theta^n(y|x) \leqslant \pi(y)\}$. En inspectant l'expression (5.11), on peut finalement montrer que Δ est une fonction continue. En effet, $R_\theta^n(x)$ est une fonction continue de (x,θ) étant donné que $R_\theta(x)$ est aussi continue par rapport à (x,θ) par le lemme 5.3. Par le lemme 5.5, $p_\theta^n(y|x)$ est continu par rapport à (x,y,θ). Puisque $\pi(y)$ est également continu par rapport à (x,y,θ), on a que $|p_\theta^n(y|x) - \pi(y)|$ est aussi continue par rapport à (x,y,θ). Finalement, on peut montrer que l'intégrale de cette expression est également continue par rapport à (x,θ) en utilisant le corollaire 5.10. Il suffit de borner $|p_\theta^n(y|x) - \pi(y)|$ uniformément en (x,θ) par une fonction intégrable. Par l'inégalité du triangle, on a

$$|p_\theta^n(y|x) - \pi(y)| \leqslant p_\theta^n(y|x) + \pi(y)$$

Par la preuve du lemme 5.5, on sait que $p_\theta^n(y|x)$ est uniformément borné par $K^n \cdot (Q^+(y))^n$. Enfin,

$$|p_\theta^n(y|x) - \pi(y)| \leqslant K^n \cdot (Q^+(y))^n + \pi(y),$$

qui est intégrable par les conditions du lemme 5.5 et par le fait que π est une densité. Les conditions du corollaire 5.10 sont alors toutes vérifiées, ce qui montre la continuité de $\int_{\mathcal{X}} |p_\theta^n(y|x) - \pi(y)|\lambda(\,\mathrm{d}y)$ par rapport à (x,θ). Puisque $\Delta(x,\theta,n)$ est une somme de fonctions continues, on conclut que $(x,\theta) \mapsto \Delta(x,\theta,n)$ est continu pour tout $n \geqslant 1$. □

Le théorème 5.6 comporte plusieurs conditions sur les différents éléments de l'algorithme aMTM. On s'attarde alors à comprendre ce que signifient ces conditions ainsi qu'à trouver les situations où ces conditions sont satisfaites.

En supposant que le support $\mathcal{X}$ de π est compact, plusieurs conditions sont directement vérifiées. En effet, si π est une densité continue et positive sur $\mathcal{X}$ compact, alors π est uniformément borné par le haut et par le bas sur $\mathcal{X}$. De plus, $\{X_n\}_{n\geqslant 1}$ est donc borné en probabilité puisque borné par le support $\mathcal{X}$. Également, les différentes fonctions servant de bornes uniformes dans les lemmes 5.3 et 5.5 pourront être facilement choisies intégrables lorsque $\mathcal{X}$ est compact. Le cas où $\mathcal{X}$ est non-borné est considéré à la section 5.5.1.3.

Les différentes structures de corrélations considérées à la section 5.3.4 affectent la vérification des conditions des résultats préalables au théorème 5.6. La section 5.5.1.2 contient une discussion à ce sujet. De manière générale, toutes les structures de corrélation vont requérir une certaine condition de continuité de la fonction de poids et de la fonction de probabilité d'acceptation.

Notons finalement que le théorème 5.6 ne concerne que la famille de transitions $\{P_\theta\}_{\theta\in\Theta}$ et ne dépend donc pas de la méthode de mise à jour des paramètres ni de la fréquence d'adaptation. Ainsi, toutes les déclinaisons de l'algorithme présentées à la section 5.3 peuvent être couvertes par ce résultat en supposant certaines des conditions mentionnées dans la présente discussion. Cependant, ce résultat requiert minimalement la compacité de l'espace des paramètres Θ. Pour assurer cette propriété, il est possible d'encapsuler l'une ou l'autre des fonctions de mise à jour dans une fonction de reprojection sur $\widetilde{\Theta}$ compact. Par exemple, si la fonction de mise à jour produisait un paramètre $\theta \notin \widetilde{\Theta}$, alors le nouveau paramètre serait plutôt choisi comme le point de $\widetilde{\Theta}$ le plus près de θ par rapport à la distance

euclidienne. Notons de plus que l'espace qui doit être compact est l'espace des K covariances : lorsqu'un paramètre d'échelle $\lambda^{(k)}$ est utilisé (e.g. mise à jour ASWAM) alors on doit s'assurer que le produit $\lambda^{(k)}\Sigma^{(k)}$ fait bien partie de l'espace compact. Pour ce faire, on peut imposer que $\Sigma^{(k)}$ soit dans un espace compact et que $\lambda^{(k)}$ soit dans un intervalle compact.

5.4.1.2 Adaptation diminuante

La vérification de la condition d'adaptation diminuante 3.1 s'effectue mieux dans le contexte des algorithmes MCMC par approximation stochastique. En effet, la proposition 3.30 permet de vérifier cette propriété en supposant certaines conditions sur l'approximation stochastique.

En inspectant la preuve de la proposition 3.30, on s'aperçoit que la condition 3.3 n'est utilisée que pour montrer que la fonction test V est telle que $\{V(X_n)\}_{n\geqslant 1}$ est borné en probabilité. Étant donné qu'on suppose $\{X_n\}_{n\geqslant 1}$ borné en probabilité pour montrer la convergence bornée dans le théorème 5.6, il nous est donc possible de choisir $V \equiv 1$ pour vérifier cette condition.

Théorème 5.7 *Soit π une densité cible à support $\mathcal{X} \subset \mathbb{R}^d$ compact. Soit $\{P_\theta\}_{\theta\in\Theta}$ une famille de transitions MTM à densités instrumentales marginales gaussiennes, où Θ est compact et qui satisfait la condition 3.24. Un algorithme aMTM telle que la fonction de mise à jour H_θ satisfait à la condition*

$$\sup_{\theta\in\mathcal{K}} \|H_\theta\|_V < \infty \tag{5.12}$$

avec $V \equiv 1$, vérifie alors la condition d'adaptation diminuante 3.1.

Démonstration. On vérifie les conditions de la proposition 3.30. Tel que mentionné, la condition 3.3 est vérifiée par l'utilisation de la fonction test $V \equiv 1$ et la compacité de $\mathcal{X}$. Puisque l'espace des paramètres est supposé compact, alors $\{\theta_n\}_{n\geqslant 1}$ est borné et donc borné en probabilité. La condition 3.24 et la condition sur H_θ sont vérifiées par hypothèse. Puisque toutes les conditions sont vérifiées, l'algorithme satisfait l'adaptation diminuante. □

Ainsi, il reste à vérifier les conditions 3.24 et (5.12) pour montrer que les différentes déclinaisons de l'algorithme aMTM satisfont à la condition d'adaptation diminuante.

5.4.1.3 Condition de transitions lipschitziennes 3.24

Dans le contexte où Θ est supposé compact et où la fonction test est choisie pour être $V \equiv 1$, la condition 3.24 exige la propriété lipschitzienne suivante,

$$\|P_\theta f - P_{\theta'} f\|_1 \leqslant C\|f\|_1\|\theta - \theta'\|_2,$$

pour toute paire θ,$\theta' \in \Theta$, pour un certain $C < \infty$ et toute fonction f telle que $\|f\|_1 < \infty$.

La vérification d'une telle condition lipschitzienne au sein d'une famille de transitions MTM ne peut se faire d'une manière générale. En effet, certaines des manipulations requises nécessitent une connaissance relativement explicite de la densité instrumentale conditionnelle $Q_\theta^{(-k)}$. Celle-ci est cependant fortement dépendante du type de candidats utilisés de sorte qu'on doit procéder d'une manière cas par cas. Heureusement, une grande partie de la démonstration est indépendante du type de candidats. Ainsi, on est en mesure d'obtenir le résultat général suivant.

Proposition 5.8 *Soit $\{P_\theta\}_{\theta\in\Theta}$ une famille de transitions MTM à densités instrumentales marginales gaussiennes où chacune des covariances $\Sigma^{(k)}$, $k = 1, \ldots, K$, est comprise dans un sous-ensemble*

$\mathcal{K}$ compact de $\mathcal{C}_d^+$. Supposons la condition lipschitzienne suivante sur la probabilité d'acceptation intégrée : il existe $L < \infty$ tel que pour tout x,$y \in \mathcal{X}$,

$$\left| A_\theta^{(k)}(y|x) - A_{\theta'}^{(k)}(y|x) \right| \leqslant L \left| \theta - \theta' \right|_2,$$

Alors, il existe $C < \infty$ tel que pour toute fonction f telle que $\|f\|_1 < \infty$,

$$\|P_\theta f - P_{\theta'} f\|_1 \leqslant C \|f\|_1 \left| \theta - \theta' \right|_2, \qquad \forall\, \theta,\theta' \in \mathcal{K}\,. \tag{5.13}$$

En particulier, la condition 3.24 est satisfaite pour $V^r \equiv 1$.

Démonstration. Par définition, on a

$$\|P_\theta f - P_{\theta'} f\|_1 = \sup_{x \in \mathcal{X}} |P_\theta f(x) - P_{\theta'} f(x)|,$$

et $\|f\|_1 < \infty$ implique que

$$\frac{|f(x)|}{\|f\|_1} \leqslant 1, \qquad \forall\, x \in \mathcal{X}\,. \tag{5.14}$$

Soit $\|f\|_1 < \infty$. On considère d'abord le développement de $P_\theta f(x) - P_{\theta'} f(x)$ suivant :

$$\begin{aligned}
P_\theta f(x) - P_{\theta'} f(x) &= \int_{\mathcal{X}} f(y) P_\theta(y|x) \lambda(\mathrm{d}y) - \int_{\mathcal{X}} f(y) P_{\theta'}(y|x) \lambda(\mathrm{d}y) \\
&= \int_{\mathcal{X}} f(y) \left[P_\theta(y|x) - P_{\theta'}(y|x) \right] \lambda(\mathrm{d}y) \\
&= \int_{\mathcal{X}} f(y) \left[R_\theta(x)\delta_x(y) + p_\theta(y|x) - R_{\theta'}(x)\delta_x(y) - p_{\theta'}(y|x) \right] \lambda(\mathrm{d}y) \\
&= \int_{\mathcal{X}} f(y) \left[\left(R_\theta(x) - R_{\theta'}(x) \right) \delta_x(y) + \left(p_\theta(y|x) - p_{\theta'}(y|x) \right) \right] \lambda(\mathrm{d}y).
\end{aligned}$$

Ensuite, par les propriétés des intégrales, l'inégalité du triangle et par (5.14), on obtient

$$\begin{aligned}
\frac{|P_\theta f(x) - P_{\theta'} f(x)|}{\|f\|_1} &\leqslant \int_{\mathcal{X}} \frac{f(y)}{\|f\|_1} \left[|R_\theta(x) - R_{\theta'}(x)| \delta_x(y) + |p_\theta(y|x) - p_{\theta'}(y|x)| \right] \lambda(\mathrm{d}y) \\
&\leqslant \int_{\mathcal{X}} \left[|R_\theta(x) - R_{\theta'}(x)| \delta_x(y) + |p_\theta(y|x) - p_{\theta'}(y|x)| \right] \lambda(\mathrm{d}y) \\
&= |R_\theta(x) - R_{\theta'}(x)| + \int_{\mathcal{X}} |p_\theta(y|x) - p_{\theta'}(y|x)| \lambda(\mathrm{d}y).
\end{aligned}$$

Ensuite, on note que

$$\begin{aligned}
|R_\theta(x) - R_{\theta'}(x)| &= \left| 1 - \int_{\mathcal{X}} p_\theta(y|x) \lambda(\mathrm{d}y) - 1 + \int_{\mathcal{X}} p_{\theta'}(y|x) \lambda(\mathrm{d}y) \right| \\
&= \left| \int_{\mathcal{X}} \left[p_{\theta'}(y|x) - p_\theta(y|x) \right] \lambda(\mathrm{d}y) \right| \\
&\leqslant \int_{\mathcal{X}} |p_\theta(y|x) - p_{\theta'}(y|x)| \lambda(\mathrm{d}y).
\end{aligned}$$

Ainsi, on peut borner

$$|P_\theta f(x) - P_{\theta'} f(x)| \leqslant 2 \|f\|_1 \int_{\mathcal{X}} |p_\theta(y|x) - p_{\theta'}(y|x)| \lambda(\mathrm{d}y). \tag{5.15}$$

Maintenant, on écrit

$$\begin{aligned} p_\theta(y|x) - p_{\theta'}(y|x) &= \sum_{k=1}^{K} A_\theta^{(k)}(y|x) Q_\theta^{(k)}(y|x) - \sum_{k=1}^{K} A_{\theta'}^{(k)}(y|x) Q_{\theta'}^{(k)}(y|x) \\ &= \sum_{k=1}^{K} \left[A_\theta^{(k)} Q_\theta^{(k)} - A_{\theta'}^{(k)} Q_{\theta'}^{(k)} \right](y|x) \\ &= \sum_{k=1}^{K} \left[A_\theta^{(k)} Q_\theta^{(k)} - A_\theta^{(k)} Q_{\theta'}^{(k)} + A_\theta^{(k)} Q_{\theta'}^{(k)} - A_{\theta'}^{(k)} Q_{\theta'}^{(k)} \right](y|x) \\ &= \sum_{k=1}^{K} \left[A_\theta^{(k)} \left(Q_\theta^{(k)} - Q_{\theta'}^{(k)} \right) + \left(A_\theta^{(k)} - A_{\theta'}^{(k)} \right) Q_{\theta'}^{(k)} \right](y|x). \end{aligned} \tag{5.16}$$

Dans cette dernière expression, il est possible de borner directement le premier terme. En effet, $A_\theta^{(k)} \leqslant 1$ et $Q_\theta^{(k)} - Q_{\theta'}^{(k)}$ peut être borné par la proposition 5.16 :

$$\begin{aligned} \int_{\mathcal{X}} \left| \sum_{k=1}^{K} A_\theta^{(k)} \left(Q_\theta^{(k)} - Q_{\theta'}^{(k)} \right)(y|x) \right| \lambda(\,\mathrm{d}y) &\leqslant \sum_{k=1}^{K} \int_{\mathcal{X}} 1 \cdot \left| Q_\theta^{(k)}(y|x) - Q_{\theta'}^{(k)}(y|x) \right| \lambda(\,\mathrm{d}y) \\ &\leqslant \sum_{k=1}^{K} \int_{\mathcal{X}} |\varphi_{\Sigma^{(k)}}(z) - \varphi_{\Sigma'^{(k)}}(z)| \lambda(\,\mathrm{d}z) \\ &\leqslant \frac{Kd}{\lambda_{\min}} \left\| \Sigma^{(k)} - \Sigma'^{(k)} \right\|_F, \\ &\leqslant \frac{Kd}{\lambda_{\min}} \left\| \theta - \theta' \right\|_2, \end{aligned} \tag{5.17}$$

où $\lambda_{\min}$ est la valeur propre minimale de l'ensemble $\mathcal{K}$. Le second terme de (5.16) peut quand à lui est borné par l'hypothèse lipschitzienne sur $A_\theta^{(k)}$. En effet, on obtient

$$\int_{\mathcal{X}} \left| A_\theta^{(k)} - A_{\theta'}^{(k)} \right| Q_{\theta'}^{(k)}(y|x) \lambda(\,\mathrm{d}y) \leqslant \int_{\mathcal{X}} L \left\| \theta - \theta' \right\|_2 Q_{\theta'}^{(k)}(y|x) \lambda(\,\mathrm{d}y) = L \left\| \theta - \theta' \right\|_2. \tag{5.18}$$

En combinant les inégalités (5.17) et(5.18), on peut désormais borner l'intégrale dans (5.15). En effet, on obtient

$$\begin{aligned} \int_{\mathcal{X}} |p_\theta(y|x) - p_{\theta'}(y|x)| \lambda(\,\mathrm{d}y) &\leqslant \sum_{k=1}^{K} \int_{\mathcal{X}} A_\theta^{(k)} \left| Q_\theta^{(k)} - Q_{\theta'}^{(k)} \right|(y|x) \lambda(\,\mathrm{d}y) \\ &\quad + \sum_{k=1}^{K} \int_{\mathcal{X}} \left| A_\theta^{(k)} - A_{\theta'}^{(k)} \right| Q_{\theta'}^{(k)}(y|x) \lambda(\,\mathrm{d}y) \\ &\leqslant \frac{Kd}{\lambda_{\min}} \left\| \theta - \theta' \right\|_2 + KL \left\| \theta - \theta' \right\|_2 \\ &\leqslant K \left(\frac{d}{\lambda_{\min}} + L \right) \left\| \theta - \theta' \right\|_2. \end{aligned}$$

ce qui conclut la preuve. □

La condition lipschitzienne (5.13) sur la probabilité d'acceptation intégrée $A_\theta^{(k)}$ exige alors un traitement particulier selon les différentes variantes de l'algorithme. La section 5.5.2 contient une discussion complète ; en voici les grandes lignes.

Le proposition 5.8 s'applique à tous les types de mise à jour ainsi qu'à toute fréquence d'adaptation. En effet, il ne s'agit que d'un résultat sur la famille de densités de transition et seule la collection des covariances $\Sigma^{(k)}$, $k = 1, , \ldots, K$, est utilisée pour définir la transition. Il faut cependant s'assurer que les covariances demeurent dans un ensemble compact lorsqu'un facteur d'échelle est utilisé ; ceci peut être accompli en imposant au facteur d'échelle de demeurer dans un certain intervalle compact.

Une fonction de poids indépendante de θ (par exemple, des poids proportionnels à la densité

cible) ne requière aucune condition supplémentaire. Pour des poids plus généraux, il est nécessaire de supposer une condition lipschitzienne en θ sur les poids de sélection. Par exemple, les poids par importance vérifient une telle condition dès que π admet un gradient continu et a un support $\mathcal{X}$ compact.

Des candidats indépendants ne requièrent aucune condition spécifique. Pour ce qui est des autres structures de corrélation, il est nécessaire de vérifier une condition lipschitzienne sur les poids de sélection ainsi que sur la probabilité d'acceptation. Ces conditions peuvent être vérifiées en supposant à nouveau que π admet un gradient continu et a un support $\mathcal{X}$ compact.

5.4.1.4 Condition de mises à jour bornées

Lorsque la fonction test est choisie comme $V \equiv 1$, la condition (5.12) exige tout simplement que l'incrément de la fonction de mise-à-jour H_θ soit borné lorsque θ est contenu dans un espace compact $\mathcal{K} \subseteq \Theta$. L'expression de H_θ dépend évidemment du type de mise à jour choisie ainsi que de la fréquence d'adaptation. Par contre, la supposition de compacité sur $\mathcal{X}$ ainsi que celle sur $\mathcal{K}$ font en sorte que toutes les quantités utilisées dans la mise à jour seront directement bornées uniformément, et ce, peu importe les particularités de l'adaptation. La section 5.5.3 contient la vérification explicite dans chacun des cas, mais aucune supposition supplémentaire ne doit être faite pour y arriver.

5.4.1.5 Généralisations

La grande majorité des résultats présentés dans cette section supposaient la compacité de l'espace des paramètres Θ ainsi que de l'espace d'états $\mathcal{X}$. Sous ces suppositions, il a été possible de vérifier l'ergodicité en variation totale de l'algorithme aMTM ainsi que de toutes ses déclinaisons considérées à la section 5.3 (en supposant à l'occasion des conditions supplémentaires relativement faibles). Par contre, la généralisation à des espaces de paramètres non-borné ou à des densités cibles à support non-borné n'est pas une simple tâche. En effet, les techniques utilisées dans les différentes preuves ne peuvent pas être aisément étendues à des espaces plus généraux.

Par exemple, l'utilisation de la fonction test $V \equiv 1$ dans la proposition 3.30 n'est possible qu'en supposant $\mathcal{X}$ compact. Si $\mathcal{X}$ n'est pas compact, alors un autre choix de fonction doit être fait et les dérivations dans les preuves des différents résultats (e.g. proposition 5.8) ne sont donc plus valides. De plus, la supposition que Θ soit compact permet de borner les valeurs propres des matrices de covariances $\Sigma^{(k)}$, ce qui permet alors de borner uniformément certaines intégrales via la proposition 5.16. Si Θ est non-borné, alors une toute autre technique doit être employée pour obtenir des bornes uniformes.

Dans le contexte où une transition MTM s'apparente fortement à une transition MH, on peut s'attendre à ce que les résultats d'ergodicité d'algorithmes Metropolis adaptatifs puissent être étendus aux algorithme MTM adaptatifs généraux (c.-à-d., ne supposant pas la compacité de Θ ou de $\mathcal{X}$). On inspecte donc certains des résultats d'ergodicité de l'algorithme AM afin de voir s'il est possible d'obtenir des résultats semblables dans le cas MTM.

La proposition 3.27 émet des conditions sous lesquelles un algorithme Metropolis adaptatif satisfait la convergence bornée. D'abord, la densité cible doit être régulière (condition 3.13) et fortement décroissante (condition 3.14) et doit avoir des ailes super-exponentielles (condition 3.18). Ces conditions ne portent que sur la densité cible π et peuvent donc être reprises directement. Ensuite, la densité

instrumentale de chaque transition Metropolis doit être symétrique (condition 3.16) et localement positive (condition 3.15). Dans le cas MTM, ces suppositions pourrait prendre la forme suivante : que chaque densité instrumentale marginale ($Q_\theta^{(k)}$, $k = 1, \ldots, K$) soit symétrique et localement positive. Ceci se vérifie facilement si ces dernières sont choisies gaussiennes. Enfin, la condition 3.21, peut être vérifiée par le lemme 3.28 qui exige que la densité instrumentale soit uniformément bornée par le bas pour de grands pas. Pour ce faire, l'espace Θ doit être restreint de sorte à éviter que toutes les covariances tendent vers 0. Ainsi, une extension de ce résultat permettrait principalement de relâcher la condition de compacité de $\mathcal{X}$. Cependant, la démonstration d'un tel résultat n'est pas aisée. En effet, les techniques utilisées dans la preuve de la proposition 3.27 sont fortement basées sur la géométrie des régions d'acceptation et de rejet. Dans le cas Metropolis, celles-ci sont données par le rapport des densités ; dans le cas MTM, celles-ci dépendent en plus des autres candidats ainsi que des points de référence, ce qui rend l'analyse sensiblement plus complexe.

Pour relâcher la condition de compacité de Θ, il est nécessaire de considérer un algorithme adaptatif à recouvrement compact (cf. section 3.2.3.2). Dans ce cas, la convergence bornée peut être vérifiée en supposant que la famille de transitions MTM est géométriquement ergodique uniformément sur les espaces compacts (condition 3.3) et que $\{\theta_n\}_{n\in\mathbb{N}}$ est un processus borné en probabilité. Des conditions similaires à celles énoncées au paragraphe précédent devraient être suffisantes pour vérifier l'ergodicité géométrique sur tout compact puisque la proposition 3.27 vérifie en fait l'ergodicité géométrique pour Θ compact. Ensuite, pour montrer que $\{\theta_n\}_{n\in\mathbb{N}}$ est borné en probabilité, il faudra utiliser une généralisation du théorème 3.31 aux transitions MTM. Ce dernier est basé sur les conditions 3.24, 3.25, 3.26 et 3.27 le tout pour une certaine fonction test V. Dans le cas AM, le choix $V \propto \pi^{-\eta}$ pour un $\eta \in (0,1)$ permet de vérifier toutes ces conditions. La vérification de la condition de transitions lipschitziennes 3.24 dépend fortement de l'expression de la probabilité d'acceptation ainsi que des régions d'acceptation et de rejet : il n'est donc pas clair qu'il soit possible d'effectuer la vérification dans le cas MTM. La vérification de la condition de mises à jour lipschitziennes 3.25 sera pratiquement identique dans le cas MTM lorsque des mises à jour AM ou ASWAM sont utilisées (la situation est plus complexe pour les mises à jour RAM). La vérification de la condition 3.26 sur les propriétés du champ moyen pourra être effectuée en utilisant la même fonction de Lyapunov étant donné que le champ moyen MTM n'est qu'une répétition du champ moyen AM lorsque les mises à jour AM sont utilisées. Enfin, la condition 3.27 ne dépend pas du type de transition et est donc trivialement vérifiée comme dans le cas AM.

5.4.2 Théorèmes limites

Tel que mentionné à la section 3.3.1, la loi faible des grands nombres et l'ergodicité en variation totale sont deux propriétés qui se vérifient par un même ensemble de conditions pour les algorithmes MCMC adaptatifs. En effet, l'adaptation diminuante et la convergence bornée sont suffisantes pour l'ergodicité (théorème 3.12) ainsi que pour la loi faible des grands nombres pour toute fonction bornée (théorème 3.7). Ainsi, puisque les différentes déclinaisons de l'algorithme aMTM satisfont toutes (sous certaines conditions) à l'adaptation diminuante et à la convergence bornée, on a la loi faible des grands nombres pour toute fonction bornée.

Quand à elle, la loi forte des grands nombres requiert une ergodicité V-géométrique (théorème 3.10). Cette dernière serait satisfaite pour toute fonction de V-norme bornée. Cependant, comme abordé à la section 5.4.1.5, il n'est pas clair que l'algorithme aMTM satisfasse à ce mode d'ergodicité, de sorte

qu'on ne peut garantir la convergence de l'estimateur Monte Carlo $\hat{\pi}_N(f)$ vers l'espérance recherchée $\pi(f)$ lorsque f n'est pas bornée.

Enfin, les théorèmes limites centraux d'algorithmes MCMC adaptatifs exigent également une ergodicité V géométrique. Ainsi, l'estimation ponctuelle de l'espérance $\pi(f)$ ne peut être accompagnée d'une erreur standard Monte Carlo qui soit fiable étant donné que l'algorithme aMTM ne possède pas les garanties théoriques requises.

Malgré le fait que l'ergodicité géométrique des algorithmes aMTM ne puisse être démontrée dans le contexte actuel, il demeure que la forte ressemblance avec les algorithmes Metropolis adaptatifs suggère que les algorithmes aMTM devraient satisfaire à ce mode de convergence sous des conditions similaires à celles requises par les algorithmes Metropolis adaptatifs. Ainsi, pour des densités cibles satisfaisant aux conditions communément supposées (régularité, régularité des contours et ailes super-exponentielles), il n'est pas déraisonnable de faire la conjecture que les algorithmes aMTM sont V-géométriquement ergodiques, où $V \propto \pi^{-r}$ pour un certain $r \geqslant 0$. Dans un tel cas, on obtiendrait la loi forte des grands nombres et un théorème limite central pour toute fonction sous-exponentielle (e.g. polynomiale). Davantage de recherche est nécessaire afin de trouver des conditions sous lesquelles on peut démontrer rigoureusement l'ergodicité V-géométrique des algorithmes aMTM.

Notons finalement que les algorithmes MTM non-adaptatifs satisfont à la loi forte des grands nombres pour toute fonction f qui est π-intégrable (corollaire 4.9). Ainsi, il est possible de considérer un algorithme aMTM à adaptation finie de sorte à produire un échantillon utilisant seulement une transition MTM fixe. De plus, le théorème 2.23 indique qu'un théorème limite central est vérifié si la variance asymptotique est finie. En pratique, on peut étudier le comportement d'un estimateur de la variance asymptotique pour voir si une telle supposition est raisonnable pour ensuite utiliser les conclusions d'un théorème limite central.

5.5 Annexes au chapitre 5

5.5.1 Suppléments à la section 5.4.1.1

La section 5.4.1.1 traite de la vérification de la condition de convergence bornée pour l'algorithme aMTM. Afin d'arriver au résultat principal – le théorème 5.6 – plusieurs résultats préalables sont requis. De plus, la vérification des différentes conditions nécessaire à ces résultats invitent à certaines discussions selon les déclinaisons de l'algorithme. Cette annexe est donc consacrée à ces considérations.

5.5.1.1 Résultats préalables au théorème 5.6

Pour vérifier la continuité de $(x,\theta) \mapsto \Delta(x,\theta,n)$, on considère d'abord un résultat théorique : le théorème de convergence dominée permet d'établir la continuité de certaines fonctions définies par des intégrales.

Théorème 5.9 (Convergence dominée de Lebesgue, Williams, 1991, section 5.9) *Soit μ une mesure, et soit $\{f_n\}_{n\geqslant 1}$, une suite de fonctions μ-intégrables qui converge ponctuellement vers une fonction f. Supposons qu'il existe une fonction g intégrable telle que $|f_n| \leqslant |g|$ pour tout $n \geqslant 1$, alors f est intégrable et*

$$\lim_{n\to\infty} \mu(f_n) = \mu(f).$$

Corollaire 5.10 *Soit $F(w) = \int_{\mathcal{T}} f(w,t)\lambda(\,\mathrm{d}t)$ une fonction $F : \mathcal{W} \to \mathbb{R}$, où $f : \mathcal{W}\times\mathcal{T} \to \mathbb{R}$ est continue par rapport à (w,t) et où λ dénote la mesure de Lebesgue. Supposons qu'il existe une fonction $g : \mathcal{T} \to \mathbb{R}$ telle que $|f(w,t)| \leqslant |g(t)|$ pour tout $(w,t) \in \mathcal{W}\times\mathcal{T}$ et telle que $\lambda(|g|) < \infty$, alors F est une fonction continue de w sur tout $\mathcal{W}$.*

Démonstration. La fonction F est continue sur $\mathcal{W}$ si et seulement si $\lim_{n\to\infty} F(w_n) = F(w)$ pour toute séquence $\{w_n\}_{n\geqslant 1}$ telle que $w_n \to w$. On considère donc $w_n \to w$ et on définit

$$f_n(t) = f(w_n,t), \qquad t \in \mathcal{T}\,, n \geqslant 1.$$

Par la continuité de f par rapport à w, on a que $f_n(t) \to f(w,t)$ pour tout $t \in \mathcal{T}$. De plus, $|f_n(t)| = |f(w_n,t)| \leqslant |g(t)|$ pour tout $n \geqslant 1$ par hypothèse. Par l'existence de g, les conditions du théorème de convergence dominée 5.9 sont satisfaites et l'on obtient que

$$\lim_{n\to\infty} F(w_n) = \lim_{n\to\infty} \int_{\mathcal{T}} f(w_n,t)\lambda(\,\mathrm{d}t) = \lim_{n\to\infty} \lambda(f_n) = \lambda(f(w,\cdot)) = F(w),$$

ce qui conclut la preuve. □

À l'aide du corollaire 5.10, plusieurs expressions faisant partie de $\Delta(x,\theta,n)$ peuvent être montrées continues.

Lemme 5.11 *Pour un noyau MTM à paramètre θ, la probabilité d'acceptation intégrée de x vers y via le k-ième candidat,*

$$A_\theta^{(k)}(y|x) = \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \alpha_\theta(y,y^{(-k)}|x,x_*^{(-k)})\bar{w}_\theta^{(k)}(y;y^{(-k)}|x) \times Q_\theta^{(-k)}(y^{(-k)}|x,y)Q_\theta^{(-k)}(x_*^{(-k)}|x,y)\,\mathrm{d}x_*^{(-k)}\,\mathrm{d}y^{(-k)},$$

est une fonction continue de (x,y,θ) à condition que les densités instrumentales conditionnelles $Q_\theta^{(-k)}$, la fonction de poids $\bar{w}_\theta^{(k)}$ et la probabilité d'acceptation α_θ soient toutes continues par rapport

à leurs arguments et à leurs paramètres et que les densités conditionnelles sont uniformément bornées par une fonction $Q^+ : \mathcal{X}^{K-1} \to \mathbb{R}_{\geqslant 0}$ telle que Q^+ est intégrable.

Démonstration. On procède par deux utilisations successives du corollaire 5.10. On écrit

$$A_\theta^{(k)}(y|x) = \int_{\mathcal{X}^{K-1}} A_\theta^{(k)}(y; y^{(-k)}|x) Q_\theta^{(-k)}(y^{(-k)}|x,y)\, \mathrm{d}y^{(-k)},$$

où

$$A_\theta^{(k)}(y; y^{(-k)}|x) = \int_{\mathcal{X}^{K-1}} \bar{w}^{(k)}(y; y^{(-k)}|x) Q_\theta^{(-k)}(x_*^{(-k)}|x,y) \alpha(y,y^{(-k)}|x,x_*^{(-k)})\, \mathrm{d}x_*^{(-k)}.$$

Premièrement, on montre que $A_\theta^{(k)}(y; y^{(-k)}|x)$ est une fonction continue de $(x,y,y^{(-k)},\theta)$. On identifie la fonction cible $F(w) = A_\theta^{(k)}(y; y^{(-k)}|x)$, la variable $w = (x,y,y^{(-k)},\theta)$ et son domaine $\mathcal{W} = \mathcal{X}^{K+1} \times \Theta$, la variable d'intégration $t = x_*^{(-k)}$ et son domaine $\mathcal{T} = \mathcal{X}^{K-1}$ et la fonction intégrée

$$f(w,t) = \bar{w}_\theta^{(k)}(y; y^{(-k)}|x) \alpha_\theta(y,y^{(-k)}|x,x_*^{(-k)}) Q_\theta^{(-k)}(x_*^{(-k)}|x,y).$$

Étant donné que chacun des termes du produit est supposé continu par rapport à toutes les variables et paramètres en jeu, on trouve que f est également continue par rapport à (w,t). Puisque les poids et la probabilité d'acceptation MH sont tous deux compris dans l'intervalle [0,1], on trouve

$$|f(w,t)| \leqslant Q_\theta^{(-k)}(x_*^{(-k)}|x,y)$$

Pour w fixé, $Q_\theta^{(-k)}(x_*^{(-k)}|x,y)$ est une densité sur $\mathcal{X}^{K-1}$. Par l'hypothèse que les densités conditionnelles sont uniformément bornées par Q^+, on trouve donc que

$$|f(w,t)| \leqslant Q^+(x_*^{(-k)}) =: g(t), \qquad w \in \mathcal{W}.$$

Finalement, puisque Q^+ est intégrable, g est directement intégrable. Par le corollaire 5.10, F est continu par rapport à w, c'est-à-dire que $A_\theta^{(k)}(y; y^{(-k)}|x)$ est continu par rapport à $(x,y,y^{(-k)},\theta)$.

Deuxièmement, on montre que $A_\theta^{(k)}(y|x)$ est une fonction continue de (x,y,θ) sachant maintenant que $A_\theta^{(k)}(y; y^{(-k)}|x)$ est une fonction continue de $(x,y,y^{(-k)},\theta)$. On identifie la fonction cible $F(w) = A_\theta^{(k)}(y|x)$, la variable $w = (x,y,\theta)$ et son domaine $\mathcal{W} = \mathcal{X}^2 \times \Theta$, la variable d'intégration $t = y^{(-k)}$ et son domaine $\mathcal{T} = \mathcal{X}^{K-1}$ et la fonction intégrée

$$f(w,t) = A_\theta^{(k)}(y; y^{(-k)}|x) Q_\theta^{(-k)}(y^{(-k)}|x,y).$$

Étant donné que $Q_\theta^{(-k)}$ est supposé continu par rapport à toutes les variables et paramètres en jeu, on trouve que f est également continue par rapport à (w,t). Puisque $A_\theta^{(k)}(y; y^{(-k)}|x) \in [0,1]$, on trouve

$$|f(w,t)| \leqslant Q_\theta^{(-k)}(y^{(-k)}|x,y)$$

Pour w fixé, $Q_\theta^{(-k)}(y^{(-k)}|x,y)$ est une densité sur $\mathcal{X}^{K-1}$. Par l'hypothèse que les densités conditionnelles sont uniformément bornées par Q^+, on trouve donc que

$$|f(w,t)| \leqslant Q^+(y^{(-k)}) =: g(t), \qquad w \in \mathcal{W}.$$

Finalement, puisque Q^+ est intégrable, g est directement intégrable. Par le corollaire 5.10, F est continu par rapport à w, c'est-à-dire que $A_\theta^{(k)}(y|x)$ est continu par rapport à (x,y,θ). □

Les différentes conditions de continuité du lemme 5.11 sont généralement simples à vérifier bien que certaines remarques sont importantes à faire à ce niveau. D'une part, les conditions de continuité par rapport au paramètre θ dépendent de la famille, mais la famille de densités gaussiennes satisfait à ces conditions. Les poids de sélection standardisés,

$$\bar{w}_\theta^{(k)}(y; y^{(-k)}|x) = \frac{w_\theta^{(k)}(y|x)}{\sum_{j=1}^K w_\theta^{(j)}(y^{(j)}|x)},$$

seront continus dès que les fonctions de poids $w^{(k)}$ sont continues et strictement positives. Tous les

choix de fonctions de poids de la section 4.1.2 satisfont à ces conditions en supposant des conditions mineures sur la densité cible ou sur les densités instrumentales marginales. Similairement, la probabilité d'acceptation,

$$\alpha_\theta(y,y^{(-k)}|x,x_*^{(-k)}) = \min\left\{1, \frac{\pi(y)Q_\theta^{(k)}(x|y)\bar{w}_\theta^{(k)}(x;x_*^{(-k)}|y)}{\pi(x)Q_\theta^{(k)}(y|x)\bar{w}_\theta^{(k)}(y;y^{(-k)}|x)}\right\}$$

sera continue sous des conditions semblables sur π et sur les $Q^{(k)}$. Enfin, lorsque les distributions conditionnelles admettent des densités par rapport à la mesure de Lebesgue (e.g. candidats indépendants), les conditions de continuité sont directes. Cependant, certains choix de structure de corrélation (e.g. quasi-Monte Carlo ou par variable aléatoire commune) produisent des distributions conditionnelles dégénérées. Lorsque des candidats extrêmement antithétiques sont utilisés, les densités conditionnelles sont des densités dans un certain sous-espace de $\mathcal{X}^{K-1}$; des arguments similaires peuvent alors être utilisés. Dans ce cas, $A_\theta^{(k)}$ n'est plus une intégrale et sa continuité découle de celle des poids standardisés et de la probabilité d'acceptation MH. Dans tous les cas, il demeure que la continuité de $A_\theta^{(k)}$ peut être vérifiée aisément sous des conditions peu restrictives sur la densité cible et sur les densités instrumentales marginales. Enfin, la borne supérieure Q^+ correspond en fait à

$$Q^+(y^{(-k)}) = \sup_{(x,y,\theta)\in\mathcal{X}^2\times\Theta} Q_\theta^{(-k)}(y^{(-k)}|x,y), \qquad y^{(-k)} \in \mathcal{X}^{K-1}.$$

Pour que Q^+ soit intégrable, il sera souvent suffisant de supposer la compacité de $\mathcal{X}$ et de Θ étant donné que $Q_\theta^{(-k)}$ varie d'une manière continue par rapport à (x,y,θ).

En utilisant à nouveau le corollaire 5.10, il nous est possible de montrer la continuité de la probabilité de rejet $R_\theta(x)$ donnée par

$$R_\theta(x) = 1 - \sum_{k=1}^{K} \overline{A}_\theta^{(k)}(x),$$
$$\overline{A}_\theta^{(k)}(x) = \int_{\mathcal{X}} A_\theta^{(k)}(y|x)Q_\theta^{(k)}(y|x)\,\mathrm{d}y.$$

Lemme 5.12 *Pour un noyau MTM à paramètre θ, la probabilité d'acceptation intégrée à partir de x via le k-ième candidat $\overline{A}_\theta^{(k)}(x)$ est une fonction continue de (x,θ) à condition que $A_\theta^{(k)}(y|x)$ soit une fonction continue de (x,y,θ) et que $Q_\theta^{(k)}$ soit une densité par rapport à la mesure de Lebesgue telle qu'il existe $Q^+ : \mathcal{X} \to \mathbb{R}_{\geqslant 0}$ intégrable avec $Q_\theta^{(k)}(y|x) \leqslant Q^+(y)$ uniformément sur (x,θ). De plus, sous ces mêmes conditions, la probabilité de rejet $R_\theta(x)$ est également une fonction continue de (x,θ).*

Démonstration. Il s'agit d'une conséquence directe de corollaire 5.10. En effet, on identifie la variable $w = (x,\theta)$ et son domaine $\mathcal{W} = \mathcal{X}\times\Theta$, la variable d'intégration $t = y$ et son domaine $\mathcal{T} = \mathcal{X}$, la fonction cible $F(w) = \overline{A}_\theta^{(k)}(x)$ et la fonction intégrée

$$f(w,t) = A_\theta^{(k)}(y|x)Q_\theta^{(k)}(y|x).$$

Puisque $A_\theta^{(k)}$ est continue par hypothèse et que $Q_\theta^{(k)}$ est continue puisque c'est une densité par rapport à la mesure de Lebesgue, on a que f est continue par rapport à $(w,t) = ((x,\theta),y)$. Puisque $A_\theta^{(k)} \in [0,1]$, on peut borner

$$f(w,t) \leqslant Q_\theta^{(k)}(y|x) \leqslant Q^+(y) =: g(t).$$

La supposition sur l'intégrabilité de Q^+ assure donc l'intégrabilité de g, ce qui vérifie toutes les conditions du corollaire 5.10. On conclut donc que F est continue par rapport à w, c'est-à-dire que $\overline{A}_\theta^{(k)}(x)$ est continu par rapport à (x,θ). Finalement, puisque $R_\theta(x)$ n'est qu'une fonction linéaire des

$\overline{A}_\theta^{(k)}(x)$, $k = 1, \dots, K$, alors $R_\theta(x)$ est aussi continu par rapport à (x,θ). □

Comme dans le cas du lemme 5.11, la condition de borne supérieure sur les $Q^{(k)}$ correspond à

$$Q^+(y) = \sup_{(x,\theta)\in\mathcal{X}\times\Theta} Q_\theta^{(k)}(y|x)$$

Son intégrabilité peut être assurée par une supposition de compacité de $\mathcal{X}$ et de Θ étant donné que $Q_\theta^{(k)}$ varie d'une manière continue par rapport à (x,θ).

Lemme 5.13 *Soit un noyau de transition de la forme Metropolis-Hastings,*

$$P(\mathrm{d}y|x) = R(x)\delta_x(\mathrm{d}y) + p(y|x)\lambda(\mathrm{d}y).$$

Alors, le noyau de transition itéré de n pas peut être écrit sous la forme

$$P^n(\mathrm{d}y|x) = R^n(x)\delta_x(\mathrm{d}y) + p^n(y|x)\lambda(\mathrm{d}y), \tag{5.10}$$

où

$$p^n(y|x) = \int p^{n-1}(y|z)p(z|x)\lambda(\mathrm{d}z), \qquad p^0(y|z) = \delta_z(y).$$

Démonstration. On procède par induction sur $n \geqslant 1$. Le cas $n = 1$ est direct :

$$\begin{aligned}
P^1(\mathrm{d}y|x) &= R^1(x)\delta_x(\mathrm{d}y) + p^1(y|x)\lambda(\mathrm{d}y) \\
&= R(x)\delta_x(\mathrm{d}y) + \int p^0(y|z)p(z|x)\lambda(\mathrm{d}z)\lambda(\mathrm{d}y) \\
&= R(x)\delta_x(\mathrm{d}y) + \int \delta_z(y)p(z|x)\lambda(\mathrm{d}z)\lambda(\mathrm{d}y) \\
&= R(x)\delta_x(\mathrm{d}y) + p(y|x)\lambda(\mathrm{d}y) \\
&= P(\mathrm{d}y|x).
\end{aligned}$$

Pour $n > 1$, on suppose l'hypothèse (5.10) vraie. Alors,

$$\begin{aligned}
P^{n+1}(\mathrm{d}y|x) &= \int P^n(\mathrm{d}y|z)P(\mathrm{d}z|x) \\
&= \int \left[R^n(z)\delta_z(\mathrm{d}y) + p^n(y|z)\lambda(\mathrm{d}y)\right]\left[R(x)\delta_x(\mathrm{d}z) + p(z|x)\lambda(\mathrm{d}z)\right] \\
&= \int R^n(z)\delta_z(\mathrm{d}y)R(x)\delta_x(\mathrm{d}z) + \int R^n(z)\delta_z(\mathrm{d}y)p(z|x)\lambda(\mathrm{d}z) \\
&\qquad + \int p^n(y|z)\lambda(\mathrm{d}y)R(x)\delta_x(\mathrm{d}z) + \int p^n(y|z)\lambda(\mathrm{d}y)p(z|x)\lambda(\mathrm{d}z) \\
&= R^{n+1}(x)\delta_x(\mathrm{d}y) + 0 + 0 + \int p^n(y|z)p(z|x)\lambda(\mathrm{d}z)\lambda(\mathrm{d}y) \\
&= R^{n+1}(x)\delta_x(\mathrm{d}y) + p^{n+1}(y|x)\lambda(\mathrm{d}y).
\end{aligned}$$

□

Lemme 5.14 *Sous les conditions du lemme 5.3 et en supposant que Q^+ est intégrable, $p_\theta^n(y|x)$ est une fonction continue de (x,y,θ) pour tout $n \geqslant 1$.*

Démonstration. On procède par induction sur $n \geqslant 1$ pour montrer que $p_\theta^n(y|x)$ est continu comme fonction de (x,y,θ) et est borné uniformément par $K^n Q^+(y)$, où $K = \sup_{y\in\mathcal{X}} Q^+(y) < \infty$
Pour $n = 1$, on a

$$p_\theta^1(y|x) = \int_{\mathcal{X}} \delta_z(y)p_\theta(z|x)\lambda(\mathrm{d}z) = p_\theta(y|x) = \sum_{k=1}^{K} A_\theta^{(k)}(y|x)Q_{\theta^{(k)}}^{(k)}(y|x),$$

qui est une somme de produits de fonctions qui sont toutes continues par hypothèse. Ainsi, $p_\theta^1 = p_\theta$

est une fonction continue de (x,y,θ). La borne uniforme est directe :

$$\left|p_\theta^1(y|x)\right| \leqslant \sum_{k=1}^{K} \left|A_\theta^{(k)}(y|x) Q_{\theta^{(k)}}^{(k)}(y|x)\right| \leqslant \sum_{k=1}^{K} Q^+(y) = K \cdot Q^+(y).$$

Pour $n > 1$, on suppose que p_θ^{n-1} est continue comme fonction de (x,y,θ). Par définition,

$$p_\theta^n(y|x) = \int p_\theta^{n-1}(y|z) p_\theta(z|x) \lambda(\,\mathrm{d}z).$$

Afin de vérifier la continuité de cette fonction, on utilise le corollaire 5.10. On identifie la fonction cible $F(w) = p_\theta^n(y|x)$, la variable $w = (x,y,\theta)$ et son domaine $\mathcal{W} = \mathcal{X}^2 \times \Theta$, la variable d'intégration $t = z$ et son domaine $\mathcal{T} = \mathcal{X}$ et la fonction intégrée

$$f(w,t) = p_\theta^{n-1}(y|z) p_\theta(z|x).$$

Étant donné que $p_\theta^{n-1}(y|z)$ est supposé continu et $p_\theta(z|x)$ est continu par le cas $n = 1$, tous deux par rapport à (x,y,z,θ), on trouve que f est également continue par rapport à (w,t). Par l'hypothèse d'induction, on trouve que

$$\left|p_\theta^{n-1}(y|z)\right| \leqslant K^{n-1} \cdot Q^+(y) \leqslant K^{n-1} \sup_{y \in \mathcal{X}} Q^+(y) \leqslant K^n.$$

Ainsi, on trouve que

$$|f(w,t)| \leqslant [K^n] \cdot \left[K \cdot Q^+(z)\right] = K^n \cdot Q^+(z) =: g(t)$$

Par la supposition supplémentaire sur Q^+, on obtient directement que g est intégrable pour tout n. Par le corollaire 5.10, F est continu par rapport à w, c'est-à-dire que $p_\theta^n(y|x)$ est continu par rapport à (x,y,θ). □

La supposition que Q^+ est intégrable peut sembler forte. Cependant, lorsque $\mathcal{X}$ est un espace compact, cette supposition est directement vérifiée puisque Q^+ est une fonction continue.

Notons finalement que les différentes fonctions servant de bornes uniformes dans les lemmes 5.3 et 5.5 pourront être facilement choisies intégrables lorsque $\mathcal{X}$ est compact. Dans chacun des deux cas, la fonction Q^+ est exprimée comme le suprémum pris sur un ensemble compact de fonctions continues. Ainsi, les Q^+ sont des fonctions continues de leur argument. Puis, étant donné que $\mathcal{X}$ est compact, on peut poser que les densités instrumentales marginales sont nulles au-delà du diamètre de $\mathcal{X}$ puisque des candidats à l'extérieur de $\mathcal{X}$ seront toujours rejetés. On trouve donc que les Q^+ peuvent être choisis positifs seulement sur un espace compact, ce qui implique leur intégrabilité.

5.5.1.2 Sur la structure de corrélation

Ensuite, la condition de continuité de $A_\theta^{(k)}(y|x)$ dans le lemme 5.3 peut être vérifiée de deux façons. D'une part, le lemme 5.11 peut être utilisé. Dans ce cas, on requiert la continuité des densités instrumentales conditionnelles, de la fonction de poids et de la probabilité d'acceptation ainsi qu'une borne uniforme sur les densités conditionnelles. La borne uniforme peut quant à elle être assurée par une supposition de compacité de $\mathcal{X}$. La continuité de la fonction de poids et de la probabilité d'acceptation sont des conditions généralement faciles à vérifier avec des suppositions faibles telles que la continuité et la positivité de la densité cible et des densités instrumentales. Puis, la continuité des densités conditionnelles dépend de la manière dont les candidats sont simultanément générés. Des candidats indépendants satisfont à cette condition comme le montre l'expression (5.9) : les densités conditionnelles sont des densités gaussiennes sur $\mathcal{X}^{K-1}$ et sont donc continues par rapport aux arguments et aux paramètres. D'autre part, la condition de continuité de $A_\theta^{(k)}(y|x)$ peut parfois être vérifiée directement. En effet, lorsque les autres candidats sont déterministes conditionnellement

à l'état actuel x et à la proposition choisie y, la densité conditionnelle $Q_\theta^{(-k)}(\cdot|x,y)$ est alors dégénérée et on obtient

$$A_\theta^{(k)} = \bar{w}^{(k)}(y; y^{(-k)}|x)\alpha(y; y^{(-k)}|x,x_*^{(-k)}),$$

où $y^{(-k)}$ et $x_*^{(-k)}$ sont entièrement déterminés par x et y. Alors, la continuité de $A_\theta^{(k)}(y|x)$ découle directement de celle de la fonction de poids et de celle de la probabilité d'acceptation ; ces propriétés peuvent, à leur tour, être aisément assurées par des conditions mineures sur π et sur les $Q_\theta^{(k)}$. En somme, toutes les structures de corrélation définies à la section 5.3.4 sont couvertes par l'un ou l'autre de ces cas.

5.5.1.3 Cas où l'espace d'états est non-borné

Si on ne suppose pas la compacité de $\mathcal{X}$, il est possible de trouver certaines conditions pour assurer que $\{X_n\}_{n\geqslant 1}$ est borné en probabilité. Dans ce cas, on doit considérer le sous-ensemble $K \subset \mathcal{X}$ et la transition fixe P à l'extérieur de K. En fait, la démonstration ne dépend pas de la famille de noyaux de transition, seulement des propriétés de P étant donné que le processus est borné lorsque les transitions adaptées sont utilisées. Ainsi, les résultats contenus dans Craiu et collab. (2015) peuvent être utilisés directement. En particulier, on peut dégager le lemme suivant.

Lemme 5.15 (Craiu et collab., 2015) *Soit $K \subseteq \mathcal{X}$, un sous-ensemble borné du support $\mathcal{X}$ de la densité cible π et soit un algorithme MCMC adaptatif à sauts bornés (3.23) sans adaptation et utilisant une certaine transition P à l'extérieur de K (i.e. l'algorithme satisfait (3.24)). Si P est borné par le haut (3.25) et satisfait la condition ε,δ (3.26), alors le processus $\{X_n\}_{n\geqslant 1}$ est borné en probabilité dans le sens*

$$\lim_{L\to\infty} \sup_{n\geqslant 1} \mathbb{P}\left(\|X_n\|_2 > L \,|\, X_0 = x_0\right) = 0.$$

Les conditions (3.25) et (3.26) sur la transition fixe P peuvent être facilement vérifiées pour une transition Metropolis-Hastings à densité instrumentale q symétrique bornée uniformément par $M < \infty$ et localement bornée par le bas ($\|y - x\|_2 < \delta$ implique $q(y|x) \geqslant \varepsilon$ pour un certain choix de δ,$\varepsilon > 0$.) En effet, soit $x \in K_D \setminus K$ et $y \in K_{2D} \setminus K_D$ où K_r est l'ensemble des points à une distance d'au plus r de K, i.e. $K_r := \{x \in \mathbb{R}^d : \text{dist}(x,K) \leqslant r\}$, alors

$$\begin{aligned} P(\,\mathrm{d}y|x) &= \alpha(y|x)q(y|x)\lambda(\,\mathrm{d}y) + R(x)\delta_x(\,\mathrm{d}y) \\ &\leqslant 1 \cdot q(y|x)\lambda(\,\mathrm{d}y) + R(x) \cdot 0 \\ &\leqslant \left(\sup_{y\in\mathcal{X}} q(y|x)\right)\lambda(\,\mathrm{d}y) =: M\lambda(\,\mathrm{d}y). \end{aligned}$$

Pour tout $\|y - x\|_2 < \delta$, on trouve

$$P(\,\mathrm{d}y|x) \geqslant \alpha(y|x)q(y|x)\lambda(\,\mathrm{d}y) \geqslant \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}\varepsilon\lambda(\,\mathrm{d}y).$$

Pour un certain rectangle J compact, on a que π est borné par le haut et par le bas sur J. Si π est positif sur J, alors le ratio $\pi(y)/\pi(x)$ est uniformément borné par le bas sur x,$y \in J$ par une constante positive, ce qui vérifie la condition (3.26).

5.5.2 Suppléments à la section 5.4.1.3

5.5.2.1 *Résultat préalable sur les densités gaussiennes*

On débute par la démonstration d'un résultat qui permet de borner l'intégrale de la différence entre deux densités gaussiennes par la différence entre les covariances. Ce type d'inégalité sera utilisé à plusieurs reprises afin de produire des bornes lipschitziennes à une différence de densités normales.

Proposition 5.16 *Soit $\{\varphi_\Sigma\}_{\Sigma\in\mathcal{S}}$ une famille de densités gaussiennes à d dimensions de moyenne $\mathbf{0}_d$ et de covariance $\Sigma \in \mathcal{S} \subseteq \mathcal{C}_d^+$, i.e.*

$$\varphi_\Sigma(z) = (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2} z^\top \Sigma^{-1} z\right\}, \qquad z \in \mathbb{R}^d .$$

Si $\mathcal{S}$ est un ensemble compact, alors

$$\int_{\mathbb{R}^d} |\varphi_\Sigma(z) - \varphi_{\Sigma'}(z)| \lambda(\mathrm{d}z) \leqslant \frac{d}{\lambda_{\min}} \left\|\Sigma - \Sigma'\right\|_F, \qquad \forall\, \Sigma, \Sigma' \in \mathcal{S},$$

où $\lambda_{\min}$ est la valeur propre d'une covariance $\Sigma \in \mathcal{S}$ et $\|\cdot\|_F$ dénote la norme de Frobenius usuelle.

Démonstration. Puisque $\mathcal{S}$ est supposé compact, il existe $\lambda_{\min} < \lambda_{\max} \in (0,\infty)$ tels que les valeurs propres de Σ soient toutes contenues dans $[\lambda_{\min},\lambda_{\max}]$ pour tout $\Sigma \in \mathcal{S}$. Inspiré d'une étape de la preuve de Haario et collab. (2001, théorème 1), on considère ensuite la combinaison convexe $\Sigma_t = \Sigma + t(\Sigma' - \Sigma) = (1-t)\Sigma + t\Sigma'$ (Haario et collab., 2001, théorème 1) qui vaut Σ lorsque $t = 0$, Σ' lorsque $t = 1$ et qui définit une matrice symétrique définie positive pour tout $t \in [0,1]$. En effet, Σ_t est évidemment symétrique puisque combinaison linéaire de matrices symétriques et définie positive : pour tout $t \in [0,1]$ et tout $x \in \mathbb{R}^d$,

$$x^\top \Sigma_t x = \underbrace{(1-t)}_{\geqslant 0} \underbrace{x^\top \Sigma x}_{\geqslant 0} + \underbrace{t}_{\geqslant 0} \underbrace{x^\top \Sigma' x}_{\geqslant 0} \geqslant 0.$$

Ainsi, φ_{Σ_t} est une densité gaussienne bien définie pour tout $t \in [0,1]$ et tout $\Sigma, \Sigma' \in \mathcal{C}_d^+$. L'intérêt de cette définition est l'identité suivante

$$\int_0^1 \left(\frac{\partial}{\partial t} \varphi_{\Sigma_t}(z)\right) \mathrm{d}t = \varphi_{\Sigma_t}(z)\Big|_{t=0}^{t=1} = \varphi_\Sigma(z) - \varphi_{\Sigma'}(z),$$

à condition que $\varphi_{\Sigma_t}(z)$ soit dérivable par rapport à t (ce qui sera montré implicitement par les calculs qui suivent). Il sera alors possible de relier plus aisément la différence $\varphi_\Sigma(z) - \varphi_{\Sigma'}(z)$ à la différence $\Sigma - \Sigma'$ étant donné que $\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t = \Sigma' - \Sigma$. On procède alors à la dérivation logarithmique pour trouver

$$\begin{aligned}
\frac{\partial}{\partial t} \varphi_{\Sigma_t}(z) &= \varphi_{\Sigma_t}(z) \frac{\partial}{\partial t} \log \varphi_{\Sigma_t}(z) \\
&= -\frac{1}{2} \varphi_{\Sigma_t}(z) \frac{\partial}{\partial t} \left[d \log(2\pi) + \log\det(\Sigma_t) + z^\top \Sigma_t^{-1} z\right],
\end{aligned}$$

où (Petersen et Pedersen, 2008)

$$\begin{aligned}
\frac{\partial}{\partial t} \log\det(\Sigma_t) &= \mathrm{tr}\left(\Sigma_t^{-1} \frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t\right) = \mathrm{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma)\right), \\
\frac{\partial}{\partial t} z^\top \Sigma_t^{-1} z &= -z^\top \left(\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t^{-1}\right) z \\
&= -z^\top \Sigma_t^{-1} \left(\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t\right) \Sigma_t^{-1} z \\
&= -z^\top \Sigma_t^{-1} (\Sigma' - \Sigma) \Sigma_t^{-1} z \\
&= \mathrm{tr}\left(-z^\top \Sigma_t^{-1} (\Sigma' - \Sigma) \Sigma_t^{-1} z\right) \\
&= \mathrm{tr}\left(-\Sigma_t^{-1} z z^\top \Sigma_t^{-1} (\Sigma' - \Sigma)\right).
\end{aligned}$$

La dépendance en la différence $\Sigma - \Sigma'$ est désormais évidente :

$$\frac{\partial}{\partial t}\varphi_{\Sigma_t}(z) = -\frac{1}{2}\varphi_{\Sigma_t}(z)\operatorname{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma) - \Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right)$$

Par l'inégalité du triangle, on trouve

$$\left|\frac{\partial}{\partial t}\log\varphi_{\Sigma_t}(z)\right| \leqslant \left|\operatorname{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma)\right)\right| + \left|\operatorname{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right)\right|.$$

Par l'inégalité de Hölder sur les normes matricielles de Schatten, i.e. $|\operatorname{tr}(AB)| \leqslant \|A\|_F\|B\|_F$ où la norme de Frobenius correspond à la norme de Schatten d'ordre 2, on peut borner

$$\left|\operatorname{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma)\right)\right| \leqslant \left\|\Sigma_t^{-1}\right\|_F\left\|\Sigma' - \Sigma\right\|_F.$$

Similairement, on peut borner

$$\begin{aligned}
\left|\operatorname{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right)\right| &\leqslant \left\|\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\right\|_F\left\|\Sigma' - \Sigma\right\|_F \\
&\leqslant \sqrt{\operatorname{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\right)}\left\|\Sigma' - \Sigma\right\|_F \\
&\leqslant \sqrt{\operatorname{tr}\left(\Sigma_t^{-1}z(z^\top\Sigma_t^{-2}z)z^\top\Sigma_t^{-1}\right)}\left\|\Sigma' - \Sigma\right\|_F \\
&\leqslant \sqrt{(z^\top\Sigma_t^{-2}z)\operatorname{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\right)}\left\|\Sigma' - \Sigma\right\|_F \\
&\leqslant \sqrt{(z^\top\Sigma_t^{-2}z)\operatorname{tr}\left(z^\top\Sigma_t^{-2}z\right)}\left\|\Sigma' - \Sigma\right\|_F \\
&\leqslant \sqrt{(z^\top\Sigma_t^{-2}z)^2}\left\|\Sigma' - \Sigma\right\|_F \\
&\leqslant \left|z^\top\Sigma_t^{-2}z\right|\left\|\Sigma' - \Sigma\right\|_F
\end{aligned}$$

Notons ensuite que, si A est symétrique, alors

$$x^\top A^2x = x^\top AAx = x^\top A^\top Ax = (Ax)^\top Ax = \|Ax\|_2 \geqslant 0.$$

Ainsi, $z^\top\Sigma_t^{-2}z > 0$ pour tout z, ce qui implique

$$\left|z^\top\Sigma_t^{-2}z\right| = z^\top\Sigma_t^{-2}z.$$

On peut donc borner

$$\begin{aligned}
\left|\frac{\partial}{\partial t}\log\varphi_{\Sigma_t}(z)\right| &\leqslant \left\|\Sigma_t^{-1}\right\|_F\left\|\Sigma' - \Sigma\right\|_F + \left\|\Sigma' - \Sigma\right\|_F z^\top\Sigma_t^{-2}z \\
&= \left\|\Sigma' - \Sigma\right\|_F\left(\left\|\Sigma_t^{-1}\right\|_F + z^\top\Sigma_t^{-2}z\right).
\end{aligned}$$

Par la théorie des formes quadratiques, on note que si $Z \sim \mathcal{N}_d(\mathbf{0},\Sigma)$, alors $\mathbb{E}\{Z^\top AZ\} = \operatorname{tr}(A\Sigma)$, et donc

$$\int_{\mathcal{X}}(z^\top\Sigma_t^{-2}z)\varphi_{\Sigma_t}(z)\lambda(\,\mathrm{d}z) = \operatorname{tr}\left(\Sigma_t^{-2}\Sigma_t\right) = \operatorname{tr}\left(\Sigma_t^{-1}\right).$$

Ainsi, on trouve

$$\int_{\mathcal{X}}\left(\left\|\Sigma_t^{-1}\right\|_F + z^\top\Sigma_t^{-2}z\right)\varphi_{\Sigma_t}(z)\lambda(dz) = \left\|\Sigma_t^{-1}\right\|_F + \operatorname{tr}\left(\Sigma_t^{-1}\right).$$

Maintenant, la norme $\left\|\Sigma_t^{-1}\right\|_F$ ainsi que $\operatorname{tr}\left(\Sigma_t^{-1}\right)$ peuvent être bornées uniformément. En effet, puisque Σ_t est une combinaison convexe de Σ et de Σ', toutes deux aux valeurs propres dans $[\lambda_{\min},\lambda_{\max}]$, alors les valeurs propres de Σ_t seront également dans le même intervalle (conséquence de Bhatia, 1997, théorème III.2.1). Ainsi,

$$\begin{aligned}
\left\|\Sigma_t^{-1}\right\|_F^2 &= \sum_{i=1}^d\lambda_i^2(\Sigma_t^{-1}) = \sum_{i=1}^d\lambda_i^{-2}(\Sigma_t) \leqslant d\lambda_{\min}^{-2}, \\
\operatorname{tr}\left(\Sigma_t^{-1}\right) &= \sum_{i=1}^d\lambda_i(\Sigma_t^{-1}) \leqslant d\lambda_{\min}^{-1}.
\end{aligned}$$

Enfin, on trouve la borne

$$\int_{\mathcal{X}} \left(\left\| \Sigma_t^{-1} \right\|_F + z^\top \Sigma_t^{-2} z \right) \varphi_{\Sigma_t}(z) \lambda(dz) \leqslant \sqrt{d} \lambda_{\min}^{-1} + d\lambda_{\min}^{-1} \leqslant 2d\lambda_{\min}^{-1}.$$

Ceci qui permet de conclure

$$\begin{aligned}
\int_{\mathcal{X}} |\varphi_\Sigma(z) - \varphi_{\Sigma'}(z)| \lambda(\mathrm{d}z) &= \int_{\mathcal{X}} \left| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} \varphi_{\Sigma_t}(z) \, \mathrm{d}t \right| \lambda(\mathrm{d}y) \\
&\leqslant \int_{\mathcal{X}} \int_0^1 \frac{1}{2} \varphi_{\Sigma_t}(z) \left| \frac{\mathrm{d}}{\mathrm{d}t} \log \varphi_{\Sigma_t}(z) \right| \mathrm{d}t \lambda(\mathrm{d}y) \\
&= \frac{1}{2} \int_0^1 \int_{\mathcal{X}} \left| \frac{\mathrm{d}}{\mathrm{d}t} \log \varphi_{\Sigma_t}(z) \right| \varphi_{\Sigma_t}(z) \lambda(\mathrm{d}y) \, \mathrm{d}t \\
&\leqslant \frac{1}{2} \int_0^1 2d\lambda_{\min}^{-1} \left\| \Sigma' - \Sigma \right\|_F \mathrm{d}t \\
&= \frac{d}{\lambda_{\min}} \left\| \Sigma' - \Sigma \right\|_F.
\end{aligned}$$

□

5.5.2.2 Poids indépendants et candidats indépendants

D'abord, il nous est possible de vérifier la condition lipschitzienne (5.13) pour un algorithme aMTM dont les candidats sont indépendants et dont les poids de sélection sont proportionnels à la densité cible.

Proposition 5.17 *Soit $\{P_\theta\}_{\theta \in \Theta}$ une famille de transitions MTM à densités instrumentales marginales gaussiennes où chacune des covariances $\Sigma^{(k)}$, $k = 1, \ldots, K$, est comprise dans un sous-ensemble $\mathcal{K}$ compact de $\mathcal{C}_d^+$, où la fonction de poids est indépendante de θ (e.g. proportionnelle à la densité cible) et où les candidats sont indépendants. Alors, la condition (5.13) est satisfaite.*

Démonstration. On note d'abord que $\bar{w}^{(k)}$ est indépendant de θ étant donné que les poids $w^{(k)}$ le sont également. De plus, la probabilité d'acceptation est aussi indépendante de θ sachant que les poids le sont aussi et que les densités marginales sont symétriques puisque gaussiennes :

$$\begin{aligned}
\alpha^{(k)}(y; y^{(-k)} | x; x_*^{(-k)}) &= \min \left\{ 1, \frac{\pi(y) Q_\theta^{(k)}(x|y) \bar{w}^{(k)}(x; x_*^{(-k)}|y)}{\pi(x) Q_\theta^{(k)}(y|x) \bar{w}^{(k)}(y; y^{(-k)}|x)} \right\} \\
&= \min \left\{ 1, \frac{\pi(y) \bar{w}^{(k)}(x; x_*^{(-k)}|y)}{\pi(x) \bar{w}^{(k)}(y; y^{(-k)}|x)} \right\}.
\end{aligned}$$

On développe donc l'expression de la différence $A_\theta^{(k)} - A_{\theta'}^{(k)}$ en utilisant ces indépendances :

$$\begin{aligned}
\left[A_\theta^{(k)} - A_{\theta'}^{(k)} \right](y|x) = \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} & \bar{w}^{(k)}(y; y^{(-k)}|x) \alpha^{(k)}(y; y^{(-k)}|x; x_*^{(-k)}) \\
& \times \Big[Q_\theta^{(-k)}(x_*^{(-k)}|y,x) Q_\theta^{(-k)}(y^{(-k)}|x,y) \\
& - Q_{\theta'}^{(-k)}(x_*^{(-k)}|y,x) Q_{\theta'}^{(-k)}(y^{(-k)}|x,y) \Big] \, \mathrm{d}y^{(-k)} \, \mathrm{d}x_*^{(-k)}.
\end{aligned}$$

Puisque $0 \leqslant \bar{w}^{(k)} \leqslant 1$ et $0 \leqslant \alpha^{(k)} \leqslant 1$, on peut borner

$$\begin{aligned}
\left| A_\theta^{(k)} - A_{\theta'}^{(k)} \right|(y|x) \leqslant \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} & \left| Q_\theta^{(-k)}(x_*^{(-k)}|y,x) Q_\theta^{(-k)}(y^{(-k)}|x,y) \right. \\
& \left. - Q_{\theta'}^{(-k)}(x_*^{(-k)}|y,x) Q_{\theta'}^{(-k)}(y^{(-k)}|x,y) \right| \mathrm{d}y^{(-k)} \, \mathrm{d}x_*^{(-k)}.
\end{aligned} \tag{5.19}$$

Puisque les candidats sont indépendants, on trouve que le produit de densités instrumentales conditionnelles

est en fait le produit de $2(K-1)$ densités gaussiennes indépendantes :

$$Q_\theta^{(-k)}(x_*^{(-k)}|y,x)Q_\theta^{(-k)}(y^{(-k)}|x,y) = \prod_{j\neq k}\varphi_{\Sigma^{(j)}}(y^{(j)}-x)\prod_{j\neq k}\varphi_{\Sigma^{(j)}}(x_*^{(j)}-y).$$

Ainsi, les valeurs propres de la covariance globale de cette distribution, données par

$$\operatorname{diag}\left(\Sigma^{(1)},\dots,\Sigma^{(k-1)},\Sigma^{(k+1)},\dots,\Sigma^{(K)},\Sigma^{(1)},\dots,\Sigma^{(k-1)},\Sigma^{(k+1)},\dots,\Sigma^{(K)}\right),$$

sont toujours bornées par les mêmes valeurs qui bornent les valeurs propres de $\Sigma^{(k)}$, $k=1,\dots,K$. Ainsi, il est possible d'appliquer la proposition 5.16 :

$$\left|A_\theta^{(k)}-A_{\theta'}^{(k)}\right|(y|x) \leqslant \frac{4d(K-1)}{\lambda_{\min}}\sum_{j\neq k}\left\|\Sigma^{(j)}-\Sigma'^{(j)}\right\|_F \leqslant \frac{4d(K-1)}{\lambda_{\min}}\left\|\theta-\theta'\right\|_2,$$

ce qui conclut la preuve. □

5.5.2.3 Poids généraux et candidats indépendants

La proposition 5.17 requiert aussi l'indépendance de la fonction de poids par rapport au paramètre θ. Cette supposition peut cependant être relâchée en modifiant la preuve. L'indépendance n'est utilisée que pour simplifier la borne (5.19), ce qui permet d'utiliser la proposition 5.16 sur les intégrales de différences de densités gaussiennes. Il est cependant possible d'exiger des conditions lipschitziennes sur les poids et sur la probabilité d'acceptation pour arriver aux mêmes conclusions.

Proposition 5.18 *Soit $\{P_\theta\}_{\theta\in\Theta}$ une famille de transitions MTM à densités instrumentales marginales gaussiennes où chacune des covariances $\Sigma^{(k)}$, $k=1,\dots,K$, est comprise dans un sous-ensemble $\mathcal{K}$ compact de $\mathcal{C}_d^+$, où les candidats sont indépendants et où les conditions lipschitziennes suivantes sont satisfaites. Il existe $L_w<\infty$ et $L_\alpha<\infty$ tels que*

$$|\bar{w}_\theta-\bar{w}_{\theta'}| \leqslant L_w\left\|\theta-\theta'\right\|_2, \qquad |\alpha_\theta-\alpha_{\theta'}| \leqslant L_\alpha\left\|\theta-\theta'\right\|_2,$$

uniformément sur les arguments $(y,y^{(-k)},x,x_*^{(-k)}) \in \mathcal{X}^{2K}$ et pour tout $(\theta,\theta') \in \Theta^2$. Alors, la condition (5.13) est satisfaite.

Démonstration. Pour alléger la notation des manipulations, on omet les arguments pour écrire

$$A_\theta^{(k)} = \iint \bar{w}_\theta^{(k)}\alpha_\theta^{(k)}Q_\theta^{(-k)}Q_\theta^{(-k)}.$$

On peut donc écrire

$$\begin{aligned}
A_\theta^{(k)}-A_{\theta'}^{(k)} &= \iint \bar{w}_\theta^{(k)}\alpha_\theta^{(k)}Q_\theta^{(-k)}Q_\theta^{(-k)} - \iint \bar{w}_{\theta'}^{(k)}\alpha_{\theta'}^{(k)}Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)} \\
&= \iint \bar{w}_\theta^{(k)}\alpha_\theta^{(k)}Q_\theta^{(-k)}Q_\theta^{(-k)} - \iint \bar{w}_\theta^{(k)}\alpha_\theta^{(k)}Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)} \\
&\quad + \iint \bar{w}_\theta^{(k)}\alpha_\theta^{(k)}Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)} - \iint \bar{w}_{\theta'}^{(k)}\alpha_{\theta'}^{(k)}Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)} \\
&= \iint \bar{w}_\theta^{(k)}\alpha_\theta^{(k)}\left(Q_\theta^{(-k)}Q_\theta^{(-k)}-Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)}\right) \\
&\quad + \iint \left(\bar{w}_\theta^{(k)}\alpha_\theta^{(k)}-\bar{w}_{\theta'}^{(k)}\alpha_{\theta'}^{(k)}\right)Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)}.
\end{aligned}$$

La première intégrale peut être bornée exactement comme dans la preuve de la proposition 5.17 :

$$\begin{aligned}\iint \left|\bar{w}_\theta^{(k)}\alpha_\theta^{(k)}\left(Q_\theta^{(-k)}Q_\theta^{(-k)} - Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)}\right)\right| &\leqslant \iint \left|Q_\theta^{(-k)}Q_\theta^{(-k)} - Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)}\right| \\ &\leqslant \frac{4d(K-1)}{\lambda_{\min}}\sum_{j\neq k}\left\|\Sigma^{(j)} - \Sigma'^{(j)}\right\|_F \\ &\leqslant \frac{4d(K-1)}{\lambda_{\min}}\left\|\theta-\theta'\right\|_2.\end{aligned}$$

Pour la seconde intégrale, on considère la décomposition suivante,

$$\begin{aligned}\bar{w}_\theta^{(k)}\alpha_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'}^{(k)} &= \bar{w}_\theta^{(k)}\alpha_\theta^{(k)} - \bar{w}_\theta^{(k)}\alpha_{\theta'}^{(k)} + \bar{w}_\theta^{(k)}\alpha_{\theta'}^{(k)} - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'}^{(k)} \\ &= \bar{w}_\theta^{(k)}\left[\alpha_\theta^{(k)} - \alpha_{\theta'}^{(k)}\right] + \left[\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right]\alpha_{\theta'}^{(k)}.\end{aligned}$$

Étant donné les deux conditions lipschitziennes, on trouve

$$\begin{aligned}\left|\bar{w}_\theta^{(k)}\alpha_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'}^{(k)}\right| &\leqslant 1\cdot\left|\alpha_\theta^{(k)} - \alpha_{\theta'}^{(k)}\right| + \left|\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right|\cdot 1 \\ &\leqslant L_\alpha\left\|\theta-\theta'\right\|_2 + L_w\left\|\theta-\theta'\right\|_2,\end{aligned}$$

et ce, uniformément sur $(y,y^{(-k)},x,x_*^{(-k)}) \in \mathcal{X}^{2K}$ et pour tout $(\theta,\theta') \in \Theta^2$. C'est donc dire que

$$\begin{aligned}\iint \left|\left(\bar{w}_\theta^{(k)}\alpha_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'}^{(k)}\right)Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)}\right| &\leqslant \iint \left|\bar{w}_\theta^{(k)}\alpha_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'}^{(k)}\right|Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)} \\ &\leqslant \iint (L_\alpha + L_w)\left\|\theta-\theta'\right\|_2 Q_{\theta'}^{(-k)}Q_{\theta'}^{(-k)} \\ &= (L_\alpha + L_w)\left\|\theta-\theta'\right\|_2,\end{aligned}$$

et donc que

$$\begin{aligned}\left|A_\theta^{(k)} - A_{\theta'}^{(k)}\right| &\leqslant \frac{4d(K-1)}{\lambda_{\min}}\left\|\theta-\theta'\right\|_2 + (L_\alpha + L_w)\left\|\theta-\theta'\right\|_2 \\ &= \left(\frac{4d(K-1)}{\lambda_{\min}} + L_\alpha + L_w\right)\left\|\theta-\theta'\right\|_2.\end{aligned}$$

□

Les conditions lipschitziennes sur les poids standardisés et la probabilité d'acceptation peuvent quant à elles être vérifiées en supposant la continuité des gradients de $\bar{w}_\theta^{(k)}$ et de $\alpha_\theta^{(k)}$, la compacité de $\mathcal{X}$ et de Θ et la convexité de Θ.

Lemme 5.19 *Soit $f : \Theta \to \mathbb{R}$ une fonction à gradient continu sur Θ compact et convexe. Alors, f est lipschitz sur Θ.*

Démonstration. On veut montrer qu'il existe $C < \infty$ tel que, pour tout $\theta,\theta' \in \Theta^2$,

$$\left|f(\theta) - f(\theta')\right| \leqslant C\cdot\left\|\theta-\theta'\right\|_2$$

Pour ce faire, on considère la combinaison convexe $\theta_t = t\theta + (1-t)\theta'$ où $t \in [0,1]$. Puisque Θ est convexe, alors $\theta_t \in \Theta$ pour tout $t \in [0,1]$. Ainsi, f admet un gradient continu en θ_t pour tout $t \in [0,1]$. On trouve donc, par le théorème fondamental du calcul,

$$\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}f(\theta_t)\,\mathrm{d}t = f(\theta_t)\Big|_0^1 = f(\theta_1) - f(\theta_0) = f(\theta) - f(\theta').$$

On peut alors considérer la borne suivante,

$$\begin{aligned}
\left|f(\theta) - f(\theta')\right| &\leqslant \left|\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} f(\theta_t)\,\mathrm{d}t\right| \\
&\leqslant \int_0^1 \left|\frac{\mathrm{d}}{\mathrm{d}t} f(\theta_t)\right| \mathrm{d}t \\
&= \int_0^1 \left|\nabla_\theta f(\theta_t) \cdot \frac{\mathrm{d}}{\mathrm{d}t}\theta_t\right| \mathrm{d}t \\
&\leqslant \int_0^1 \|\nabla_\theta f(\theta_t)\|_2 \left\|\frac{\mathrm{d}}{\mathrm{d}t}\theta_t\right\|_2 \mathrm{d}t \\
&= \int_0^1 \|\nabla_\theta f(\theta_t)\|_2 \left\|\theta' - \theta\right\|_2 \mathrm{d}t \\
&= \left\|\theta' - \theta\right\|_2 \int_0^1 \|\nabla_\theta f(\theta_t)\|_2 \,\mathrm{d}t.
\end{aligned}$$

Finalement, puisque $\nabla_\theta f$ est continu et que Θ est compact, on trouve que $\|\nabla_\theta f(\theta_t)\|_2$ est uniformément borné sur Θ par $C < \infty$, d'où

$$\left|f(\theta) - f(\theta')\right| \leqslant \left\|\theta' - \theta\right\|_2 \int_0^1 C\,\mathrm{d}t \leqslant C \cdot \left\|\theta' - \theta\right\|_2,$$

ce qui complète la preuve puisque (θ,θ') sont arbitraires. □

Proposition 5.20 *Supposons que $\bar{w}_\theta^{(k)}$ (respectivement $\alpha_\theta^{(k)}$) est une fonction continue de θ et de $(y,y^{(-k)},x)$ (respectivement de $(y,y^{(-k)},x,x_*^{(-k)})$) et que $\nabla_\theta \bar{w}_\theta^{(k)}$ (respectivement $\nabla_\theta \alpha_\theta^{(k)}$) est une fonction continue de θ pour tout $(y,y^{(-k)},x)$ (respectivement de $(y,y^{(-k)},x,x_*^{(-k)})$) fixé. Supposons de plus que $\mathcal{X}$ et Θ sont des espaces compacts et que Θ est convexe. Alors, $\bar{w}_\theta^{(k)}$ et $\alpha_\theta^{(k)}$ satisfont à la condition lipschitzienne de la proposition 5.18.*

Démonstration. Pour chaque $(y,y^{(-k)},x)$ fixé, $\bar{w}_\theta^{(k)}(y,y^{(-k)}|x)$ admet un gradient continu. Puisque Θ est compact et convexe, alors $\bar{w}_\theta^{(k)}(y,y^{(-k)}|x)$ est une fonction lipschitzienne de θ étant donné le lemme 5.19. Il existe donc $L_w(y,y^{(-k)},x) < \infty$ tel que

$$\left|\bar{w}_\theta^{(k)}(y,y^{(-k)}|x) - \bar{w}_{\theta'}^{(k)}(y,y^{(-k)}|x)\right| \leqslant L_w(y,y^{(-k)},x)\left\|\theta - \theta'\right\|_2, \forall\,(\theta,\theta') \in \Theta^2.$$

En particulier, on peut choisir

$$L_w(y,y^{(-k)},x) = \sup_{\theta\neq\theta'} \left\|\theta - \theta'\right\|_2^{-1} \left|\bar{w}_\theta^{(k)}(y,y^{(-k)}|x) - \bar{w}_{\theta'}^{(k)}(y,y^{(-k)}|x)\right|.$$

Maintenant, $L_w(y,y^{(-k)},x)$ est une fonction continue de $(y,y^{(-k)},x)$ étant donné qu'il s'agit du suprémum d'une fonction continue. Puis, vu que $\mathcal{X}$ est compact, on trouve la constante lipschitzienne nécessaire

$$L_w := \sup_{\mathcal{X}^{K+1}} L_w(y,y^{(-k)},x) < \infty,$$

qui est telle que la condition lipschitzienne de la proposition 5.18 est satisfaite. Finalement, pour ce qui est de $\alpha_\theta^{(k)}$, la preuve est identique en ajoutant seulement la dépendance en $x_*^{(-k)}$ là où nécessaire. □

5.5.2.4 Candidats extrêmement antithétiques

La dernière restriction de la proposition 5.16 sur l'algorithme aMTM est que les candidats doivent être générés indépendamment. Cette propriété est utilisée pour borner l'intégrale suivante,

$$\iint_{\mathcal{X}^{K-1}\times\mathcal{X}^{K-1}} \left|Q_\theta^{(-k)}(x_*^{(-k)}|y,x)Q_\theta^{(-k)}(y^{(-k)}|x,y) - Q_{\theta'}^{(-k)}(x_*^{(-k)}|y,x)Q_{\theta'}^{(-k)}(y^{(-k)}|x,y)\right| \mathrm{d}y^{(-k)}\,\mathrm{d}x_*^{(-k)},$$

en utilisant le fait que les densités conditionnelles sont des produits de densités gaussiennes pour des candidats indépendants. Le produit des densités gaussiennes est alors à nouveau une densité gaussienne de dimension $2(K-1)$ donc les valeurs propres sont à nouveau comprises dans un intervalle compact. Il est donc possible d'appliquer la proporition 5.16 pour borner l'intégrale.

Si les candidats sont plutôt générés par la méthode extrêmement antithétique, les densités instrumentales conditionnelles sont également gaussiennes, mais seulement sur un certain sous-espace de $\mathcal{X}^{K-1}$. Le résultat suivant, ainsi que sa preuve, contiennent des précisions quant à ce sous-espace et les propriétés de la densités conditionnelle extrêmement antithétique.

Proposition 5.21 *Soit $\Sigma_\theta^{(-k)}$ la covariance de la densité conditionnelle extrêmement antithétique où $\theta \in \Theta$. Si Θ est un sous-ensemble compact de $(\mathcal{C}_d^+)^K$, alors les valeurs propres non-nulles de $\Sigma_\theta^{(-k)}$ sont comprises dans un intervalle compact pour tout $\theta \in \Theta$.*

Démonstration. D'abord, puisque Θ est compact, il existe $\lambda_{\min} < \lambda_{\max}$ tels que toutes les valeurs propres de $\Sigma_\theta^{(k)}$ sont contenues dans l'intervalle $[\lambda_{\min},\lambda_{\max}]$ pour tout $k = 1,\dots,K$. Puis, pour la décomposition $\Sigma_\theta^{(k)} = S_\theta^{(k)} S_\theta^{(k)\top}$, on a que les valeurs propres de $S_\theta^{(k)}$ sont données par la racine carrée des valeurs propres de $\Sigma_\theta^{(k)}$, de sorte qu'elles sont comprises dans $[\sqrt{\lambda}_{\min},\sqrt{\lambda}_{\max}]$.
Pour simplifier les calculs, on supposera que $k = K$, mais toutes les étapes peuvent être reproduites en interchangeant certaines lignes ou certaines colonnes. La covariance de la densité conditionnelle est donnée par

$$\Sigma_\theta^{(-K)} := S_\theta^{(-K)} \Sigma_{EA}^{(-K)} S_\theta^{(-K)},$$

où $\Sigma_{EA}^{(-K)}$ est la matrice définie semi-positive suivante (cf. section 5.3.4.2)

$$\Sigma_{EA}^{(-K)} := (1-\rho) \begin{pmatrix} (1+\rho)I_d & \cdots & \rho I_d \\ \vdots & \ddots & \vdots \\ \rho I_d & \cdots & (1+\rho)I_d \end{pmatrix} = (1-\rho) \begin{pmatrix} 1+\rho & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1+\rho \end{pmatrix} \otimes I_d,$$

avec $\rho = -1/(K-1)$, et où

$$S_\theta^{(-K)} := \operatorname{diag}\left(S_\theta^{(1)},\dots,S_\theta^{(K-1)}\right)$$

est la matrice des racines carrées de $\Sigma_\theta^{(-1)},\dots,\Sigma_\theta^{(K-1)}$. La matrice $S_\theta^{(-K)}$ ne cause pas problème puisque ses valeurs propres sont toutes comprises dans $[\sqrt{\lambda_{\min}},\sqrt{\lambda_{\max}}]$. La matrice $\Sigma_{EA}^{(-K)}$ pose quant à elle problème puisqu'elle admet 0 comme valeur propre pour les vecteurs propres de la forme suivante

$$\mathbf{u}_j = \mathbf{1}_{K-1} \otimes \mathbf{e}_j, \qquad j = 1,\dots,d,$$

où $\mathbf{e}_j$ est le j-ième vecteur unitaire en dimension d ; par exemple,

$$\mathbf{u}_1 = \Big(\underbrace{1,0,\dots,0}_{=\mathbf{e}_1^\top}, \underbrace{1,0,\dots,0}_{=\mathbf{e}_1^\top}, \dots, \underbrace{1,0,\dots,0}_{=\mathbf{e}_1^\top} \Big)^\top.$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{K-1 \text{ fois}}$$

En effet, une ligne de $\Sigma_{EA}^{(-K)}$ contient typiquement une composante $(1+\rho)$ ainsi que $K-2$ composantes ρ toutes à intervalles de d composantes (le tout multiplié par $1-\rho$) de sorte que leur somme donne

$$(1+\rho) + (K-2)\rho = 1 + \frac{-1}{K-1} + (K-2)\frac{-1}{K-1} = 1 + (K-1)\frac{-1}{K-1} = 0.$$

On peut désormais déterminer les vecteurs propres de $\Sigma_\theta^{(-K)}$ associés à la valeur propre 0 à partir de ceux de $\Sigma_{EA}^{(-K)}$. On trouve directement que $\mathbf{v}_{\theta,j} = [S_\theta^{(-K)}]^{-1}\mathbf{u}_j$ satisfait à cette condition puisque

$$\Sigma_\theta^{(-K)} \mathbf{v}_{\theta,j} = S_\theta^{(-K)} \Sigma_{EA}^{(-K)} S_\theta^{(-K)} [S_\theta^{(-K)}]^{-1} \mathbf{u}_j = S_\theta^{(-K)} \Sigma_{EA}^{(-K)} \mathbf{u}_j = \mathbf{0}, \qquad j = 1,\dots,d.$$

C'est donc dire que la densité conditionnelle est une densité gaussienne non-singulière sur le complément orthogonal de l'espace engendré par la collection de ces vecteurs propres, $\operatorname{Vect}(\mathbf{v}_{\theta,1},\dots,\mathbf{v}_{\theta,d})$. En

termes d'algèbre linéaire, cet espace est le sous-espace nul de $\Sigma_\theta^{(-K)}$ et son complément orthogonal est alors l'image de $\Sigma_\theta^{(-K)}$, dénotée $\text{Im}(\Sigma_\theta^{(-K)})$. Maintenant, on doit s'attarder aux autres valeurs propres de $\Sigma_{EA}^{(-K)}$ afin de borner les valeurs propres non-nulles de $\Sigma_\theta^{(-K)}$. Les vecteurs propres de $\Sigma_{EA}^{(-K)}$ peuvent tous être choisis de la forme $\mathbf{a} \otimes \mathbf{e}_j$ pour un certain vecteur $\mathbf{a} \in \mathbb{R}^{K-1}$ et un $j \in \{1, \dots, d\}$ étant donné l'orthogonalité de I_d. Pour cette raison, les valeurs propres de $\Sigma_{EA}^{(-K)}$ auront toutes une multiplicité de d comme ce fut le cas pour la valeur propre 0. On se concentre donc sur la matrice

$$\begin{pmatrix} 1+\rho & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1+\rho \end{pmatrix}.$$

Par inspection, on trouve que les vecteurs de la forme $\mathbf{1}_{K-1} - (K-1)\mathbf{e}_k$ pour un $k \in \{1, \dots, K-1\}$ sont tous vecteurs propres pour la valeur propre 1. En effet,

$$\begin{aligned}
\begin{pmatrix} 1+\rho & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1+\rho \end{pmatrix} (\mathbf{1}_{K-1} - (K-1)\mathbf{e}_k) &= \begin{pmatrix} (1+\rho) + (K-2)\rho - (K-1)(\mathbb{1}(k=1)+\rho) \\ \vdots \\ (1+\rho) + (K-2)\rho - (K-1)(\mathbb{1}(k=K-1)+\rho) \end{pmatrix} \\
&= \begin{pmatrix} 1 + (1+K-2-K+1)\rho - (K-1)\,\mathbb{1}(k=1) \\ \vdots \\ 1 + (1+K-2-K+1)\rho - (K-1)\,\mathbb{1}(k=K-1) \end{pmatrix} \\
&= \begin{pmatrix} 1 - (K-1)\,\mathbb{1}(k=1) \\ \vdots \\ 1 - (K-1)\,\mathbb{1}(k=K-1) \end{pmatrix} \\
&= \mathbf{1}_{K-1} - (K-1)\mathbf{e}_k.
\end{aligned}$$

Notons que ces $K-1$ vecteurs propres sont linéairement dépendants puisque leur somme sur $k = 1, \dots, K-1$ donne le vecteur nul. Cependant, exclure n'importe quel de ces vecteurs produit un ensemble linéairement indépendant de sorte qu'on obtient $K-2$ vecteurs propres linéairement indépendants qui forment dès lors une base de l'espace propre associé à la valeur propre 1. Ceci signifie que les valeurs propres non-nulles de $\Sigma_{EA}^{(-K)}$ sont toutes données par

$$1 - \rho = 1 - \frac{-1}{K-1} = 1 + \frac{1}{K-1} = \frac{K}{K-1},$$

qui est compris dans l'intervalle (1,2] pour $K \in \{2,3,\dots\}$. Enfin, puisque $\Sigma_\theta^{(-K)}$ est le produit de trois matrices dont les valeurs propres sont respectivement comprises dans les intervalles $[\sqrt{\lambda_{\min}}, \sqrt{\lambda_{\max}}]$, (1,2] et $[\sqrt{\lambda_{\min}}, \sqrt{\lambda_{\max}}]$, alors les valeurs propres de $\Sigma_\theta^{(-K)}$ sont comprises dans l'intervalle $[\lambda_{\min}, 2\lambda_{\max}]$. □

Proposition 5.22 *Soit $\{P_\theta\}_{\theta\in\Theta}$ une famille de transitions MTM à densités instrumentales marginales gaussiennes où chacune des covariances $\Sigma^{(k)}$, $k = 1, \dots, K$, est comprise dans un sous-ensemble $\mathcal{K}$ compact de $\mathcal{C}_d^+$, où les candidats sont extrêmement antithétiques. Supposons qu'il existe $L_w < \infty$ et $L_\alpha < \infty$ tels que*

$$\left|\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right| \leqslant L_w \left\|\theta - \theta'\right\|_2, \qquad \left|\alpha_\theta^{(k)} - \alpha_{\theta'}^{(k)}\right| \leqslant L_\alpha \left\|\theta - \theta'\right\|_2,$$

uniformément sur les arguments respectifs, en prenant en compte la dépendance en θ induite par le sous-espace et pour tout $(\theta,\theta') \in \Theta^2$. Alors, la condition (5.13) est satisfaite.

Démonstration. La proposition 5.21 montre que les densités $Q_\theta^{(-k)}$ sont en fait des densités gaussiennes restreintes à un sous-espace de dimension $d(K-2)$ et dont la matrice de covariance restreinte au sous-espace est telle que ses valeurs propres sont contenues dans l'intervalle $[\lambda_{\min}, 2\lambda_{\max}]$. Cependant, pour $\theta \neq \theta'$, les sous-espaces respectifs seront différents et l'intersection sera de dimension strictement inférieure. Alors, l'approche utilisée dans la preuve de la proposition 5.17 ne peut être directement reproduite puisque la valeur absolue de la différence des produits de densités,

$$\left| Q_\theta^{(-k)}(x_*^{(-k)}|y,x) Q_\theta^{(-k)}(y^{(-k)}|x,y) - Q_{\theta'}^{(-k)}(x_*^{(-k)}|y,x) Q_{\theta'}^{(-k)}(y^{(-k)}|x,y) \right|,$$

sera égale à l'un ou l'autre des termes selon la région : l'intégration sera donc bornée par 2 et non

d'une manière lipschitzienne.

Pour y arriver, on cherche à paramétrer les deux sous-espaces d'une manière commune. Les développements de la proposition 5.21 montrent que le sous-espace associé au paramètre θ est donné par l'espace engendré par les vecteurs $S_\theta^{(-k)}\mathbf{w}_j^{(l)}$, où

$$\mathbf{w}_j^{(l)} = [\mathbf{1}_{K-1} - (K-1)\mathbf{e}_l] \otimes \mathbf{e}_j, \qquad j = 1, \ldots, d, l = 1, \ldots, K.$$

Une base de ce sous-espace peut alors être construite en retirant tous les vecteurs pour un l donné. Ainsi, on peut paramétrer le sous-espace à partir des vecteurs $\mathbf{w}_j^{(l)}$, $j = 1, \ldots, d$, et $l = 1, \ldots, K-1$. Ce sous-espace correspond en fait à l'image de la matrice $\Sigma_{EA}^{(-k)}$ que l'on dénote $\mathrm{Im}(\Sigma_{EA}^{(-k)})$. Puisque $S^{(-k)}$ est non-singulière, alors cette transformation est bijective. Pour un certain $z^{(-k)} \in \mathrm{Im}(\Sigma_{EA}^{(-k)})$, on trouve le point équivalent dans le sous-espace donné par

$$y^{(-k)} := \mathbf{1}_{K-1} \otimes y - \rho S_\theta^{(-k)}[S_\theta^{(k)}]^{-1}(y - x) + S_\theta^{(-k)} z^{(-k)}.$$

Similairement, pour les points de référence, on trouve, pour un $z_*^{(-k)} \in \mathrm{Im}(\Sigma_{EA}^{(-k)})$, on obtient

$$x_*^{(-k)} := \mathbf{1}_{K-1} \otimes x - \rho S_\theta^{(-k)}[S_\theta^{(k)}]^{-1}(x - y) + S_\theta^{(-k)} z_*^{(-k)}.$$

Ceci permet d'utiliser une variable d'intégration commune $(z^{(k)}, z_*^{(k)})$ aux deux produits de densités à condition de multiplier chacun des termes par le déterminant du Jacobien de la transformation linéaire : celui-ci est donné par $\det(S^{(-k)})^2$. Il est donc possible d'écrire la probabilité d'acceptation intégrée de la façon suivante en gardant en tête que $y^{(k)}$ et $x_*^{(-k)}$ sont fonction de x, de y, de θ, de $z^{(k)}$ et de $z_*^{(k)}$:

$$\begin{aligned}
A_\theta^{(k)}(y|x) &= \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \bar{w}_\theta^{(k)}(y; y^{(-k)}|x)\alpha_\theta(y, y^{(-k)}|x, x_*^{(-k)}) \\
&\qquad \times Q_\theta^{(-k)}(y^{(-k)}|x, y) Q_\theta^{(-k)}(x_*^{(-k)}|x, y)\, \mathrm{d}x_*^{(-k)}\, \mathrm{d}y^{(-k)} \\
&= \int_{\mathrm{Im}(\Sigma^{(-k)})} \int_{\mathrm{Im}(\Sigma^{(-k)})} \bar{w}_\theta^{(k)}(y; y^{(-k)}|x)\alpha_\theta(y, y^{(-k)}|x, x_*^{(-k)}) \\
&\qquad \times Q_\theta^{(-k)}(y^{(-k)}|x, y) Q_\theta^{(-k)}(x_*^{(-k)}|x, y)\, \mathrm{d}x_*^{(-k)}\, \mathrm{d}y^{(-k)} \\
&= \int_{\mathrm{Im}(\Sigma_{EA}^{(-k)})} \int_{\mathrm{Im}(\Sigma_{EA}^{(-k)})} \bar{w}_\theta^{(k)}(y; y^{(-k)}|x)\alpha_\theta(y, y^{(-k)}|x, x_*^{(-k)}) \left(\det(S^{(-k)})\right)^2 \\
&\qquad \times Q_\theta^{(-k)}(y^{(-k)}|x, y) Q_\theta^{(-k)}(x_*^{(-k)}|x, y)\, \mathrm{d}z_*^{(-k)}\, \mathrm{d}z^{(-k)}.
\end{aligned}$$

Soit $Q_{EA}^{(-k)}$ la densité de la distribution normale d'espérance nulle et de covariance $\Sigma_{EA}^{(-k)}$ qui est positive seulement sur $\mathrm{Im}(\Sigma_{EA}^{(-k)})$. Alors, on trouve

$$Q_\theta^{(-k)}(y^{(-k)}|x, y) Q_\theta^{(-k)}(x_*^{(-k)}|x, y) = Q_{EA}^{(-k)}(z^{(-k)}) Q_{EA}^{(-k)}(z_*^{(-k)}),$$

et donc,

$$\begin{aligned}
A_\theta^{(k)}(y|x) = \int_{\mathrm{Im}(\Sigma_{EA}^{(-k)})} \int_{\mathrm{Im}(\Sigma_{EA}^{(-k)})} &\bar{w}_\theta^{(k)}(y; y^{(-k)}|x)\alpha_\theta(y, y^{(-k)}|x, x_*^{(-k)}) \det(S^{(-k)})^2 \\
&\times Q_{EA}^{(-k)}(z^{(-k)}) Q_{EA}^{(-k)}(z_*^{(-k)})\, \mathrm{d}z_*^{(-k)}\, \mathrm{d}z^{(-k)}.
\end{aligned}$$

Étant donné que $Q_{EA}^{(-k)}$ ne dépend plus de θ, on trouve que la différence entre deux probabilités d'acceptation intégrée peut s'écrire sous une seule et même intégrale :

$$\begin{aligned}
A_\theta^{(k)}(y|x) - A_{\theta'}^{(k)}(y|x) = \int_{\mathrm{Im}(\Sigma_{EA}^{(-k)})} \int_{\mathrm{Im}(\Sigma_{EA}^{(-k)})} &\Big[\bar{w}_\theta^{(k)}(y; y^{(-k)}|x)\alpha_\theta(y, y^{(-k)}|x, x_*^{(-k)}) \det(S^{(-k)})^2 \\
&- \bar{w}_{\theta'}^{(k)}(y; y'^{(-k)}|x)\alpha_{\theta'}(y, y'^{(-k)}|x, x_*'^{(-k)}) \det(S'^{(-k)})^2\Big] \\
&\times Q_{EA}^{(-k)}(z^{(-k)}) Q_{EA}^{(-k)}(z_*^{(-k)})\, \mathrm{d}z_*^{(-k)}\, \mathrm{d}z^{(-k)},
\end{aligned}$$

où $y'^{(-k)}$ et $x_*'^{(-k)}$ correspondent à $y^{(-k)}$ et $x_*^{(-k)}$, mais en utilisant θ' dans leur calcul à partir de $z^{(-k)}$ et de $z_*^{(-k)}$ respectivement. Une condition lipschitzienne sur la fonction $A_\theta^{(k)}$ peut donc être

trouvée par une condition lipschitzienne sur la fonction

$$\bar{w}_\theta^{(k)}(y; y^{(-k)}|x)\alpha_\theta(y,y^{(-k)}|x,x_*^{(-k)})\det(S^{(-k)})^2.$$

Pour ce faire, on requiert une condition lipschitzienne pour chacun des trois termes et on note que ces trois termes sont bornés uniformément. En effet, considérons la décomposition suivante où le passage de θ à θ' s'effectue une fonction à la fois :

$$\begin{aligned}
&\bar{w}_\theta^{(k)}\alpha_\theta \det(S^{(-k)})^2 - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'} \det(S'^{(-k)})^2 \\
&\quad = \bar{w}_\theta^{(k)}\alpha_\theta \det(S^{(-k)})^2 - \bar{w}_{\theta'}^{(k)}\alpha_\theta \det(S^{(-k)})^2 \\
&\quad + \bar{w}_{\theta'}^{(k)}\alpha_\theta \det(S^{(-k)})^2 - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'} \det(S^{(-k)})^2 \\
&\quad + \bar{w}_{\theta'}^{(k)}\alpha_{\theta'} \det(S^{(-k)})^2 - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'} \det(S'^{(-k)})^2.
\end{aligned}$$

On trouve alors la borne suivante,

$$\begin{aligned}
&\left|\bar{w}_\theta^{(k)}\alpha_\theta \det(S^{(-k)})^2 - \bar{w}_{\theta'}^{(k)}\alpha_{\theta'} \det(S'^{(-k)})^2\right| \\
&\quad \leqslant \left|\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right|\alpha_\theta \det(S^{(-k)})^2 \\
&\qquad + \bar{w}_{\theta'}^{(k)}|\alpha_\theta - \alpha_{\theta'}| \det(S^{(-k)})^2 \\
&\qquad + \bar{w}_{\theta'}^{(k)}\alpha_{\theta'}\left|\det(S^{(-k)})^2 - \det(S'^{(-k)})^2\right| \\
&\quad \leqslant \left|\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right|M + |\alpha_\theta - \alpha_{\theta'}|M + \left|\det(S^{(-k)})^2 - \det(S'^{(-k)})^2\right|,
\end{aligned}$$

où $M = \sup_\Theta \det(S'^{(-k)})^2 < \infty$ puisque Θ est compact. On voit donc bien que si chacune des trois fonctions est lipschitzienne en θ alors la probabilité d'acceptation intégrée $A_\theta^{(k)}$ le sera également. Notons qu'une condition lipschitzienne sur le déterminant au carré est vérifiée via le lemme 5.19 étant donné que le gradient est continu et que $\mathcal{K}$ est compact. □

Maintenant, $\bar{w}_\theta^{(k)}$ et α_θ dépendent de θ de deux manières : d'abord dans l'expression elle-même (si la fonction de poids contient les densités instrumentales marginales) et ensuite à travers le calcul de $y^{(-k)}$ et de $x_*^{(k)}$. Une condition lipschitzienne peut alors être obtenue en montrant que le gradient par rapport à θ de ces fonctions est continu (lemme 5.19). Par composition, on requiert donc que $\bar{w}_\theta^{(k)}$ et α_θ aient un gradient continu par rapport à θ pour $y^{(-k)}$ avec $x_*^{(k)}$ fixé ainsi qu'un un gradient par rapport à $y^{(-k)}$ et à $x_*^{(k)}$ continu. Enfin, on requiert un jacobien continu dans les transformations $\theta \mapsto y^{(-k)}$ et $\theta \mapsto x_*^{(-k)}$. Lorsque θ est restreint à un espace compact, ces conditions sont vérifiées.

5.5.2.5 Candidats déterministes

Si les candidats sont déterministes comme dans les cas quasi-Monte Carlo et par variable aléatoire commune, alors les densités instrumentales conditionnelles sont dégénérées. Dans ce cas, on trouve

$$A_\theta^{(k)}(y|x) = \bar{w}_\theta^{(k)}(y; y^{(-k)}|x)\alpha_\theta^{(k)}(y; y^{(-k)}|x; x_*^{(-k)}),$$

où $y^{(-k)}$ et $x_*^{(-k)}$ sont entièrement déterminés sachant x et y et pour un θ fixé. Par contre, lorsque θ varie, alors $y^{(-k)}$ et $x_*^{(-k)}$ varient également même si x et y restent fixes. On peut donc voir $y^{(-k)}$ et $x_*^{(-k)}$ comme dépendants de θ seulement pour x et y fixés. Dans ce cas, $\bar{w}_\theta^{(k)}$ et $\alpha_\theta^{(k)}$ ne sont que des fonctions de θ pour x et y fixés.

On pourrait penser que les poids de sélection proportionnels à la densité cible simplifient la situation en retirant la dépendance en θ de $\bar{w}_\theta^{(k)}$ et de $\alpha_\theta^{(k)}$. Cependant, ce n'est pas le cas puisque $y^{(-k)}$ et $x_*^{(-k)}$ dépendent toujours de θ de sorte que la différence $A_\theta^{(k)} - A_{\theta'}^{(k)}$ n'est pas nulle. Il faut donc plutôt considérer des conditions de continuité lipschitzienne sur $\bar{w}_\theta^{(k)}$ et $\alpha_\theta^{(k)}$ par rapport à θ

pour trouver une borne lipschitzienne de $A_\theta^{(k)}$. Des conditions similaires à celles de la proposition 5.18 doivent donc être considérées.

Proposition 5.23 *Soit $\{P_\theta\}_{\theta\in\Theta}$ une famille de transitions MTM à densités instrumentales marginales gaussiennes où chacune des covariances $\Sigma^{(k)}$, $k = 1, \ldots, K$, est comprise dans un sous-ensemble $\mathcal{K}$ compact de $\mathcal{C}_d^+$ et où les candidats sont déterministes. On définit*

$$\bar{w}_\theta^{(k)}(y|x) := \bar{w}_\theta^{(k)}(y; y^{(-k)}|x),$$
$$\alpha_\theta^{(k)}(y|x) := \alpha_\theta^{(k)}(y; y^{(-k)}|x; x_*^{(-k)}),$$

où $y^{(-k)}$ et $x_*^{(-k)}$ sont entièrement déterminés par x, y et θ fixés. Supposons qu'il existe $L_w < \infty$ et $L_\alpha < \infty$ tels que

$$\left|\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right| \leqslant L_w \left\|\theta - \theta'\right\|_2, \qquad \left|\alpha_\theta^{(k)} - \alpha_{\theta'}^{(k)}\right| \leqslant L_\alpha \left\|\theta - \theta'\right\|_2,$$

uniformément sur les arguments $(y,x) \in \mathcal{X}^2$ et pour tout $(\theta,\theta') \in \Theta^2$. Alors, la condition (5.13) est satisfaite.

Démonstration. La preuve est sensiblement la même que celle de la proposition 5.18 à l'exception qu'aucune intégration n'est requise. En omettant les arguments x et y, on écrit

$$A_\theta^{(k)} = \bar{w}_\theta^{(k)} \alpha_\theta^{(k)}.$$

Alors, on a la décomposition

$$\begin{aligned} A_\theta^{(k)} - A_{\theta'}^{(k)} &= \bar{w}_\theta^{(k)} \alpha_\theta^{(k)} - \bar{w}_{\theta'}^{(k)} \alpha_{\theta'}^{(k)} \\ &= \bar{w}_\theta^{(k)} \alpha_\theta^{(k)} - \bar{w}_\theta^{(k)} \alpha_{\theta'}^{(k)} + \bar{w}_\theta^{(k)} \alpha_{\theta'}^{(k)} - \bar{w}_{\theta'}^{(k)} \alpha_{\theta'}^{(k)} \\ &= \bar{w}_\theta^{(k)} \left(\alpha_\theta^{(k)} - \alpha_{\theta'}^{(k)}\right) + \left(\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right) \alpha_{\theta'}^{(k)}. \end{aligned}$$

Puisque $\bar{w}_\theta^{(k)}$ et $\alpha_\theta^{(k)}$ sont tous deux compris entre 0 et 1, on trouve directement la borne suivante,

$$\begin{aligned} \left|A_\theta^{(k)} - A_{\theta'}^{(k)}\right| &\leqslant \left|\bar{w}_\theta^{(k)}\right| \left|\alpha_\theta^{(k)} - \alpha_{\theta'}^{(k)}\right| + \left|\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right| \left|\alpha_{\theta'}^{(k)}\right| \\ &\leqslant \left|\alpha_\theta^{(k)} - \alpha_{\theta'}^{(k)}\right| + \left|\bar{w}_\theta^{(k)} - \bar{w}_{\theta'}^{(k)}\right|. \end{aligned}$$

Les deux conditions lipschitziennes donnent finalement

$$\left|A_\theta^{(k)} - A_{\theta'}^{(k)}\right| \leqslant L_w \left\|\theta - \theta'\right\|_2 + L_\alpha \left\|\theta - \theta'\right\|_2 = (L_w + L_\alpha) \left\|\theta - \theta'\right\|_2, \qquad \forall\, x,y \in \mathcal{X}^2\,, (\theta,\theta') \in \Theta^2.$$

Le reste de la preuve est identique en utilisant cette borne plutôt que celle originale. □

La vérification des deux conditions lipschitziennes de la proposition 5.23 est malheureusement plus complexe que dans le cas des candidats indépendants. En effet, la dépendance de $\bar{w}_\theta^{(k)}$ et de $\alpha_\theta^{(k)}$ en θ est plus compliquée étant donné qu'elle se situe au niveau de la fonction qui relie $y^{(-k)}$ et $x_*^{(-k)}$ respectivement à x et à y. Pour des densités marginales gaussiennes, cette transformation prend la forme de transformations linéaires. Par exemple, sachant x et $y = y^{(k)}$ et pour des candidats par variable aléatoire commune, on trouve $z = [S^{(k)}]^{-1}(y - x)$ puis $y^{(j)} = x + S^{(j)}z$ pour $j \neq k$. Alors, les autres candidats sont donnés par

$$y^{(j)} = x + S^{(j)}[S^{(k)}]^{-1}(y - x), \qquad j \neq k,$$

où la dépendance en θ est donc via les matrices $S^{(j)}[S^{(k)}]^{-1}$, $j \neq k$. Selon le choix de fonction de poids, $y^{(-k)}$ et $x_*^{(-k)}$ apparaîtront à divers endroits. Notamment, la fonction de poids et la fonction d'acceptation comprennent généralement des termes $\pi(y^{(j)})$ et $\pi(x_*^{(j)})$, $j \neq k$. Dans ce cas, les gradients

de $\bar{w}_\theta^{(k)}$ et de $\alpha_\theta^{(k)}$ comprendront des termes tels que

$$\left(D_\theta y^{(j)}\right)^\top \nabla_z \pi(z)\Big|_{z=y^{(j)}},$$

où $D_\theta y^{(j)}$ dénote le jacobien de la fonction $\theta \mapsto y^{(j)}$ (pour x et y fixés.) Lorsque les matrices $\Sigma^{(k)}$, $k = 1, \ldots, K$, sont définies positives, il est possible de montrer que le jacobien est une fonction continue de θ de sorte qu'on ne requiert que la continuité de $\nabla\pi$. Similairement, des termes tels que $Q_\theta^{(j)}(y^{(j)}|x)$ peuvent faire partie de l'expression de $\bar{w}_\theta^{(k)}$: dans ce cas, on nécessite la continuité de $\nabla_z Q_\theta^{(j)}(z|x)$ qui est facilement vérifiée pour une densité gaussienne.

5.5.3 Suppléments à la section 5.4.1.4

Afin de compléter la démonstration de la condition d'adaptation diminuante 3.1 de l'algorithme aMTM, la vérification de la condition (5.12) est essentielle. Cette condition ne traite que de l'adaptation des paramètres : seul le choix de fonction de mise à jour et de fréquence d'adaptation importera. C'est donc dire que les résultats qui suivent seront généraux par rapport à la fonction de poids et à la structure de corrélation.

Les différentes méthodes d'adaptation proposées à la section 5.3 exigent des traitements distincts. Dans tous les cas, l'adaptation est définie par deux composantes : la fréquence d'adaptation, qui décrit quels paramètres sont adaptés, et la fonction de mise à jour, qui décrit comment les paramètres adaptés sont calculés. Afin de vérifier toutes les combinaisons possibles, on introduit la notation suivante qui permettra un traitement plus unifié par la suite. Lors que la n-ième itération MCMC, la chaîne se trouve à l'état x_n. À l'étape de sélection MTM, on définit K_n la variable aléatoire prenant des valeurs dans $\{1, \ldots, K\}$ avec probabilité $\bar{w}_\theta^{(k)}(y; y^{(-k)}|x)$, $k = 1, \ldots, K$, c'est-à-dire $K_n = k_n$ lorsque le candidat k_n est sélectionné. Afin de couvrir le cas où le taux de sélection est utilisé pour adapter le paramètre d'échelle, on introduit les paramètres $s^{(k)} \in [0,1]$, $k = 1, \ldots, K$, correspondant au taux de sélection courant du k-ième candidat. Ces derniers sont maintenus par les récursions suivantes :

$$s_{n+1}^{(k)} = s_n^{(k)} + \gamma_{n+1}\left(\mathbb{1}(\{k_n = k\}) - s_n^{(k)}\right), \qquad k = 1, \ldots, K. \tag{5.20}$$

Lorsqu'un paramètre d'échelle $\lambda^{(k)}$ est utilisé et adapté, on considérera plutôt les récursions sur le logarithme $l^{(k)} := \log \lambda^{(k)}$ afin de simplifier les expressions et d'assurer la positivité du paramètre d'échelle. Ainsi, le paramètre adapté complet est défini par

$$\theta = \left(\theta^{(1)}, \ldots, \theta^{(K)}\right),$$

où

$$\theta^{(k)} = \left(\mu^{(k)}, \Sigma^{(k)}, l^{(k)}, s^{(k)}\right), \qquad k = 1, \ldots, K.$$

Selon les spécificités de l'adaptation, un ou plusieurs des éléments de $\theta^{(k)}$ ne sont pas adaptés. Par exemple, si le taux de sélection n'est pas utilisé pour mettre à jour l'échelle, alors $s^{(k)}$ n'est pas adapté. Similairement, si les mises à jour AM sont utilisées, le paramètre $l^{(k)}$ n'est pas adapté. Pour conserver une uniformité dans la notation, le traitement des différents cas utilisera ces définitions : lorsqu'un paramètre n'est pas adapté, on aura simplement que la portion de H_θ adaptant ce paramètre

sera nulle et donc directement bornée. Finalement, il est possible de décrire la forme générale de H_θ : cette fonction dépendra, selon le cas, certaines des informations produites au cours de l'échantillonnage MTM, i.e. $(y,y^{(-k)},x_*^{(-k)},k_n,x_{n+1})$, ainsi que de l'état actuel x_n. On dénote l'ensemble de ces informations par $\Xi_n = \left(x_n,x_{n+1},k_n,y,y^{(-k)},x_*^{(-k)}\right)$. On écrit donc généralement

$$H_\theta\left(\Xi_n\right) = \begin{pmatrix} H_\theta^{(1)}\left(\Xi_n\right) \\ \vdots \\ H_\theta^{(K)}\left(\Xi_n\right) \end{pmatrix},$$

où $H_\theta^{(k)}$ correspond à l'adaptation des paramètres de $\theta^{(k)}$. Puisque les paramètres $\mu^{(k)}$ et $\Sigma^{(k)}$ sont adaptés simultanément (soit les deux sont adaptés conjointement ou soit seul $\Sigma^{(k)}$ est adapté), on trouve la forme générale suivante de $H_\theta^{(k)}$:

$$H_\theta^{(k)}\left(\Xi_n\right) = \begin{pmatrix} H_{\mu,\Sigma}^{(k)}\left(\Xi_n\right) \\ H_{l,\alpha}^{(k)}\left(\Xi_n\right) + H_{l,s}^{(k)}\left(\Xi_n\right) \\ H_s^{(k)}\left(\Xi_n\right) \end{pmatrix}$$

où $H_{\mu,\Sigma}^{(k)}$ correspond à l'une ou l'autre des fonctions de mise à jour de la covariance, où $H_{l,\alpha}^{(k)}$ correspond à la mise à jour de $l^{(k)}$ via la probabilité d'acceptation (mise à jour ASWAM), où $H_{l,s}^{(k)}$ correspond à la mise à jour de $l^{(k)}$ via le taux de sélection lorsque le k-ième candidat n'est pas sélectionné (algorithme 5.7) et où $H_s^{(k)}$ correspond à la mise à jour du taux de sélection par (5.20). Tel que mentionné, certaines de ces fonctions seront identiquement nulles selon les particularités de l'algorithme. Afin de borner uniformément H_θ, il suffit donc de borner chacune de ces fonctions, ce qui peut être fait indépendamment l'une de l'autre.

Les fonctions $H_{l,\alpha}^{(k)}$, $H_{l,s}^{(k)}$ et $H_s^{(k)}$ se bornent aisément. Pour les mises à jour ASWAM, on trouve

$$H_{l,\alpha}^{(k)}\left(\Xi_n\right) = \mathbb{1}\left(\{k_n = k\}\right)\left[\alpha_\theta\left(y;y^{(-k)}|x_n;x_*^{(-k)}\right) - \alpha_*\right]$$

qui est borné par

$$\left|H_{l,\alpha}^{(k)}\left(\Xi_n\right)\right| \leqslant \mathbb{1}\left(\{k_n = k\}\right)\left|\alpha_\theta\left(y;y^{(-k)}|x_n;x_*^{(-k)}\right) - \alpha_*\right| \leqslant 1.$$

Dans tous les autres cas, $H_{l,\alpha}^{(k)} = 0$. Lorsque le taux de sélection est utilisé, on a

$$H_{l,s}^{(k)}\left(\Xi_n\right) = \mathbb{1}\left(\{k_n \neq k\}\right)\left[s_n^{(k)} - s_*\right]$$

qui est borné par

$$\left|H_{l,s}^{(k)}\left(\Xi_n\right)\right| \leqslant \mathbb{1}\left(\{k_n \neq k\}\right)\left|s_n^{(k)} - s_*\right| \leqslant 1.$$

De plus,

$$H_s^{(k)}\left(\Xi_n\right) = \mathbb{1}\left(\{k_n = k\}\right) - s_n^{(k)}$$

qui est borné par

$$\left|H_s^{(k)}\left(\Xi_n\right)\right| = \left|\mathbb{1}\left(\{k_n = k\}\right) - s_n^{(k)}\right| \leqslant 1.$$

Dans tous les autres cas, $H_{l,s}^{(k)} = 0$ et $H_s^{(k)} = 0$.

Le travail principal à effectuer est donc sur les fonctions $H_{\mu,\Sigma}^{(k)}$ puisque celles-ci ont des images dépendantes de Ξ_n et de θ et puisque plusieurs déclinaisons existent. Les différents cas à traiter sont les suivants : les mises à jours AM ou ASWAM (toutes deux ont la même fonction $H_{\mu,\Sigma}^{(k)}$), leurs équivalents locaux, les mises à jour RAM et l'utilisation d'une composante globale (algorithme 5.6).

L'utilisation d'une composante globale signifie seulement que jusqu'à deux covariances sont adaptées à chaque itération plutôt qu'une seule. Pour chaque, $k = 2, \ldots, K$ ceci ne produit aucune différence ; pour $k = 1$, ceci a pour effet de retirer la condition d'adaptation $\mathbb{1}\left(\{k_n = k\}\right)$. La borne résultant n'est donc pas affectée par l'utilisation d'une composante globale. Les versions locales des mises à jour AM et ASWAM se traitent d'une manière pratiquement identique à la version originale de sorte qu'un même résultat peut être produit pour les deux situations.

Lemme 5.24 *Soit $H_{\mu,\Sigma}^{(k)}$ la fonction de mise à jour AM, ASWAM ou leurs versions locales des paramètres $(\mu^{(k)},\Sigma^{(k)})$. Alors,*

$$\sup_{\theta \in \mathcal{K}} \left\| H_{\mu,\Sigma}^{(k)} \right\|_1 < \infty. \tag{5.21}$$

Démonstration. Les mises à jour AM et ASWAM sont telles que

$$H_{\mu,\Sigma}^{(k)}\left(\Xi_n\right) = \mathbb{1}(\{k_n = k\}) \begin{pmatrix} x_{n+1} - \mu^{(k)} \\ (x_{n+1} - \mu^{(k)})(x_{n+1} - \mu^{(k)})^\top - \Sigma^{(k)} \end{pmatrix}.$$

Ainsi, $H_{\mu,\Sigma}^{(k)}$ ne dépend que de θ, de k_n et de x_{n+1}.

Par définition, on a

$$\left\| H_{\mu,\Sigma}^{(k)} \right\|_1 = \sup_{(x_{n+1},k_n) \in \mathcal{X} \times \{1,\ldots,K\}} \left\| H_{\mu,\Sigma}^{(k)}\left(\Xi_n\right) \right\|_2.$$

On note que le suprémum sur $k_n \in \{1, \ldots, K\}$ sera atteint pour $k_n = k$ étant donné le terme $\mathbb{1}\left(\{k_n = k\}\right)$. Alors, on trouve la borne suivante

$$\begin{aligned} \left\| H_{\mu,\Sigma}^{(k)} \right\|_2 &\leqslant \left\| x_{n+1} - \mu^{(k)} \right\|_2 + \left\| (x_{n+1} - \mu^{(k)})(x_{n+1} - \mu^{(k)})^\top - \Sigma^{(k)} \right\|_F \\ &\leqslant \left\| x_{n+1} \right\|_2 + \left\| \mu^{(k)} \right\|_2 + \left\| x_{n+1} x_{n+1}^\top \right\|_F + 2\left\| \mu^{(k)} x_{n+1}^\top \right\|_F + \left\| \mu^{(k)} \mu^{(k)\top} \right\|_2 + \left\| \Sigma^{(k)} \right\|_F. \end{aligned} \tag{5.22}$$

Puisque $\mathcal{X}$ et $\mathcal{K}$ sont compacts, alors x_{n+1}, $\mu^{(k)}$ et $\Sigma^{(k)}$ sont bornés de sorte que $\left\| H_{\mu,\Sigma}^{(k)} \right\|_2$ est uniformément borné sur $(x,k) \in \mathcal{X} \times \{1, \ldots, K\}$ et sur $\theta \in \mathcal{K}$.

Pour ce qui est des mises à jour locales, on doit substituer x_n à $\mu^{(k)}$ dans les derniers développements et considérer le suprémum sur $x_n \in \mathcal{X}$ en plus de sur x_n et sur k_n. On trouvera la même borne que celle dans l'expression (5.22) en substituant x_n à $\mu^{(k)}$; dans ce cas la compacité de $\mathcal{X}$ permet à nouveau de trouver une borne uniforme. □

Pour ce qui est des mises à jour RAM, on réécrit (5.7) selon une expression du type récursion de

Robbins-Monro :

$$\begin{aligned}\Sigma_{n+1}^{(k)} &= S_n^{(k)}\left(I_d + \gamma_{n+1}\,\mathbb{1}\left(\{k_n = k\}\right)\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right)S_n^{(k)\top}\\ &= S_n^{(k)}S_n^{(k)\top} + \gamma_{n+1}\,\mathbb{1}\left(\{k_n = k\}\right)S_n^{(k)}\left(\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right)S_n^{(k)\top}\\ &=: \Sigma_n^{(k)} + \gamma_{n+1}H_{\Sigma_n}^{(k)}\left(\Xi_n\right),\end{aligned}$$

où

$$H_{\Sigma_n}^{(k)}\left(\Xi_n\right) = \mathbb{1}\left(\{k_n = k\}\right)S_n^{(k)}\left(\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right)S_n^{(k)\top},$$

avec $u_{n+1} = y - x_n$ et $\Sigma^{(k)} = S^{(k)}S^{(k)\top}$.

Lemme 5.25 *Soit $H_\Sigma^{(k)}$ la fonction de mise à jour RAM de $\Sigma^{(k)}$. Alors,*

$$\sup_{\theta\in\mathcal{K}}\left\|H_\Sigma^{(k)}\right\|_1 < \infty. \tag{5.23}$$

Démonstration. La norme de $H_\Sigma^{(k)}$ peut être bornée ainsi :

$$\begin{aligned}\left\|H_\Sigma^{(k)}\left(\Xi_n\right)\right\|_2 &\leqslant \left\|S^{(k)}\left(\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right)S^{(k)\top}\right\|_2\\ &\leqslant \left\|S^{(k)}\right\|_2\left\|\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right\|_2\left\|S^{(k)\top}\right\|_2\\ &\leqslant \left\|S^{(k)}\right\|_2\frac{\left\|u_{n+1}u_{n+1}^\top\right\|_2}{\|u_{n+1}\|_2^2}\left\|S^{(k)}\right\|_2\\ &\leqslant \left\|S^{(k)}\right\|_2\frac{\|u_{n+1}\|_2^2}{\|u_{n+1}\|_2^2}\left\|S^{(k)}\right\|_2 = \left\|S^{(k)}\right\|_2^2.\end{aligned}$$

Puisque cK est compact, alors $\left\|S^{(k)}\right\|_2$ est uniformément borné, ce qui implique que $\left\|H_\Sigma^{(k)}\right\|_1$ est uniformément borné sur $\theta \in \mathcal{K}$ ainsi que pour tout Ξ_n. C'est donc dire que $\sup_{\theta\in\mathcal{K}}\left\|H_\Sigma^{(k)}\right\|_1 < \infty$. □

Chapitre 6

Études numériques

Dans ce chapitre, on s'interroge sur la performance empirique de l'algorithme aMTM. Pour ce faire, la procédure est implémentée dans un *package* `R` présenté à la section 6.1, qui contient également un exemple d'utilisation. Ensuite, à la section 6.2, on effectue un rappel et une discussion des différentes mesures de performance présentées à la section 2.4.3. Puis, à l'aide de trois distributions cibles aux caractéristiques différentes, on étudie empiriquement le comportement de l'algorithme appliqué à diverses situations (section 6.3). Enfin, on discute de l'ensemble des résultats à la section 6.4.

6.1 Implémentation

6.1.1 Description du progiciel `aMTM`

L'algorithme aMTM décrit à la section 5.3 est implémenté dans un progiciel `R` (R Core Team, 2013) du même nom qui est disponible sur GitHub via la commande `R` suivante [1] :

```
devtools::install_github("fontaine618/aMTM/aMTM")
```

Le *package* `aMTM` contient trois fonctions. La première, `aMTM`, est la fonction principale qui effectue l'échantillonnage selon l'algorithme aMTM. Toutes les déclinaisons décrites à la section 5.3 sont implémentées parmi les options de la fonction. Il est possible de spécifier la valeur initiale de la chaîne, les paramètres initiaux des différentes densités instrumentales, le type de candidats, le type de poids, les particularités de l'adaptation (fréquence, fonction de mise à jour, pas d'adaptation, taux d'acceptation cible) et l'utilisation d'une période de chauffe. Cette fonction est en fait l'habillage d'une fonction écrite en `C++`, appelée `aMTMsample`, qui effectue l'échantillonnage et l'adaptation. Bien que cette dernière ne puisse seulement évaluer des densités cibles écrites en `R`, la structure du code permet des généralisations faciles à d'autres langages. Au moment d'effectuer l'adaptation des matrices de covariance, la fonction `aMTMsample` fait appel à une fonction écrite par Helske (2016) qui effectue une **mise à jour de Cholesky de rang un**.

1. En supposant le progiciel `devtools` (Wickham et collab., 2018) préalablement installé.

Définition 6.1 (Mise à jour de Cholesky de rang un) *Soit $\Sigma \in \mathcal{C}_d^+$ une matrice $d \times d$ symétrique et définie positive et soit $u \in \mathbb{R}^d$. Alors, les matrices $\Sigma' = \Sigma + uu^\top$ ou $\Sigma' = \Sigma - uu^\top$ admettent chacune une décomposition de Cholesky $\Sigma' = S'S'^\top$ qui peut être calculée à partir de la décomposition de Cholesky de Σ donnée par $\Sigma = SS^\top$.*

L'algorithme de Helske (2016) permet le calcul de S' à partir de S et de u sans passer par Σ ni par Σ', ce qui accélère les calculs. De plus, l'échantillonnage s'effectue plus aisément étant donné le facteur S que la covariance Σ (voir tableau 5.1). Les mises à jour AM, ASWAM (et leurs versions locales) et RAM peuvent toutes être réécrites dans le format d'une mise à jour de Cholesky de rang un. Par exemple, les mises à jour AM et ASWAM correspondent à

$$\begin{aligned}
\Sigma' &= \Sigma + \gamma\left[(x-\mu)(x-\mu)^\top - \Sigma\right] \\
&= (1-\gamma)\Sigma + \gamma(x-\mu)(x-\mu)^\top, \\
\Rightarrow \qquad \frac{1}{1-\gamma}\Sigma' &= \left(\sqrt{\frac{1}{1-\gamma}}S'\right)\left(\sqrt{\frac{1}{1-\gamma}}S'\right)^\top \\
&= \Sigma + \frac{\gamma}{1-\gamma}(x-\mu)(x-\mu)^\top,
\end{aligned}$$

où l'on peut mettre en relation S'', S', S et u de la façon suivante

$$S''S''^\top = \left(\sqrt{\frac{1}{1-\gamma}}S'\right)\left(\sqrt{\frac{1}{1-\gamma}}S'\right)^\top = SS^\top + uu^\top = SS^\top + \left(\sqrt{\frac{\gamma}{1-\gamma}}(x-\mu)\right)\left(\sqrt{\frac{\gamma}{1-\gamma}}(x-\mu)\right)^\top. \tag{6.1}$$

L'algorithme produit S'', la décomposition de $\Sigma'/(1-\gamma)$, ce qui permet alors de trouver la décomposition de Cholesky de $\Sigma' = S'S'^\top$ par le calcul de $S' = S''\sqrt{1-\gamma}$. Les versions locales sont semblables en substituant μ par l'état actuel de la chaîne dans (6.1). La mise à jour RAM peut être écrite de la façon suivante :

$$\begin{aligned}
SS' &= S\left(I_d - \gamma(\alpha(y|x)-\alpha_*)\frac{uu^\top}{\|u\|_2^2}\right)S^\top \\
&= SS^\top - \gamma(\alpha(y|x)-\alpha_*)\frac{Suu^\top S^\top}{\|u\|_2^2} \\
&= SS^\top - \operatorname{sgn}(\alpha(y|x)-\alpha_*)\gamma|\alpha(y|x)-\alpha_*|\frac{Suu^\top S^\top}{\|u\|_2^2} \\
&= SS^\top \pm vv^\top,
\end{aligned}$$

où

$$v = \frac{\sqrt{\gamma|\alpha(y|x)-\alpha_*|}}{\|u\|_2}Su.$$

Lorsque la mise à jour de rang un est positive, c.-à-d., $\Sigma' = \Sigma + uu^\top$, alors la matrice résultante est systématiquement définie positive. Lorsque la mise à jour de rang un est négative, c.-à-d., $\Sigma' = \Sigma - uu^\top$, alors la matrice résultante peut ne plus être définie positive. Cependant, seule la mise à jour RAM utilise ce genre de mise à jour et la matrice Σ' sera toujours définie positive par construction lorsque $\gamma \in [0,1)$.

L'algorithme aMTM contient des algorithmes plus simples comme cas spéciaux. En effet, lorsque $K = 1$ on retrouve les algorithmes AM, ASWAM et RAM. De plus, lorsqu'aucune adaptation

n'est effectuée, on retrouve les différentes variantes de l'algorithme MTM. Étant donné que tous ces algorithmes sont compris dans la même implémentation, il est alors possible de comparer les temps de calculs respectifs sans ce soucier de l'optimalité de l'implémentation.

L'implémentation de la fonction `aMTM` s'assure de la compatibilité de la sortie avec d'autres *packages* `R` couramment utilisés. En particulier, la chaîne produite par l'algorithme est de classe `mcmc` afin de pouvoir utiliser les fonctionnalités du *package* `coda` (Plummer et collab., 2006).

La seconde fonction du package `aMTM` est la fonction `plot.aMTM` qui permet certains affichages de la chaîne produite par l'algorithme. D'abord, cette fonction permet d'afficher le tracé de la chaîne dans chacune des d composantes accompagnée de la densité empirique lissée associée. Ensuite, cette fonction permet d'afficher les graphiques en nuage de toutes les paires de composantes de la chaîne. Il est également possible de superposer les densités instrumentales finales ou de colorer les points selon la proposition choisie afin d'étudier le comportement de l'adaptation.

La troisième fonction du package `aMTM` est la fonction `stats.aMTM` qui permet le calcul de certaines statistiques associées à la chaîne résultante : l'ESEJD et l'ESJD (section 2.4.3.3), la norme de Frobenius de l'ACT de la chaîne (c.-à-d., pour la fonction identité) et l'ESS multivarié (section 2.4.3.4).

6.1.2 Exemple d'utilisation

Comme premier exemple d'utilisation de l'algorithme aMTM, nous revenons sur l'exemple de la distribution bimodale en dimension $d = 2$ présenté à la section 5.1. La densité cible, donnée par

$$\pi = w_1 \mathcal{N}_2(\mu_1,\Sigma_1) + w_2 \mathcal{N}_2(\mu_2,\Sigma_2),$$

$$\begin{aligned} w_1 &= 0.3, & \mu_1 &= (0,20)^\top, & \Sigma_1 &= \text{diag}(9,1); \\ w_2 &= 0.7, & \mu_2 &= (8,0)^\top, & \Sigma_2 &= \text{diag}(1,9), \end{aligned}$$

peut être programmée de la façon suivante, en notant que la fonction `aMTM` accepte la log-densité cible accompagnée de certains paramètres déterminant la densité [2] :

```
#les deux moyennes
mu1 <- c(20,0); mu2 <- c(0,8)
#les deux matrices de covariances
Sig1 <- matrix(c(9,0,0,1),2,2,T)
Sig2 <- matrix(c(1,0,0,9),2,2,T)
#le calcul des matrices inverses
S1 <- solve(Sig1); S2 <- solve(Sig2)
#le calcul des termes constants
d1 <- (det(2*3.1416*Sig1))^-0.5
d2 <- (det(2*3.1416*Sig2))^-0.5
#les poids du melange
w1 <- 0.3; w2 <- 1-w1
#la fonction de log-densite vectorisee
logp <- function(X, p){
   if(is.vector(X)) X <-t(X)
   X <- as.matrix(X)
   apply(X, 1, function(x){
```

2. Tout le code inclus dans cette section se trouve dans le *package* `aMTM` à l'adresse suivante : https://github.com/fontaine618/aMTM/blob/master/aMTM/tests/motivation.R

```
     log(p$w1*p$d1*exp(-t(x-p$mu1)%*%p$S1%*%(x-p$mu1)/2)+
         p$w2*p$d2*exp(-t(x-p$mu2)%*%p$S2%*%(x-p$mu2)/2))
   })
}
#la liste des parametres
p <- list(w1=w1,w2=w2,d1=d1,d2=d2,mu1=mu1,mu2=mu2,S1=S1,S2=S2)
```

L'algorithme aMTM comporte plusieurs options qui doivent être spécifiées. On choisit les mises à jour ASWAM locales sans adaptation globale ni de l'échelle ; le pas d'adaptation est donné par $n^{-0.7}$; les candidats seront indépendants ; le taux d'acceptation cible sera de $\alpha_* = 0.50$; les poids de sélection seront proportionnels à la densité cible (certains de ces choix sont en fait les valeurs par défaut de la fonction `aMTM`). Ces choix sont pour l'instant relativement arbitraires, mais une étude plus approfondie est présentée à la section 6.3. La chaîne est initialisée à l'état $x_0 = (0{,}10)^\top$ et les covariances initiales des $K = 2$ propositions sont des multiples différents de la matrice identité. La chaîne sera d'une longueur de $N = 10{,}000$ afin de comparer les résultats avec ceux de la section 5.1.

```
K<-2; d<-2
#covariances initiales
sig0 = array(0, dim = c(d,d,K))
for(k in seq(K)) sig0[,,k] <- diag(d)*10^k
#echantillonnage aMTM
library(aMTM)
set.seed(1)
mcmc <- aMTM(target=logp, N=1e4, K=K, x0=c(0,0), parms=p,
             sig0 = sig0, adapt = 2, local = T, accrate = 0.5)
# statistiques de l'echantillon
mean(mcmc$X[,1]>5)
# [1] 0.2897
mcmc$acc.rate
# [1] 0.5016
mcmc$sel.prop
#      1      2
# 0.9800 0.0200
```

On peut inspecter certaines des propriétés de la chaîne produite afin d'étudier le comportement de l'algorithme. Par exemple, on trouve que la chaîne évalue la proportion de points avec une première composante supérieure à 5 à 0,2897, ce qui est près de la valeur exacte donnée par 0,3. En fait, seuls des échantillons i.i.d. directement de la densité cible(0,3027, figure 5.1) et MTM à $K = 2$ propositions bien ajustées (0,3097, figure 5.1) ont une estimation du poids du deuxième mode plus près de la réalité. De plus, le taux d'acceptation (0,5016) s'approche sensiblement de la probabilité recherchée (0,5), indiquant que l'algorithme peut atteindre le taux recherché tout en trouvant les deux modes, contrairement à l'algorithme AM (voir e.g. figure 5.1 (c) et (d)). Les résultats obtenus sont plutôt de l'ordre de ceux des algorithmes MTM à $K = 2$ ou $K = 3$ (respectivement figure 5.1 (e) et (f)), mais sans la nécessité de spécifier parfaitement les paramètres des densités instrumentales.

Ensemble, ces deux caractéristiques montrent déjà une certaine supériorité de l'algorithme aMTM par rapport aux algorithmes utilisés à la section 5.1 : l'algorithme produit les bons poids des modes et a un taux d'acceptation élevé et près de l'optimalité. En effet, les résultats d'échelle optimale des algorithmes MTM (tableau 4.1) montrent qu'asymptotiquement le taux d'acceptation devrait être de 0,46 pour $K = 2$ propositions et les résultats d'échelle optimale en dimension finie (tableau 2.1) montrent que le taux optimal augmente lorsque d diminue vers 1. Évidemment, ces résultats ne sont pas nécessairement valides pour la distribution cible en question, mais on peut tout de même les utiliser pour guider l'optimisation de l'algorithme.

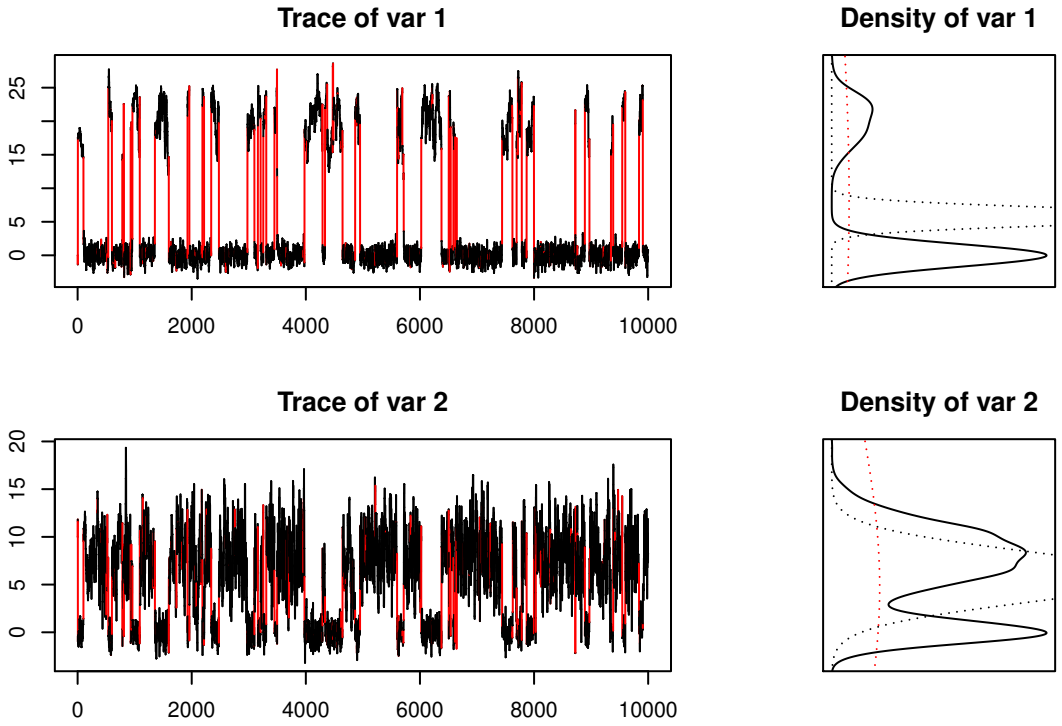


Figure 6.1 *Sorties unidimensionnelles de la chaîne produite par l'algorithme aMTM avec $K = 2$ propositions pour l'exemple bimodal en $d = 2$ dimensions. De haut en bas, on trouve les deux dimensions de l'espace d'état. Dans la colonne de gauche, on trouve le graphique des tracés de la chaîne : en noir, on trouve les sauts issus de la proposition 1 et, en rouge, on trouve les sauts issus de la proposition 2. Dans la colonne de droite, on trouve la densité empirique lissée (ligne pleine) ainsi que les densités instrumentales finales marginales (lignes pointillées noire et rouge).*

Un algorithme à une seule proposition, même adaptatif, ne pourra jamais être simultanément efficace relativement à ces deux aspects puisqu'une augmentation du taux d'acceptation empêchera systématiquement la découverte du second mode. La flexibilité des essais multiples est donc nécessaire dans ce cas et l'adaptation permet de trouver un ensemble de paramètres relativement optimal, sans connaître les propriétés de la densité cible.

Ensuite, les proportions de sélection montrent que la proposition 1 est sélectionnée dans 98% des itérations et que la seconde proposition n'est sélectionnée que dans 2% des itérations. En inspectant les figures 6.1 et 6.2, qui affichent respectivement les tracés uni- et bidimensionnels de la chaîne, on s'aperçoit que la seconde proposition s'assure de produire des sauts entre les deux modes alors que la première proposition effectue l'échantillonnage local. Il s'agit en fait du comportement souhaité de l'algorithme et ce résultat s'est produit automatiquement.

Enfin, la sortie de la fonction `aMTM` contient la chaîne produite et celle-ci est de classe `mcmc` afin d'assurer la compatibilité avec d'autres *packages* `R`. Par exemple, le *package* `coda` permet d'effectuer des diagnostiques, de calculer des statistiques et de tracer des graphiques. L'ensemble de ces fonctionnalités facilite l'étude de la sortie de l'algorithme. Le diagnostique de Geweke (section 2.4) peut être calculé pour chacune des composantes ; ce test est basé sur le fait qu'une chaîne ayant atteint la stationnarité devrait être similaire du début à fin.

```
(diag <- coda::geweke.diag(mcmc$X))
# Fraction in 1st window = 0.1
```

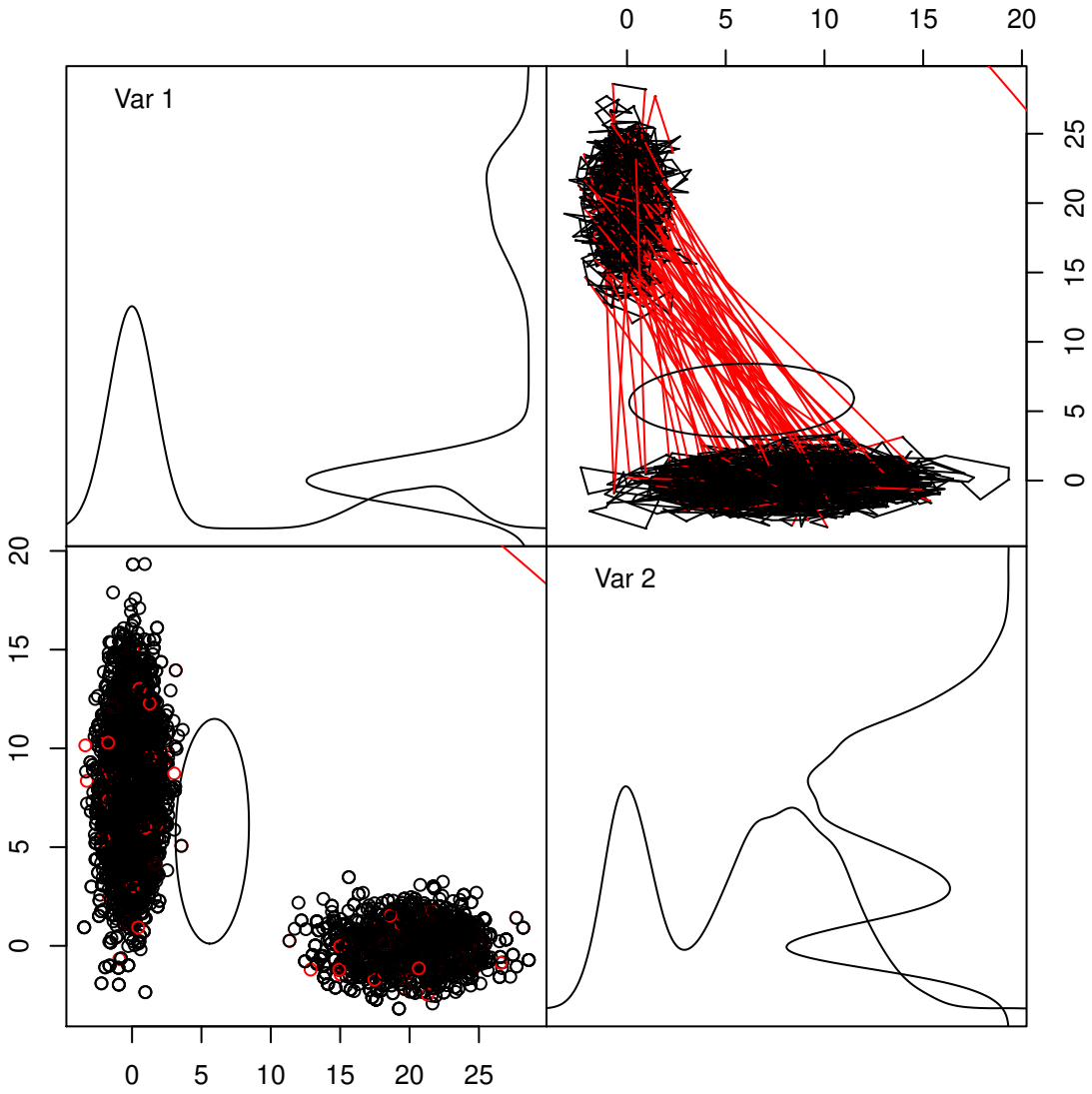


Figure 6.2 *Sorties bidimensionnelles de la chaîne produite par l'algorithme aMTM avec $K = 2$ propositions pour l'exemple bimodal en $d = 2$ dimensions. Sur la diagonale, on trouve les densités empiriques marginales lissées de chaque composante de la chaîne. Sur le triangle supérieur, on trouve le graphique du tracé dans l'espace de la chaîne, où les lignes correspondent au saut et la couleur indique la proposition choisie. Sur la diagonale inférieure, on trouve le graphique en nuage de la chaîne où les couleurs indiquent la proposition choisie pour produire le point. Les ellipses montrent les densités instrumentales finales. Notons que la seconde covariance instrumentale (en rouge) est large est n'apparait donc que très peu dans le cadre des graphes.*

```
# Fraction in 2nd window = 0.5
#
#    var1    var2
# -0.1348 -0.2013
pnorm(abs(diag$z),lower.tail=FALSE)*2
#      var1      var2
# 0.8927779 0.8404725
```

La sortie de ce test est un score Z pour chaque variable qui permet donc de calculer des niveaux critiques pour chacune des composantes. Étant donné que les niveaux critiques sont très élevés, il n'est pas possible de rejeter l'hypothèse de stationnarité. Un second diagnostique peut être effectué à l'aide de chaînes multiples. On produit donc une collection de dix chaînes pour ensuite produire le test de Gelman (section 2.4).

```
mcmclist <- lapply(seq(10), function(i){
   set.seed(i)
   mcmclist[[i]] <- aMTM(target=logp,N=1e4,K=K,x0=c(0,10),
                         parms=p,sig0=sig0,adapt='ASWAM',
                         local=T,accrate=0.5)$X
}
coda::gelman.diag(mcmclist)
# Potential scale reduction factors:
#
#      Point est. Upper C.I.
# [1,]       1.02       1.03
# [2,]       1.01       1.02
#
# Multivariate psrf
#
# 1.02
```

Puisque les PSFR univariés et le MPSFR sont tous près de 1, on ne trouve pas d'évidence que les chaînes n'ont pas toutes convergé vers la même distribution limite. En effet, ces statistiques, résultantes du diagnostique de Gelman (section 2.4), comparent les variances asymptotiques des différentes chaînes entre elles et sont minimisées à 1 lorsque toutes les chaînes partagent la même variance asymptotique. Ensemble, les résultats négatifs à ces deux tests permettent à l'utilisateur d'avoir une confiance accrue en la (ou les) chaîne(s) produite(s). Il est ensuite possible de produire les statistiques d'ESS pour chacune des composantes de la chaîne,

```
coda::effectiveSize(mcmc$X)
#     var1      var2
# 79.17247 113.43820
```

ou bien les intervalles de crédibilité à 95% marginales :

```
coda::HPDinterval(mcmc$X)
#           lower    upper
# var1 -2.279166 23.46645
# var2 -1.637213 12.95471
```

Finalement, il est possible de produire le graphique de la fonction d'autocorrélation pour chacune des composantes (figure 6.3).

```
coda::acfplot(mcmc$X, lag.max=200)
```

Le *package* `mcmcse` étend certaines des fonctionnalités du *package* `coda` dans le contexte des chaînes multivariées. La fonction `stats.aMTM` utilise ce *package* afin de calculer l'ACT multivarié de la chaîne ainsi que l'ESS multivarié, en plus de produire les statistique de sauts moyens (ESEJD et

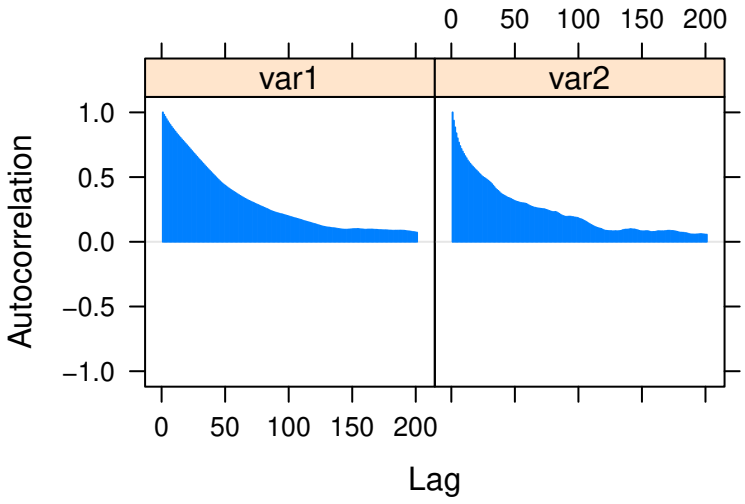


Figure 6.3 *Fonctions d'autocorrélation de chacune des composantes de la chaîne produite par l'algorithme aMTM avec $K = 2$ propositions pour l'exemple bimodal en $d = 2$ dimensions. Les graphiques sont produits à partir de la fonction* `acfplot` *du package* `coda`.

ESJD). L'ACT étant une matrice lorsque la fonction cible est multivariée, la fonction `stats.aMTM` produit donc la norme de Frobenius de cette matrice comme résumé de l'ACT.

```
stats.aMTM(mcmc$X)
# msejd    msjd     act     ess
# 1.109   0.325  66.332 409.908
```

Ces statistiques ne sont pas particulièrement parlantes en soit ; elles doivent être utilisées dans un contexte de comparaison avec la sortie d'autres algorithmes. Elle permettront donc de comparer la performance des algorithmes entre eux. Finalement, le *package* `mcmcse` permet l'estimation de la matrice de covariance asymptotique d'estimateurs Monte Carlo pour des fonctions multivariées. Par exemple, pour l'estimation de l'espérance de la distribution cible bimodale (donc l'espérance de la fonction identité en deux dimensions), on trouve l'estimé suivant accompagné d'un estimé de la covariance de l'estimateur :

```
mcmcse::mcse.multi(mcmc$X)$est
#      var1     var2
# 5.834973 5.781332
mcmcse::mcse.multi(mcmc$X)$cov/N
#             [,1]       [,2]
# [1,]  0.5686868 -0.2347880
# [2,] -0.2347880  0.1035699
```

6.2 Mesures de performance

Avant de se lancer dans de multiples simulations numériques permettant l'étude de la performance de l'algorithme aMTM, nous discutons des différentes mesures de performance qui peuvent être utilisées à cette fin. Évidemment, aucune mesure ne peut définitivement déterminer la supériorité d'un algorithme comparativement à un autre : pour cette raison, plusieurs mesures, évaluant chacune un aspect différent de la performance, doivent être considérées afin de bien comprendre le comportement des algorithmes.

Chaque expérience doit fixer la densité cible π ainsi que la fonction f dont l'espérance sous π est recherchée. À ce niveau, on dégage deux types d'expérience. D'abord, il y a les situations où il nous est possible d'obtenir une vraie solution : par exemple, le calcul de $\pi(f)$ peut être fait analytiquement ou bien un échantillon i.i.d. de π peut être obtenu facilement. Ces cas n'adviennent typiquement que par construction et ne sont donc pas représentatifs d'applications réelles, mais permettent heureusement d'évaluer concrètement la performance d'algorithmes MCMC. Ensuite, il y a les situations s'approchant davantage de situations réelles, où le calcul de $\pi(f)$ ne peut être effectué directement et où il est impossible d'obtenir un échantillon i.i.d. Dans ces cas, il n'est plus possible de comparer la sortie d'un algorithme MCMC à la vraie solution ; nous ne pouvons que comparer les algorithmes entre eux sans savoir si un (ou plusieurs) des algorithmes considérés est réellement près de la réalité. Ainsi, les expériences présentées dans ce chapitre seront du premier type étant donné que des conclusions plus fortes pourront en être tirées.

Dépendant du type d'expérience menée (c.-à-d., avec ou sans échantillon i.i.d.), différentes mesures d'efficacité peuvent être obtenues. En effet, lorsqu'un calcul analytique de $\pi(f)$ est possible ou qu'un échantillon i.i.d peut être obtenu, davantage de mesures de performance peuvent être construites. Enfin, tel que discuté à la section 4.3, la question de l'efficacité des algorithmes MCMC doit aussi tenir compte de la complexité du calcul. Il est effectivement possible de conclure qu'un algorithme est supérieur à un autre, en termes absolus, mais inférieur à ce même algorithme lorsque le temps de calcul est pris en compte.

6.2.1 Mesures comparatives

D'une manière générale, plusieurs mesures de performance peuvent être calculées à partir de la sortie d'un algorithme MCMC. Ces mesures sont toutes reliées, de près ou de loin, à l'efficacité d'estimation de l'algorithme via la covariance asymptotique. En effet, les mesures décrites à la section 2.4.3 cherchent à comparer la covariance asymptotique de l'estimateur Monte Carlo. Parmi ces mesures, on note qu'il sera toujours possible d'estimer l'ACT, l'ESS ainsi que l'ESEJD et l'ESJD. Dans le cas où l'espace d'état est multivarié, l'ACT et l'ESS doivent être résumés en une mesure univariée pour des fins comparatives : la norme de Frobenius de l'ACT et l'extension multivariée de l'ESS (section 2.4.3.4) seront considérées. Ces mesures ne sont généralement pas informatives en soit : elles doivent être utilisées d'un point de vue comparatif. Lorsqu'un ensemble d'algorithmes est utilisé pour échantillonner une même densité cible π, l'ordre induit par ces mesures permet de tirer des conclusions sur la performance relative des algorithmes.

Notons que l'ESJD, qui requiert la connaissance de la covariance de la densité cible, ne peut être estimé lorsque cette covariance est inconnnue ; il pourrait être tentant d'estimer cette covariance par la covariance de l'échantillon, mais cette dernière peut ne pas être représentative de la vraie covariance cible. Pour des raisons similaire, l'ESS, qui utilise également un estimé de la covariance, peut s'avérer une mesure erronée lorsque l'estimé n'est pas valide.

Le taux d'acceptation MH peut également informer sur la performance de l'algorithme étant donné les résultats d'optimalité des sections 2.5 et 4.2.3. Pour une densité cible donnée, il est possible d'ajuster les paramètres de l'algorithme de sorte à atteindre une probabilité d'acceptation qui soit optimale (selon le type d'algorithme et selon la dimension de l'espace d'états). Par contre, les adaptations ASWAM ou RAM dans l'algorithme aMTM cherchent à produire un taux d'acceptation précis de sorte que le taux d'acceptation observé sera généralement proche de celui recherché (c.-à-d., dicté par les résultats d'optimalité). Néanmoins, cette mesure peut identifier que des algorithmes n'utilisant pas un taux d'acceptation cible (e.g. MTM et AM) ne sont pas optimaux.

Notons que les simulations dans les expériences seront toutes répétées afin d'évaluer la variation des mesures due à la randomisation inhérente aux algorithmes MCMC. Des chaînes parallèles et indépendantes sont donc produites, ce qui permet de calculer certaines mesures exigeant une telle réplication. Par exemple, le diagnostique de Gelman-Rubin-Brooks (section 2.4.1) est une mesure qui évalue la ressemblance entre des chaînes multiples. Cette mesure permet donc d'évaluer la propension d'un algorithme à produire des résultats différents selon la randomisation : un MPSRF élevé peut indiquer que l'ensemble des chaînes n'est pas uniforme. Par exemple, si π est multimodale et que les chaînes trouvent des nombres de modes différents, alors le MPSRF sera élevé puisque les échantillons seront considérablement différents. Ce type de mesure s'inscrit dans l'évaluation de la qualité de l'exploration que l'on retrouve dans des études telles que Ballnus et collab. (2017).

6.2.2 Mesures absolues

Lorsque la densité cible permet le calcul analytique de certaines espérances $\pi(f)$, alors celles-ci peuvent servir de références pour évaluer l'efficacité d'un algorithme MCMC. Similairement, s'il est possible d'obtenir un échantillon i.i.d. de la densité cible, alors on peut obtenir des estimés fiables de certaines espérances $\pi(f)$, qui serviront de référence à leur tour.

Les expériences de simulation sont généralement définies de sorte à évaluer un aspect particulier de la performance plutôt que la performance générale de l'algorithme. Si on veut étudier le comportement de l'algorithme pour une distribution cible multimodale, par exemple, alors on choisira une densité cible avec des modes distincts. Un échantillon i.i.d. ou un calcul direct pourra nous informer de la probabilité associée à chaque mode, qu'on pourra ensuite comparer à la proportion d'observations dans chacune des régions. Cette comparaison peut être interprétée comme le biais d'estimation d'une fonction f, où f est la fonction indicatrice d'une des régions. Une autre interprétation est via l'estimation de la variation totale. En effet, la norme en variation totale (définition 2.16) considère la différence entre la distribution cible et la distribution empirique et est donnée par le suprémum, pris sur toute région de l'espace d'états, de cette différence. Une distribution empirique qui sous-évalue un mode aura une distance en variation totale pratiquement égale à la probabilité du mode manquant. Ainsi, le poids de régions critiques – difficiles à échantillonner – est un choix logique de mesure de performance. C'est exactement la justification de l'utilisation de la mesure $\mathbb{P}(X_1 > 5)$ considérée à la section 5.1 pour montrer

l'inadéquation de l'algorithme AM pour une densité cible bimodale. Notons que si l'appartenance à un mode est connue via l'échantillonnage i.i.d., les différentes méthodes de classification, telle que l'analyse discriminante linéaire (Fisher, 1936), peuvent aider à la définition des régions.

Similairement, le poids d'autres types de régions peut être considéré comme mesure de performance. Lorsque la densité cible comporte des ailes particulièrement lourdes, on peut considérer la fonction indicatrice de régions relativement éloignées mais à probabilité positive. Un algorithme ayant de la difficulté à échantillonner une densité à ailes lourdes affichera donc un biais élevé dans l'estimation de l'espérance de cette fonction. Pour une densité cible dont la géométrie est fortement irrégulière, certaines régions peuvent être difficilement accessibles par une majorité d'algorithmes MCMC, ce qui peut indiquer une région intéressante pour définir une mesure. Il n'y a pas de méthode générale pour définir ces régions dont l'échantillonnage est particulièrement ardu, mais l'information additionnelle, connue par construction de l'expérience, est cruciale pour y arriver.

Plus généralement, le biais de n'importe quelle fonction f peut constituer une mesure de performance : le choix de f dictera quel aspect de la performance est évalué. Cependant, le biais d'estimation ne mesure que la différence moyenne d'estimation. Par extension, l'erreur absolue (MAE : *Mean Absolute Error*) ou carrée (MSE : *Mean Squared Error*) entre l'estimation MCMC et la vraie valeur de l'espérance (ou de l'estimé i.i.d.) peuvent également constituer des mesures de performance intéressantes. Une estimation sans biais n'est généralement pas suffisante puisque l'estimé peut être près de la vraie valeur en moyenne, mais sa variabilité peut être grande. Le MAE et le MSE ne souffrent pas de ce problème. Toutes ces mesures sont calculés d'une manière similaire : la statistique est calculée pour chacune des chaînes et la moyenne de ces statistique produit la mesure de performance.

Finalement, lorsque la covariance Σ_π de la densité peut être calculée directement ou estimée avec confiance par un échantillon i.i.d., le saut quadratique (ESJD, section 2.4.3.3) devient une option intéressante de mesure de performance. Cette mesure ne souffre pas du problème fondamental de l'ESEJD où des variances extrêmes (grandes ou petites) dans certaines directions peuvent considérablement compromettre l'interprétation de la mesure : l'ESJD est défini spécifiquement pour corriger cette problématique, mais requiert la connaissance de Σ_π.

6.2.3 Mesures ajustées pour la complexité de calcul

Une inspection rapide de la complexité de l'algorithme aMTM permet de relever que l'augmentation du temps de calcul, par rapport à des algorithmes plus simples, est due principalement aux multiples évaluations de la densité cible à chaque itération. Grossièrement, pour une dimension d fixe, la complexité de l'algorithme aMTM est de la forme

$$a + N[b + c(2K - 1)],$$

où a est le coût computationnel fixe à l'extérieur des itérations MCMC, b est le coût fixe de chaque itération (échantillonnage, sélection, adaptation, etc.) et c est le coût associé à l'évaluation de la densité cible. Pour N fixé, la complexité tendra donc vers $cN(2K-1)$ lorsque K augmente. Lorsque a et b sont négligeables par rapport à c, c.-à-d., lorsque l'évaluation de la densité cible est coûteuse, la complexité sera donc très près de $cN(2K-1)$. La transformation d'un algorithme AM, ASWAM ou RAM à sa version MTM dans l'algorithme aMTM ajoute principalement du temps de calcul par l'évaluation de la densité cible : d'une seule évaluation on passe à $2K-1$ évaluations pour un algorithme à K

candidats. Le passage d'un algorithme MTM à une version adaptative ajoute du temps de calcul à l'étape d'adaptation. Celle-ci est indépendante du nombre de candidats et tend à devenir insignifiante dès que le coût computationnel d'une évaluation devient élevé.

Deux quantités permettent de mesurer le coût computationnel. D'abord, le temps de calcul, dénoté CPU, correspond au temps pris par la machine utilisée pour compléter l'échantillonnage. Cette mesure dépend de l'implémentation utilisée, qui n'est possiblement pas optimisée pour chacun des algorithmes. Ainsi, le temps mesuré pour les algorithmes plus simples ne sera possiblement pas représentatif du temps réellement requis. De plus, cette mesure dépend du temps requis pour évaluer la densité cible. Une densité cible peu coûteuse peut faire en sorte que la majeure partie du calcul se produise à l'adaptation alors qu'une densité cible très coûteuse peut masquer le temps de calcul requis par l'adaptation. Ensuite, le nombre d'évaluations, dénoté NbEval, de la densité cible peut constituer une seconde mesure de la complexité de l'algorithme. Il s'agit du facteur d'impact principal par rapport aux algorithmes AM, ASWAM ou RAM, dans lesquels la quantité d'adaptation demeure fixe. Cette mesure ne fait cependant pas distinction entre un algorithme MTM et sa version adaptative, mais cette distinction peut être superflue dans le cas où l'adaptation est peu coûteuse par rapport à l'évaluation de la densité cible.

À l'aide d'une expérience, on montre ici la pertinence de chacune de ces mesures. On cherche à analyser le lien entre le temps de calcul total, le nombre d'évaluations et le temps de calcul d'une évaluation. On considère donc deux densités cibles en $d = 5$ dimensions à coût computationnel respectivement faible et élevé. Deux séries d'algorithmes sont considérées : l'algorithme MTM (incluant l'algorithme Metropolis lorsque $K = 1$) et l'algorithme aMTM (incluant l'algorithme ASWAM lorsque $K = 1$). On fait varier $K = 1, \ldots, 10$ pour étudier la relation entre le temps de calcul et le nombre d'évaluations de la densité cible. Dans le cas où l'évaluation est peu coûteuse, on s'attend à voir une influence des coût fixes a et b sur le temps de calcul en fonction de NbEval. Dans le cas où l'évaluation est coûteuse, on s'attend plutôt à voir disparaître l'effet de a et de b de sorte à obtenir une relation près de $\text{CPU} = cN \cdot \text{NbEval}$.

La densité cible peu coûteuse est une densité normale alors que la densité coûteuse est un mélange de 100 densités normales. On trouve qu'une évaluation du mélange est environ 22 fois plus longue qu'une évaluation de la densité normale. Pour inspecter l'effet du temps d'évaluation, on trace le graphique du temps de calcul en fonction du nombre d'évaluations et l'on trace la droite $\text{CPU} = \hat{c}N \cdot \text{NbEval}$, où $\hat{c}$ est le temps de calcul moyen d'une évaluation. Les résultats pour les deux densités se trouvent à la figure 6.4. On trouve une influence importante des composantes fixes a et b lorsque le temps de calcul d'une évaluation est petit : en effet, la courbe du temps de calcul est translatée vers le haut par rapport au temps de calcul des évaluations seulement. Pour ce qui est de la densité plus coûteuse, on trouve plutôt que la courbe de temps de calcul se confond avec la droite de temps de calcul des évaluations seulement. Ainsi, pour des densités coûteuses, on trouve bien que la complexité de calcul est principalement capturée par le coût d'évaluation. Notons de plus que l'adaptation ne semble pas augmenter significativement le temps de calcul de sorte que l'amélioration des algorithmes MTM à leurs versions adaptatives est pratiquement gratuite.

Dans le cas d'expériences où produire un échantillon i.i.d. est accessible, il arrivera souvent que la densité cible soit rapide à évaluer. Il serait alors imprudent d'ajuster les mesures d'efficacité par le temps de calcul puisque celui-ci ne reflète pas les situations réelles. Cette observation est d'autant plus importante lorsque K est petit puisque la différence relative entre temps de calcul et temps d'évaluation y est énorme. Dans ce cas, il sera préférable d'utiliser NbEval comme mesure de la complexité pour

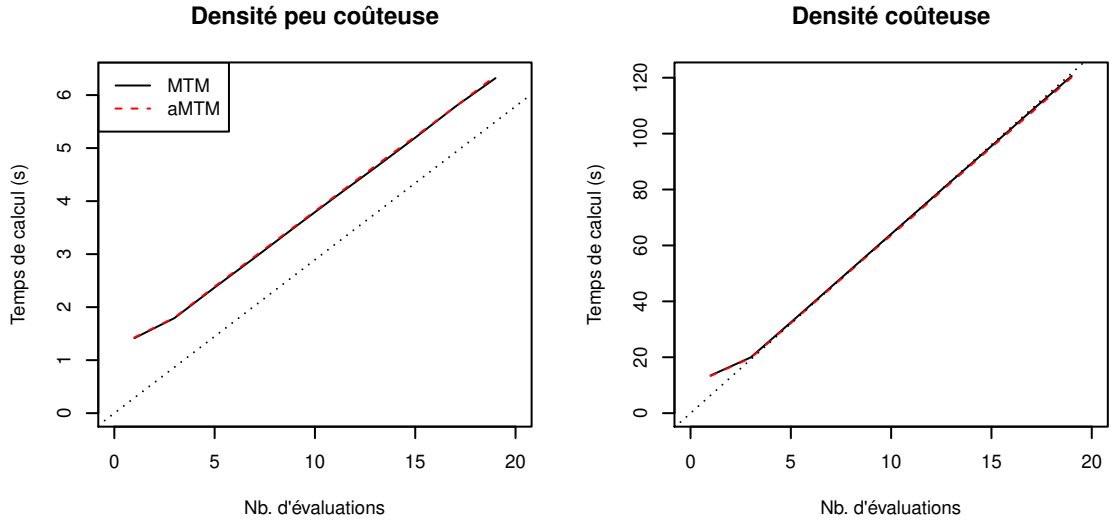


Figure 6.4 *Comparaison du temps de calcul (en secondes) en fonction du nombre d'évaluations de la densité cible par itération MCMC (c.-à-d., $2K-1$ où K est le nombre de candidats). À gauche : une densité cible peu coûteuse, à droite : une densité cible 22 fois plus coûteuse que celle de gauche. Le tracé plein noir correspond à l'algorithme MTM (donc sans adaptation) pour $K = 1, \ldots, 10$ candidats ; le tracé rouge hachuré correspond à l'algorithme aMTM (mises à jour ASWAM) pour $K = 1, \ldots, 10$ candidats. Le tracé pointillé noir correspond au temps de calcul des évaluations des densités cibles seulement.*

transposer les conclusions obtenues aux situations réelles où les densités cibles sont généralement plus coûteuses à évaluer. La mesure CPU sera quant à elle préférée lorsque la densité cible est plus coûteuse puisqu'elle mesure plus fidèlement la complexité complète de l'algorithme.

Dans tous les cas, l'algorithme aMTM sera plus coûteux en temps de calcul que ses équivalents non-adaptatifs ou à un seul candidat. Cependant, ces calculs supplémentaires amélioreront généralement l'efficacité de l'algorithme et les mesures de performance refléterons la supériorité de cet algorithme par rapport à des algorithmes plus simples. Afin de bien évaluer la performance, il est important d'ajuster les mesures de performance pour le temps de calcul additionnel. La section 4.3.2 comporte quelques exemples dans la littérature de tels ajustements.

L'interprétation de la mesure de l'ESS est particulièrement pratique pour définir des mesures ajustées. En effet, puisque l'ESS correspond à la taille échantillonnale i.i.d. équivalente, le ratio par rapport au temps de calcul ou au nombre d'évaluations produit une mesure ajustée. Par exemple, l'ESS/CPU correspond au nombre de points i.i.d. équivalents produits à chaque unité de calcul ; l'ESS/NbEval correspond au nombre de points i.i.d. équivalents produits par évaluation de densité cible. Les autres mesures de performance ne jouissent malheureusement pas d'une telle interprétation linéaire et tout ajustement pour le temps de calcul devient arbitraire.

6.3 Expériences de simulation

Afin de comparer les différentes déclinaisons de l'algorithme aMTM entre elles ainsi qu'à d'autres algorithmes, on considère une série d'expériences numériques. L'algorithme aMTM peut être vu à la fois comme une extension adaptative des algorithmes MTM ou bien une extension à essais multiples des algorithmes adaptatifs AM, ASWAM ou RAM. Ainsi, on explore la pertinence de l'algorithme aMTM sous ces deux angles en comparant sa performance à ses versions comportant un seul candidat (AM, ASWAM et RAM avec $K = 1$) et non-adaptative (MTM pour $K \geqslant 1$).

6.3.1 Densité multimodale en basses dimensions

6.3.1.1 Description de l'expérience

Dans cette première expérience de simulations, nous considérons une densité bimodale en $d = 5$ dimensions. Il s'agit d'un mélange de deux densités gaussiennes avec les paramètres

$$\mu_1 = \begin{pmatrix} -5, & -5, & 5, & 0, & 0 \end{pmatrix}^\top, \qquad \Sigma_1 = \operatorname{diag}(9{,}5{,}5{,}5{,}5),$$
$$\mu_2 = \begin{pmatrix} 0, & 0, & -5, & -5, & -5 \end{pmatrix}^\top, \qquad \Sigma_2 = \operatorname{diag}(5{,}9{,}5{,}5{,}5),$$

et où les deux densités ont le même poids. La figure 6.5 contient un graphique par paires de composantes à partir d'un échantillon i.i.d. de taille $N = 1\,000$. Les deux gaussiennes ont des supports effectifs pratiquement distincts ; en effet, un classificateur linéaire permet de classifier correctement 99,946% des points d'un échantillon de taille $N = 100\,000$. Ainsi, il existe donc une région séparant les deux modes où la densité est virtuellement nulle.

Les algorithmes MCMC de type marche aléatoire ont généralement de la difficulté à échantillonner de telles distributions puisque le passage entre les modes doit s'effectuer par un grand saut. Une densité instrumentale permettant de grands sauts aura cependant de la difficulté à échantillonner chaque mode efficacement. Les algorithmes à essais multiples peuvent remédier à cette problématique en utilisant des densités instrumentales d'échelles variées : certaines échantillonnant localement chaque mode, certaines permettant le saut entre les modes.

Dans une série d'expériences sur cette densité cible, on explore les différentes variantes de l'algorithmes aMTM afin d'identifier certains choix étant plus adaptés au cas multimodal. De plus, on compare les différentes variantes aux versions non-adaptatives (c.-à-d., à l'algorithme MTM) ainsi qu'aux versions à un seul essai (c.-à-d., aux algorithmes AM, ASWAM et RAM).

On considère quelques mesures de performances afin d'effectuer les comparaisons. D'abord, l'efficacité d'estimation sera étudiée à l'aide des mesures du MSEJD, de l'ESS et du taux d'acceptation. L'efficacité empirique sera quant à elle étudiée par l'ESS/NbEval. Ensuite, on produit une borne inférieure sur la distance en variation totale entre l'échantillon et la densité cible en considérant les deux régions définies par un classificateur linéaire (construit par l'analyse discriminante de Fisher) : chacune des deux régions devrait contenir 50% des points de l'échantillon. Cet estimé permet surtout d'identifier

des algorithmes qui n'arrivent pas à bien échantillonner de la vraie densité cible : cette mesure sera particulièrement élevée dès qu'un mode est manqué. Enfin, on considère le biais d'estimation de la moyenne de la troisième composante ; cette composante sépare le mieux les deux modes et rater un des modes sera reflété par un fort biais.

Dans tous les cas, $N = 50\,000$ points sont produits après une période de chauffe de 10% (c.-à-d., 5555 points rejetés initialement). La chaîne est initialisés à $x_0 = (0,0,0,0,0)^\top$ et les covariances des densités instrumentales sont définies comme des multiples de la matrice identité et les multiples sont régulièrement espacés dans l'intervalle [10,1 000] suivant une progression logarithmique. Par exemple, pour $K = 3$, les trois covariances initiales sont $10I_5$, $100I_5$ et $1000I_5$. Enfin pour toute variante de l'algorithme, 100 réplications sont produites pour étudier la distribution des différentes statistiques.

Expérience 1A. Dans cette situation, on investigue les différentes possibilités de l'algorithme MTM au sein de l'algorithme aMTM. D'abord, quatre types de propositions sont possibles : indépendantes, par variable aléatoire commune, quasi-Monte Carlo randomisées et extrêmement antithétiques. Ensuite, deux fonctions de poids sont considérées : par importance et proportionnels à la densité cible. Les quatre algorithmes principaux sont considérés (MTM, AM, ASWAM et RAM) et les autres caractéristiques des algorithmes sont fixées aux valeurs suivantes :

— $K = 5$ propositions ;

— Sans adaptation d'une densité globale ;

— Sans adaptation de l'échelle lorsqu'une densité n'est pas sélectionnée ;

— Sans adaptation locale ;

— Une acceptation cible de $\alpha_* = 0{,}50$ pour les algorithmes ASWAM et RAM ;

— Un pas d'adaptation de $n^{-\gamma}$ avec $\gamma = 0{,}70$.

Les résultats de cette expérience seront utilisés pour choisir un type de candidat et une fonction de poids qui semble mieux adaptées à la densité cible. La variante choisie sera utilisée pour les expériences subséquentes.

Expérience 1B. Dans cette situation, on étudie les différentes variantes possibles dans l'adaptation. En plus des quatre algorithmes principaux, on considère les déclinaisons définies par l'adaptation d'une densité globale, l'adaptation des échelles, l'adaptation locale ainsi que l'interaction entre toutes ces possibilités. Le type de candidat et la fonction de poids sont déterminés par les résultats de l'expérience 1A. Les autres caractéristiques des algorithmes sont fixées aux valeurs suivantes :

— $K = 5$ propositions ;

— Une acceptation cible de $\alpha_* = 0{,}50$ pour les algorithmes ASWAM et RAM ;

— Un pas d'adaptation de $n^{-\gamma}$ avec $\gamma = 0{,}70$.

Notons que les mises à jour RAM ne sont pas affectées par la variante de mises à jour locales. De plus, évidemment, l'algorithme MTM sans adaptation n'est pas modifié par l'une ou l'autre de ces variantes.

Expérience 1C. Dans cette situation, on étudie l'effet du nombre de propositions sur la performance de l'algorithme. On considère donc $K \in \{1,2,\ldots,10\}$ pour chacun des quatre algorithmes principaux (MTM, AM, ASWAM, RAM) et pour le type de proposition et la fonction de poids trouvés à l'expérience 1A. Les autres caractéristiques des algorithmes sont fixées aux valeurs suivantes :

— Une adaptation régulière (sans proposition globale, sans adaptation des échelles, sans mises à jour locales) ;

— Une acceptation cible de $\alpha_* = 0{,}50$ pour les algorithmes ASWAM et RAM ;

— Un pas d'adaptation de $n^{-\gamma}$ avec $\gamma = 0{,}70$.

Cette expérience permet des comparaisons absolues (MSEJD, ESS, etc.) ainsi qu'une comparaison modulée pour le temps de calcul. Puisque la densité cible est peu coûteuse computationnellement, on considère donc l'ESS/NbEval.

Expérience 1D. Dans cette situation, on étudie l'effet du taux d'acceptation cible sur la performance des algorithmes à mise à jour ASWAM et RAM. On considère $\alpha_* \in \{0{,}20; 0{,}25; \ldots; 0{,}80\}$ et on utilisera les résultats de l'expérience 1A pour déterminer les détails de l'échantillonnage MTM. Les autres caractéristiques des algorithmes sont fixées aux valeurs suivantes :

— Une adaptation régulière (sans proposition globale, sans adaptation des échelles, sans mises à jour locales) ;

— $K = 5$ propositions ;

— Un pas d'adaptation de $n^{-\gamma}$ avec $\gamma = 0{,}70$.

Expérience 1E. Dans cette situation, on étudie l'effet du paramètre de pas d'adaptation. Pour chacun des quatre algorithmes principaux (en choisissant une variante spécifique à l'aide des autres expériences), on produit les mesures de performance pour chaque $\gamma \in \{0{,}30; 0{,}35; \ldots; 1{,}00\}$ dans le pas d'adaptation $\gamma_n = n^{-\gamma}$. Les garanties théoriques sur l'algorithme aMTM sont valides pour tout $\gamma \leqslant 1$ (pour assurer $\sum_{n \geqslant 1} \gamma_n < \infty$), mais notons que $\gamma < 0.5$ peut produire des comportements non-souhaités. Notamment, les paramètres des densités instrumentales peuvent ne pas converger pour $\gamma < 0.5$ (voir e.g. condition 3.27 et théorème 3.31). Les autres caractéristiques des algorithmes sont fixées aux valeurs suivantes :

— Une adaptation régulière (sans proposition globale, sans adaptation des échelles, sans mises à jour locales) ;

— $K = 5$ propositions ;

— Une acceptation cible de $\alpha_* = 0{,}50$ pour les algorithmes ASWAM et RAM.

6.3.1.2 Résultats et analyse

Expérience 1A. Les résultats de l'expérience 1A concernant le type de proposition et la fonction de poids se trouvent au tableau 6.1. Voici les observations qu'on peut en tirer :

MSEJD

— Les algorithmes adaptatifs produisent systématiquement des chaînes à valeur de MSEJD plus élevée ; les mises à jour ASWAM sont légèrement plus performantes à ce niveau ;

— Les candidats par variable aléatoire commune sont particulièrement moins efficaces que les autres types de candidats ; alors que les candidats extrêmement antithétiques semblent légèrement supérieurs aux autres pour cette mesure ;

— Dans la plupart des cas, les poids par importance produisent une valeur de MSEJD plus élevée que les poids proportionnels à la densité cible ;

— La meilleure combinaison est la même pour les quatre mises à jour : candidats extrêmement antithétiques avec poids par importance.

ESS

— Les mises à jour AM produisent des mesures d'ESS plus élevées que les trois autres choix de mise à jour et les mises à jour ASWAM se place au second rang ;

— Les candidats par variable aléatoire commune sont particulièrement moins efficaces que les trois autres types de candidat qui sont relativement équivalents au niveau de l'ESS ;

— Les poids proportionnels à la densité cible semblent performer un peu mieux que les poids par importance ;

— Pour trois des quatre types de mise à jour, la meilleure option est d'utiliser les candidats QMCR avec les poids proportionnels à la densité cible.

Distance TV

— Les mises à jour AM et ASWAM produisent des chaînes dont les distributions empiriques sont plus près de la distribution cible que la mise à jour RAM ou sans mise à jour ;

— Les candidats par variable aléatoire commune semblent être moins efficaces pour s'approcher de la distribution cible ;

— Il n'y a pas d'indication claire que le choix de poids influence cette mesure ;

— La meilleure combinaison est la même pour les quatre mises à jour : candidats extrêmement antithétiques avec poids par importance.

Général

— Pour chacun des quatre algorithmes principaux, les propositions par variable aléatoire commune performent significativement moins bien que tous les autres types de proposition, et ce, pour les trois mesures considérées ;

— Pour ce qui est des autres types de proposition, les candidats extrêmement antithétiques sont à l'occasion mieux que les candidats indépendants ou QMCR ;

— On ne peut identifier un choix de poids clairement supérieur à l'autre ;

— Les mises à jour AM et ASWAM performent uniformément mieux que les mises à jour RAM ;

— Ne pas procéder à l'adaptation des densité instrumentales est systématiquement pire que l'inverse, et ce, peu importe le type de mise à jour.

Ainsi, pour les expériences subséquentes, on utilisera des propositions extrêmement antithétiques et des poids par importance.

Les propositions extrêmement antithétiques sont conçues pour maximiser la distance euclidienne entre les candidats : il n'est donc pas surprenant que ce type de propositions produise un MSEJD plus élevé. Les propositions par variable aléatoire commune sont généralement moins efficaces : il arrive souvent que plusieurs des densités instrumentales convergent vers une même densité, et les candidats générés se trouvent alors dans une même région de l'espace d'état. On n'améliore donc pas l'exploration de l'espace en utilisant ce type de proposition. Les candidats indépendants et QMCR répartissent mieux les points dans l'espace que les candidats par variables aléatoire commune, mais pas aussi optimalement que les candidats EA, ce qui explique leur performance légèrement moins bonne.

Les poids par importance favorisent les grands sauts comparativement aux poids proportionnels à la densité cible. Ainsi, il n'est également pas surprenant d'observer un MSJED un peu plus élevé pour

ce choix de fonction de poids. Il n'est pas clair que ce type de poids soit systématiquement supérieur aux poids proportionnels à la densité cible puisque l'ESS est généralement plus élevé pour ce choix de fonction. Tel qu'observé par plusieurs études (section 4.1.2), ces deux types de poids sont relativement équivalents en pratique, ce que nos résultats semblent confirmer à nouveau.

Expérience 1B. Les résultats de l'expérience 1B sur les variantes d'adaptation se trouvent au tableau 6.2. Voici les observations qu'on peut en tirer :

MSEJD

— Les algorithmes adaptatifs surclassent significativement l'algorithme MTM sans adaptation pour tout choix de variante (à quelques exceptions près) ;

— Les mises à jour ASWAM et RAM semblent produire un MSEJD légèrement plus élevé que les mises à jour AM ;

— L'adaptation d'une densité instrumentale globale ne semble pas affecter significativement la performance de l'algorithme puisqu'on observe certaines augmentations et certaines diminutions de MSEJD ;

— L'adaptation des échelles à toutes les itérations semble nuire aux mises à jour AM et ASWAM, mais aider les mises à jour RAM ;

— L'utilisation de mises à jour locales diminue presque systématiquement le MSJED des chaînes produites par les algorithmes AM et ASWAM.

ESS

— Les trois algorithmes adaptatifs atteignent des mesures d'ESS particulièrement plus élevées que l'algorithme sans adaptation et les trois atteignent des valeurs similaires ;

— L'adaptation d'une densité globale n'affecte pas la performance des algorithmes considérés ;

— Il n'y a pas de différence claire dans l'ESS des algorithmes AM et ASWAM lors de l'adaptation ou non des échelles ; les mises à jour RAM, par contre, profitent bien de cette variante ;

— L'utilisation de mises à jour locales diminuent généralement l'ESS des algorithmes AM et ASWAM.

Distance TV

— Les algorithmes AM et ASWAM atteignent les estimés de la distance en variation totale les plus bas parmi les quatre algorithmes ;

— L'utilisation d'une densité globale n'affecte pas la performance des algorithmes ;

— L'adaptation des échelles diminue la performance des algorithmes AM et ASWAM alors quelle augmente celle de l'algorithme RAM ;

— L'utilisation de mises à jour locales diminue systématiquement la performance des algorithmes AM et ASWAM.

Général

— Les trois types de mise à jour adaptative performent significativement mieux qu'un algorithme sans adaptation ;

— L'adaptation d'une densité globale ne semble pas nuire ni aider à la performance des algorithmes aMTM ;

— L'adaptation des échelles n'est utile que pour l'algorithme RAM ;

— Les mises à jour locales nuisent à la performance des algorithmes AM et ASWAM.

Aucunes de ces variante ne semble considérablement améliorer la performance des algorithmes considérés à l'exception des mises à jour RAM, qui tirent bien profit de l'adaptation des échelles. Ce type de mise à jour n'adapte pas directement l'échelle de ses propositions comme le fait l'algorithme ASWAM.

Expérience 1C. Les résultats de l'expérience 1C sur le nombre de propositions se trouvent à la figure 6.6. Sans surprise, les mesures de performance (MSEJD, ESS, Distance TV et Biais d'estimation de x_3) s'améliorent lorsque K augmente. Cependant, la croissance ralentit rapidement et il semble possible de trouver un nombre idéal de candidats.

Pour ce qui est du MSEJD, l'augmentation semble se stabiliser après $K = 5$ candidats. Les trois algorithmes adaptatifs battent significativement l'algorithme sans adaptation pour tout nombre de candidats et l'algorithme ASWAM semble être légèrement mieux que les deux autres pour cette mesure.

La majeure partie de l'augmentation de l'ESS est atteinte vers $K = 7$ candidats. Au-delà de $K = 4$ candidats, les algorithmes adaptatifs performent tous mieux que l'algorithme MTM sans adaptation et l'algorithme AM surclasse tous les algorithmes sur l'ensemble du domaine.

L'algorithme ASWAM atteint bien le taux d'acceptation cible de $\alpha_* = 0{,}5$ alors que l'algorithme RAM tend à s'en éloigner lorsque K augmente (notons que l'algorithme RAM ne garantit pas l'atteindre du taux cible, bien que ce taux soit utilisé dans le calcul des mises à jour). L'algorithme AM a un taux d'acceptation croissant avec K. Tous les algorithmes ont un taux d'acceptation plus élevé que l'algorithme MTM sans adaptation et atteignent des valeurs qui peuvent être considérées comme étant meilleures (cf. tableau 4.1).

Lorsque le temps de calcul est pris en compte (via le nombre d'évaluations), il semble qu'un seul candidat soit optimal pour tous les algorithmes, mais ce nombre ne produit pas des chaînes qui représentent bien la distribution cible. Pour tous les algorithmes adaptatifs, on note de légers maximums locaux pour $K = 3$ et pour $K = 5$. Pour l'algorithme MTM sans adaptation, $K = 2$ candidats semble optimal.

Finalement, pour ce qui est de l'ajustement à la distribution cible mesuré par la distance TV et le biais d'estimation de x_3, on note que les algorithmes MTM et RAM ne s'améliorent plus au-delà de $K = 2$ candidats alors que les algorithmes AM et ASWAM semblent optimaux au-delà de $K = 5$ candidats. Ces deux algorithmes ont des mesures significativement meilleures que les algorithmes RAM et MTM.

Globalement, il semble que l'algorithme MTM ne requiert que $K = 2$ candidats afin d'être à son meilleur pour cette distribution cible alors que les algorithmes aMTM requièrent plutôt $K = 5$ candidats pour être optimaux. Les algorithmes AM et ASWAM performent significativment mieux que les algorithmes MTM et RAM.

La pertinence des essais multiples est claire pour cette distribution cible. Intuitivement, les essais multiples permettent de sauter entre les modes en plus de bien effectuer l'exploration de chacun des modes. Empiriquement, lorsqu'un seul candidat est produit ($K = 1$), on trouve que les algorithmes ne produisent pas des chaînes qui soient représentatives de la densité cible. En effet, la distance en variation totale et le biais d'estimation de x_3 sont tous deux minimisés pour une plus grande quantité de candidats.

Expérience 1D. Les résultats de l'expérience 1D sur l'effet du taux d'acceptation cible se trouvent à la figure 6.7. Notons que les algorithmes AM et MTM ne sont pas affectés par un changement de taux d'acceptation cible.

L'algorithme ASWAM atteint bien le taux d'acceptation cible tel que prévu par construction. Les différentes mesures indiquent différentes valeurs de taux cible optimal :

— MSEJD : $\alpha_* \in [0{,}40; 0{,}65]$;

— ACT et ESS : $\alpha_* \in [0{,}20; 0{,}50]$;

— Distance TV et Biais : $\alpha_* \in [0{,}20; 0{,}70]$.

Il semble donc qu'un taux cible entre 0,40 et 0,50 produise un algorithme ASWAM qui soit optimal uniformément par rapport à la mesure de performance.

L'algorithme RAM n'atteint pas quant à lui le taux d'acceptation cible (on rappelle qu'il ne s'agit pas d'une garantie de l'algorithme). Cet algorithme est optimal pour un taux cible dans les intervalles suivants :

— MSEJD : $\alpha_* \in [0{,}40; 0{,}75]$;

— ACT et ESS : $\alpha_* \in [0{,}20; 0{,}50]$;

— Distance TV et Biais : $\alpha_* \in [0{,}20; 0{,}80]$.

Il semble donc qu'un taux cible entre 0,40 et 0,50 produise un algorithme RAM qui soit optimal uniformément par rapport à la mesure de performance.

Expérience 1E. Les résultats de l'expérience 1E sur l'effet du pas d'adaptation se trouvent à la figure 6.8. Évidemment, l'algorithme MTM n'est pas affecté par le rythme de décroissance du pas d'adaptation étant donné que cet algorithme n'effectue pas d'adaptation.

L'algorithme ASWAM atteint bien le taux d'acceptation cible pour $\gamma \in [0{,}35; 0{,}65]$; au-delà de 0,65, l'algorithme n'arrive pas à atteindre le taux cible et on comprend donc que le rythme d'adaptation est trop lent. Cette observation est confirmée par l'étude du MSEJD, de l'ACT et de l'ESS qui sont tous optimisés pour $\gamma \in [0{,}40; 0{,}65]$.

L'algorithme RAM atteint le taux d'acceptation cible pour $\gamma \in [0{,}35; 0{,}50]$ ce qui pourrait indiquer une région où cet algorithme est possiblement mieux adapté. Le MSEJD, l'ACT et l'ESS indiquent conjointement que l'algorithme RAM est optimisé pour $\gamma \in [0{,}35; 0{,}50]$.

Finalement, l'algorithme AM semble performer de mieux en mieux en diminuant la valeur du paramètre γ. L'expérience s'arrête à $\gamma = 0{,}30$ de sorte qu'on ne peut spéculer sur le comportement de l'algorithme pour de plus petites valeurs. Cependant, on rappelle que $\gamma < 0{,}5$ n'est pas recommandable pour des raisons de convergence de l'adaptation.

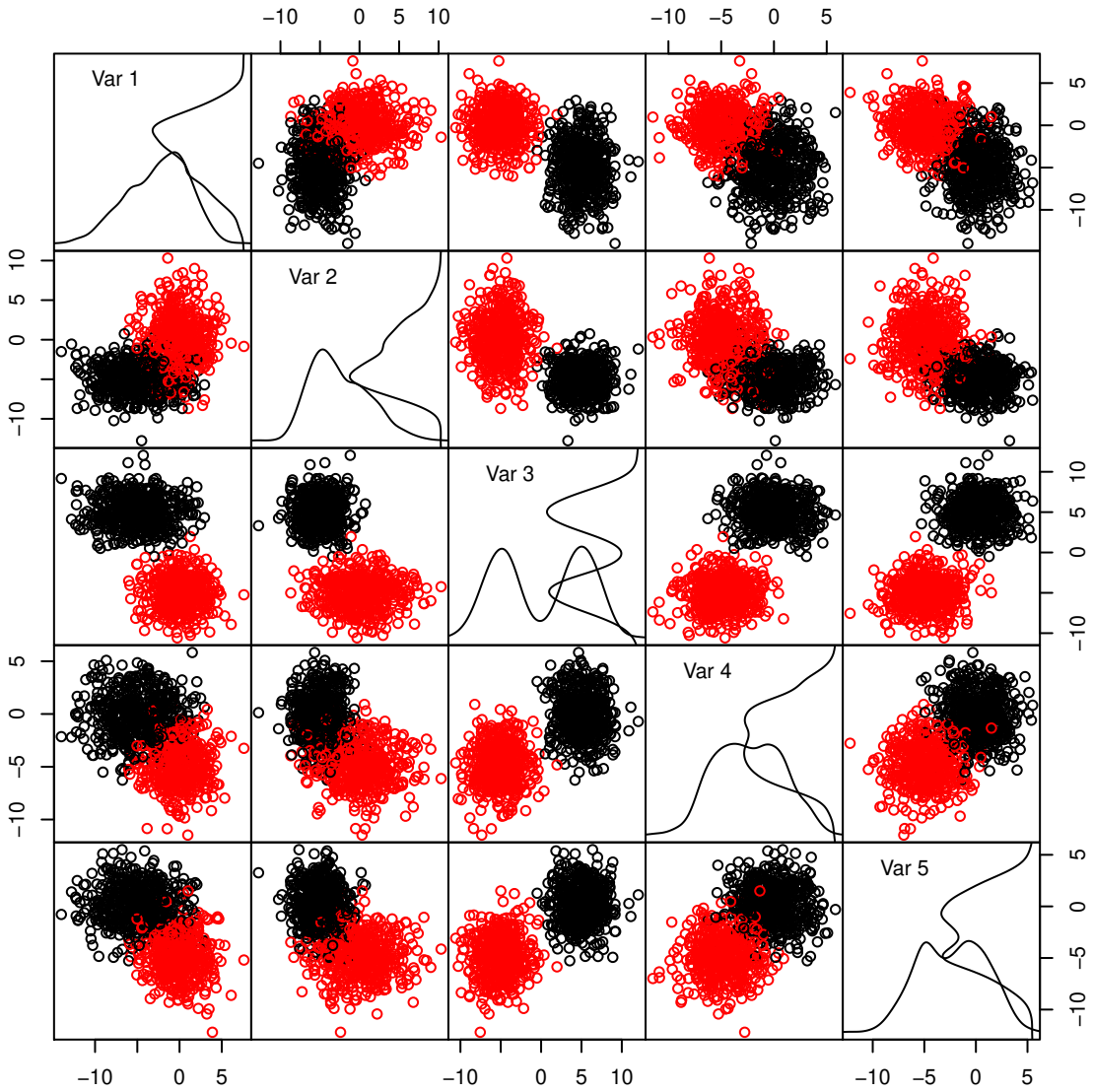


Figure 6.5 *Graphique par paires d'un échantillon i.i.d. de la densité bimodale utilisée dans l'expérience 1. Sur la diagonale se trouve une estimation non-paramétrique des densités marginales. Les couleurs représentent la composante qui a généré le point : les point noirs correspondent à la première composante alors que les points rouges proviennent de la seconde composante.*

Expérience 1A - Statistiques en fonction du type de proposition et de poids							
		MSEJD		ESS		Distance TV	
Proposition	Poids	Moyenne	Err. std.	Moyenne	Err. std.	Moyenne	Err. std.
Sans adaptation (MTM)							
Indépendante	Importance	0.342	(0.001)	1784	(8)	0.022	(0.002)
	Prop. cible	0.376	(0.001)	2086	(8)	0.023	(0.002)
V.A. Commune	Importance	0.259	(0.001)	1608	(6)	0.033	(0.002)
	Prop. cible	0.310	(0.001)	1742	(8)	0.029	(0.002)
QMCR	Importance	0.347	(0.001)	1844	(8)	0.023	(0.001)
	Prop. cible	0.380	(0.001)	**2122**	(8)	0.027	(0.002)
EA	Importance	**0.421**	(0.001)	1956	(9)	**0.017**	(0.001)
	Prop. cible	0.372	(0.001)	2099	(9)	0.024	(0.002)
Mises à jour AM							
Indépendante	Importance	0.768	(0.003)	3586	(23)	0.007	(0.000)
	Prop. cible	0.636	(0.003)	3756	(19)	**0.007**	(0.001)
V.A. Commune	Importance	0.315	(0.000)	1833	(10)	0.012	(0.001)
	Prop. cible	0.321	(0.000)	1902	(9)	0.010	(0.001)
QMCR	Importance	0.791	(0.003)	3876	(26)	**0.006**	(0.001)
	Prop. cible	0.662	(0.002)	**4026**	(22)	**0.007**	(0.001)
EA	Importance	**0.917**	(0.003)	3960	(26)	**0.006**	(0.000)
	Prop. cible	0.629	(0.003)	3775	(24)	0.008	(0.001)
Mises à jour ASWAM							
Indépendante	Importance	0.876	(0.004)	2185	(24)	0.009	(0.001)
	Prop. cible	0.611	(0.004)	2420	(31)	0.012	(0.001)
V.A. Commune	Importance	0.399	(0.001)	1294	(7)	0.026	(0.002)
	Prop. cible	0.446	(0.001)	1506	(11)	0.015	(0.001)
QMCR	Importance	0.892	(0.004)	2432	(26)	0.009	(0.001)
	Prop. cible	0.639	(0.005)	2738	(41)	0.011	(0.001)
EA	Importance	**1.060**	(0.005)	**3350**	(37)	**0.007**	(0.001)
	Prop. cible	0.604	(0.004)	2426	(32)	0.013	(0.001)
Mises à jour RAM							
Indépendante	Importance	0.729	(0.002)	1897	(9)	0.023	(0.002)
	Prop. cible	0.622	(0.001)	2159	(10)	0.031	(0.003)
V.A. Commune	Importance	0.428	(0.001)	1900	(9)	0.037	(0.003)
	Prop. cible	0.567	(0.000)	1689	(7)	0.026	(0.002)
QMCR	Importance	0.755	(0.002)	2099	(9)	0.024	(0.002)
	Prop. cible	0.639	(0.001)	**2323**	(11)	0.029	(0.002)
EA	Importance	**0.837**	(0.003)	2191	(11)	**0.019**	(0.001)
	Prop. cible	0.623	(0.001)	2214	(12)	0.031	(0.002)

Tableau 6.1 *(Expérience 1A) Statistiques en fonction du type de candidat et de la fonction de poids pour les quatre algorithmes principaux. Les statistiques en gras indiquent le meilleur choix par type de mise à jour et celles soulignées indiquent le meilleur choix global.*

Expérience 1B - Statistiques en fonction du type d'adaptation								
Adaptation			MSEJD		ESS		Distance TV	
Globale	Échelle	Locale	Moyenne	Err. std.	Moyenne	Err. std.	Moyenne	Err. std.
Sans adaptation (MTM)								
–	–	–	0.421	(0.001)	1956	(8)	0.017	(0.001)
Mises à jour AM								
Non	Non	Non	0.916	(0.003)	**3973**	(22)	**<u>0.006</u>**	(0.000)
		Oui	0.863	(0.011)	1242	(29)	0.033	(0.002)
	Oui	Non	0.873	(0.009)	2147	(59)	0.011	(0.001)
		Oui	0.144	(0.022)	3217	(594)	0.373	(0.019)
Oui	Non	Non	0.926	(0.004)	**4034**	(28)	**<u>0.006</u>**	(0.000)
		Oui	**0.988**	(0.013)	1563	(37)	0.027	(0.002)
	Oui	Non	0.898	(0.009)	2446	(72)	0.009	(0.001)
		Oui	0.152	(0.022)	2742	(778)	0.347	(0.020)
Mises à jour ASWAM								
Non	Non	Non	**1.052**	(0.004)	**3296**	(35)	**<u>0.006</u>**	(0.000)
		Oui	0.866	(0.016)	1518	(76)	0.025	(0.002)
	Oui	Non	0.977	(0.007)	2836	(49)	**<u>0.007</u>**	(0.001)
		Oui	0.663	(0.019)	757	(46)	0.066	(0.008)
Oui	Non	Non	1.007	(0.008)	3058	(48)	0.008	(0.000)
		Oui	0.989	(0.012)	2090	(65)	0.023	(0.002)
	Oui	Non	0.976	(0.008)	2918	(64)	**<u>0.007</u>**	(0.000)
		Oui	0.705	(0.019)	877	(56)	0.051	(0.007)
Mises à jour RAM								
Non	Non	–	0.836	(0.002)	2207	(11)	0.020	(0.001)
	Oui	–	**<u>1.068</u>**	(0.011)	**<u>4444</u>**	(41)	**0.013**	(0.001)
Oui	Non	–	0.624	(0.002)	1740	(8)	0.022	(0.002)
	Oui	–	1.001	(0.016)	4053	(47)	**0.012**	(0.001)

Tableau 6.2 *(Expérience 1B) Statistiques en fonction du type d'adaptation (proposition globale, adaptation de l'échelle, mise à jour locale) pour les quatre algorithmes principaux. Les statistiques en gras indiquent le meilleur choix par type de mise à jour et celles soulignées indiquent le meilleur choix global.*

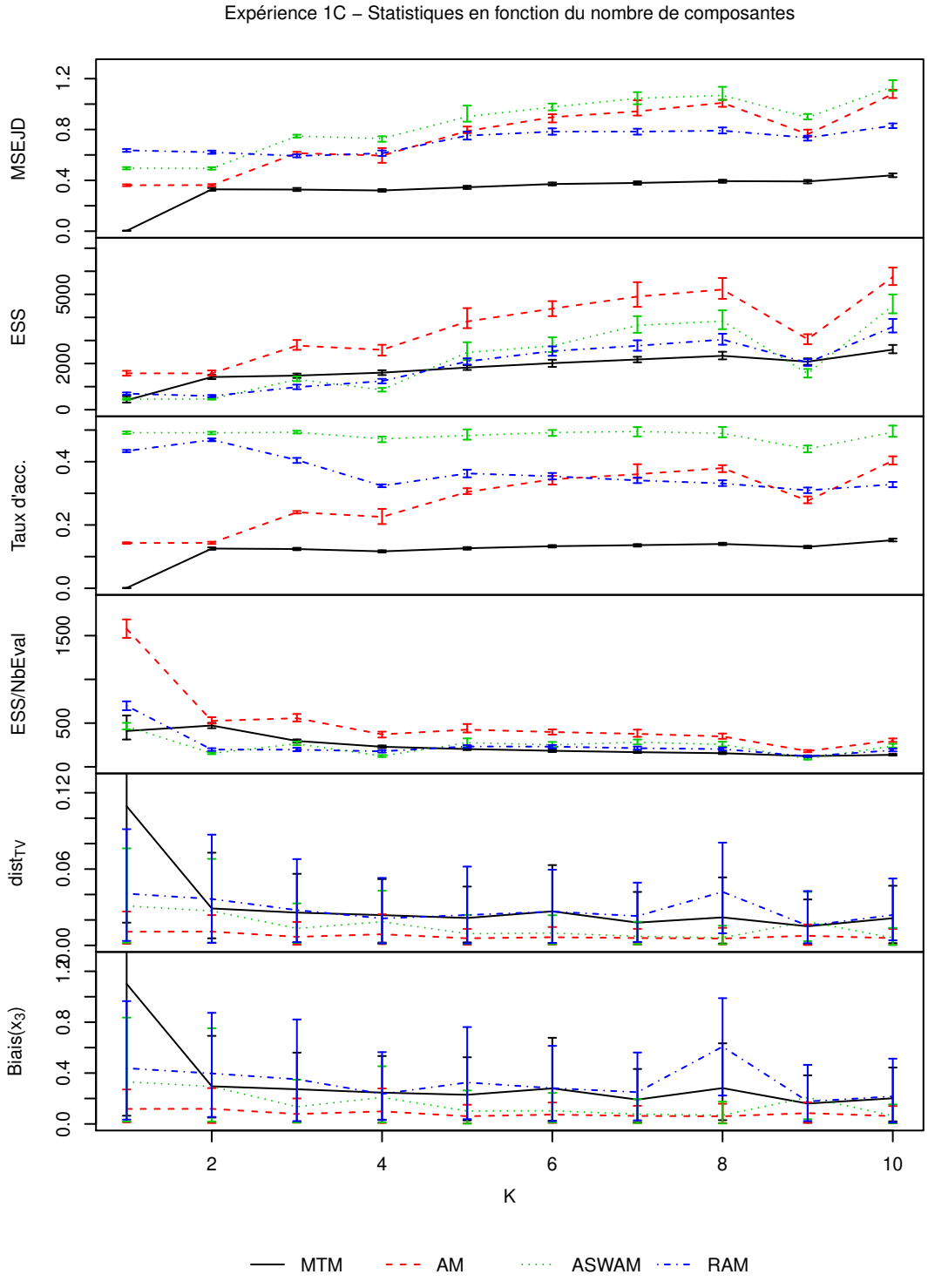


Figure 6.6 *(Expérience 1C) Statistiques en fonction du nombre de propositions pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

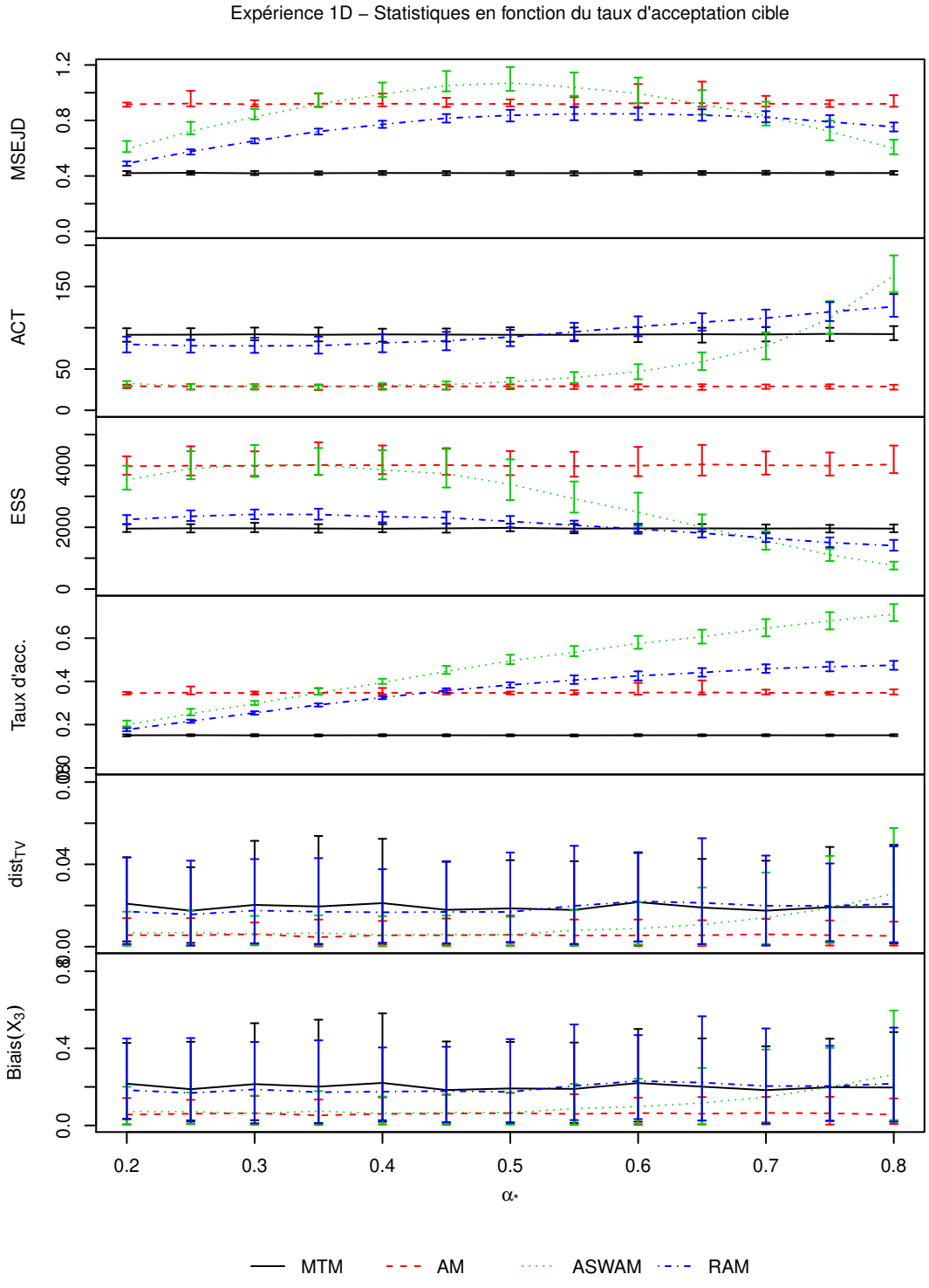


Figure 6.7 *(Expérience 1D) Statistiques en fonction du taux d'acceptation cible pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

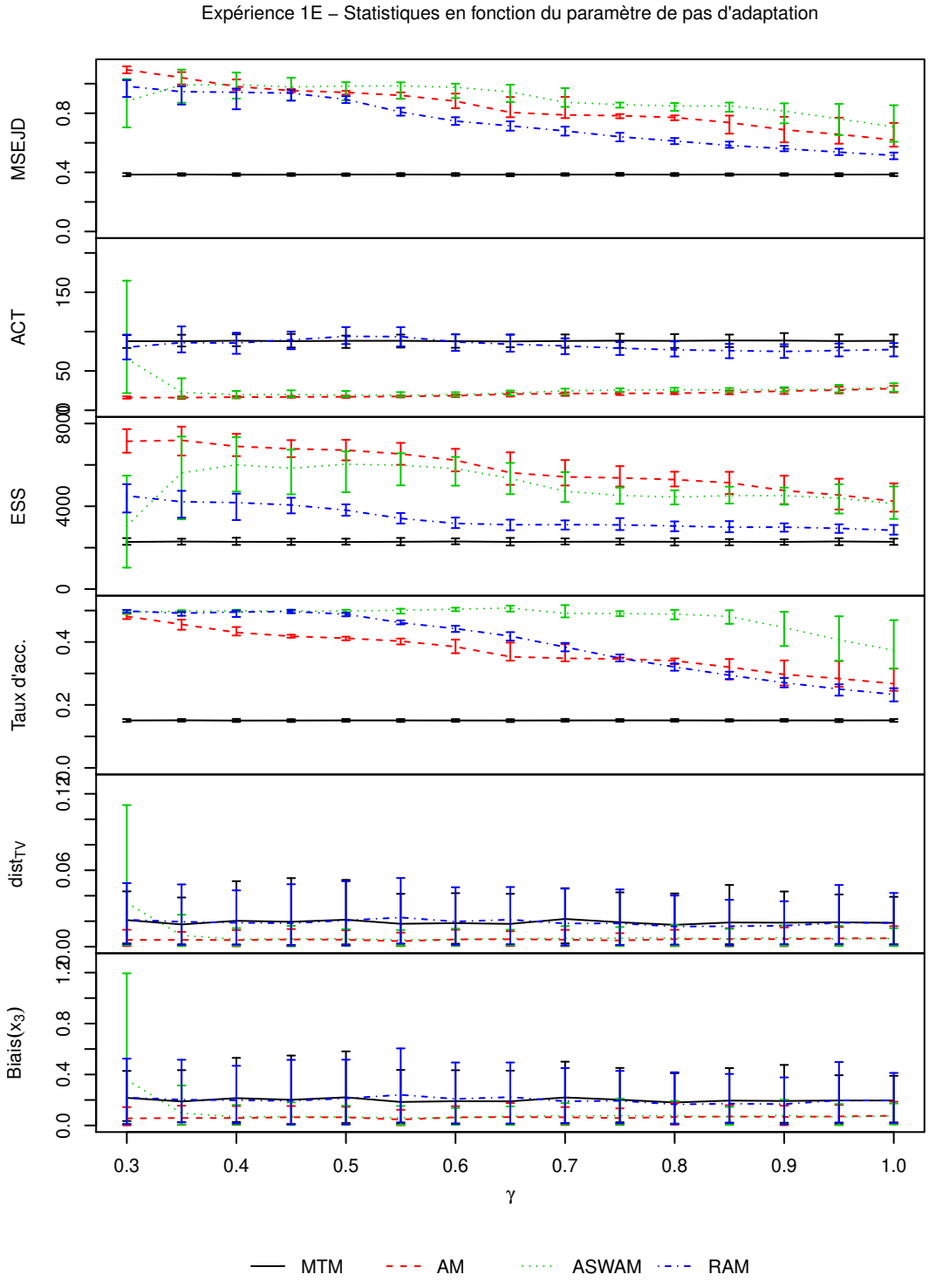


Figure 6.8 *(Expérience 1E) Statistiques en fonction du paramètre de pas d'adaptation pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

6.3.2 Densité à géométrie complexe

6.3.2.1 *Description de l'expérience*

Dans cette seconde expérience, on considère une densité cible qui possède un support effectif particulièrement difforme. La densité cible prend la forme d'une banane et est souvent employée lors d'expériences de simulation (MCMC adaptatifs : Haario et collab., 1999, Haario et collab., 2001, Haario et collab., 2006, Andrieu et Thoms, 2008, Roberts et Rosenthal, 2009, Sejdinovic et collab., 2014, Tran et collab., 2016 ; MCMC à essais multiples : Martino et Read, 2013, Yang et collab., 2019). Les détails de cette construction se trouvent à l'appendice 6.5.1. On considère une densité en $d = 5$ dimensions avec trois blocs de composantes indépendantes. Les deux blocs formés respectivement des deux premières et des deux dernières composantes forment chacun une banane ; la troisième composante est une densité normale standard. L'expression de la densité est donnée par

$$\log p\,(x) \propto -\frac{1}{2}\left[\frac{x_1^2}{a_1^2} + \left(-B_1 x_1^2 + x_2 - B_1 a_1^2\right)^2 + x_3^2 + \frac{x_4^2}{a_2^2} + \left(-B_2 x_4^2 + x_5 - B_2 a_2^2\right)^2\right],$$

où $a_1 = 1$, $a_2 = 1$, $B_1 = 3$ et $B_2 = 1$. La figure 6.9 contient un graphique par paires de composantes, obtenu à partir d'un échantillon i.i.d. de taille $N = 10\,000$.

La difficulté dans l'échantillonnage de cette densité se trouve dans les queues des bananes. En effet, ces régions ont des covariances particulièrement différentes de celle près du mode. Ainsi, la mesure de la performance d'un algorithme doit tenir compte de ces régions. On produit donc un estimé de la distance en variation totale en séparant l'espace d'états en cinq régions : une pour chacune des queues des deux bananes et une pour le reste du support. La définition de ces régions se trouve au tableau 6.3 et la figure 6.9 identifie les points dans chacune des régions. De plus, on considère le biais d'estimation de x_2^2 qui sera élevé dès que les queues de la banane formée par les composantes 1 et 2 ne sont pas biens représentées dans l'échantillon.

Région	x_1	x_2	x_4	x_5	Proportion
1	dans aucune autre région				82%
2	–	–	> 0	> 5	3%
3	–	–	< 0	> 5	3%
4	> 0	> 10	–	–	6%
5	< 0	> 10	–	–	6%

Tableau 6.3 *(Expérience 2) Définition des régions utilisées pour l'estimation de la distance en variation totale.*

La méthodologie de l'expérience 1 (section 6.3.1) est reprise intégralement : les mêmes cinq sous-expériences sont effectuées afin d'étudier les différentes variantes des algorithmes et de comparer l'algorithme aMTM aux algorithmes plus simples. La seule différence est que les échelles des covariances instrumentales sont initialisées selon une suite régulière logarithmique sur l'intervalle $[0{,}1; 10\,000]$.

6.3.2.2 *Résultats et analyse*

Expérience 2A. Les résultats de l'expérience 2A sur le type de proposition et sur la fonction de poids se trouvent au tableau 6.4. Voici les observations qu'on peut en tirer :

MSEJD

- Les quatre types de mise à jour se classent dans l'ordre suivant : ASWAM, RAM, MTM et AM ;
- À l'exception des candidats par variable aléatoire commune qui performent systématiquement moins bien, les autres types de candidats sont relativement équivalents ;
- Les poids proportionnels à la densité cible produisent des chaînes à MSEJD plus élevé pour les algorithmes MTM, AM et RAM alors que c'est l'inverse pour les mises à jour ASWAM ;
- Les algorithmes MTM, AM et RAM sont tous trois optimisés pour des candidats QMCR et des poids proportionnels à la densité cible, mais les valeurs de MSEJD les plus élevées sont obtenues par l'algorithme ASWAM avec des candidats extrêmement antithétiques et des poids par importance.

ESS

- Dans l'ordre, les valeurs d'ESS les plus élevées sont obtenues pour les algorithmes AM, ASWAM, RAM et finalement MTM ;
- Les trois types de candidat sont relativement équivalents pour cette mesure si ce n'est d'une légère baisse pour les candidats par variable aléatoire commune ;
- Tous les algorithmes sont optimisés pour des candidats QMCR, bien que la différence soit très faible dans certains cas ;
- Les poids proportionnels à la cible augmentent significativement l'ESS des chaînes produites dans la grande majorité des variantes considérées ; en particulier, tous les algorithmes sont optimisés pour ce choix de poids ;

Distance TV

- Les mises à jour AM produisent des estimés de la distance en variation totale systématiquement moins élevés que les autres algorithmes ;
- Les meilleurs ajustements à la distribution cible sont généralement obtenus en utilisant des candidats indépendants ou QMCR ;
- Les poids proportionnels à la densité cible améliorent significativement cette mesure comparativement aux poids par importance, et ce, pour toutes les variantes considérées.

Général

- Les algorithmes adaptatifs performent mieux que l'algorithme MTM sans adaptation, mais aucune des mises à jour n'est clairement supérieure ;
- Les candidats par variable aléatoire commune sont significativement moins efficaces que les autres types de candidats et les candidats QMCR sont parfois un peu plus efficaces que les candidats indépendants ou extrêmement antithétiques ;
- Les poids proportionnels à la densité cible sont préférables aux poids par importance pour ce choix de densité cible.

Les conclusions qu'il a été possible de soulever à partir de cette expérience nous renseignent bien sur le comportement de notre algorithme. En effet, les quatre ailes de la densité cible sont particulièrement difficiles à atteindre. L'adaptation permet de bien ajuster la covariance et l'échelle des densités instrumentales pour être en mesure de trouver les directions pertinentes. De plus, les types de candidat mieux répartis dans l'espace (QMCR et EA) facilitent l'exploration du support. Enfin, il semblerait qu'il soit plus important de viser dans la bonne direction (poids proportionnels à la densité cible) que de viser loin (poids par importance).

On utilisera donc les candidats QMCR et les poids proportionnels à la densité cible pour le reste des expériences.

Expérience 2B. Les résultats de l'expérience 2B sur les variantes d'adaptation se trouvent au tableau 6.5. Voici les observations qu'on peut en tirer :

MSEJD

— Les plus hautes valeurs de MSEJD sont obtenues, en ordre, par les algorithmes RAM, ASWAM, MTM et AM ;

— L'adaptation d'une densité globale ne semble pas affecter significativement cette mesure ;

— L'adaptation des échelles nuit à la performance des algorithmes AM et ASWAM alors qu'elle améliore celle de l'algorithme RAM ;

— L'utilisation de mises à jour locales est parfois mieux et parfois pire en terme de MSEJD que l'inverse ;

ESS

— Les algorithmes adaptatifs produisent des chaînes à ESS particulièrement supérieurs à ceux de l'algorithme MTM ; les algorithmes AM et ASWAM sont légèrement mieux que l'algorithme RAM à ce niveau ;

— L'adaptation d'une densité globale améliore parfois cette mesure pour les algorithmes AM et ASWAM ;

— L'adaptation des échelles améliore grandement la performance de l'algorithme RAM alors que l'effet de cette variante sur les autres types de mise à jour varie d'un cas à l'autre ;

— Les mises à jour locales nuisent systématiquement aux mises à jour ASWAM alors que cette variante est parfois positive pour les mises à jour AM.

Distance TV

— Les algorithmes adaptatifs atteignent des estimés de la distance en variation totale plus bas que ceux de l'algorithme MTM sans adaptation ;

— L'adaptation d'une densité globale semble avoir un léger impact négatif sur cette mesure ;

— L'adaptation des échelles améliore généralement la performance des algorithmes à ce niveau pour les algorithmes ASWAM et RAM alors que l'effet sur les mises à jour AM est moins clair ;

— L'utilisation de mises à jour locales diminue la performance de l'algorithme AM, mais augmente celle de l'algorithme ASWAM.

Général

— Les algorithmes adaptatifs arrivent systématiquement à battre la performance d'un algorithme non-adaptatif ;

— L'adaptation d'une densité globale n'a pas un effet important sur la performance des algorithmes ;

— L'adaptation des échelles améliore la performance des mises à jour RAM, mais l'effet de cette variante sur les algorithmes AM et ASWAM n'est pas constamment dans la même direction ;

— L'utilisation des mises à jour locales n'influence pas clairement la performance des algorithmes AM ou ASWAM d'une manière positive ou négative.

Expérience 2C. Les résultats de l'expérience 2C sur le nombre de propositions se trouvent à la figure 6.10. Tel qu'attendu, toutes les mesures montrent une augmentation de l'efficacité des algorithmes lorsque le nombre de propositions augmente. En inspectant les graphiques, on cherche à trouver une valeur de K qui soit aussi petite que possible tout en donnant une performance optimale.

Pour ce qui est du MSEJD, on note un comportement similaire entre les quatre algorithmes : la croissance initiale s'atténue avec K croissant pour atteindre un plateau. La relation entre les algorithmes est constante en fonction de K, alors qu'on observe l'ordre suivant : ASWAM, RAM, MTM et puis AM.

L'ESS croît légèrement avec K, mais peu d'augmentation est observée au-delà de $K = 8$. On trouve l'ordre suivant : AM, ASWAM, RAM et MTM.

Au niveau du taux d'acceptation, on remarque que l'algorithme ASWAM arrive bien au taux cible, et ce, peu importe le nombre de candidats. Les autres algorithmes, sans surprise, ont des taux d'acceptation empiriques croissant avec K étant donné que la sélection parmi un plus grand nombre de candidats fournira généralement des propositions de meilleure qualité.

La mesure d'ESS ajustée pour le temps de calcul est monotonement décroissante avec K. Les meilleures valeurs de cette mesure sont obtenues par les algorithmes AM et ASWAM.

La distance en variation totale permet d'identifier des algorithmes qui n'arrivent pas à bien représenter la distribution cible. Au-delà de $K = 5$ pour les algorithmes RAM et MTM et au-delà de $K = 8$ pour les algorithmes AM et ASWAM, cette mesure ne diminue plus significativement. Les plus basses valeurs de distance TV sont obtenues par les algorithmes AM et ASWAM. L'étude du biais d'estimation de x_2^2 engendre des observations similaires sur la performance des algorithmes par rapport au nombre de candidats.

D'une manière générale, il semblerait donc que les algorithmes AM et ASWAM sont plus performants que les algorithmes RAM et MTM, de façon uniforme par rapport au nombre de candidats. Tous ces algorithmes nécessitent un certain nombre de candidats (au moins 5 ou même 8) afin d'obtenir une performance optimale, ce qui appuie à nouveau la pertinence des algorithmes à essais multiples en comparaison à leurs équivalents à un seul candidat. En effet, les quatre algorithmes n'arrivent pas à bien échantillonner de la densité cible lorsqu'ils n'utilisent qu'un seul candidat, comme on peut le voir à l'aide des mesures de distance TV et de biais d'estimation de x_2^2.

Expérience 2D. Les résultats de l'expérience 2D sur le taux d'acceptation cible se trouvent à la figure 6.11. Notons que les algorithmes AM et MTM ne sont pas affectés par un changement de taux d'acceptation cible.

L'algorithme ASWAM atteint bien le taux d'acceptation cible tel que prévu par construction. Les différentes mesures indique différentes valeurs de taux cible optimal :

— MSEJD : $\alpha_* \in [0{,}40; 0{,}70]$;

— ACT et ESS : $\alpha_* \in [0{,}20; 0{,}55]$;

— Distance TV et Biais : $\alpha_* \in [0{,}20; 0{,}55]$.

Il semble donc qu'un taux cible entre 0,40 et 0,55 produise un algorithme ASWAM qui soit optimal uniformément par rapport à la mesure de performance.

L'algorithme RAM n'atteint pas quant à lui le taux d'acceptation cible (on rappelle que ceci n'est pas surprenant vu qu'il ne s'agit pas d'une garantie de l'algorithme). Cet algorithme est optimal pour un taux cible dans les intervalles suivants :

— MSEJD : $\alpha_* \in [0{,}30; 0{,}55]$;

— ACT et ESS : $\alpha_* \in [0{,}20; 0{,}50]$;

— Distance TV et Biais : $\alpha_* \in [0{,}20; 0{,}80]$.

Il semble donc qu'un taux cible entre 0,30 et 0,55 produise un algorithme RAM qui soit optimal uniformément par rapport à la mesure de performance.

Expérience 2E. Les résultats de l'expérience 2E sur l'effet du pas d'adaptation se trouvent à la figure 6.12. Évidemment, l'algorithme MTM n'est pas affecté par le rythme de décroissance du pas d'adaptation étant donné que cet algorithme n'effectue pas d'adaptation.

Les algorithmes AM et ASWAM affichent un comportement similaire par rapport au paramètre γ : tous deux produisent des chaînes plus efficaces lorsque γ décroît. En effet, les valeurs du MSEJD, de l'ACT et de l'ESS empirent lorsque γ augmente ; l'algorithme ASWAM atteint le taux d'acceptation cible pour $\gamma \in [0{,}35; 0{,}65]$ seulement ; le taux d'acceptation de l'algorithme AM augmente lorsque γ décroît. Cependant, les mesures vérifiant que les chaînes représentent bien la distribution cible, c.-à-d., le biais d'estimation de x_2^2 et la distance TV, indiquent plutôt que de petites valeurs de γ ne sont pas souhaitables. Il semble que $\gamma > 0{,}85$ soit nécessaire pour que la distribution empirique soit le plus près possible de la distribution cible.

L'algorithme RAM, quant à lui, montre une performance relativement stable sur le domaine considéré. On observe une légère augmentation de l'efficacité (MSEJD surtout) ainsi qu'un meilleur ajustement à la distribution cible (distance TV) pour de petites valeurs de γ.

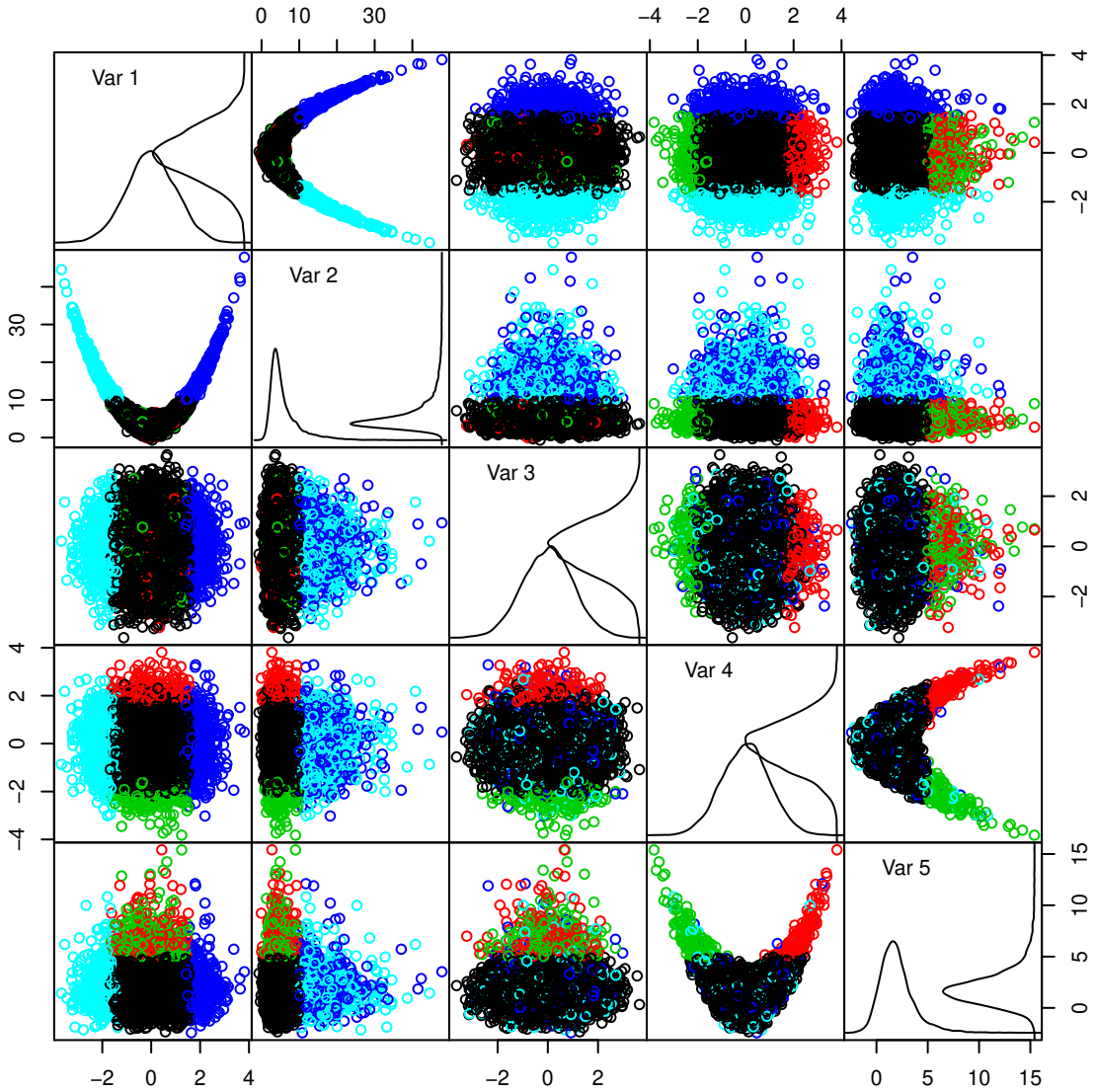


Figure 6.9 *Graphique par paires d'un échantillon i.i.d. de la densité à géométrie complexe utilisée dans l'expérience 2. Sur la diagonale se trouve une estimation non-paramétrique des densités marginales. Les couleurs représentent les cinq régions définies pour l'estimation de la distance en variation totale.*

Expérience 2A - Statistiques en fonction du type de proposition et de poids							
		MSEJD		ESS		Distance TV	
Proposition	Poids	Moyenne	Err. std.	Moyenne	Err. std.	Moyenne	Err. std.
Sans adaptation (MTM)							
Indépendante	Importance	0.263	(0.002)	312	(9)	0.182	(0.007)
	Prop. cible	0.324	(0.001)	715	(19)	**0.046**	(0.003)
V.A. Commune	Importance	0.239	(0.001)	**752**	(21)	0.047	(0.002)
	Prop. cible	0.303	(0.001)	696	(18)	0.048	(0.003)
QMCR	Importance	0.265	(0.002)	355	(11)	0.155	(0.007)
	Prop. cible	**0.331**	(0.001)	**764**	(16)	**0.044**	(0.002)
EA	Importance	0.302	(0.002)	340	(9)	0.163	(0.006)
	Prop. cible	0.324	(0.001)	737	(20)	0.050	(0.004)
Mises à jour AM							
Indépendante	Importance	0.218	(0.001)	781	(12)	0.071	(0.001)
	Prop. cible	0.255	(0.001)	1368	(18)	**0.035**	(0.001)
V.A. Commune	Importance	0.206	(0.001)	1312	(15)	0.045	(0.001)
	Prop. cible	0.211	(0.001)	1308	(18)	0.048	(0.002)
QMCR	Importance	0.229	(0.001)	826	(12)	0.070	(0.001)
	Prop. cible	**0.272**	(0.001)	**1489**	(19)	0.039	(0.001)
EA	Importance	0.242	(0.001)	729	(9)	0.093	(0.001)
	Prop. cible	0.255	(0.001)	1412	(19)	0.039	(0.001)
Mises à jour ASWAM							
Indépendante	Importance	0.390	(0.003)	404	(7)	0.127	(0.002)
	Prop. cible	0.358	(0.003)	1100	(17)	**0.045**	(0.002)
V.A. Commune	Importance	0.283	(0.001)	1005	(14)	0.069	(0.002)
	Prop. cible	0.288	(0.001)	956	(14)	0.056	(0.002)
QMCR	Importance	0.433	(0.003)	505	(8)	0.100	(0.002)
	Prop. cible	0.386	(0.002)	**1192**	(18)	0.052	(0.001)
EA	Importance	**0.459**	(0.004)	505	(9)	0.119	(0.002)
	Prop. cible	0.362	(0.003)	**1177**	(18)	0.055	(0.002)
Mises à jour RAM							
Indépendante	Importance	0.262	(0.002)	260	(8)	0.180	(0.008)
	Prop. cible	0.368	(0.001)	799	(19)	0.041	(0.002)
V.A. Commune	Importance	0.247	(0.001)	579	(16)	0.058	(0.005)
	Prop. cible	0.325	(0.001)	675	(16)	0.045	(0.003)
QMCR	Importance	0.262	(0.002)	260	(8)	0.184	(0.009)
	Prop. cible	**0.378**	(0.001)	**833**	(16)	**0.035**	(0.002)
EA	Importance	0.301	(0.002)	277	(9)	0.176	(0.007)
	Prop. cible	0.368	(0.001)	808	(18)	0.043	(0.003)

Tableau 6.4 *(Expérience 2A) Statistiques en fonction du type de candidat et de la fonction de poids pour les quatre algorithmes principaux. Les statistiques en gras indiquent le meilleur choix par type de mise à jour et celles soulignées indiquent le meilleur choix global.*

Expérience 2B - Statistiques en fonction du type d'adaptation								
Adaptation			MSEJD		ESS		Distance TV	
Globale	Échelle	Locale	Moyenne	Err. std.	Moyenne	Err. std.	Moyenne	Err. std.
Sans adaptation (MTM)								
–	–	–	0.331	(0.001)	788	(17)	0.046	(0.002)
Mises à jour AM								
Non	Non	Non	0.271	(0.001)	1504	(22)	**0.040**	(0.001)
		Oui	0.221	(0.002)	573	(13)	0.052	(0.004)
	Oui	Non	0.214	(0.002)	660	(11)	0.048	(0.002)
		Oui	0.036	(0.002)	1283	(70)	0.226	(0.019)
Oui	Non	Non	**0.279**	(0.001)	**1603**	(22)	0.048	(0.001)
		Oui	0.209	(0.002)	459	(10)	0.073	(0.004)
	Oui	Non	0.216	(0.002)	938	(14)	0.087	(0.002)
		Oui	0.032	(0.002)	**1697**	(133)	0.225	(0.021)
Mises à jour ASWAM								
Non	Non	Non	0.385	(0.002)	1212	(18)	0.054	(0.001)
		Oui	0.405	(0.003)	929	(19)	**0.030**	(0.002)
	Oui	Non	0.457	(0.001)	1387	(21)	0.035	(0.001)
		Oui	0.220	(0.002)	733	(16)	0.038	(0.002)
Oui	Non	Non	0.410	(0.001)	1419	(21)	0.069	(0.001)
		Oui	**0.472**	(0.002)	1043	(21)	**0.030**	(0.002)
	Oui	Non	0.454	(0.001)	**1551**	(21)	0.044	(0.001)
		Oui	0.225	(0.002)	831	(18)	0.034	(0.002)
Mises à jour RAM								
Non	Non	–	0.377	(0.001)	827	(21)	0.041	(0.002)
	Oui	–	0.567	(0.002)	**1215**	(33)	**0.038**	(0.002)
Oui	Non	–	0.383	(0.001)	848	(21)	0.043	(0.002)
	Oui	–	**0.581**	(0.002)	1174	(31)	0.041	(0.003)

Tableau 6.5 *(Expérience 2B) Statistiques en fonction du type d'adaptation (proposition globale, adaptation de l'échelle, mise à jour locale) pour les quatre algorithmes principaux. Les statistiques en gras indiquent le meilleur choix par type de mise à jour et celles soulignées indiquent le meilleur choix global.*

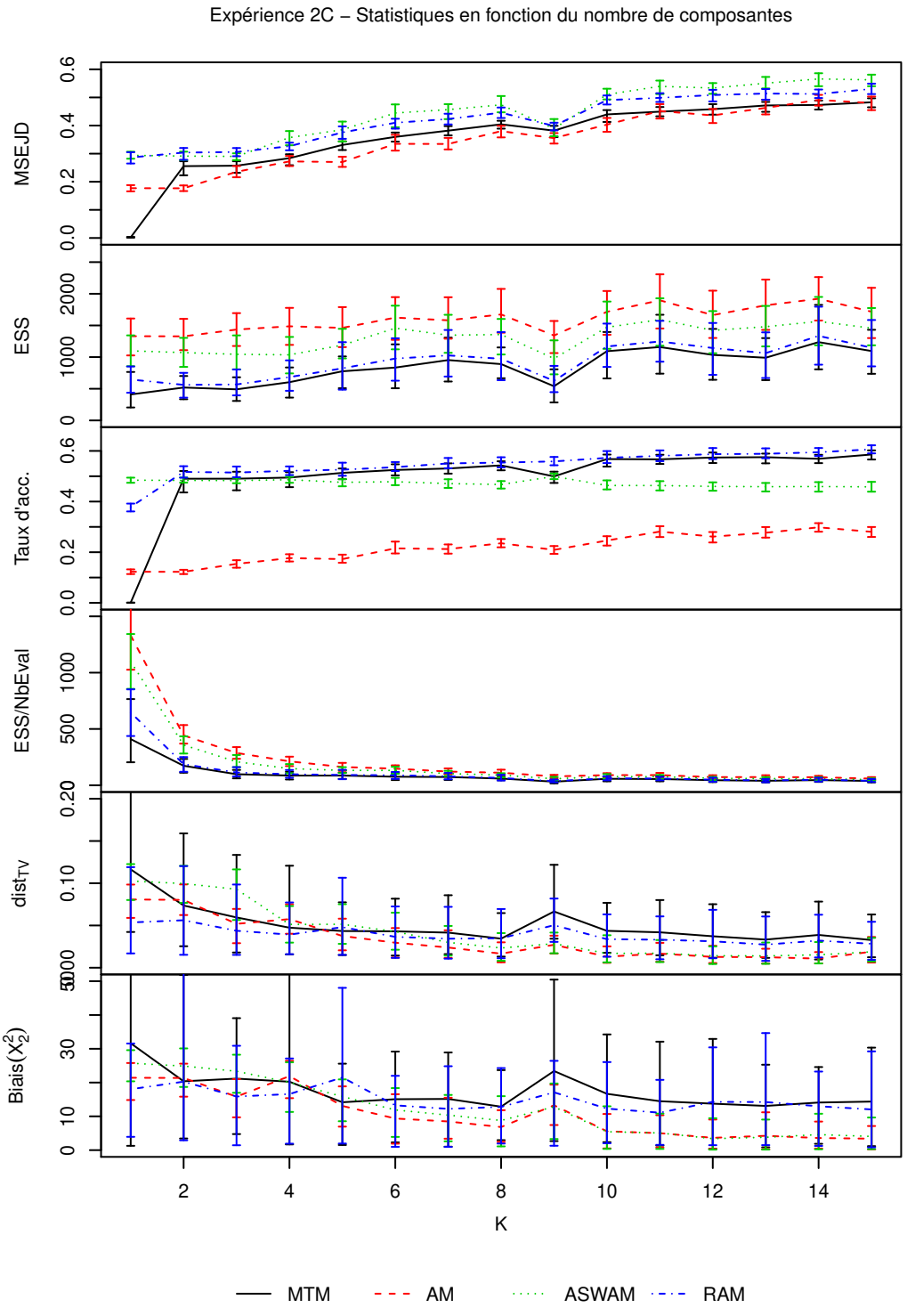


Figure 6.10 *(Expérience 2C) Statistiques en fonction du nombre de propositions pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

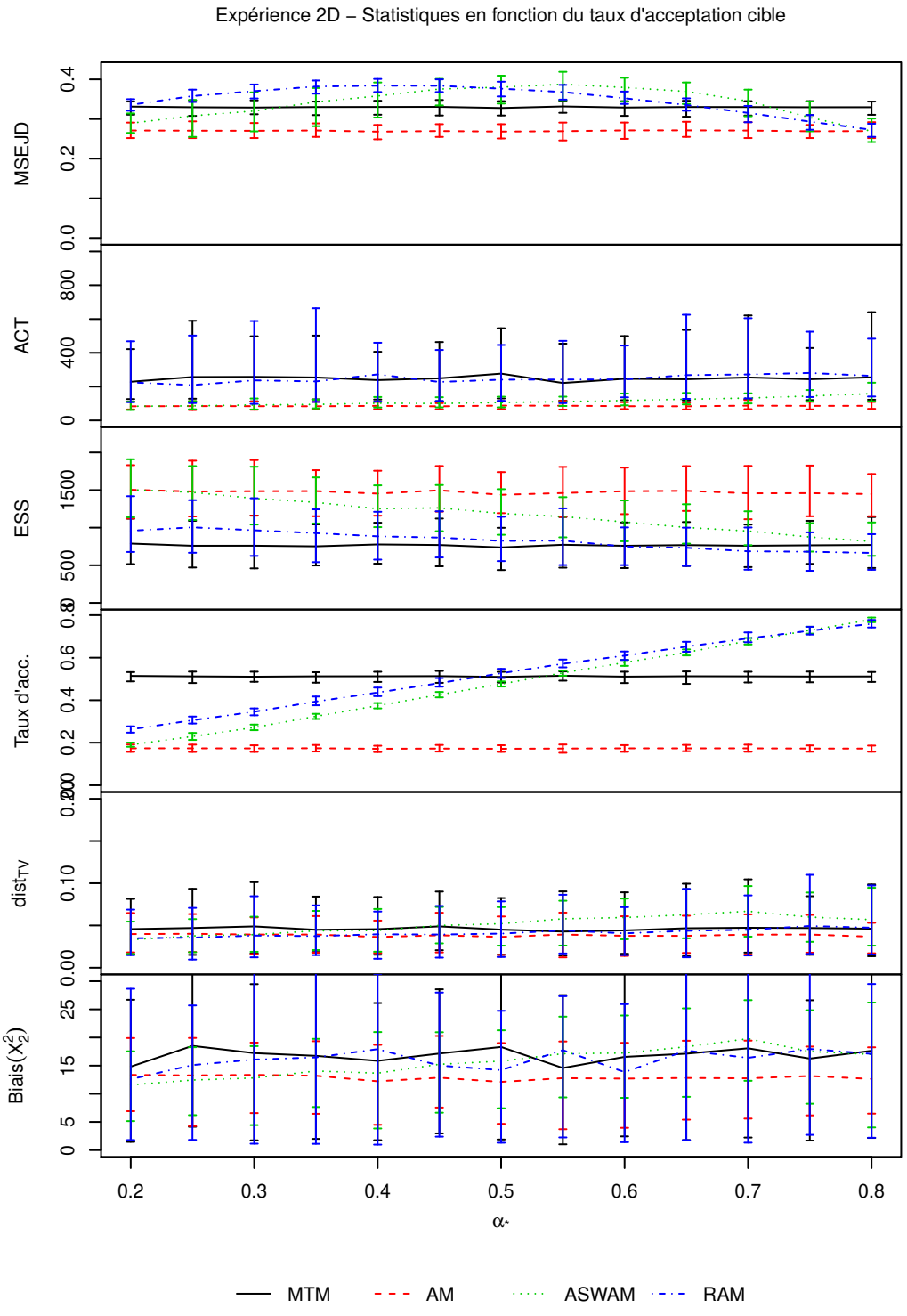


Figure 6.11 *(Expérience 2D) Statistiques en fonction du taux d'acceptation cible pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

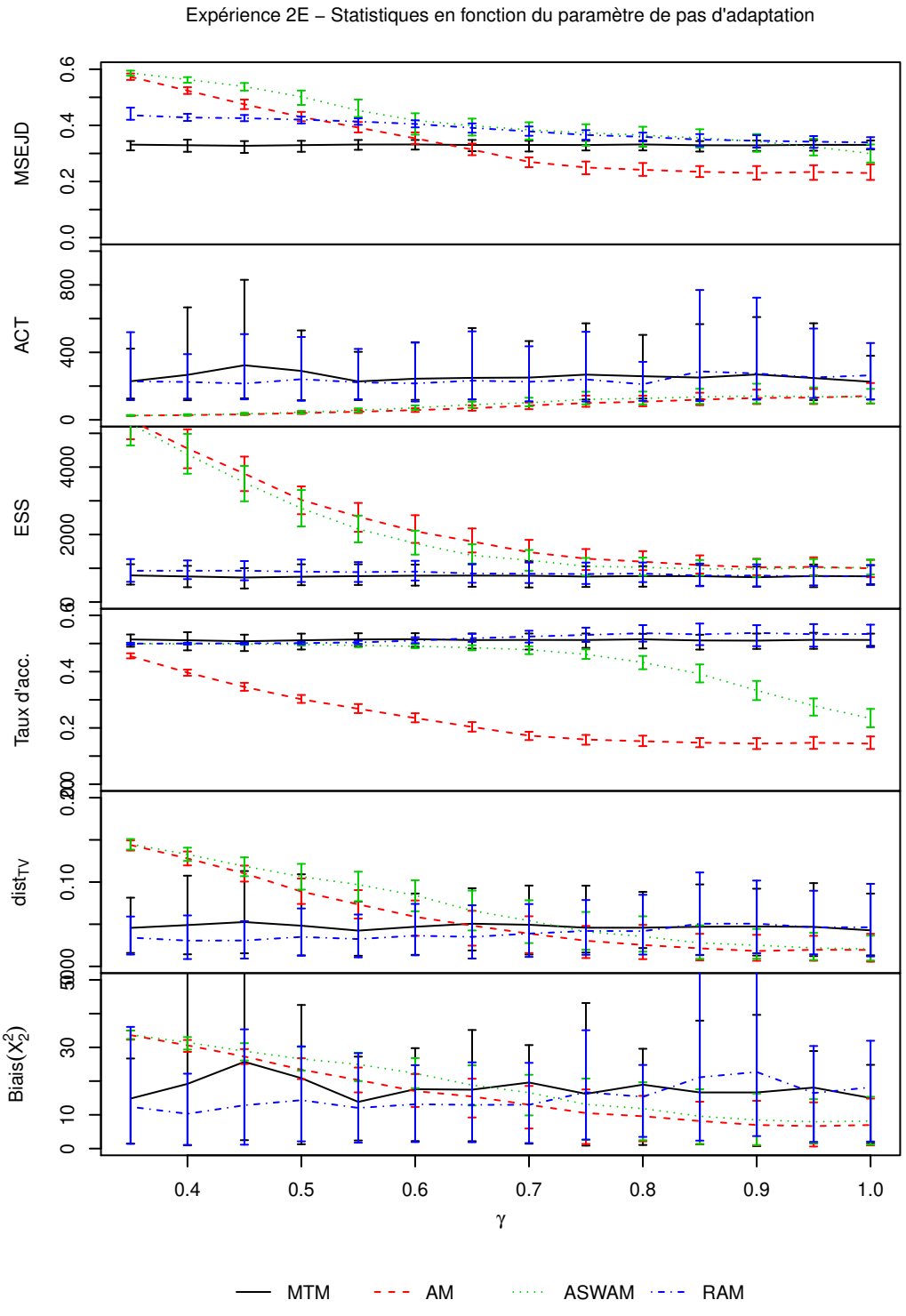


Figure 6.12 *(Expérience 2E) Statistiques en fonction du paramètre de pas d'adaptation pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

6.3.3 Densité multimodale en hautes dimensions

6.3.3.1 Description de l'expérience

Dans cette troisième expérience de simulations, nous étudions la capacité de l'algorithme aMTM à s'adapter à une distribution cible en dimensions relativement plus élevées. Dû à la fameuse malédiction de la dimensionalité, il s'agit d'un problème ardu pour tout algorithme MCMC ou, plus généralement, pour toute méthode d'intégration numérique. Même pour un nombre de dimensions modérément élevé, des méthodes par composante (e.g algorithme *Metropolis-within-Gibbs* 2.7 et ses nombreuses variantes) sont souvent préférées, mais on explore tout de même le comportement de l'algorithme aMTM dans cette situation.

En grandes dimensions, il est difficile d'identifier la ou les directions pertinentes vers lesquelles proposer de nouveaux candidats : la majorité des directions pointent vers des régions de faible densité. L'adaptation d'une covariance complète est donc tout aussi ardue puisque certaines régions à visiter peuvent être difficile à découvrir ou à atteindre. Il est donc primordial d'initialiser les covariances instrumentales en se basant sur de grandes échelles afin de faciliter la découverte des régions pertinentes.

La distribution cible qui sera utilisée est un mélange de trois densités normales en $d = 20$ dimensions où les moyennes et les covariances des trois composantes sont générées aléatoirement pour produire trois modes qui sont distincts, mais relativement près les uns des autres. Les trois modes définissent trois régions qui peuvent être décrites par des classificateurs linéaires. Cet ensemble de classificateurs arrive à bien prévoir la composante de 64,4% des points d'un échantillon i.i.d., ce qui indique que les trois modes sont bien distincts, mais que ceux-ci ne sont pas complètement séparés par des régions de faible densité. La figure 6.13 montre la projection de l'échantillon i.i.d. sur le plan effectuant la meilleure séparation linéaire des composantes du mélange. Ces mêmes régions seront utilisées pour produire un estimé de la distance en variation totale entre la distribution empirique d'un échantillon et la distribution cible.

La distribution cible a une covariance qui contient des corrélations faibles : la plus forte corrélation, en valeur absolue, est de 0,36. De plus les échelles des différentes dimensions varient très peu : les variances marginales se trouvent dans l'intervalle $[14.7; 31.2]$. Ainsi, l'initialisation des covariances instrumentales à un multiple de l'identité est relativement bien spécifiée de sorte qu'un algorithme sans adaptation n'est pas fortement défavorisé par l'expérience.

La méthodologie des expériences 1 et 2 (sections 6.3.1 et 6.3.2) est reprise intégralement : les cinq mêmes sous-expériences sont effectuées afin d'étudier les différentes variantes des algorithmes et de comparer l'algorithme aMTM aux algorithmes plus simples. L'échelle des covariances initiales s'étend sur l'intervalle $[0{,}1; 10\,000]$ à l'exception de l'expérience 3C où l'on considère l'intervalle [100,10000] afin de ne pas trop désavantager les algorithmes avec peu de composantes.

6.3.3.2 Résultats et analyse

Expérience 3A. Les résultats de l'expérience 3A sur le type de proposition et sur la fonction de poids se trouvent au tableau 6.6. Voici les observations qu'on peut en tirer :

MSEJD

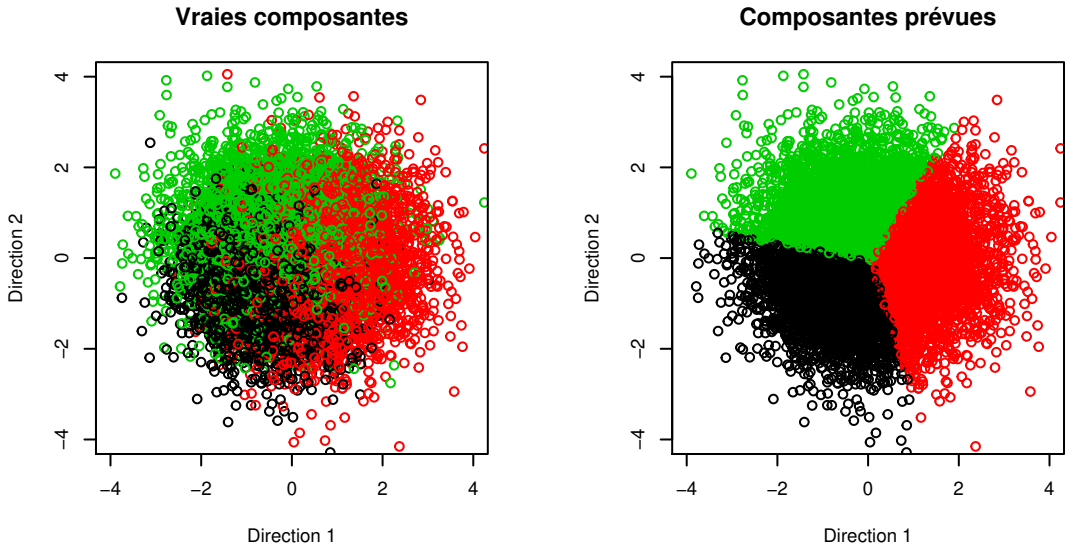


Figure 6.13 *(Expérience 3) Définition des régions pour l'estimation de la distance en variation totale. À gauche, le graphique des vraies composantes utilisées pour générer l'échantillon i.i.d. et, à droite, les composantes prévues par l'analyse discriminante linéaire. Les axes représentent les deux directions séparant le mieux les trois composantes du mélange.*

— Les algorithmes AM et ASWAM présentent des valeurs de MSEJD plus élevées que celles des algorithmes MTM et RAM ;

— À l'exception des candidats par variable aléatoire commune qui performent systématiquement moins bien, les autres types de candidats sont relativement équivalents, et ce, pour les quatre algorithmes ;

— Les poids par importance produisent des chaînes à MSEJD plus élevés chez les mises à jour AM et ASWAM alors qu'on observe des résultats mixtes pour les algorithmes MTM et RAM ;

— Les quatre algorithmes sont optimisés pour des candidats extrêmement antithétiques et des poids par importance.

ESS

— L'ESS ne semble pas une bonne mesure comparative pour cette expérience puisqu'une chaîne où un mode est sous-représenté obtiens une meilleure valeur d'ESS ; en effet, il est plus facile d'échantillonner une densité qui comporte seulement un ou deux modes. Ceci est particulièrement visible en observant le lien direct entre un ESS élevé et une distance en variation totale élevée (e.g. candidats par variable aléatoire commune avec poids par importance pour les quatre algorithmes).

Distance TV

— Dans l'ordre, les algorithmes produisant les chaînes les plus représentatives de la densité cible sont l'algorithme RAM, MTM, AM et ASWAM.

— À l'exception des candidats par variable aléatoire commune, les trois autres types de candidats sont relativement équivalents pour cette mesure ;

— Les poids par importance sont préférables pour les algorithmes AM et ASWAM, alors que les poids proportionnels à la densité cible semblent un meilleur choix pour les algorithmes MTM et RAM.

Général

Cette expérience montre bien que le choix de mesure de performance influence grandement l'interprétation qu'on peut faire de l'efficacité relative d'algorithmes MCMC. Un premier critère devrait d'abord être la représentativité de la chaîne vis-à-vis la densité cible. La distance en variation totale estimée permet d'identifier si une chaîne échantillonne bel et bien de la bonne densité cible ou bien si un mode est sous représenté, voire absent, de l'échantillon. Ensuite, une fois que l'on s'assure que la distribution empirique est suffisamment similaire à la distribution cible, on peut alors comparer l'efficacité d'estimation à l'aide de mesures telles que le MSEJD ou l'ESS. Malgré cela, il est tout de même possible de dégager certaines généralités :

— D'abord, l'algorithme RAM semble le plus efficace afin d'échantillonner l'ensemble du support de la distribution cible ;

— Ensuite, les candidats indépendants, quasi-Monte Carlo randomisés et extrêmement antithétiques sembles tous trois plus efficaces que les candidats par variable aléatoire commune ;

— Enfin, il n'est pas clair, à nouveau, qu'un choix de poids soit systématiquement supérieur à l'autre.

Expérience 3B. Les résultats de l'expérience 3B sur les variantes d'adaptation se trouvent au tableau 6.7. Voici les observations qu'on peut en tirer :

MSEJD

— Les plus hautes valeurs de MSEJD sont obtenues, dans l'ordre, par les algorithmes ASWAM, AM, RAM et MTM ;

— L'adaptation d'une densité globale semble très légèrement améliorer cette mesure ;

— L'adaptation des échelles nuit à la performance des algorithmes AM et ASWAM alors qu'elle améliore celle de l'algorithme RAM ;

— L'utilisation de mises à jour locales réduit significativement la valeur de cette mesure pour les algorithmes AM et ASWAM.

ESS

— À nouveau, on observe que l'ESS n'est pas un bonne mesure comparative étant donné que des valeurs élevées d'ESS sont liées à de grandes valeurs de distance en variation totale ;

— Cependant, pour les variantes affichant de petites valeurs de distance TV (inférieures à 0,124), on trouve que les valeurs les plus élevées s'observent pour l'algorithme ASWAM avec mises à jour locales et sans adaptation des échelles.

Distance TV

— Aucune variante de l'algorithme AM n'arrive à produire des chaînes affichant des estimés de la distance TV suffisamment basses pour être comparées ;

— L'algorithme ASWAM produit de faibles valeurs de distance TV seulement pour des mises à jour locales et sans adaptation des échelles ; l'adaptation d'une densité globale ne semble pas avoir d'effet ;

— L'algorithme RAM semble profiter légèrement de l'adaptation des échelles ainsi que de l'adaptation d'une densité globale.

Général

— Tout comme à l'expérience 3A, les différentes mesures nous mènent à des conclusions bien variées. En considérant d'abord la distance en variation totale, on trouve que l'algorithme ASWAM peut battre les algorithmes MTM et RAM dans certains cas alors que l'algorithme RAM est systématiquement supérieur à l'algorithme MTM.

— Les expériences 1B et 2B ont soulevé une observation importante sur l'algorithme RAM qui semble être confirmée à nouveau. En effet, ce type de mise à jour semble être moins rapide pour modifier les densités instrumentales, notamment les échelles, de sorte que l'adaptation des échelles améliore souvent cet algorithme. Dans l'expérience actuelle, ce type d'adaptation (plus lente) semble être profitable pour l'algorithme : une convergence trop rapide des densités instrumentales peut empêcher la découverte de régions d'intérêt dans le support de la densité cible.

Expérience 3C. Les résultats de l'expérience 3C sur le nombre de propositions se trouvent à la figure 6.14. De façon générale, toutes les mesures montrent une augmentation de l'efficacité des algorithmes lorsque le nombre de propositions augmente.

Les quatre algorithmes produisent des chaînes à distance en variation totale relativement constantes à partir de $K = 4$ candidats. L'algorithme AM est cependant beaucoup moins efficace que les autres algorithmes pour produire des chaînes qui représentent bien la distribution cible. Le biais d'estimation de x_7 semble montrer aussi que $K = 4$ candidats est suffisant pour atteindre l'optimalité. En particulier, n'utiliser qu'un seul candidat est particulièrement inefficace pour bien explorer l'ensemble de la distribution cible.

Le MSEJD augmente bien avec K pour les algorithmes MTM, ASWAM et RAM, mais l'augmentation ralentit avec K de sorte qu'un plateau est atteint dès $K = 6$ pour l'algorithme ASWAM et $K = 10$ pour les algorithmes MTM et RAM.

Sur les domaines à faible distance TV, l'ESS est relativement constant pour tous les algorithmes. Quant à l'ESS ajusté pour le temps de calcul, c.-à-d., l'ESS/NbEval, on note une décroissance monotone avec K de sorte qu'on cherche à minimiser le nombre de candidats utilisés tout en s'assurant que la distance en variation totale soit aussi minimisée.

Aucune mesure ne peut distinctement indiquer un nombre idéal de candidats à utiliser, mais il semble qu'environ $K = 4$ candidats soit recommandable pour cette distribution cible.

Expérience 3D. Les résultats de l'expérience 3D sur le taux d'acceptation cible se trouvent à la figure 6.15. Notons que les algorithmes AM et MTM ne sont pas affectés par un changement de taux d'acceptation cible.

ASWAM

— Les mises à jour ASWAM atteignent bien le taux d'acceptation cible sur l'ensemble du domaine considéré ;

— En inspectant la distance en variation totale, on remarque que les mises à jour ASWAM sont particulièrement inefficaces au-delà de $\alpha_* = 0{,}40$;

— Le MSEJD est maximal sur l'intervalle $\alpha_* \in [0{,}25; 0{,}45]$;

— L'ACT est minimal sur l'intervalle $\alpha_* \in [0{,}10; 0{,}30]$ (en considérant seulement le domaine où la distance TV est acceptable) ;

— L'ESS est maximal au-delà de $\alpha_* = 0{,}40$, mais la chaîne ne représente pas bien la distribution cible sur cet intervalle. L'ESS est relativement constant en-deçà de ce taux cible ;

— Le biais d'estimation de x_7 est minimal sur $\alpha_* \in [0{,}10; 0{,}25]$;

— En combinant ces observations, on trouve que l'algorithme ASWAM semble optimal, pour cette distribution cible, pour un taux d'acceptation cible de $\alpha_* \approx 0{,}25$.

RAM

— On note d'abord que les mises à jour RAM n'atteignent pas leur taux d'acceptation cible, mais présentent tout de même une augmentation du taux observé suivant l'augmentation du taux cible ;

— La distance en variation totale est constante sur le domaine ;

— Le MSEJD croît légèrement avec α_* ;

— L'ACT croît légèrement avec α_* ;

— L'ESS diminue légèrement avec α_* ;

— Le biais d'estimation de x_7 semble être légèrement moindre pour de petites valeurs de α_* ;

— En combinant ces observations, les mises à jour RAM ne semble pas grandement influencées par le taux d'acceptation cible et une valeur médiane telle que $\alpha_* = 0{,}5$ semble bien convenir.

Expérience 3E. Les résultats de l'expérience 3E sur l'effet du paramètre du pas d'adaptation se trouvent à la figure 6.16. Évidemment, l'algorithme MTM n'est pas affecté par le rythme de décroissance du pas d'adaptation étant donné que cet algorithme n'effectue pas d'adaptation.

AM

— En inspectant la distance en variation totale, l'algorithme AM ne semble produire des chaînes représentatives que pour $\gamma \in [0{,}80; 1{,}00]$;

— Le MSEJD est maximal sur l'intervalle $\gamma \in [0{,}65; 0{,}85]$;

— L'ACT augmente avec γ croissant, mais affiche un comportement erratique pour $\gamma \in [0{,}40; 0{,}50]$;

— L'ESS diminue avec γ croissant ;

— Le taux d'acceptation observé n'a pas de valeur optimale connue pour cette distribution cible, mais notons que les mises à jour AM s'approche de la valeur de 0,25 jugée optimale pour l'algorithme ASWAM lorsque γ s'approche de 1,00 ;

— Le biais d'estimation de x_7 est minimal sur $\gamma \in [0{,}75; 1{,}00]$;

— En combinant ces observations, on trouve que l'algorithme AM semble optimal, pour cette distribution cible, pour $\gamma \approx 0{,}85$.

ASWAM

— Au niveau de la distance en variation totale, l'algorithme ASWAM affiche un comportement variable sur le domaine considéré, mais semble être moins efficace sur l'intervalle $\gamma \in [0{,}50; 0{,}60]$ et affiche sa meilleure performance pour $\gamma \approx 0{,}90$;

— Le MSEJD est maximal sur l'intervalle $\gamma \in [0{,}40; 0{,}55]$;

— L'ACT est minimal sur l'intervalle $\gamma \in [0{,}40; 0{,}55]$;

— Pour les régions à faible distance en variation totale, l'algorithme ASWAM a une mesure d'ESS élevée pour $\gamma \in [0{,}40; 0{,}45]$ et pour $\gamma \approx 0{,}80$;

— Les mises à jour ASWAM atteignent bien le taux d'acceptation cible pour $\gamma \in [0{,}40; 0{,}70]$ et produisent un taux observé relativement plus élevé au-delà de $\gamma = 0{,}70$;

— Le biais d'estimation de x_7 est élevé sur l'ensemble du domaine mais diminue légèrement pour $\gamma \in [0{,}90; 1{,}00]$;

— Notons qu'un taux cible de $\alpha_* = 0{,}50$ n'est pas optimal pour cet algorithme et cette densité cible (voir expérience 3D) de sorte que ces résultats sont faussés pour cette raison. L'algorithme semble toutefois plus robuste lorsque γ est près de 1,00.

RAM

— Les mises à jour RAM ont une faible distance en variation totale sur l'ensemble du domaine ;

— Le MSEJD est plus élevé pour $\gamma \in [0{,}40; 0{,}60]$;

— L'ACT est légèrement plus faible pour $\gamma \in [0{,}50; 1{,}00]$;

— L'ESS est sensiblement constant sur le domaine ;

— Le taux d'acceptation cible s'approche du taux cible lorsque γ diminue est l'atteint presque pour $\gamma \approx 0{,}40$;

— Le biais d'estimation de x_7 semble constant sur le domaine ;

— En combinant ces observations, les mises à jour RAM ne semblent pas grandement influencées par le pas d'adaptation, mais une diminution plus lente (faibles γ) semble être légèrement plus optimale. En particulier, $\gamma = 0{,}50$ est uniformément optimal par rapport à la mesure considérée.

Expérience 3A - Statistiques en fonction du type de proposition et de poids							
		MSEJD		ESS		Distance TV	
Proposition	Poids	Moyenne	Err. std.	Moyenne	Err. std.	Moyenne	Err. std.
Sans adaptation							
Indépendante	Importance	0.239	(0.001)	156	(3)	**0.145**	(0.005)
	Prop. cible	0.254	(0.002)	355	(4)	**0.129**	(0.006)
V.A. Commune	Importance	0.016	(0.000)	**703**	(40)	0.241	(0.011)
	Prop. cible	0.185	(0.002)	370	(5)	0.151	(0.008)
QMCR	Importance	0.239	(0.001)	151	(3)	0.143	(0.006)
	Prop. cible	0.255	(0.002)	331	(4)	**0.125**	(0.007)
EA	Importance	**0.291**	(0.001)	150	(2)	0.140	(0.005)
	Prop. cible	0.255	(0.002)	363	(3)	**0.123**	(0.006)
Mises à jour AM							
Indépendante	Importance	1.390	(0.036)	1218	(54)	**0.171**	(0.005)
	Prop. cible	0.784	(0.023)	1899	(102)	0.239	(0.005)
V.A. Commune	Importance	0.531	(0.006)	**2654**	(51)	0.301	(0.004)
	Prop. cible	0.409	(0.006)	1448	(36)	0.259	(0.005)
QMCR	Importance	1.534	(0.032)	1385	(46)	0.188	(0.005)
	Prop. cible	0.791	(0.021)	1772	(68)	0.234	(0.005)
EA	Importance	**1.741**	(0.043)	1338	(59)	**0.173**	(0.005)
	Prop. cible	0.790	(0.021)	1837	(79)	0.231	(0.005)
Mises à jour ASWAM							
Indépendante	Importance	1.255	(0.015)	859	(20)	0.180	(0.004)
	Prop. cible	0.918	(0.020)	2604	(107)	0.256	(0.004)
V.A. Commune	Importance	0.624	(0.002)	**3154**	(30)	0.345	(0.004)
	Prop. cible	0.512	(0.006)	1555	(38)	0.273	(0.005)
QMCR	Importance	1.226	(0.017)	797	(20)	0.182	(0.004)
	Prop. cible	0.867	(0.019)	2291	(98)	0.249	(0.005)
EA	Importance	**1.550**	(0.027)	980	(34)	**0.174**	(0.005)
	Prop. cible	0.927	(0.018)	2678	(120)	0.255	(0.004)
Mises à jour RAM							
Indépendante	Importance	0.249	(0.001)	156	(3)	0.142	(0.007)
	Prop. cible	0.278	(0.003)	364	(4)	0.112	(0.006)
V.A. Commune	Importance	0.026	(0.001)	**482**	(9)	0.194	(0.008)
	Prop. cible	0.196	(0.002)	355	(3)	0.120	(0.006)
QMCR	Importance	0.249	(0.001)	149	(3)	0.155	(0.007)
	Prop. cible	0.269	(0.002)	330	(3)	**0.105**	(0.005)
EA	Importance	**0.299**	(0.001)	146	(2)	0.137	(0.005)
	Prop. cible	0.276	(0.002)	365	(3)	**0.108**	(0.006)

Tableau 6.6 *(Expérience 3A) Statistiques en fonction du type de candidat et de la fonction de poids pour les quatre algorithmes principaux. Les statistiques en gras indiquent le meilleur choix par type de mise à jour et celles soulignées indiquent le meilleur choix global.*

Expérience 3B - Statistiques en fonction du type d'adaptation								
Adaptation			MSEJD		ESS		Distance TV	
Globale	Échelle	Locale	Moyenne	Err. std.	Moyenne	Err. std.	Moyenne	Err. std.
Sans adaptation								
–	–	–	0.254	(0.002)	367	(4)	0.123	(0.006)
Mises à jour AM								
Non	Non	Non	0.757	(0.021)	1737	(85)	0.232	(0.005)
		Oui	0.021	(0.001)	1104	(82)	0.295	(0.014)
	Oui	Non	0.281	(0.008)	747	(32)	**0.203**	(0.009)
		Oui	0.050	(0.001)	1786	(63)	0.384	(0.014)
Oui	Non	Non	**0.901**	(0.016)	**2386**	(92)	0.256	(0.004)
		Oui	0.033	(0.002)	2227	(591)	0.264	(0.012)
	Oui	Non	0.333	(0.006)	1271	(36)	0.263	(0.008)
		Oui	0.054	(0.001)	1814	(68)	0.367	(0.015)
Mises à jour ASWAM								
Non	Non	Non	0.920	(0.017)	2633	(120)	0.256	(0.004)
		Oui	0.387	(0.004)	423	(4)	**0.089**	(0.005)
	Oui	Non	0.374	(0.007)	762	(33)	0.188	(0.009)
		Oui	0.157	(0.002)	630	(25)	0.245	(0.010)
Oui	Non	Non	**0.983**	(0.015)	**3312**	(114)	0.278	(0.003)
		Oui	0.375	(0.005)	428	(5)	0.109	(0.006)
	Oui	Non	0.433	(0.008)	1191	(46)	0.251	(0.007)
		Oui	0.143	(0.002)	547	(23)	0.213	(0.010)
Mises à jour RAM								
Non	Non	–	0.276	(0.002)	366	(3)	0.118	(0.005)
	Oui	–	0.398	(0.003)	**396**	(3)	0.101	(0.006)
Oui	Non	–	0.274	(0.002)	366	(3)	0.103	(0.005)
	Oui	–	**0.409**	(0.002)	390	(2)	**0.094**	(0.005)

Tableau 6.7 *(Expérience 3B) Statistiques en fonction du type d'adaptation (proposition globale, adaptation de l'échelle, mise à jour locale) pour les quatre algorithmes principaux. Les statistiques en gras indiquent le meilleur choix par type de mise à jour et celles soulignées indiquent le meilleur choix global.*

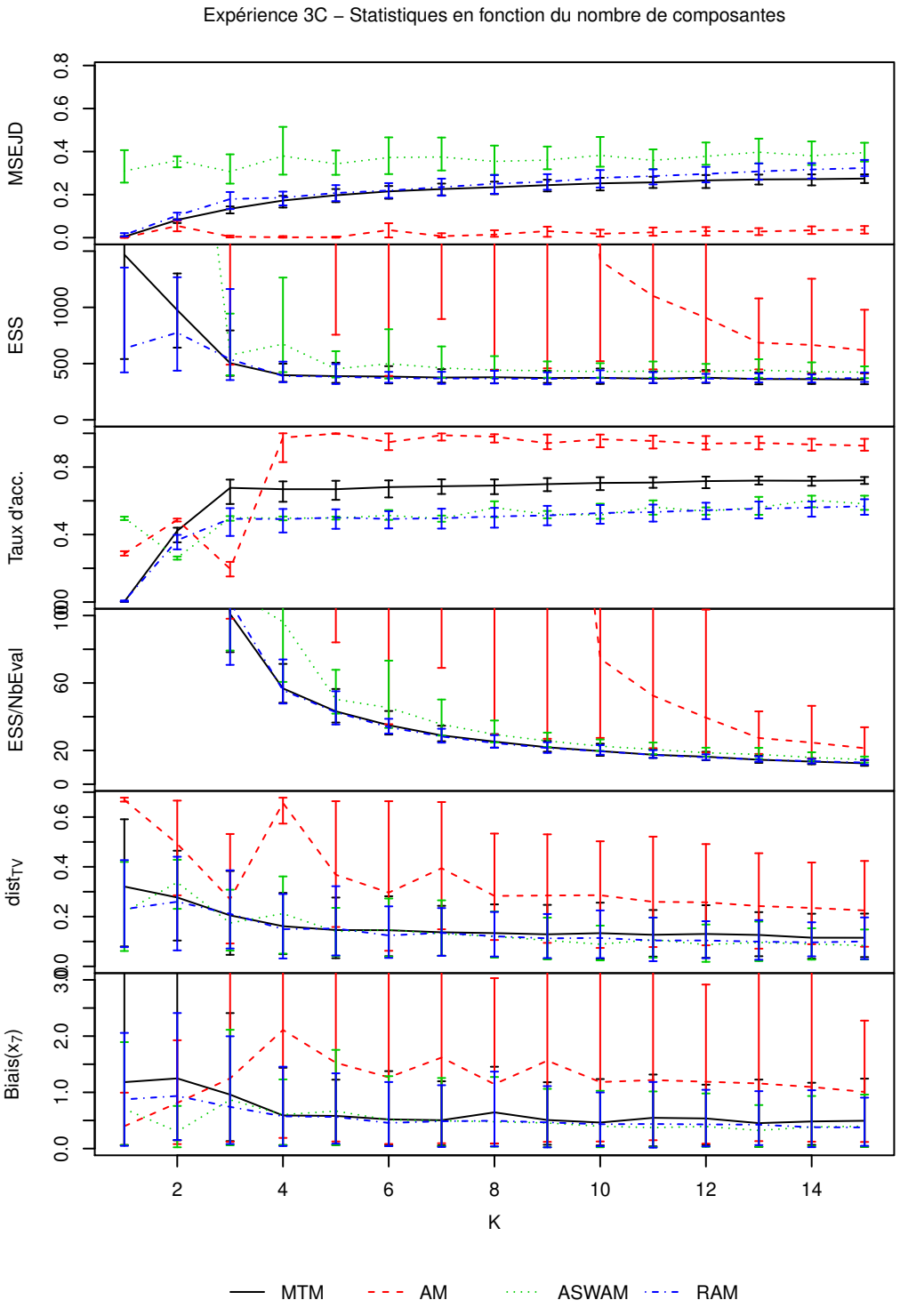


Figure 6.14 *(Expérience 3C) Statistiques en fonction du nombre de propositions pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

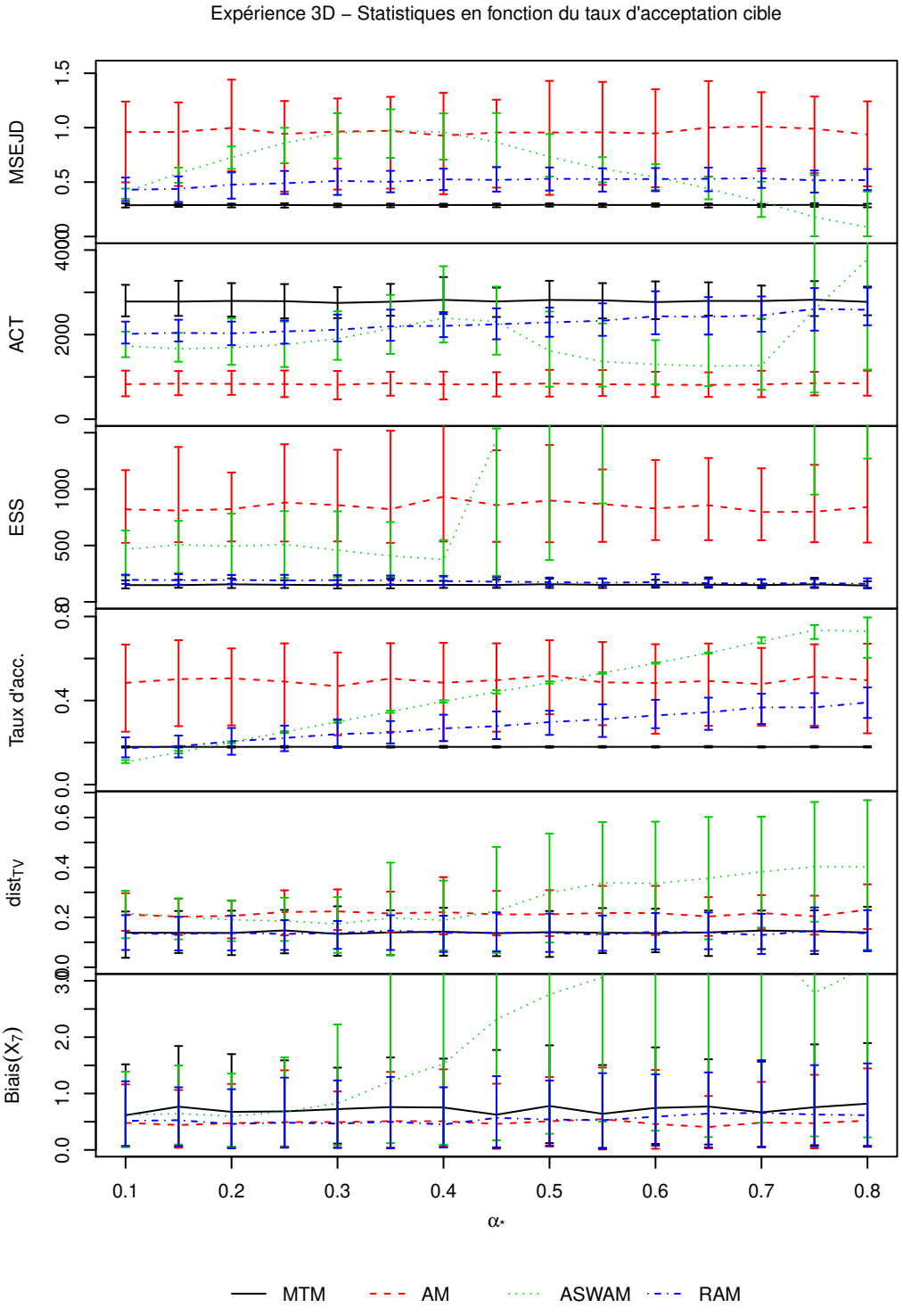


Figure 6.15 *(Expérience 3D) Statistiques en fonction du taux d'acceptation cible pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

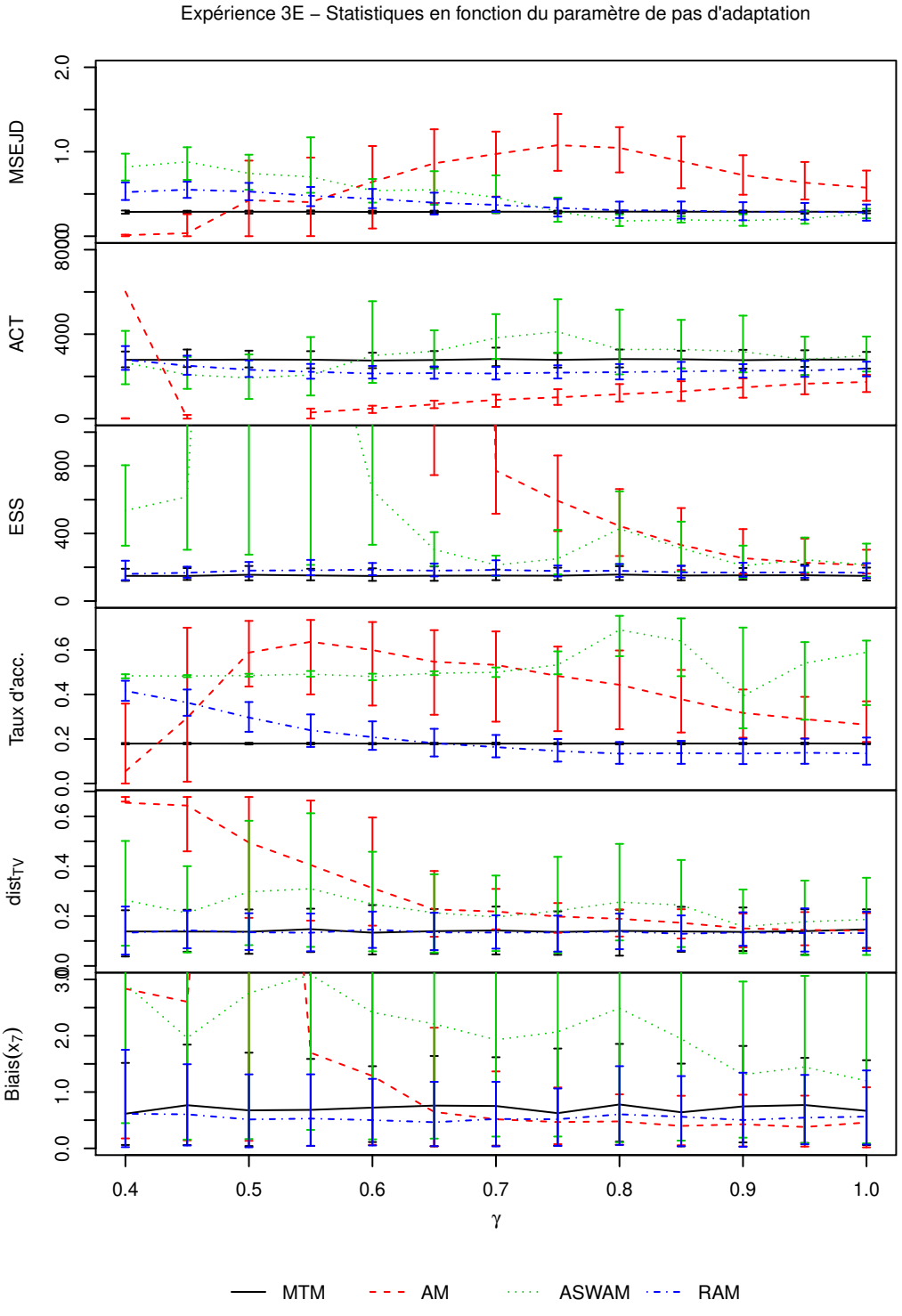


Figure 6.16 *(Expérience 3E) Statistiques en fonction du paramètre de pas d'adaptation pour les quatre algorithmes principaux. Les barres verticales représentent les quantiles 5% et 95% de la statistique sur les 100 réplications.*

6.4 Discussion

Les trois expériences présentées à la section 6.3 ont permis d'étudier le comportement et la performance de l'algorithme aMTM (et de ses multiples variantes) face à différentes situations et en comparaison à ses versions plus simples (MTM sans adaptation ; AM, ASWAM et RAM à un seul candidat). Dans cette section, on synthétise certaines des observations effectuées pour mieux cerner les charactéristiques de l'algorithme aMTM.

6.4.1 Sur l'échantillonnage MTM

D'abord, les résultats des expériences montrent clairement que les candidats par variable aléatoire commune sont significativement moins efficaces que les trois autres types de candidats. Une explication possible est que les densités instrumentales tendent alors à se ressembler, de sorte que les propositions seront pratiquement identiques. Ainsi, l'utilisation de candidats additionnels apporte peu d'information additionnelle sur le support de la distribution cible et l'efficacité de l'estimation ne s'en trouve pas augmentée.

Ensuite, les trois autres types de proposition semblent relativement équivalents empiriquement. Bien que seuls les candidats extrêmement antithétiques et quasi-Monte Carlo randomisés sont construits spécifiquement pour bien se répartir dans l'espace, les candidats indépendants ont un comportement similaire en moyenne, surtout lorsque K augmente. Il n'est donc pas surprenant que ces candidats arrivent à produire des algorithmes relativement aussi performants.

Enfin, comme il fut mentionné à maintes reprises, le choix de la fonction de poids (c.-à-d., les poids par importance ou proportionnels à la densité cible) ne semble pas avoir un impact important sur la performance des algorithmes. On a tout de même observé de faibles différences, qui pourraient être attribuées à la distribution cible. En effet, une densité multimodale requiert de grands sauts et ceux-ci sont favorisés par les poids par importance. À l'opposé, une distribution à support complexe profite un peu plus des poids proportionnels à la densité cible puisque ceux-ci favorisent les régions de haute densité, ce qui aide à trouver les directions pertinentes à l'échantillonnage.

6.4.2 Sur l'adaptation

Dans tout les cas, les algorithmes adaptatifs atteignent des niveaux de performance supérieurs à ceux des algorithmes non-adaptatifs. Notons cependant que l'algorithme AM a démontré une faible performance dans la troisième expérience, ce qui peut être dû au fait qu'une échelle optimale est difficile à trouver en grande dimension et que cet algorithme n'adapte pas directement l'échelle.

Pour ce qui est de l'utilisation d'une densité instrumentale adaptée à toutes les itérations, les expériences ont montré qu'il n'y a pas vraiment d'avantage ni de désavantage à le faire. On se doute, en fait, que plusieurs des densités instrumentales s'adaptent déjà vers la covariance globale de la distribution cible de sorte qu'aucune différence n'est observable.

L'adaptation des échelles par le taux de sélection semble être favorable aux mises à jour RAM. Une hypothèse expliquant de ce phénomène est que l'algorithme RAM s'adapte plus lentement que les deux autres algorithmes adaptatifs ; l'adaptation de l'échelle accélère alors l'adaptation. Cette même adaptation semble cependant nuire aux mises à jour AM et ASWAM. Dans le cas ASWAM, on se doute que l'interaction entre l'adaptation par le taux d'acceptation et par le taux de sélection peut causer problème.

L'utilisation de mises à jour locales ne semble pas aider les algorithmes AM et ASWAM et leur nuit même à l'occasion. Ce type de mise à jour fut proposée pour augmenter l'ajustement local à la densité cible. Par contre, ceci implique des covariances plus petites, donc des sauts plus petits et une autocorrélation plus grande : ceci engendre alors des mesures de performance moindres.

6.4.3 Sur le nombre de candidats

Toutes les expériences ont produit des résultats similaires quant à l'évolution de la performance par rapport au nombre de candidats utilisés. En effet, la performance tend toujours à augmenter avec K, mais atteint rapidement un plateau. D'une manière sommaire, un faible nombre de candidats (entre 5 et 10) a été suffisant pour que tous les algorithmes atteignent leur optimalité respective.

De plus, ces expériences ont bien révélé l'utilité des essais multiples : il n'a jamais été possible de bien échantillonner à partir de la distribution cible en n'utilisant qu'un seul candidat. Dans tous les cas, la distance en variation totale entre la distribution cible et la distribution empirique était très élevée pour $K = 1$ et particulièrement moindre dès $K \approx 5$.

Il devient alors très simple de comparer les algorithmes au niveau de leur efficacité empirique ajustée pour le temps de calcul, étant donné qu'il est nécessaire d'utiliser des essais multiples pour obtenir des chaînes qui représentent bien la distribution cible. Heureusement, la performance se stabilise assez rapidement avec l'augmentation de K de sorte que la quantité de calcul supplémentaire pour obtenir des chaînes satisfaisantes n'est pas énorme.

6.4.4 Sur le taux d'acceptation cible

Les résultats d'échelle optimale ne peuvent pas être directement appliqués aux expériences étudiées dans ce chapitre étant donné que les distributions cibles sont loin de satisfaire les conditions requises. Cependant, les résultats obtenus semblent confirmer une certaine robustesse des résultats théoriques.

Par exemple, le tableau 4.1 nous indique que, pour $K = 5$ candidats et asymptotiquement pour $d \to \infty$, le taux d'acceptation optimal est de 0,55 pour des candidats extrêmement antithétiques et de 0,41 pour des candidats indépendants. De plus, les résultats en dimensions finies pour $K = 1$ candidat (cf. tableau 2.1) montrent que le taux optimal augmente lorsque d diminue vers 1. Maintenant, pour des distributions cibles moins régulières, comme celles considérées dans les expériences, on peut s'attendre à des taux optimaux un peu plus bas afin que l'échantillonneur favorise des points plus éloignés. Ceci aura pour effet d'améliorer l'exploration de l'espace.

Ainsi, trouver des taux cibles optimaux dans les environs de l'intervalle $[0{,}25; 0{,}50]$ semble bien correspondre avec la théorie. De plus, on remarque que l'algorithme ASWAM requiert un taux cible

plus faible dans la troisième expérience, qui était en $d = 20$ dimensions. Ceci confirme également l'intuition que la probabilité d'acceptation doit être moindre en plus grandes dimensions pour favoriser l'exploration de l'espace qui y est plus difficile. On note finalement que l'algorithme RAM produit quant à lui des chaînes dont le taux d'acceptation observé est généralement plus bas que le taux cible ; ce comportement semble cependant être dû à une adaptation plus lente : des pas d'adaptation décroissants plus lentement semblent rapprocher le taux observé du taux cible.

6.4.5 Sur le pas d'adaptation

Le paramètre γ déterminant le vitesse de décroissance des pas d'adaptation influence grandement la vitesse à laquelle les algorithmes adaptatifs adaptent leur densités instrumentales. Une adaptation plus rapide peut être souhaitable afin d'atteindre à un ensemble de densités optimales plus rapidement, mais une adaptation plus lente peut être souhaitable afin d'augmenter la robustesse de l'algorithme.

Il semble que l'algorithme RAM requiert une valeur de γ moindre pour contrebalancer l'adaptation plus lente produite par les mises à jour. Dans les trois expériences, une décroissance donnée par $\gamma = 0{,}50$ produisait des chaînes légèrement plus performantes.

Certaines régions particulièrement difficiles à explorer peuvent prendre du temps (en nombre d'itérations MCMC) avant d'être découvertes par l'algorithme. Si l'adaptation est trop rapide, l'algorithme n'arrivera peut-être jamais à récupérer de la convergence vers un sous-ensemble du support seulement. On a observé ce phénomène à la seconde et à la troisième expérience alors que des valeurs de γ plus près de 1 (décroissance plus rapide, pas plus petits, adaptation plus lente) produisaient des chaînes significativement plus représentatives des distributions cibles respectives.

6.5 Annexes au chapitre 6

6.5.1 Suppléments à la section 6.3.2

On produit une densité cible à géométrie complexe en s'inspirant d'un exemple très répandu dans la littérature : la « banane ». Soit $y_1,y_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ et soit la transformation

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = g\left(y_1,y_2\right) = \begin{pmatrix} ay_1 \\ y_2 + B\left(a^2y_1^2 + a^2\right) \end{pmatrix}, \tag{6.2}$$

où a et B sont des paramètres contrôlant respectivement l'échelle et la « bananicité », c.-à-d., la courbure du support.

Par les propriétés des transformations de variables aléatoires, on a que la densité conjointe de (x_1,x_2) est donnée, sous certaines conditions satisfaites par notre choix de transformation g, par

$$f_X\left(x_1,x_2\right) = \left|\frac{\partial}{\partial x} g^{-1}\left(x_1,x_2\right)\right| f_Y\left(g^{-1}\left(x_1,x_2\right)\right),$$

où $g^{-1}\left(x_1,x_2\right) = \left(y_1,y_2\right)$ est la fonction inverse de g. Puis,

$$f_Y\left(y_1,y_2\right) = \varphi\left(y_1\right)\varphi\left(y_2\right)$$

est la densité conjointe de (y_1,y_2). En isolant (y_1,y_2) dans (6.2), on trouve $y_1 = \frac{1}{a}x_1$ et, puis,

$$\begin{aligned} y_2 &= x_2 - Bx_1^2 - Ba^2 \\ &= -Bx_1^2 + x_2 - Ba^2. \end{aligned}$$

Ainsi, le Jacobien de la transformation est donné par

$$\frac{\partial}{\partial x} g^{-1}\left(x_1,x_2\right) = \begin{pmatrix} \frac{\partial}{\partial x_1}y_1 & \frac{\partial}{\partial x_2}y_1 \\ \frac{\partial}{\partial x_1}y_2 & \frac{\partial}{\partial x_2}y_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{a} & 0 \\ -2Bx_1 & 1 \end{pmatrix}$$

de sorte que le déterminant est tout simplement $1/a$. C'est donc dire que la densité conjointe de (x_1,x_2) est donnée par

$$\begin{aligned} f_X\left(x_1,x_2\right) &= \frac{1}{a}\varphi\left(\frac{1}{a}x_1\right)\varphi\left(-Bx_1^2 + x_2 - Ba^2\right) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{a^2}x_1^2 + \left(-Bx_1^2 + x_2 - Ba^2\right)^2\right]\right\}. \end{aligned}$$

Étant donné qu'il s'agit d'une transformation de varaibles aléatoires gaussiennes, il est possible de produire un échantillon i.i.d. de cette densité. La figure 6.17 contient des échantillons i.i.d. pour divers choix de valeurs des paramètres a et B.

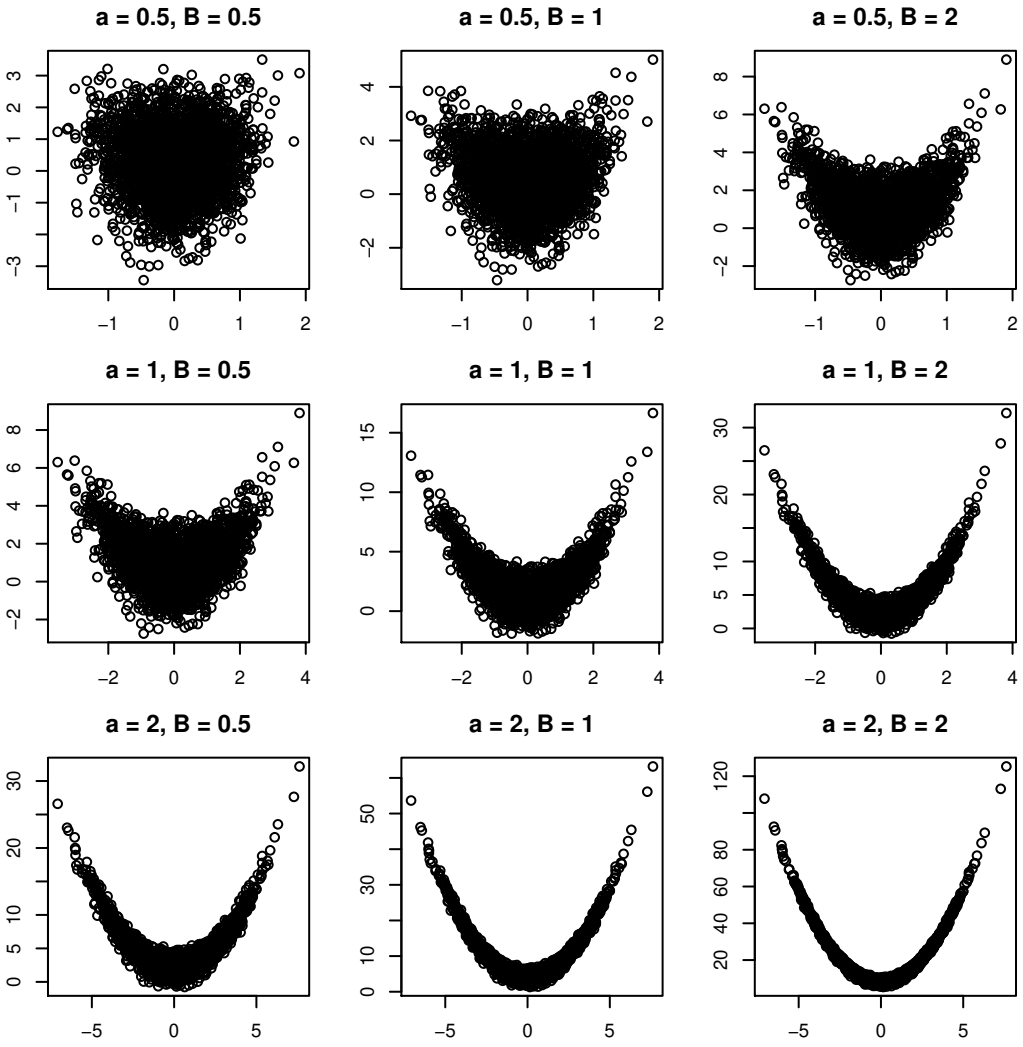


Figure 6.17 *Échantillon i.i.d. d'une densité de type « banane » en deux dimensions pour différentes valeurs des paramètres d'échelle a et de « bananicité » B.*

Chapitre 7

Conclusion

L'algorithme Metropolis à essais multiples adaptatif (aMTM) proposé dans ce mémoire incorpore une composante d'adaptation au sein de l'algorithme Metropolis à essais multiples (MTM). L'adaptation implémentée permet une sélection automatique des paramètres des densités instrumentales utilisées lors de l'échantillonnage MTM. Cette automatisation a deux conséquences importantes. D'abord, elle permet une réduction de l'intervention de l'utilisateur dans la mise au point de l'algorithme. Ensuite, l'adaptation permet de choisir un ensemble de paramètres qui soit supérieur ou même optimal en termes d'efficacité de l'algorithme. Conjointement, ces deux caractéristiques impliquent que l'utilisateur de l'aMTM bénéficera d'un algorithme à haute performance sans avoir à chercher activement les paramètres produisant cette haute performance.

Du point de vue théorique, il a été possible de dériver des ensembles de conditions suffisantes à l'ergodicité de l'algorithme aMTM. Les conditions énoncées sont très semblables à celles généralement utilisées pour montrer l'ergodicité de l'algorithme AM : les suppositions faites sur la distribution cible et sur l'espace des paramètres sont en fait les mêmes ; les conditions sur l'adaptation sont similaires étant donné que le même genre de mise à jour est utilisé. L'ergodicité, qui est la garantie minimale requise de tout algorithme MCMC, assure que la distribution marginale produite par l'échantillonnage converge vers la distribution cible. Les conditions permettant la vérification de l'ergodicité sont également suffisantes à la vérification d'une loi faible des grands nombres pour toute fonction bornée, ce qui garantit que l'estimé Monte Carlo de l'espérance de telles fonctions converge en probabilité vers l'espérance recherchée. Des garanties plus fortes, telles qu'une loi forte des grands nombres ou même un théorème limite central, en particulier pour des fonctions non-bornées, ne peuvent être vérifiées dans le contexte de preuve actuel. Alors que la section 5.4.1.5 contient une discussion quant à la généralisation des résultats obtenus, ces considérations théoriques sont laissées à de futures recherches.

Du côté empirique, l'algorithme aMTM a bien démontré sa supériorité par rapport aux algorithmes qu'il cherche à améliorer. Que ce soit pour une distribution cible multimodale ou à géométrie complexe, l'algorithme aMTM surclasse, d'une part, l'algorithme AM (c.-à-d., à un seul essai) par son exploration de l'espace d'états plus efficace due à l'utilisation d'essais multiples et, d'autre part, l'algorithme MTM (c.-à-d., sans adaptation) par son efficacité d'estimation supérieure induite par un meilleur choix de paramètres.

Les algorithmes MTM et aMTM sont cependant très coûteux computationnellement étant donné que de nombreuses évaluations de la densité cible sont requises. Ainsi, le gain en performance doit être suffisamment élevé pour contrebalancer l'augmentation du temps de calcul. Les expériences considérées

n'ont pas permis une étude approfondie de ce compromis, par exemple en étudiant la performance par rapport au nombre d'essais, puisque les densités choisies étaient si complexes que des algorithmes utilisant peu d'essais (e.g. $K < 5$ essais) n'arrivaient tout simplement pas à bien échantillonner le support.

De plus, les expériences effectuées ne comparent pas l'algorithme aMTM à d'autres algorithmes MCMC qui cherchent à échantillonner des densités complexes et il serait donc pertinent d'étudier la performance de l'algorithme aMTM dans un contexte comparatif plus général. On pense, par exemple, aux algorithme TINCA (Craiu et collab., 2009), RAPTOR (Bai et collab., 2011a) et AIMTM (Casarin et collab., 2013) et à l'échantillonneur Équi-énergie (Kou et collab., 2006) qui peuvent tous être vus comme des compétiteurs à l'algorithme aMTM.

Bibliographie

Andrieu, C. et Y. F. Atchadé. 2007, «On the efficiency of adaptive MCMC algorithms», *Electron. Comm. Probab.*, vol. 12, doi :10.1214/ECP.v12-1320, p. 336–349, ISSN 1083-589X. URL https://doi.org/10.1214/ECP.v12-1320.

Andrieu, C., A. Jasra, A. Doucet et P. Del Moral. 2011, «On nonlinear Markov chain Monte Carlo», *Bernoulli*, vol. 17, n° 3, doi :10.3150/10-BEJ307, p. 987–1014, ISSN 1350-7265. URL https://doi.org/10.3150/10-BEJ307.

Andrieu, C. et É. Moulines. 2006, «On the ergodicity properties of some adaptive MCMC algorithms», *Ann. Appl. Probab.*, vol. 16, n° 3, doi :10.1214/105051606000000286, p. 1462–1505, ISSN 1050-5164. URL https://doi.org/10.1214/105051606000000286.

Andrieu, C., É. Moulines et P. Priouret. 2005, «Stability of stochastic approximation under verifiable conditions», *SIAM Journal on Control and Optimization*, vol. 44, n° 1, p. 283–312.

Andrieu, C. et C. P. Robert. 2001, *Controlled MCMC for Optimal Sampling*, INSEE. URL https://pdfs.semanticscholar.org/8186/0afd9ffaf9eca3e9c52f5b3f629ba0d4771e.pdf.

Andrieu, C. et J. Thoms. 2008, «A tutorial on adaptive MCMC», *Stat. Comput.*, vol. 18, n° 4, doi :10.1007/s11222-008-9110-y, p. 343–373, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-008-9110-y.

Andrieu, C. et M. Vihola. 2014, «Markovian stochastic approximation with expanding projections», *Bernoulli*, vol. 20, n° 2, doi :10.3150/12-BEJ497, p. 545–585, ISSN 1350-7265. URL https://doi.org/10.3150/12-BEJ497.

Atchadé, Y. et G. Fort. 2010, «Limit theorems for some adaptive MCMC algorithms with subgeometric kernels», *Bernoulli*, vol. 16, n° 1, doi :10.3150/09-BEJ199, p. 116–154, ISSN 1350-7265. URL https://doi.org/10.3150/09-BEJ199.

Atchadé, Y., G. Fort, É. Moulines et P. Priouret. 2011, «Adaptive Markov chain Monte Carlo : Theory and methods», dans *Bayesian time series models*, Cambridge Univ. Press, Cambridge, p. 32–51.

Atchadé, Y. F. 2006, «An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift», *Methodol. Comput. Appl. Probab.*, vol. 8, n° 2, doi :10.1007/s11009-006-8550-0, p. 235–254, ISSN 1387-5841. URL https://doi.org/10.1007/s11009-006-8550-0.

Atchadé, Y. F. 2009, «Resampling from the past to improve on MCMC algorithms», *Far East J. Theor. Stat.*, vol. 27, n° 1, p. 81–99, ISSN 0972-0863.

Atchadé, Y. F. 2010, «A cautionary tale on the efficiency of some adaptive Monte Carlo schemes», *The Annals of Applied Probability*, vol. 20, n° 3, p. 841–868.

Atchadé, Y. F. 2011, «Kernel estimators of asymptotic variance for adaptive Markov chain Monte Carlo», *Ann. Statist.*, vol. 39, n° 2, doi :10.1214/10-AOS828, p. 990–1011, ISSN 0090-5364. URL https://doi.org/10.1214/10-AOS828.

Atchadé, Y. F., G. O. Roberts et J. S. Rosenthal. 2011, «Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo», *Statistics and Computing*, vol. 21, n° 4, p. 555–568.

Atchadé, Y. F. et J. S. Rosenthal. 2005, «On adaptive Markov chain Monte Carlo algorithms», *Bernoulli*, vol. 11, n° 5, p. 815–828.

Bai, Y. 2009a, «An adaptive directional Metropolis-within-Gibbs algorithm», *Preprint*.

Bai, Y. 2009b, «Simultaneous drift conditions for adaptive Markov chain Monte Carlo algorithms», *Preprint*. URL http://www.probability.ca/jeff/ftpdir/yanbai2.pdf.

Bai, Y., R. V. Craiu et A. F. Di Narzo. 2011a, «Divide and conquer : A mixture-based approach to regional adaptation for MCMC», *J. Comput. Graph. Statist.*, vol. 20, n° 1, doi :10.1198/jcgs.2010.09035, p. 63–79, ISSN 1061-8600. URL https://doi.org/10.1198/jcgs.2010.09035, supplementary material available online.

Bai, Y., G. O. Roberts et J. S. Rosenthal. 2011b, «On the containment condition for adaptive Markov chain Monte Carlo algorithms», *Adv. Appl. Stat.*, vol. 21, n° 1, p. 1–54, ISSN 0972-3617.

Ballnus, B., S. Hug, K. Hatz, L. Görlitz, J. Hasenauer et F. J. Theis. 2017, «Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems», *BMC systems biology*, vol. 11, n° 1, p. 63.

Baragatti, M., A. Grimaud et D. Pommeret. 2013, «Parallel tempering with equi-energy moves», *Statistics and Computing*, vol. 23, n° 3, p. 323–339.

Baxendale, P. H. 2005, «Renewal theory and computable convergence rates for geometrically ergodic Markov chains», *The Annals of Applied Probability*, vol. 15, n° 1B, p. 700–738.

Bédard, M. 2007, «Weak convergence of Metropolis algorithms for non-*IID* target distributions», *The Annals of Applied Probability*, vol. 17, n° 4, p. 1222–1244.

Bédard, M. 2008, «Efficient sampling using Metropolis algorithms : Applications of optimal scaling results», *Journal of Computational and Graphical Statistics*, vol. 17, n° 2, p. 312–332.

Bédard, M. 2008, «Optimal acceptance rates for Metropolis algorithms : Moving beyond 0.234», *Stochastic Process. Appl.*, vol. 118, n° 12, doi :10.1016/j.spa.2007.12.005, p. 2198–2222, ISSN 0304-4149. URL https://doi.org/10.1016/j.spa.2007.12.005.

Bédard, M., R. Douc et É. Moulines. 2012, «Scaling analysis of multiple-try MCMC methods», *Stochastic Processes and their Applications*, vol. 122, n° 3, p. 758–786.

Bédard, M., R. Douc et É. Moulines. 2014, «Scaling analysis of delayed rejection MCMC methods», *Methodol. Comput. Appl. Probab.*, vol. 16, n° 4, doi :10.1007/s11009-013-9326-y, p. 811–838, ISSN 1387-5841. URL https://doi.org/10.1007/s11009-013-9326-y.

Bédard, M. et M. Mireuta. 2013, «On the empirical efficiency of local MCMC algorithms with pools of proposals», *Canadian Journal of Statistics*, vol. 41, n° 4, p. 657–678.

Bédard, M. et J. S. Rosenthal. 2008, «Optimal scaling of Metropolis algorithms : Heading toward general target ddistributions», *Canadian Journal of Statistics*, vol. 36, n° 4, p. 483–503.

Benveniste, A., M. Métivier et P. Priouret. 1987, *Algorithmes Adaptatifs Et Approximations Stochastiques : Théorie Et Applications à L'identification, Au Traitement Du Signal Et à La Reconnaissance Des Formes*, Masson.

Beskos, A., N. Pillai, G. Roberts, J.-M. Sanz-Serna et A. Stuart. 2013, «Optimal tuning of the hybrid Monte Carlo algorithm», *Bernoulli*, vol. 19, n° 5A, doi :10.3150/12-BEJ414, p. 1501–1534, ISSN 1350-7265. URL https://doi.org/10.3150/12-BEJ414.

Beskos, A., G. Roberts et A. Stuart. 2009, «Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions», *Ann. Appl. Probab.*, vol. 19, n° 3, doi :10.1214/08-AAP563, p. 863–898, ISSN 1050-5164. URL https://doi.org/10.1214/08-AAP563.

Bhatia, R. 1997, *Matrix Analysis*, *Graduate Texts in Mathematics*, vol. 169, Springer-Verlag, New York, ISBN 0-387-94846-5, xii+347 p., doi :10.1007/978-1-4612-0653-8. URL https://doi.org/10.1007/978-1-4612-0653-8.

Billingsley, P. 2012, *Probability and Measure*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, ISBN 978-1-118-12237-2, xviii+624 p.. Anniversary edition [of MR1324786], With a foreword by Steve Lalley and a brief biography of Billingsley by Steve Koppes.

Breyer, L. A., M. Piccioni et S. Scarlatti. 2004, «Optimal scaling of MaLa for nonlinear regression», *Ann. Appl. Probab.*, vol. 14, n° 3, doi :10.1214/105051604000000369, p. 1479–1505, ISSN 1050-5164. URL https://doi.org/10.1214/105051604000000369.

Breyer, L. A. et G. O. Roberts. 2000, «From Metropolis to diffusions : Gibbs states and optimal scaling», *Stochastic Process. Appl.*, vol. 90, n° 2, doi :10.1016/S0304-4149(00)00041-7, p. 181–206, ISSN 0304-4149. URL https://doi.org/10.1016/S0304-4149(00)00041-7.

Brockwell, A. E. et J. B. Kadane. 2005, «Identification of regeneration times in MCMC simulation, with application to adaptive schemes», *J. Comput. Graph. Statist.*, vol. 14, nº 2, doi :10.1198/106186005X47453, p. 436–458, ISSN 1061-8600. URL https://doi.org/10.1198/106186005X47453.

Brooks, S. P., P. Dellaportas et G. O. Roberts. 1997, «An approach to diagnosing total variation convergence of MCMC algorithms», *J. Comput. Graph. Statist.*, vol. 6, nº 3, doi :10.2307/1390732, p. 251–265, ISSN 1061-8600. URL https://doi.org/10.2307/1390732.

Brooks, S. P. et A. Gelman. 1998, «General methods for monitoring convergence of iterative simulations», *J. Comput. Graph. Statist.*, vol. 7, nº 4, doi :10.2307/1390675, p. 434–455, ISSN 1061-8600. URL https://doi.org/10.2307/1390675.

Brooks, S. P. et G. O. Roberts. 1998, «Convergence assessment techniques for Markov chain Monte Carlo», *Statistics and Computing*, vol. 8, nº 4, doi :10.1023/A:1008820505350, p. 319–335. URL https://doi.org/10.1023/A:1008820505350.

Cai, B., R. Meyer et F. Perron. 2008, «Metropolis-Hastings algorithms with adaptive proposals», *Stat. Comput.*, vol. 18, nº 4, doi :10.1007/s11222-008-9051-5, p. 421–433, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-008-9051-5.

Cappé, O. et É. Moulines. 2009, «On-line expectation-maximization algorithm for latent data models», *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 71, nº 3, p. 593–613.

Casarin, R., R. Craiu et F. Leisen. 2013, «Interacting multiple try algorithms with different proposal distributions», *Stat. Comput.*, vol. 23, nº 2, doi :10.1007/s11222-011-9301-9, p. 185–200, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-011-9301-9.

Casella, G. et C. P. Robert. 1998, «Post-processing accept-reject samples : Recycling and rescaling», *J. Comput. Graph. Statist.*, vol. 7, nº 2, doi :10.2307/1390810, p. 139–157, ISSN 1061-8600. URL https://doi.org/10.2307/1390810.

Chauveau, D. et P. Vandekerkhove. 2002, «Improving convergence of the Hastings-Metropolis algorithm with an adaptive proposal», *Scand. J. Statist.*, vol. 29, nº 1, doi :10.1111/1467-9469.00064, p. 13–29, ISSN 0303-6898. URL https://doi.org/10.1111/1467-9469.00064.

Chauveau, D. et P. Vandekerkhove. 2014, «The nearest neighbor entropy estimate : An adequate tool for adaptive MCMC evaluation», *Preprint*. URL https://hal.archives-ouvertes.fr/docs/01/06/80/81/PDF/Chauveau_NN_AMCMC.pdf.

Chen, H. F., G. Lei et A. J. Gao. 1988, «Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds», *Stochastic Process. Appl.*, vol. 27, nº 2, doi :10.1016/0304-4149(87)90039-1, p. 217–231, ISSN 0304-4149. URL https://doi.org/10.1016/0304-4149(87)90039-1.

Chen, Y., D. Keyes, K. J. H. Law et H. Ltaief. 2016, «Accelerated dimension-independent adaptive Metropolis», *SIAM J. Sci. Comput.*, vol. 38, nº 5, doi :10.1137/15M1026432, p. S539–S565, ISSN 1064-8275. URL https://doi.org/10.1137/15M1026432.

Chimisov, C., K. Latuszynski et G. Roberts. 2018, «Air Markov chain Monte Carlo», *arXiv preprint arXiv :1801.09309*. URL https://arxiv.org/pdf/1801.09309.

Chung, K. L. 2001, *A Course in Probability Theory*, 3e éd., Academic Press, Inc., San Diego, CA, ISBN 0-12-174151-6, xviii+419 p..

Cowles, M. K. et B. P. Carlin. 1996, «Markov chain Monte Carlo convergence diagnostics : A comparative review», *J. Amer. Statist. Assoc.*, vol. 91, nº 434, doi :10.2307/2291683, p. 883–904, ISSN 0162-1459. URL https://doi.org/10.2307/2291683.

Craiu, R. V., L. Gray, K. Ł atuszyński, N. Madras, G. O. Roberts et J. S. Rosenthal. 2015, «Stability of adversarial Markov chains, with an application to adaptive MCMC algorithms», *Ann. Appl. Probab.*, vol. 25, nº 6, doi :10.1214/14-AAP1083, p. 3592–3623, ISSN 1050-5164. URL https://doi.org/10.1214/14-AAP1083.

Craiu, R. V. et C. Lemieux. 2007, «Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling», *Stat. Comput.*, vol. 17, nº 2, doi :10.1007/s11222-006-9009-4, p. 109–120, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-006-9009-4.

Craiu, R. V., J. Rosenthal et C. Yang. 2009, «Learn from thy neighbor : Parallel-chain and regional adaptive MCMC», *J. Amer. Statist. Assoc.*, vol. 104, n° 488, doi :10.1198/jasa.2009.tm08393, p. 1454–1466, ISSN 0162-1459. URL https://doi.org/10.1198/jasa.2009.tm08393.

Dai, N. et G. L. Jones. 2017, «Multivariate initial sequence estimators in Markov chain Monte Carlo», *J. Multivariate Anal.*, vol. 159, doi :10.1016/j.jmva.2017.05.009, p. 184–199, ISSN 0047-259X. URL https://doi.org/10.1016/j.jmva.2017.05.009.

Fisher, R. A. 1936, «The use of multiple measurements in taxonomic problems», *Annals of Eugenics*, vol. 7, n° 2, doi :10.1111/j.1469-1809.1936.tb02137.x, p. 179–188. URL https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Flegal, J. M. et G. L. Jones. 2010, «Batch means and spectral variance estimators in Markov chain Monte Carlo», *Ann. Statist.*, vol. 38, n° 2, doi :10.1214/09-AOS735, p. 1034–1070, ISSN 0090-5364. URL https://doi.org/10.1214/09-AOS735.

Fort, G. et E. Moulines. 2000a, «Computable bounds for subgeometrical and geometrical ergodicity», dans *IN STOCHASTIC PROCESSES APPL*, doi :10.1.1.36.4309. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.4309.

Fort, G. et É. Moulines. 2000b, «V-subgeometric ergodicity for a Hastings-Metropolis algorithm», *Statistics & Probability Letters*, vol. 49, n° 4, p. 401–410.

Fort, G., É. Moulines et P. Priouret. 2011, «Convergence of adaptive and interacting Markov chain Monte Carlo algorithms», *Ann. Statist.*, vol. 39, n° 6, doi :10.1214/11-AOS938, p. 3262–3289, ISSN 0090-5364. URL https://doi.org/10.1214/11-AOS938.

Fort, G., É. Moulines, P. Priouret et P. Vandekerkhove. 2014, «A central limit theorem for adaptive and interacting Markov chains», *Bernoulli*, vol. 20, n° 2, doi :10.3150/12-BEJ493, p. 457–485, ISSN 1350-7265. URL https://doi.org/10.3150/12-BEJ493.

Frenkel, D. et B. Smit. 1996, *Understanding Molecular Simulation : From Algorithms to Applications*, Academic Press.

Garren, S. T. et R. L. Smith. 2000, «Estimating the second largest eigenvalue of a Markov transition matrix», *Bernoulli*, vol. 6, n° 2, doi :10.2307/3318575, p. 215–242, ISSN 1350-7265. URL https://doi.org/10.2307/3318575.

Garthwaite, P. H., Y. Fan et S. A. Sisson. 2016, «Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process», *Comm. Statist. Theory Methods*, vol. 45, n° 17, doi :10.1080/03610926.2014.936562, p. 5098–5111, ISSN 0361-0926. URL https://doi.org/10.1080/03610926.2014.936562.

Gåsemyr, J. 2003, «On an adaptive version of the Metropolis-Hastings algorithm with independent proposal distribution», *Scand. J. Statist.*, vol. 30, n° 1, doi :10.1111/1467-9469.00324, p. 159–173, ISSN 0303-6898. URL https://doi.org/10.1111/1467-9469.00324.

Gelman, A., G. O. Roberts et W. R. Gilks. 1996, «Efficient Metropolis jumping rules», , p. 599–607.

Gelman, A. et D. B. Rubin. 1992, «Inference from iterative simulation using multiple sequences», *Statistical Science*, vol. 7, n° 4, p. 457–472. URL https://projecteuclid.org/euclid.ss/1177011136.

Geweke, J. 1992, «Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments», dans *Bayesian statistics, 4 (Peñíscola, 1991)*, Oxford Univ. Press, New York, p. 169–193.

Geyer, C. J. 1991, «Markov chain Monte Carlo maximum likelihood», URL https://pdfs.semanticscholar.org/07d9/dd5c25c944bf009256cdcb622feda53dabba.pdf.

Geyer, C. J. 1992, «Practical Markov chain Monte Carlo», *Statistical Science*, vol. 7, n° 4, p. 473–483. URL https://projecteuclid.org/euclid.ss/1177011137.

Geyer, C. J. 2011, «Introduction to Markov chain Monte Carlo», dans *Handbook of Markov chain Monte Carlo*, édité par S. Brooks, A. Gelman, G. L. Jones et X.-L. Meng, chap. 1, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton, FL, ISBN 978-1-4200-7941-8, p. 3–48, doi :10.1201/b10905. URL https://doi.org/10.1201/b10905.

Geyer, C. J. et E. A. Thompson. 1995, «Annealing Markov chain Monte Carlo with applications to ancestral inference», *Journal of the American Statistical Association*, vol. 90, n° 431, p. 909–920. URL https://www.jstor.org/stable/2291325.

Gilks, W. R., N. Best et K. Tan. 1995, «Adaptive rejection Metropolis sampling within Gibbs sampling», *Applied Statistics*, vol. 44, n° 4, p. 455–472. URL https://www.jstor.org/stable/2986138#metadata_info_tab_contents.

Gilks, W. R., G. O. Roberts et E. I. George. 1994, «Adaptive direction sampling», *The Statistician*, vol. 43, n° 3, p. 179–189. URL https://www.jstor.org/stable/2348942#metadata_info_tab_contents.

Gilks, W. R., G. O. Roberts et S. K. Sahu. 1998, «Adaptive Markov chain Monte Carlo through regeneration», *J. Amer. Statist. Assoc.*, vol. 93, n° 443, doi :10.2307/2669848, p. 1045–1054, ISSN 0162-1459. URL https://doi.org/10.2307/2669848.

Giordani, P. et R. Kohn. 2010, «Adaptive independent Metropolis-Hastings by fast estimation of mixtures of normals», *J. Comput. Graph. Statist.*, vol. 19, n° 2, doi :10.1198/jcgs.2009.07174, p. 243–259, ISSN 1061-8600. URL https://doi.org/10.1198/jcgs.2009.07174.

Haario, H., M. Laine, A. Mira et E. Saksman. 2006, «DRAM : Efficient adaptive MCMC», *Stat. Comput.*, vol. 16, n° 4, doi :10.1007/s11222-006-9438-0, p. 339–354, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-006-9438-0.

Haario, H., E. Saksman et J. Tamminen. 1999, «Adaptive proposal distribution for random walk Metropolis algorithm», *Computational Statistics*, vol. 14, n° 3, doi :10.1007/s001800050022, p. 375–396. URL https://doi.org/10.1007/s001800050022.

Haario, H., E. Saksman et J. Tamminen. 2001, «An adaptive metropolis algorithm», *Bernoulli*, vol. 7, n° 2, doi :10.2307/3318737, p. 223–242, ISSN 1350-7265. URL https://doi.org/10.2307/3318737.

Haario, H., E. Saksman et J. Tamminen. 2005, «Componentwise adaptation for high dimensional MCMC», *Comput. Statist.*, vol. 20, n° 2, doi :10.1007/BF02789703, p. 265–273, ISSN 0943-4062. URL https://doi.org/10.1007/BF02789703.

Halmos, P. R. 2013, *Measure Theory*, vol. 18, Springer.

Hastings, W. K. 1970, «Monte Carlo sampling methods using Markov chains and their applications», *Biometrika*, vol. 57, n° 1, doi :10.1093/biomet/57.1.97, p. 97–109, ISSN 0006-3444. URL https://doi.org/10.1093/biomet/57.1.97.

Heidelberger, P. et P. D. Welch. 1983, «Simulation run length control in the presence of an initial transient», *Operations Research*, vol. 31, n° 6, p. 1109–1144.

Helske, J. 2016, *ramcmc : Robust Adaptive Metropolis Algorithm*. URL http://github.com/helske/ramcmc, r package version 0.1.0.

Holden, L., R. Hauge et M. Holden. 2009, «Adaptive independent Metropolis-Hastings», *Ann. Appl. Probab.*, vol. 19, n° 1, doi :10.1214/08-AAP545, p. 395–413, ISSN 1050-5164. URL https://doi.org/10.1214/08-AAP545.

Jacob, P. E., J. O'Leary et Y. F. Atchadé. 2017, «Unbiased Markov chain Monte Carlo with couplings», *arXiv preprint arXiv :1708.03625*.

Jarner, S. F. et G. O. Roberts. 2002, «Polynomial convergence rates of Markov chains», *Annals of Applied Probability*, p. 224–247.

Jarner, S. r. F. et E. Hansen. 2000, «Geometric ergodicity of Metropolis algorithms», *Stochastic Process. Appl.*, vol. 85, n° 2, doi :10.1016/S0304-4149(99)00082-4, p. 341–361, ISSN 0304-4149. URL https://doi.org/10.1016/S0304-4149(99)00082-4.

Jones, G. L., M. Haran, B. S. Caffo et R. Neath. 2006, «Fixed-width output analysis for Markov chain Monte Carlo», *J. Amer. Statist. Assoc.*, vol. 101, n° 476, doi :10.1198/016214506000000492, p. 1537–1547, ISSN 0162-1459. URL https://doi.org/10.1198/016214506000000492.

Keith, J. M., D. P. Kroese et G. Y. Sofronov. 2008, «Adaptive independence samplers», *Stat. Comput.*, vol. 18, n° 4, doi :10.1007/s11222-008-9070-2, p. 409–420, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-008-9070-2.

Kosorok, M. R. 2000, «Monte Carlo error estimation for multivariate Markov chains», *Statist. Probab. Lett.*, vol. 46, n° 1, doi :10.1016/S0167-7152(99)00090-5, p. 85–93, ISSN 0167-7152. URL https://doi.org/10.1016/S0167-7152(99)00090-5.

Kou, S. C., Q. Zhou et W. H. Wong. 2006, «Equi-energy sampler with applications in statistical inference and statistical mechanics», *Ann. Statist.*, vol. 34, n° 4, doi :10.1214/009053606000000515, p. 1581–1652, ISSN 0090-5364. URL https://doi.org/10.1214/009053606000000515, with discussions and a rejoinder by the authors.

Ł atuszyński, K., G. O. Roberts et J. S. Rosenthal. 2013, «Adaptive Gibbs samplers and related MCMC methods», *Ann. Appl. Probab.*, vol. 23, n° 1, doi :10.1214/11-AAP806, p. 66–98, ISSN 1050-5164. URL https://doi.org/10.1214/11-AAP806.

Łatuszyński, K. et W. Niemiro. 2011, «Rigorous confidence bounds for MCMC under a geometric drift condition», *J. Complexity*, vol. 27, n° 1, doi :10.1016/j.jco.2010.07.003, p. 23–38, ISSN 0885-064X. URL https://doi.org/10.1016/j.jco.2010.07.003.

Levine, R. A. et G. Casella. 2006, «Optimizing random scan Gibbs samplers», *J. Multivariate Anal.*, vol. 97, n° 10, doi :10.1016/j.jmva.2006.05.008, p. 2071–2100, ISSN 0047-259X. URL https://doi.org/10.1016/j.jmva.2006.05.008.

Liu, C., J. Liu et D. Rubin. 1993, «A control variable for assessment the convergence of the Gibbs sampler», dans *Proceedings of the Statistical Computing Section of the American Statistical Association*, p. 74–78.

Liu, J. S., F. Liang et W. H. Wong. 2000, «The multiple-try method and local optimization in Metropolis sampling», *J. Amer. Statist. Assoc.*, vol. 95, n° 449, doi :10.2307/2669532, p. 121–134, ISSN 0162-1459. URL https://doi.org/10.2307/2669532.

Luengo, D. et L. Martino. 2013, «Fully adaptive Gaussian mixture Metropolis-Hastings algorithm», dans *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, IEEE, p. 6148–6152. URL https://arxiv.org/abs/1212.0122.

Marinari, E. et G. Parisi. 1992, «Simulated tempering : A new Monte Carlo scheme», *EPL (Europhysics Letters)*, vol. 19, n° 6, doi :10.1209/0295-5075/19/6/002, p. 451. URL https://arxiv.org/abs/hep-lat/920501.

Marshall, T. et G. Roberts. 2012, «An adaptive approach to Langevin MCMC», *Stat. Comput.*, vol. 22, n° 5, doi :10.1007/s11222-011-9276-6, p. 1041–1057, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-011-9276-6.

Martino, L., R. Casarin, F. Leisen et D. Luengo. 2018, «Adaptive independent sticky MCMC algorithms», *EURASIP Journal on Advances in Signal Processing*, vol. 2018, n° 1, p. 5. URL https://arxiv.org/abs/1308.3779.

Martino, L., V. P. Del Olmo et J. Read. 2012, «A multi-point Metropolis scheme with generic weight functions», *Statist. Probab. Lett.*, vol. 82, n° 7, doi :10.1016/j.spl.2012.04.008, p. 1445–1453, ISSN 0167-7152. URL https://doi.org/10.1016/j.spl.2012.04.008.

Martino, L. et J. Read. 2013, «On the flexibility of the design of multiple try Metropolis schemes», *Comput. Statist.*, vol. 28, n° 6, doi :10.1007/s00180-013-0429-2, p. 2797–2823, ISSN 0943-4062. URL https://doi.org/10.1007/s00180-013-0429-2.

Martino, L., J. Read et D. Luengo. 2015, «Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling», ISSN 1053-587X, p. 3123–3138, doi :10.1109/TSP.2015.2420537. URL https://doi.org/10.1109/TSP.2015.2420537.

Mattingly, J. C., N. S. Pillai et A. M. Stuart. 2012, «Diffusion limits of the random walk Metropolis algorithm in high dimensions», *Ann. Appl. Probab.*, vol. 22, n° 3, doi :10.1214/10-AAP754, p. 881–930, ISSN 1050-5164. URL https://doi.org/10.1214/10-AAP754.

Meketon, M. S. et B. Schmeiser. 1984, «Overlapping batch means : Something for nothing ?», dans *Proceedings of the 16th conference on Winter simulation*, IEEE Press, p. 226–230.

Mengersen, K. L., C. P. Robert et C. Guihenneuc-Jouyaux. 1999, «MCMC convergence diagnostics : A review», *Bayesian Statistics*, vol. 6, p. 415–440.

Mengersen, K. L. et R. L. Tweedie. 1996, «Rates of convergence of the Hastings and Metropolis algorithms», *Ann. Statist.*, vol. 24, n° 1, doi :10.1214/aos/1033066201, p. 101–121, ISSN 0090-5364. URL https://doi.org/10.1214/aos/1033066201.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller et E. Teller. 1953, «Equation of state calculations by fast computing machines», *The Journal of Chemical Physics*, vol. 21, n° 6, doi :10.1063/1.1699114, p. 1087–1092.

Metropolis, N. et S. Ulam. 1949, «The Monte Carlo method», *Journal of the American Statistical Association*, vol. 44, n° 247, p. 335–341.

Meyn, S. et R. L. Tweedie. 2009, *Markov Chains and Stochastic Stability*, 2e éd., Cambridge University Press, Cambridge, ISBN 978-0-521-73182-9, xxviii+594 p., doi :10.1017/CBO9780511626630. URL https://doi.org/10.1017/CBO9780511626630, with a prologue by Peter W. Glynn.

Meyn, S. P. et R. L. Tweedie. 1994, «Computable bounds for geometric convergence rates of Markov chains», *Ann. Appl. Probab.*, vol. 4, n° 4, p. 981–1011, ISSN 1050-5164. URL http://links.jstor.org/sici?sici=1050-5164(199411)4:4<981:CBFGCR>2.0.CO;2-U&origin=MSN.

Miasojedow, B. a., É. Moulines et M. Vihola. 2013, «An adaptive parallel tempering algorithm», *J. Comput. Graph. Statist.*, vol. 22, n° 3, doi :10.1080/10618600.2013.778779, p. 649–664, ISSN 1061-8600. URL https://doi.org/10.1080/10618600.2013.778779.

Mira, A. 2001a, «On Metropolis-Hastings algorithms with delayed rejection», *Metron*, vol. 59, n° 3-4, p. 231–2412 002, ISSN 0026-1424.

Mira, A. 2001b, «Ordering and improving the performance of Monte Carlo Markov chains», *Statist. Sci.*, vol. 16, n° 4, doi :10.1214/ss/1015346319, p. 340–350, ISSN 0883-4237. URL https://doi.org/10.1214/ss/1015346319.

Nummelin, E. 1984, *General Irreducible Markov Chains and Nonnegative Operators*, *Cambridge Tracts in Mathematics*, vol. 83, Cambridge University Press, Cambridge, ISBN 0-521-25005-6, xi+156 p., doi :10.1017/CBO9780511526237. URL https://doi.org/10.1017/CBO9780511526237.

Pandolfi, S., F. Bartolucci et N. Friel. 2010, «A generalization of the multiple-try Metropolis algorithm for Bayesian estimation and model selection», dans *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, p. 581–588. URL http://proceedings.mlr.press/v9/pandolfi10a.html.

Pasarica, C. et A. Gelman. 2010, «Adaptively scaling the Metropolis algorithm using expected squared jumped distance», *Statist. Sinica*, vol. 20, n° 1, p. 343–364, ISSN 1017-0405.

Peskun, P. H. 1973, «Optimum Monte-Carlo sampling using Markov chains», *Biometrika*, vol. 60, doi :10.1093/biomet/60.3.607, p. 607–612, ISSN 0006-3444. URL https://doi.org/10.1093/biomet/60.3.607.

Petersen, K. B. et M. S. Pedersen. 2008, «The matrix cookbook», *Technical University of Denmark*, vol. 7, n° 15, p. 510. URL https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf.

Pillai, N. S., A. M. Stuart et A. H. Thiéry. 2012, «Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions», *Ann. Appl. Probab.*, vol. 22, n° 6, doi :10.1214/11-AAP828, p. 2320–2356, ISSN 1050-5164. URL https://doi.org/10.1214/11-AAP828.

Plummer, M., N. Best, K. Cowles et K. Vines. 2006, «CODA : Convergence diagnosis and output analysis for MCMC», *R News*, vol. 6, n° 1, p. 7–11. URL https://journal.r-project.org/archive/.

Pompe, E., C. Holmes et K. Łatuszyński. 2018, «A framework for adaptive MCMC targeting multimodal distributions», URL https://arxiv.org/abs/1812.02609.

Qin, Z. S. et J. S. Liu. 2001, «Multipoint Metropolis method with application to hybrid Monte Carlo», *J. Comput. Phys.*, vol. 172, n° 2, doi :10.1006/jcph.2001.6860, p. 827–840, ISSN 0021-9991. URL https://doi.org/10.1006/jcph.2001.6860.

R Core Team. 2013, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robert, C. P. et G. Casella. 2004, *Monte Carlo Statistical Methods*, 2e éd., Springer Texts in Statistics, Springer-Verlag, New York, ISBN 0-387-21239-6, xxx+645 p., doi :10.1007/978-1-4757-4145-2. URL https://doi.org/10.1007/978-1-4757-4145-2.

Roberts, G. 1996, «Methods for estimating L2 convergence of Markov chain Monte Carlo techniques», *Bayesian Analysis in Statistics and Econometrics : Essays in Honor of Arnold Zellner*, vol. 309, p. 373.

Roberts, G. O., A. Gelman et W. R. Gilks. 1997, «Weak convergence and optimal scaling of random walk Metropolis algorithms», *Ann. Appl. Probab.*, vol. 7, nº 1, doi :10.1214/aoap/1034625254, p. 110–120, ISSN 1050-5164. URL https://doi.org/10.1214/aoap/1034625254.

Roberts, G. O. et W. R. Gilks. 1994, «Convergence of adaptive direction sampling», *J. Multivariate Anal.*, vol. 49, nº 2, doi :10.1006/jmva.1994.1028, p. 287–298, ISSN 0047-259X. URL https://doi.org/10.1006/jmva.1994.1028.

Roberts, G. O. et J. S. Rosenthal. 1997, «Geometric ergodicity and hybrid Markov chains», *Electron. Comm. Probab.*, vol. 2, doi :10.1214/ECP.v2-981, p. no. 2, 13–25, ISSN 1083-589X. URL https://doi.org/10.1214/ECP.v2-981.

Roberts, G. O. et J. S. Rosenthal. 1998, «Optimal scaling of discrete approximations to Langevin diffusions», *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 60, nº 1, doi :10.1111/1467-9868.00123, p. 255–268, ISSN 1369-7412. URL https://doi.org/10.1111/1467-9868.00123.

Roberts, G. O. et J. S. Rosenthal. 2001, «Optimal scaling for various Metropolis-Hastings algorithms», *Statist. Sci.*, vol. 16, nº 4, doi :10.1214/ss/1015346320, p. 351–367, ISSN 0883-4237. URL https://doi.org/10.1214/ss/1015346320.

Roberts, G. O. et J. S. Rosenthal. 2004, «General state space Markov chains and MCMC algorithms», *Probab. Surv.*, vol. 1, doi :10.1214/154957804100000024, p. 20–71, ISSN 1549-5787. URL https://doi.org/10.1214/154957804100000024.

Roberts, G. O. et J. S. Rosenthal. 2007, «Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms», *J. Appl. Probab.*, vol. 44, nº 2, doi :10.1239/jap/1183667414, p. 458–475, ISSN 0021-9002. URL https://doi.org/10.1239/jap/1183667414.

Roberts, G. O. et J. S. Rosenthal. 2009, «Examples of adaptive MCMC», *J. Comput. Graph. Statist.*, vol. 18, nº 2, doi :10.1198/jcgs.2009.06134, p. 349–367, ISSN 1061-8600. URL https://doi.org/10.1198/jcgs.2009.06134.

Roberts, G. O. et J. S. Rosenthal. 2014, «Minimising MCMC variance via diffusion limits, with an application to simulated tempering», *Ann. Appl. Probab.*, vol. 24, nº 1, doi :10.1214/12-AAP918, p. 131–149, ISSN 1050-5164. URL https://doi.org/10.1214/12-AAP918.

Roberts, G. O., J. S. Rosenthal et P. O. Schwartz. 1998, «Convergence properties of perturbed Markov chains», *J. Appl. Probab.*, vol. 35, nº 1, doi :10.1017/s0021900200014625, p. 1–11, ISSN 0021-9002. URL https://doi.org/10.1017/s0021900200014625.

Roberts, G. O. et R. L. Tweedie. 1996a, «Exponential convergence of Langevin distributions and their discrete approximations», *Bernoulli*, vol. 2, nº 4, doi :10.2307/3318418, p. 341–363, ISSN 1350-7265. URL https://doi.org/10.2307/3318418.

Roberts, G. O. et R. L. Tweedie. 1996b, «Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms», *Biometrika*, vol. 83, nº 1, doi :10.1093/biomet/83.1.95, p. 95–110, ISSN 0006-3444. URL https://doi.org/10.1093/biomet/83.1.95.

Rosenthal, J. S. 2011, «Optimal proposal distributions and adaptive MCMC», vol. 4.

Rosenthal, J. S. 2017, «Simple confidence intervals for MCMC without CLTs», *Electron. J. Stat.*, vol. 11, nº 1, doi :10.1214/17-EJS1224, p. 211–214, ISSN 1935-7524. URL https://doi.org/10.1214/17-EJS1224.

Rosenthal, J. S. et J. Yang. 2018, «Ergodicity of combocontinuous adaptive MCMC algorithms», *Methodol. Comput. Appl. Probab.*, vol. 20, nº 2, doi :10.1007/s11009-017-9574-3, p. 535–551, ISSN 1387-5841. URL https://doi.org/10.1007/s11009-017-9574-3.

Sahu, S. K. et A. A. Zhigljavsky. 2003, «Self-regenerative Markov chain Monte Carlo with adaptation», *Bernoulli*, vol. 9, nº 3, doi :10.3150/bj/1065444811, p. 395–422, ISSN 1350-7265. URL https://doi.org/10.3150/bj/1065444811.

Saksman, E. et M. Vihola. 2010, «On the ergodicity of the adaptive Metropolis algorithm on unbounded domains», *Ann. Appl. Probab.*, vol. 20, nº 6, doi :10.1214/10-AAP682, p. 2178–2203, ISSN 1050-5164. URL https://doi.org/10.1214/10-AAP682.

Schmeiser, B. 1982, «Batch size effects in the analysis of simulation output», *Oper. Res.*, vol. 30, nº 3, doi : 10.1287/opre.30.3.556, p. 556–568, ISSN 0030-364X. URL https://doi.org/10.1287/opre.30.3.556.

Sejdinovic, D., H. Strathmann, M. L. Garcia, C. Andrieu et A. Gretton. 2014, «Kernel adaptive Metropolis-Hastings», dans *International Conference on Machine Learning*, p. 1665–1673.

Shaby, B. et M. T. Wells. 2010, «Exploring an adaptive Metropolis algorithm», cahier de recherche, University of Berkeley.

Sherlock, C. 2018, «Reversible markov chains : Variational representations and ordering», *arXiv preprint arXiv :1809.01903*.

Sherlock, C. et G. Roberts. 2009, «Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets», *Bernoulli*, vol. 15, nº 3, doi :10.3150/08-BEJ176, p. 774–798, ISSN 1350-7265. URL https://doi.org/10.3150/08-BEJ176.

Sinharay, S. 2003, «Assessing convergence of the Markov chain Monte Carlo algorithms : A review», *ETS Research Report Series*, vol. 2003, nº 1, doi :10.1002/j.2333-8504.2003.tb01899.x, p. i–52. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.2003.tb01899.x.

So, M. K. P. 2006, «Bayesian analysis of nonlinear and non-Gaussian state space models via multiple-try sampling methods», *Stat. Comput.*, vol. 16, nº 2, doi :10.1007/s11222-006-6891-8, p. 125–141, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-006-6891-8.

Stormark, K. 2006, *Multiple Proposal Strategies for Markov Chain Monte Carlo*, mémoire de maîtrise, Institutt for matematiske fag.

Tak, H., X.-L. Meng et D. A. van Dyk. 2018, «A repelling-attracting Metropolis algorithm for multimodality», *J. Comput. Graph. Statist.*, vol. 27, nº 3, doi :10.1080/10618600.2017.1415911, p. 479–490, ISSN 1061-8600. URL https://doi.org/10.1080/10618600.2017.1415911.

Ter Braak, C. J. F. 2006, «A Markov chain Monte Carlo version of the genetic algorithm differential evolution : Easy Bayesian computing for real parameter spaces», *Stat. Comput.*, vol. 16, nº 3, doi :10.1007/s11222-006-8769-1, p. 239–249, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-006-8769-1.

Thompson, M. B. 2010, «A comparison of methods for computing autocorrelation time», .

Tierney, L. 1994, «Markov chains for exploring posterior distributions», *Ann. Statist.*, vol. 22, nº 4, doi :10.1214/aos/1176325750, p. 1701–1762, ISSN 0090-5364. URL https://doi.org/10.1214/aos/1176325750, with discussion and a rejoinder by the author.

Tierney, L. 1998, «A note on Metropolis-Hastings kernels for general state spaces», *Ann. Appl. Probab.*, vol. 8, nº 1, doi :10.1214/aoap/1027961031, p. 1–9, ISSN 1050-5164. URL https://doi.org/10.1214/aoap/1027961031.

Tran, M.-N., M. K. Pitt et R. Kohn. 2016, «Adaptive Metropolis-Hastings sampling using reversible dependent mixture proposals», *Stat. Comput.*, vol. 26, nº 1-2, doi :10.1007/s11222-014-9509-6, p. 361–381, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-014-9509-6.

Vats, D., J. M. Flegal et G. L. Jones. 2018, «Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo», *Bernoulli*, vol. 24, nº 3, doi :10.3150/16-BEJ914, p. 1860–1909, ISSN 1350-7265. URL https://doi.org/10.3150/16-BEJ914.

Vats, D., J. M. Flegal et G. L. Jones. 2019, «Multivariate output analysis for Markov chain Monte Carlo», *Biometrika*, vol. 106, nº 2, doi :10.1093/biomet/asz002, p. 321–337, ISSN 0006-3444. URL https://doi.org/10.1093/biomet/asz002.

Vihola, M. 2011a, «Can the adaptive Metropolis algorithm collapse without the covariance lower bound ?», *Electron. J. Probab.*, vol. 16, doi :10.1214/EJP.v16-840, p. no. 2, 45–75, ISSN 1083-6489. URL https://doi.org/10.1214/EJP.v16-840.

Vihola, M. 2011b, «On the stability and ergodicity of adaptive scaling Metropolis algorithms», *Stochastic Process. Appl.*, vol. 121, nº 12, doi :10.1016/j.spa.2011.08.006, p. 2839–2860, ISSN 0304-4149. URL https://doi.org/10.1016/j.spa.2011.08.006.

Vihola, M. 2012, «Robust adaptive Metropolis algorithm with coerced acceptance rate», *Stat. Comput.*, vol. 22, nº 5, doi :10.1007/s11222-011-9269-5, p. 997–1008, ISSN 0960-3174. URL https://doi.org/10.1007/s11222-011-9269-5.

Warnes, G. R. 2000, *The Normal Kernel Coupler : An Adaptive Markov Chain Monte Carlo Method for Efficiently Sampling from Multi-modal Distributions*, thèse de doctorat, University of Washington.

Wickham, H., J. Hester et W. Chang. 2018, *devtools : Tools to Make Developing R Packages Easier*. URL https://CRAN.R-project.org/package=devtools, r package version 1.13.6.

Williams, D. 1991, *Probability with Martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, ISBN 0-521-40455-X ; 0-521-40605-6, xvi+251 p., doi :10.1017/CBO9780511813658. URL https://doi.org/10.1017/CBO9780511813658.

Woodard, D. B., S. C. Schmidler et M. Huber. 2009, «Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions», *Ann. Appl. Probab.*, vol. 19, n° 2, doi :10.1214/08-AAP555, p. 617–640, ISSN 1050-5164. URL https://doi.org/10.1214/08-AAP555.

Yang, C. 2008a, «On the weak law of large numbers for unbounded functionals for adaptive MCMC», *Preprint*. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.1033&rep=rep1&type=pdf.

Yang, C. 2008b, «Recurrent and ergodic properties of adaptive MCMC», doi :10.1.1.144.6221.

Yang, C. 2009, *Ergodicity of Adaptive MCMC and Its Applications*, ProQuest LLC, Ann Arbor, MI, ISBN 978-0494-59167-3, 148 p.. URL http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:NR59167, thesis (Ph.D.)–University of Toronto (Canada).

Yang, J., E. Levi, R. V. Craiu et J. S. Rosenthal. 2019, «Adaptive Component-Wise Multiple-Try Metropolis Sampling», *J. Comput. Graph. Statist.*, vol. 28, n° 2, doi :10.1080/10618600.2018.1513365, p. 276–289, ISSN 1061-8600. URL https://doi.org/10.1080/10618600.2018.1513365.

Yu, B. 1995, «Estimating L1 error of kernel estimator : Monitoring convergence of Markov samplers», cahier de recherche, Technical report, Dept. of Statistics, University of California, Berkeley.

Yu, B. et P. Mykland. 1998, «Looking at Markov samplers through CUSUM path plots : A simple diagnostic idea», *Statistics and Computing*, vol. 8, n° 3, p. 275–286.

Zhang, H., Y. Wu, L. Cheng et I. Kim. 2016, «Hit and run ARMS : Adaptive rejection Metropolis sampling with hit and run random direction», *J. Stat. Comput. Simul.*, vol. 86, n° 5, doi :10.1080/00949655.2015.1046074, p. 973–985, ISSN 0094-9655. URL https://doi.org/10.1080/00949655.2015.1046074.