Université de Montréal

Efficacité des distributions instrumentales en équilibre dans un algorithme de type Metropolis-Hastings

par

Gabriel Boisvert-Beaudry

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistiques

8 août 2019

Sommaire

Dans ce mémoire, nous nous intéressons à une nouvelle classe de distributions instrumentales informatives dans le cadre de l'algorithme Metropolis-Hastings. Ces distributions instrumentales, dites en équilibre, sont obtenues en ajoutant de l'information à propos de la distribution cible à une distribution instrumentale non informative. Une chaîne de Markov générée par une distribution instrumentale en équilibre est réversible par rapport à la densité cible sans devoir utiliser une probabilité d'acceptation dans deux cas extrêmes : le cas local lorsque la variance instrumentale tend vers 0 et le cas global lorsqu'elle tend vers l'infini.

Il est nécessaire d'approximer les distributions instrumentales en équilibre afin de pouvoir les utiliser en pratique. Nous montrons que le cas local mène au *Metropolis-adjusted Langevin algorithm* (MALA), tandis que le cas global mène à une légère modification du MALA. Ces résultats permettent de concevoir un nouvel algorithme généralisant le MALA grâce à l'ajout d'un nouveau paramètre. En fonction de celui-ci, l'algorithme peut utiliser l'équilibre local ou global ou encore une interpolation entre ces deux cas. Nous étudions ensuite la paramétrisation optimale de cet algorithme en fonction de la dimension de la distribution cible sous deux régimes : le régime asymptotique puis le régime en dimensions finies.

Diverses simulations permettent d'illustrer les résultats théoriques obtenus. De plus, une application du nouvel algorithme à un problème de régression logistique bayésienne permet de comparer son efficacité à des algorithmes existants. Les résultats obtenus sont satisfaisants autant d'un point de vue théorique que computationnel.

Mots clés : MCMC, Metropolis-Hasting, MALA, distributions instrumentales informatives, équilibre local, équilibre global

Summary

In this master's thesis, we are interested in a new class of informed proposal distributions for Metropolis-Hastings algorithms. These new proposals, called balanced proposals, are obtained by adding information about the target density to an uninformed proposal distribution. A Markov chain generated by a balanced proposal is reversible with respect to the target density without the need for an acceptance probability in two extreme cases: the local case, where the proposal variance tends to zero, and the global case, where it tends to infinity.

The balanced proposals need to be approximated to be used in practice. We show that the local case leads to the *Metropolis-adjusted Langevin algorithm* (MALA), while the global case leads to a small modification of the MALA. These results are used to create a new algorithm that generalizes the MALA by adding a new parameter. Depending on the value of this parameter, the new algorithm will use a locally balanced proposal, a globally balanced proposal, or an interpolation between these two cases. We then study the optimal choice for this parameter as a function of the dimension of the target distribution under two regimes: the asymptotic regime and a finite-dimensional regime.

Simulations are presented to illustrate the theoretical results. Finally, we apply the new algorithm to a Bayesian logistic regression problem and compare its efficiency to existing algorithms. The results are satisfying on a theoretical and computational standpoint.

Kew words: MCMC, Metropolis-Hastings, MALA, informed proposal, locally balanced, globally balanced

Table des matières

Liste des tableaux

Table des figures

Remerciements

Tel un phare dans la nuit, une étoile sur la voute céleste, Mylène Bédard a pu me guider à travers les méandres de la recherche. D'une bonne humeur inépuisable, ses conseils me permettaient de me remettre sur pied et je ressortais de son bureau toujours rempli d'espoir. Toujours attentive, mais sans jamais trop en faire, je ne pouvais demander mieux en termes de supervision. Merci également pour m'avoir donné l'opportunité de donner un cours. Ce fut une expérience mémorable et j'en suis sorti grandi.

J'aimerais remercier mes parents, Josée et Claude, pour m'avoir toujours supporté même si vous deviez bien vous demander ça fait quoi, dans la vie, un statisticien. Un merci particulier à mon frère pour avoir pu m'aider dans la gestion de ma maîtrise et pour sa patience d'ange lorsque je lui demandais conseil pour la 202^e fois sur la même démonstration.

Merci à Alexis pour ses conseils typographiques et plus généralement pour les merveilleuses conversations de couloir que j'ai pu avoir avec lui. Vous êtes, monsieur, un modèle à suivre.

Merci à Jonathan d'avoir si gentiment partagé ses connaissances sur la théorie de la mesure et d'avoir toujours su égayer mes après-midis avec ses très nombreuses pauses café ainsi que les parties de badminton endiablées.

Merci aux stateux qui m'ont accompagné durant cette maîtrise : Victoire, Vanessa, Paul, Lory-Ann et tous les autres. Merci pour les fous rires, les potins et les séances d'étude collectives toujours semi productives.

Enfin, la meilleure pour la fin, merci à Marianne pour avoir été toujours là quand il le fallait. Je peux difficilement être plus chanceux.

Introduction

En sciences, il arrive souvent que l'on doive calculer des intégrales numériquement. Cela est surtout le cas lorsque ces intégrales sont en grandes dimensions puisqu'une solution analytique est la plupart du temps très difficile, voire impossible, à déterminer. Les méthodes numériques traditionnelles sont bien souvent inefficaces en termes de temps de calcul lorsque le problème comporte beaucoup de dimensions. L'idée des méthodes Monte-Carlo est de traiter ces intégrales comme des espérances et d'utiliser la loi forte des grands nombres afin d'approximer celles-ci. Pour ce faire, il faut tirer un échantillon provenant de la distribution du vecteur aléatoire et calculer la moyenne de celui-ci. Si l'échantillon est assez grand, cette moyenne consistera en une bonne approximation de l'espérance à calculer. Cette méthode est particulièrement utile en inférence bayésienne lorsqu'on désire estimer plusieurs paramètres simultanément. Dans cette situation, l'estimateur d'intérêt est alors l'espérance *a posteriori* qui est une intégrale en grande dimension.

Malheureusement, il n'est pas toujours possible de générer un échantillon d'une distribution générique potentiellement très complexe. Pour outrepasser cette difficulté, au lieu de générer directement l'échantillon de cette distribution (dite cible), on génère un échantillon d'une distribution plus simple (dite instrumentale). Ensuite, on corrige l'échantillon pour pallier le fait que celui-ci ne provient pas de la bonne distribution. Il s'agit de l'idée derrière les méthodes Monte-Carlo par chaînes de Markov (MCMC). Plus précisément, ces méthodes génèrent une chaîne de Markov dont la distribution stationnaire est la distribution cible. Pour ce faire, les candidats générés à l'aide de la distribution instrumentale passent par une étape d'acceptation/rejet. Si ces candidats sont bons, c'est-à-dire qu'ils se retrouvent dans des endroits de haute densité cible, ils sont généralement acceptés ; sinon, ils ont de fortes chances d'être rejetés. Les candidats sont générés un après l'autre ; après plusieurs itérations,

on peut considérer que la chaîne ainsi générée consiste en un échantillon représentatif de la densité cible.

Les méthodes MCMC ont été introduites par Metropolis [**11**] dans le cadre d'une application en physique statistique et ont ensuite été généralisées par Hastings [**6**]. De nos jours, celles-ci sont utilisées autant en finance qu'en génétique et en ingénierie. Leur simplicité ainsi que leur facilité d'implémentation expliquent leur grande popularité. Le plus célèbre de ces algorithmes est le *Random Walk Metropolis* (RWM) ; la distribution instrumentale de cette méthode est symétrique, ce qui simplifie l'étape d'acceptation/rejet. Bien que cet algorithme soit très simple, ses candidats sont générés de manière aveugle. Ceci signifie que l'algorithme n'utilise aucune information sur la distribution cible afin de générer les candidats. De ce fait, ceux-ci se retrouvent souvent dans des creux de la densité cible et sont fréquemment refusés. Cela ralentit la convergence de l'algorithme, c'est-à-dire qu'il faudra un plus grand nombre d'itérations avant d'obtenir un échantillon de qualité.

Pour pallier ce défaut, plusieurs auteurs, notamment Roberts et Rosenthal [**18**] avec le *Metropolis-adjusted Langevin algorithm* (MALA) ainsi que Girolami et Calderhead [**5**] avec le Manifold MALA, ont conçu des algorithmes où la distribution instrumentale est dite informative et tient compte de la densité cible lors de la génération des candidats. Ceux-ci ont alors plus de chance d'être générés aux endroits de haute densité cible et donc d'être acceptés, ce qui accélère la convergence. Par exemple, pour le MALA, la distribution instrumentale est fonction du gradient de la (log) densité cible, ce qui permet de générer les candidats vers les modes de la densité cible. Dans ce mémoire, nous concevons un algorithme utilisant une nouvelle distribution instrumentale informative reposant sur une propriété appelée la propriété d'équilibre. Pour obtenir celle-ci, il faut ajouter de l'information provenant de la densité cible d'une manière bien précise à une distribution instrumentale aveugle. Grâce à cette propriété, les candidats qui sont générés très proche (cas local) ou très loin (cas global) de l'état actuel de la chaîne ont plus de chance d'être acceptés, ce qui accélère la convergence. Nous démontrerons qu'il est plus efficace d'utiliser le cas local en grande dimension, tandis qu'il est préférable d'utiliser le cas global dans le cas contraire.

Cette nouvelle distribution instrumentale utilise, comme le MALA, le gradient de la densité cible. En fait, cette distribution généralise celle utilisée dans le MALA en y ajoutant un paramètre supplémentaire. En faisant varier ce paramètre, il est possible d'obtenir les cas local et global, ainsi qu'une interpolation entre ces deux cas extrêmes. Nous verrons, à l'aide de diverses simulations, qu'en choisissant adéquatement la valeur de ce paramètre en fonction de la dimension du problème, l'algorithme résultant sera plus efficace que le MALA en termes de vitesse de convergence. Diverses applications permettront également d'observer que le temps de calcul par itération est similaire à celui du MALA étant donnée la forte similitude entre les deux méthodes.

Dans le chapitre 1, la théorie de base des chaînes de Markov est présentée. De plus, les principaux algorithmes qui seront utilisés dans ce mémoire, c'est-à-dire l'algorithme de Metropolis-Hastings, le MALA et le *Multiple-try-Metropolis* (MTM), sont introduits. Au chapitre 2, la propriété d'équilibre est présentée et ses ramifications sont explorées. Les équilibres local et global sont d'abord introduits. On présente ensuite la nouvelle distribution instrumentale pouvant utiliser ces deux propriétés en pratique. On justifie par la suite l'utilisation du cas local en grande dimension. Enfin, l'équilibre local est relié à la fonction de poids du MTM. Au chapitre 3, la paramétrisation optimale selon la dimension de cette nouvelle distribution instrumentale est explorée dans un régime non-asymptotique. Un critère d'efficacité est d'abord choisi. On justifie ensuite l'utilisation du cas global en faible dimension. Par la suite, la question de l'optimisation de la distribution d'intérêt dans un régime intermédiaire entre les cas local et global est abordée. Des simulations sont présentées à la fin du chapitre afin d'illustrer les différents résultats obtenus. Finalement, au chapitre 4, le nouvel algorithme est appliqué au problème de la régression logistique bayésienne avec des données réelles afin de comparer son efficacité face au RWM et au MALA. Il est trouvé que sous la paramétrisation optimale, la performance du nouvel algorithme est meilleure que celle du MALA d'un point de vue théorique pour un même coût computationnel. Par contre, la performance théorique peut être en deçà de celle du MALA lorsque l'on s'éloigne trop de la paramétrisation optimale.

Chapitre 1

Méthodes Monte-Carlo par chaînes de Markov

La première partie de ce chapitre présente le contexte d'utilisation des méthodes MCMC. La deuxième partie sert d'introduction à la théorie des chaînes de Markov. L'objectif est de présenter les éléments théoriques nécessaires à la compréhension des chapitres suivants. La présentation se veut succincte, mais le lecteur intéressé pourra retrouver plus de détails dans [**12**]. La dernière partie présente les différents algorithmes MCMC sur lesquels le présent mémoire s'attardera. Le détail du fonctionnement de ceux-ci est présenté et leurs caractéristiques principales sont explorées.

1.1. Contexte d'application des méthodes MCMC

En statistique, il arrive fréquemment que l'on désire calculer une intégrale en plusieurs dimensions de façon numérique. Deux cas se présentent habituellement. Soit l'intégrale à calculer est une espérance, comme c'est le cas la plupart du temps en inférence bayésienne, soit il est possible de réexprimer cette intégrale sous la forme d'une espérance. Dans les deux cas, la méthode Monte-Carlo permet d'approximer cette espérance de façon probabiliste. Supposons que l'on désire approximer $\mathbb{E}[h(X)]$ où X est une variable aléatoire sur l'espace échantillonnal $\mathcal{X}$ et $h : \mathcal{X} \to \mathbb{R}$ est une fonction quelconque. Si nous avons à notre disposition un échantillon indépendant et identiquement distribué (iid) $X_1, X_2, \ldots, X_n$, alors en supposant $h(X) \in L^1$, c'est-à-dire que $\int_{\mathcal{X}} h(x)\mathrm{d}x < \infty$, nous obtenons par la loi forte des grands nombres,

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{p.s} \mathbb{E}[h(X)],$$

lorsque $n \to \infty$ et où $p.s.$ dénote la convergence presque sûre. La méthode Monte-Carlo est simple à utiliser et efficace, mais une difficulté demeure. Pour l'utiliser, il faut pouvoir tirer un échantillon de la distribution de la variable X. Celle-ci est potentiellement très complexe et il n'est pas toujours possible de générer des observations directement de cette distribution. De plus, elle est parfois incomplète. Un exemple classique de ce problème est lorsque la constante de normalisation est manquante ([**20**]).

Ces deux problèmes surviennent fréquemment lorsqu'on effectue une inférence bayésienne. En effet, dans ce contexte, le paramètre à estimer, $\theta \in \mathbb{R}^d$, possède une densité *a priori*, π, que l'on désire mettre à jour après avoir observé l'échantillon $\mathbf{x}$. La vraisemblance de ce dernier est $f(\mathbf{x}|\theta)$ et la densité *a priori* est mise à jour grâce au théorème de Bayes de la manière suivante

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta) f(\mathbf{x}|\theta), \tag{1.1.1}$$

où $\pi(\theta|\mathbf{x})$ est la densité *a posteriori*. On estime alors θ grâce à l'espérance *a posteriori*, $\mathbb{E}[\theta|\mathbf{x}]$, que l'on peut approximer grâce à la méthode Monte-Carlo en générant un échantillon suivant la densité *a posteriori*. Or, en raison de la forme de l'expression (1.1.1), on ne connaît la densité qu'à une constante multiplicative près. C'est ici que les méthodes Monte-Carlo par chaîne de Markov entrent en jeu. Celles-ci permettent de générer non pas un échantillon iid, mais une chaîne de Markov de telle façon que la distribution stationnaire de cette chaîne sera la distribution d'intérêt $\pi(\theta|\mathbf{x})$.

1.2. Chaînes de Markov à espace d'états continu

Cette section définit plusieurs concepts qui forment la base de la théorie des chaînes de Markov à espace d'états continu. Les concepts présentés ici sont ceux qui seront nécessaires à la compréhension du présent mémoire, mais le lecteur intéressé peut se référer à [**12**] pour plus de détails.

Définition 1.2.1. Soit un espace probabilisé, $(\Omega, \mathcal{F}, \mathbb{P})$, T, un ensemble quelconque et $\mathcal{X}$, un espace. Un *processus stochastique* est une fonction $X : \Omega \times T \to \mathcal{X}$, où $\mathcal{X}$ est appelé l'espace d'états.

Un processus stochastique est alors un ensemble de variables aléatoires à valeurs dans $\mathcal{X}$ que l'on peut représenter par $\{X[n,\omega] : n \in T\}$ avec $\omega \in \Omega$. Afin d'alléger la notation, on notera simplement le processus par $\{X[n] : n \in T\}$ pour une réalisation ω quelconque. L'espace $\mathcal{X}$ est générique, mais dans la majorité des applications, $\mathcal{X} = \mathbb{R}^d$, avec $d \in \mathbb{N}$ et où les variables aléatoires ont une densité par rapport à la mesure de Lebesgue. L'ensemble T est souvent dénombrable et l'on pose alors $T = \mathbb{N}$; le processus devient une suite dénombrable de valeurs dans $\mathcal{X}$, soit $\{X[n] : n \in \mathbb{N}\}$. Comme c'est une suite dénombrable, il est aisé de voir cette suite de variables comme étant réalisées à tour de rôle. Ainsi, lors de la réalisation de $X[n+1]$, les n éléments précédents sont connus. Un processus stochastique tel que la prochaine valeur, $X[n+1]$, ne dépend que de $X[n]$, mais pas des autres valeurs précédentes, est appelé une chaîne de Markov.

Définition 1.2.2. Une *chaîne de Markov* est un processus stochastique, $\{X[n] : n \in \mathbb{N}\}$, défini sur l'espace d'états $\mathcal{X}$ tel que pour tout $A \subseteq \mathcal{X}$ mesurable et pour tout n :

$$\mathbb{P}(X[n+1] \in A | \, X[n], X[n-1], \ldots, X[1]) = \mathbb{P}(X[n+1] \in A | \, X[n]).$$

Cette dernière propriété est parfois appelée propriété markovienne. Une autre propriété d'intérêt est celle d'homogénéité. On dira que la chaîne est homogène dans le temps si $\mathbb{P}(X[n+1] \in A | \, X[n] = x) = \mathbb{P}(X[n] \in A | \, X[n-1] = x)$, et ce, pour tout entier n, tout sous-ensemble mesurable $A \subseteq \mathcal{X}$ et tout élément $x \in \mathcal{X}$. On supposera que toutes les chaînes considérées dans ce mémoire sont homogènes.

Définition 1.2.3. Soit une chaîne de Markov, X, définie sur l'espace d'états $\mathcal{X}$. Un *noyau de transition*

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

est une famille de mesures de probabilité conditionnelles à x. Plus précisément, il s'agit de la famille engendrée par la distribution conditionnelle de $X[n+1] \,|\, X[n]$. Cela signifie que $P(x, A) = \mathbb{P}(X[n+1] \in A | \, X[n] = x)$, et ce, pour tout $A \subseteq \mathcal{X}$ mesurable. On note également $P^k(x, A) = \mathbb{P}(X[n+k] \in A | \, X[n] = x)$, soit la $k-$ième itération du noyau P.

En raison de la propriété markovienne, il est clair qu'une chaîne de Markov est entièrement définie par sa valeur de départ et par son noyau de transition. Un autre élément important est le concept de distribution stationnaire. Le principe est que si $X[n]$ est distribué selon une

certaine distribution et qu'à partir de cette valeur de n, les éléments suivants de la chaîne, $X[n+1], X[n+2], ...$, sont distribués selon la même distribution, alors la chaîne est dite stationnaire ; la distribution associée est appelée distribution stationnaire.

Définition 1.2.4. Soit X, une chaîne de Markov à espace d'états $\mathcal{X}$ avec comme noyau de transition P et soit $\Pi(\cdot)$, une distribution de probabilité. On dit que Π est la *distribution stationnaire* de la chaîne de Markov X si pour $A \subseteq \mathcal{X}$ mesurable, on a

$$\Pi(A) = \int_{\mathcal{X}} \Pi(\mathrm{d}x) P(x, A).$$

Il existe une condition simple à valider permettant de vérifier si une chaîne est stationnaire par rapport à une certaine distribution Π. Il s'agit de la condition de réversibilité.

Définition 1.2.5. Une chaîne de Markov X à espaces d'états $\mathcal{X}$ avec noyau de transition P est dite *réversible* par rapport à une distribution Π si

$$\Pi(\mathrm{d}x) P(x, \mathrm{d}y) = \Pi(\mathrm{d}y) P(y, \mathrm{d}x), \qquad x, y \in \mathcal{X}. \tag{1.2.1}$$

Proposition 1.2.6. *Si une chaîne de Markov X à espace d'états $\mathcal{X}$ est réversible par rapport à une distribution Π, il s'agit de sa distribution stationnaire.*

Démonstration. Puisque $P(y, \cdot)$ est une mesure de probabilité sur $\mathcal{X}$, ceci implique que $\int_{\mathcal{X}} P(y, \mathrm{d}x) = 1$ pour tout $y \in \mathcal{X}$. Ainsi, pour $y \in A \subseteq \mathcal{X}$ mesurable et $x \in \mathcal{X}$, on obtient

$$\begin{aligned}
\int_{x \in \mathcal{X}} \Pi(dx) P(x, A) &= \int_{x \in \mathcal{X}} \int_{y \in A} \Pi(\mathrm{d}x) P(x, \mathrm{d}y) \\
&= \int_{x \in \mathcal{X}} \int_{y \in A} \Pi(\mathrm{d}y) P(y, \mathrm{d}x) \\
&= \int_{y \in A} \Pi(\mathrm{d}y) \int_{x \in \mathcal{X}} P(y, \mathrm{d}x) \\
&= \Pi(A).
\end{aligned}$$

□

La distribution stationnaire représente, en quelque sorte, la distribution limite des états de la chaîne. Le principe des méthodes MCMC est alors de choisir un noyau de transition particulier de telle sorte que la chaîne soit réversible par rapport à la distribution Π de laquelle on désire tirer un échantillon. Cependant, même si une chaîne possède une distribution stationnaire Π, il n'est pas vrai en général que cette chaîne va converger vers cette distribution. En d'autres mots, il n'existe pas nécessairement un n tel que $X[n], X[n+1], \ldots \sim \Pi$. La chaîne

doit également posséder différentes propriétés. La première est celle de φ-irréductibilité qui implique que l'on peut atteindre, à partir de n'importe quel point, toute région de mesure φ non nulle avec une probabilité positive.

Définition 1.2.7. Le noyau de transition P d'une chaîne de Markov X à espace d'états $\mathcal{X}$ est φ-irréductible si pour tout $x \in \mathcal{X}$ et pour tout $A \subseteq \mathcal{X}$ tel que $\varphi(A) > 0$, il existe un $n \in \mathbb{N}$ tel que $P^n(x, A) > 0$.

Dans le cas présent, il est naturel de choisir $\varphi = \Pi$. La deuxième propriété est celle d'apériodicité, c'est-à-dire que la chaîne ne se déplace pas de façon cyclique entre certains ensembles.

Définition 1.2.8. Soit X, une chaîne de Markov à espace d'états $\mathcal{X}$, noyau de transition P et distribution stationnaire Π. La chaîne X est dite *apériodique* s'il n'existe pas d'entier $k \geqslant 2$ et des ensembles disjoints $A_1, A_2, \ldots, A_k \subseteq \mathcal{X}$, où

$$P(x, A_{i+1}) = 1 \quad \forall\, x \in A_i \,(1 \leqslant i \leqslant k-1)$$

et $P(x, A_1) = 1$ si $x \in A_k$, tels que $\Pi(A_i) > 0$ pour tout $1 \leqslant i \leqslant k$. Si tel est le cas, la chaîne est *périodique* de période k.

Enfin, la dernière propriété est celle de récurrence, et plus précisément de Harris récurrence. Cette propriété implique qu'il est possible de visiter, une infinité de fois, chaque état de la chaîne.

Définition 1.2.9. Soit X, une chaîne de Markov à espace d'états $\mathcal{X}$ et noyau de transition P. On suppose que la chaîne est Π-irréductible, où Π est sa distribution stationnaire associée. Alors pour tout $A \subseteq \mathcal{X}$ tel que $\Pi(A) > 0$ et pour tout $x \in \mathcal{X}$, la chaîne est dite *récurrente* si

$$\begin{aligned} &\mathbb{P}(\{X[n+1] \in A \; i.s.\} \,|\, X[0] = x) > 0 \quad \text{pour tout } x \\ &\mathbb{P}(\{X[n+1] \in A \; i.s.\} \,|\, X[0] = x) = 1 \quad \text{pour } \Pi\text{-presque tout } x. \end{aligned} \tag{1.2.2}$$

La chaîne est *Harris récurrente* si (1.2.2) est vrai pour tout $x \in \mathcal{X}$. Ici, i.s. est une abbréviation pour infiniment souvent. Plus précisément, pour une suite d'événements $\{A_n : n \in \mathbb{N}\}$, on décrit l'événement « A_n survient infiniment souvent » comme

$$\{A_n \; i.s.\} = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n.$$

Pour que la chaîne converge vers la distribution stationnaire, il faut qu'à partir d'un certain n, $P^n(x, A) = \mathbb{P}(X[n] \in A | X[0] = x)$ soit « proche » de $\Pi(A)$, et ce, peu importe le point de départ x de la chaîne. Le théorème suivant présente les conditions nécessaires pour que ce soit le cas.

Théorème 1.2.10. *Soit une chaîne de Markov X définie sur l'espace d'états $\mathcal{X}$ avec comme noyau de transition P. On suppose que P est $\Pi-$irréductible avec Π, sa distribution stationnaire associée. De plus, on suppose que P est apériodique et Harris récurrente. Alors, pour tout $x \in \mathcal{X}$,*

$$\lim_{n\to\infty} P^n(x, A) = \Pi(A),$$

pour tout $A \subseteq \mathcal{X}$ mesurable.

Une démonstration de ce théorème se retrouve dans [**20**] (voir théorème 4). Ainsi, s'il est possible de générer une chaîne de Markov qui respecte les conditions du théorème 1.2.10, la chaîne ainsi formée convergera éventuellement vers la distribution Π. Après un certain temps n^*, appelé « burn in » en pratique, les états de la chaîne pourront être considérés comme un échantillon provenant de la distribution Π. Si la chaîne démarre directement en stationnarité, alors il n'est pas nécessaire d'attendre avant de conserver les états générés. Lorsqu'une chaîne respecte les conditions du théorème 1.2.10, une loi forte des grandes nombres « markovienne » existe et justifie l'emploi de la méthode Monte-Carlo avec cette chaîne de Markov pour approximer des intégrales. Ceci est résumé dans le théorème suivant, dont la démonstration se retrouve dans [**12**] (voir théorème 17.1.7).

Théorème 1.2.11. *Soit X, une chaîne de Markov respectant les conditions du théorème 1.2.10 et soit $h : \mathcal{X} \to \mathbb{R}$, une fonction telle que $\int_{\mathcal{X}} |h(x)|\Pi(dx) < \infty$, alors*

$$\frac{1}{n}\sum_{i=1}^{n} h\left(X[i]\right) \xrightarrow{p.s.} \int_{\mathcal{X}} h(x)\Pi(dx)$$

lorsque $n \to \infty$.

Les algorithmes de type Metropolis-Hastings sont un exemple de méthodes MCMC générant une chaîne qui respecte les conditions du théorème 1.2.10.

1.3. Algorithmes de type Metropolis-Hastings

1.3.1. Algorithme Metropolis-Hastings et RWM

Les algorithmes de type Metropolis-Hastings sont sans nul doute les méthodes les plus populaires parmi les méthodes MCMC dans la pratique. Cela est dû à leur efficacité et à leur simplicité d'implémentation. La première version a été présentée par [**11**], puis généralisée par [**6**]. Dans toutes ces versions, chaque élément de la chaîne est généré un à la fois de manière séquentielle et à chaque itération, le candidat généré passe par une étape d'acceptation/rejet. Les méthodes diffèrent habituellement dans la manière par laquelle les candidats sont générés. La version la plus simple est l'algorithme Metropolis-Hastings (MH). Étant une méthode MCMC, cet algorithme vise à échantillonner d'une distribution cible, Π, ayant π comme densité associée, c'est-à-dire que $\Pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$, avec $\mathcal{X}$ comme espace d'états et $\mathrm{d}x$, la mesure de Lebesgue. On considère que la densité cible est bornée. Il faut d'abord choisir un point de départ, x_0, de telle sorte que $\pi(x_0) > 0$. Ensuite, l'algorithme découle comme suit.

Soit x, la valeur de la chaîne au temps n, donc $X[n] = x$. Soit $Q(x, \cdot)$, une distribution quelconque avec $q(x, \cdot)$ comme densité associée, c'est-à-dire que pour tout $x, y \in \mathcal{X}$, on écrit $Q(x, \mathrm{d}y) = q(x, y)\mathrm{d}y$. La présence de x dans la distribution signifie qu'elle dépend de l'état actuel de la chaîne. Au temps $n + 1$, les étapes de l'algorithme vont comme suit :

(1) Générer un candidat $y \sim Q(x, \cdot)$ pour la prochaine valeur de la chaîne.

(2) Accepter ce candidat avec probabilité

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}. \tag{1.3.1}$$

(3) Si le candidat est accepté, poser $X[n + 1] = y$; sinon, poser $X[n + 1] = x$.

Dans l'éventualité où $\pi(x)q(x, y) = 0$ à l'étape 2, afin d'éviter toute ambiguïté, on adopte la convention que $\alpha(x, y) = 1$. Cependant, étant donnée la construction de l'algorithme, on aura toujours $\pi(x) > 0$. En effet, il n'est pas possible d'accepter des candidats y tels que $\pi(y) = 0$ étant donnés la forme de $\alpha(x, y)$ et le point de départ de la chaîne, qui est tel que $\pi(x_0) > 0$.

Remarque 1.3.1. Le noyau de transition de la chaîne de Markov formée par l'algorithme MH est, pour tout $x, y \in \mathcal{X}$,

$$P(x, \mathrm{d}y) = Q(x, \mathrm{d}y)\alpha(x, y) + \delta_x(\mathrm{d}y) \int_{z \in \mathcal{X}} (1 - \alpha(x, z))q(x, z)dz, \tag{1.3.2}$$

où δ_x est la mesure de Dirac centrée en x :

$$\delta_x(A) = \begin{cases} 1 & \text{si } x \in A, \\ 0 & \text{si } x \notin A. \end{cases}$$

La proposition suivante combine des résultats présentés dans [**10**] (lemmes 1.1 et 1.2) et [**23**] (corollaire 1). Cette proposition montre que l'algorithme MH génère bien une chaîne respectant les conditions du théorème 1.2.10 et donc que cette chaîne converge vers la distribution cible.

Proposition 1.3.2. *Soit X, la chaîne de Markov à espace d'états $\mathcal{X}$ générée par l'algorithme MH. Soit également π, la densité bornée associée à la distribution cible Π de X et q, une densité quelconque. Si la condition suivante est respectée,*

$$\pi(y) > 0 \Rightarrow q(x, y) > 0, \quad \forall\, x \in \mathcal{X}, \tag{1.3.3}$$

alors la chaîne respecte les conditions du théorème 1.2.10.

Démonstration. On démontre d'abord la réversibilité par rapport à Π, c'est-à-dire qu'il faut démontrer l'égalité (1.2.1) pour tout $x \neq y \in \mathcal{X}$, puisque que le cas où $x = y$ est trivial. Par définition, on a que $\Pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$. De plus, selon (1.3.2), lorsque $x \neq y$, on a $P(x, \mathrm{d}y) = q(x, y)\alpha(x, y)\mathrm{d}y$. En effet, pour se déplacer du point x au point y, il faut d'abord proposer le point y et ensuite l'accepter. Ainsi, ceci implique

$$\begin{aligned}
\Pi(\mathrm{d}x)P(x, \mathrm{d}y) &= \pi(x)q(x, y)\alpha(x, y)\mathrm{d}x\mathrm{d}y \\
&= \pi(x)q(x, y) \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \mathrm{d}x\mathrm{d}y \\
&= \min\{\pi(x)q(x, y), \pi(y)q(y, x)\}\mathrm{d}x\mathrm{d}y \\
&= \pi(y)q(y, x) \min\left\{1, \frac{\pi(x)q(x, y)}{\pi(y)q(y, x)}\right\} \mathrm{d}x\mathrm{d}y \\
&= \Pi(\mathrm{d}y)q(y, x)\alpha(y, x)\mathrm{d}x \\
&= \Pi(\mathrm{d}y)P(y, \mathrm{d}x).
\end{aligned}$$

Pour la $\Pi-$irréversibilité, il faut montrer que $P(x, A) > 0$ pour tout $x \in \mathcal{X}$ et pour tout $A \subseteq \mathcal{X}$ mesurable tel que $\Pi(A) > 0$. Soit un ensemble $A \subseteq \mathcal{X}$ tel que $\Pi(A) > 0$. Puisque $\Pi(A) > 0$, l'ensemble $A_0 = \{y \in A : \pi(y) > 0\}$ est tel que $\Pi(A_0) > 0$. Soit $x \in \mathcal{X}$. On considère d'abord le cas où $\pi(x) > 0$. Soit l'ensemble

$$A_0^* = \left\{ y \in A_0 : \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} < 1 \right\}.$$

En séparant l'ensemble A_0 en deux selon l'appartenance, ou non, de y à A_0^*, il est alors possible d'écrire

$$\begin{aligned}
P(x, A) &\geqslant P(x, A_0) \\
&\geqslant \int_{A_0} q(x,y)\alpha(x,y)\mathrm{d}y \\
&= \int_{A_0^*} q(x,y) \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} \mathrm{d}y + \int_{A_0 \setminus A_0^*} q(x,y) \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} \mathrm{d}y \\
&= \int_{A_0^*} \frac{\pi(y)q(y,x)}{\pi(x)} \mathrm{d}y + \int_{A_0 \setminus A_0^*} q(x,y)\mathrm{d}y \\
&= \int_{A_0^*} \frac{\pi(y)q(y,x)}{\pi(x)} \mathrm{d}y + \int_{A_0 \setminus A_0^*} \frac{\pi(y)q(x,y)}{\pi(y)} \mathrm{d}y.
\end{aligned}$$

Par la condition (1.3.3), on a que $\inf\limits_{y \in A_0} \min\{q(x,y), q(y,x)\} \geqslant \varepsilon$ pour $\varepsilon > 0$ puisque $\pi(y) > 0$ pour tout $y \in A_0$ et $\pi(x) > 0$. De plus, $m = \sup\limits_{x \in \mathcal{X}} \pi(x) < \infty$ puisque la densité π est bornée. Ainsi, il suit que

$$P(x, A) \geqslant \frac{\varepsilon}{m} \int_{A_0^*} \pi(y)\mathrm{d}y + \frac{\varepsilon}{m} \int_{A_0 \setminus A_0^*} \pi(y)\mathrm{d}y = \frac{\varepsilon}{m} \Pi(A_0) > 0.$$

Lorsque $\pi(x) = 0$, $\alpha(x,y) = 1$ par convention, ce qui implique

$$P(x, A) \geqslant P(x, A_0) \geqslant \varepsilon \Pi(A_0) > 0.$$

En ce qui concerne l'apériodicité, il est possible de trouver une contradiction si la chaîne est périodique. Supposons, sans perte de généralité, que la période est de 2. Cela signifie qu'il existe des ensembles, $A_1, A_2 \subset \mathcal{X}$, disjoints tels que $\Pi(A_1) > 0$ et $\Pi(A_2) > 0$ et où $P(x, A_2) = 1$ pour tout $x \in A_1$. Soit $x \in A_1$. Puisque $\Pi(A_1) > 0$, l'ensemble A_1 est de mesure de Lebesgue non nulle et puisque la chaîne est Π-irréductible, alors

$$P(x, A_1) \geqslant \int_{y \in A_1} q(x,y)\alpha(x,y)\mathrm{d}y > 0.$$

Ainsi, $P(x, \{A_1 \cup A_2\}) = P(x, A_1) + P(x, A_2) > 1$ ce qui est une contradiction. La chaîne est donc apériodique.

Enfin, en ce qui concerne la Harris récurrence, il s'agit d'une conséquence directe de la $\Pi-$irréversibilité tel qu'énoncée par le corollaire 1 de [**23**]. □

La distribution $Q(x, \cdot)$ est appelée la distribution instrumentale. Le choix de celle-ci influence grandement la performance de l'algorithme. Le plus souvent, une distribution normale multivariée centrée en x avec composantes indépendantes est utilisée. La même variance, σ^2, est utilisée pour toutes les composantes. Sa densité associée en dimension d est

$$q(x, y) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \|y - x\|^2\right), \tag{1.3.4}$$

où $\|\cdot\|$ est la norme euclidienne. Cette densité a la particularité d'être symétrique. L'algorithme MH utilisant une densité intrumentale symétrique est appelé *Random Walk Metropolis* (RWM) dans la littérature et est l'algorithme qui a été introduit originellement par Metropolis [**11**]. Bien que n'importe quelle densité symétrique puisse être utilisée dans un RWM, pour la suite des choses, on considérera que le RWM utilise la densité instrumentale (1.3.4) comme c'est souvent le cas en pratique. Celle-ci étant symétrique en x et en y, le ratio en (1.3.1) devient $\alpha(x, y) = \min\{1, \pi(y)/\pi(x)\}$. Le ratio (1.3.1) est appelé la probabilité d'acceptation. La forme de celle-ci n'est pas la seule possible. En fait, plus généralement, toute fonction $\alpha(x, y)$ respectant

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\alpha(y, x) \tag{1.3.5}$$

permettra d'obtenir un noyau de transition réversible par rapport à Π. Il existe plusieurs fonctions respectant ce critère, mais (1.3.1) est utilisée en pratique puisque Peskun [**15**] a démontré que cette forme menait à un algorithme optimal au sens de la variance asymptotique. Pour une fonction $h : \mathcal{X} \to \mathbb{R}$ et une probabilité de transition P, cette variance asymptotique est définie comme

$$\mathrm{var}_\pi(h, P) = \lim_{n\to\infty} \frac{1}{n}\mathrm{var}\left(\sum_{i=1}^{n} h(X[i])\right), \tag{1.3.6}$$

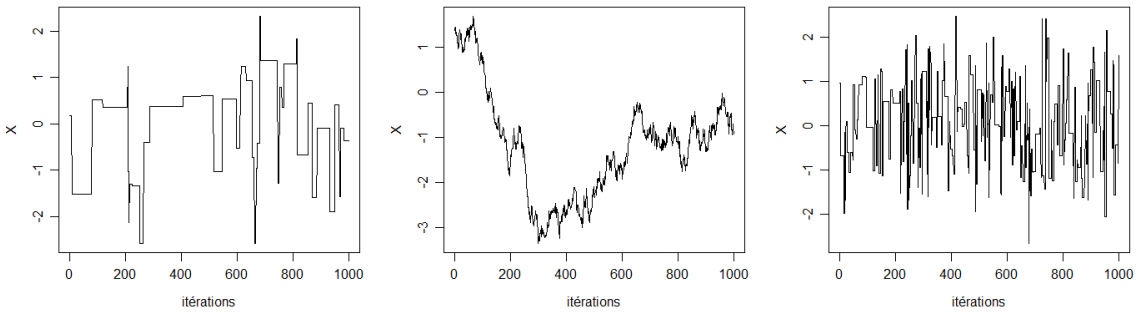
où $X[1], \ldots, X[n]$ est une chaîne de Markov avec probabilité de transition P démarrant en stationnarité ($X[1] \sim \pi$). Ainsi, plus la variance asymptotique sera petite, plus l'estimation de $\mathbb{E}_\pi[h(X)]$ sera précise. Il s'avère que parmi toutes les fonctions $\alpha(x, y)$ disponibles, (1.3.1)

permet de minimiser la variance asymptotique pour l'algorithme RWM.

Le rôle du ratio en (1.3.1) est de corriger la distribution instrumentale utilisée. Si celle-ci était la distribution cible, c'est-à-dire $q(x, y) = \pi(y)$, alors tous les candidats seraient acceptés automatiquement, puisque le ratio serait égal à 1. Cela augmenterait la vitesse à laquelle l'algorithme explore l'espace d'états, étant donné qu'il ne ferait jamais de surplace. En fait, sans ce ratio, la chaîne serait réversible par rapport à une distribution autre que la cible. Le ratio corrige donc la distribution instrumentale qui n'est pas réversible par rapport à la bonne distribution. Nous pouvons donc en déduire que le choix de la distribution instrumentale a un impact sur l'efficacité de l'algorithme. Par exemple, il est possible d'utiliser une distribution instrumentale dite informative. Les distributions de ce type utilisent de l'information fournie par Π afin de générer le candidat y. Cela permet d'explorer plus efficacement l'espace d'états, par opposition à une exploration aveugle, comme lorsque les candidats sont générés par (1.3.4). Plusieurs auteurs ont travaillé sur le design de densités instrumentales prenant en compte l'information locale afin de générer un nouveau candidat, notamment [**18**] et [**5**]. Ces algorithmes sont présentés dans la section suivante.

Enfin, lorsque la densité instrumentale (1.3.4) est choisie, le seul paramètre à ajuster est σ. Celui-ci détermine l'amplitude des pas entre les éléments de la chaîne générée. Plus σ est grand et plus les candidats générés sont éloignés de l'état actuel de la chaîne. Toutefois, si la valeur de σ est trop grande, les candidats générés risquent de trop souvent se retrouver dans des endroits où la densité cible est faible. Ils auront donc plus de chance d'être rejetés à la deuxième étape de l'algorithme. Cet aspect est particulièrement vrai lorsque la dimension, d, augmente. En pratique, il faut diminuer la valeur de σ à mesure que d croît, sous peine d'avoir un taux d'acceptation très faible. Habituellement, la variance est réduite par un facteur de $d^{-\beta}$ où la valeur de β varie selon l'algorithme utilisé. Pour le RWM, la variance utilisée est $\sigma_d^2 = \ell/d$, où ℓ est à déterminer. Le choix de ℓ est un problème du type Boucles d'or, c'est-à-dire qu'il faut savoir trouver un juste milieu pour la valeur de ℓ. Si ℓ est trop petit, l'algorithme ne fera que de petits pas et l'exploration de l'espace d'états sera lente. À l'inverse, si la valeur de ℓ est trop grande, l'algorithme aura un faible taux d'acceptation et fera donc beaucoup

de surplace menant ici aussi à une exploration de l'espace inefficace. Ceci est illustré à la figure 1.1. Le choix du σ_d idéal pour l'algorithme RWM a été étudié par Roberts, Gelman



(a) σ_d trop grand **(b)** σ_d trop petit **(c)** σ_d idéal

Figure 1.1. Les 1000 premières itérations d'un algorithme RWM sur une cible $\mathcal{N}(0,1)$ selon diverses valeurs de σ_d.

et Gilks dans [**17**] pour des distributions cibles dont les composantes sont indépendantes et identiquement distribuées (iid). Le choix de ℓ qui maximise l'efficacité de l'algorithme est celui permettant d'obtenir un taux d'acceptation asymptotique des candidats de 0,234. Le taux d'acceptation asymptotique est le taux d'acceptation des candidats lorsque $d \to \infty$. Ainsi, le choix du σ_d optimal ne dépend pas de π, ce qui facilite la tâche de l'utilisateur. À partir de ce point, on considérera que la variance σ^2 dépend toujours de la dimension d et on utilisera sans distinction les notations σ^2 et σ_d^2.

1.3.2. Algorithme MALA

Le *Metropolis-adjusted Langevin algorithm* (MALA) [**18**] est un algorithme utilisant une distribution instrumentale dite informative. Le principe est de proposer des candidats se situant dans des régions de plus grande densité en utilisant le gradient de la distribution cible afin de diriger la recherche dans la bonne direction. Les candidats ainsi proposés ont plus de chance d'être acceptés, ce qui augmente la vitesse d'exploration de l'espace d'états. L'algorithme comporte deux étapes. La première consiste à générer les candidats à l'aide d'une distribution instrumentale particulière. La deuxième est une étape d'acceptation/rejet semblable à celle de l'algorithme Metropolis-Hastings. Pour la première étape, la distribution instrumentale utilisée est basée sur la discrétisation d'un processus de diffusion de Langevin. Un processus de diffusion sert à décrire, par exemple, l'évolution de la position d'une particule

dans le temps lorsque son mouvement est aléatoire. Ce type de processus est la généralisation des chaînes de Markov en temps continu. Les éléments théoriques sur les processus de diffusion nécessaires à l'implantation du MALA sont présentés ici, mais il est possible de se référer à [**14**] pour plus de détails sur la théorie des processus stochastiques en général.

Définition 1.3.3. Un processus stochastique, $\{X[t] : t \geqslant 0\}$, à espace d'états $\mathcal{X}$ de dimension d est appelé une *diffusion* avec fonction de dérive, $\mu(x)$, et fonction de diffusion, $\beta(x)$, s'il vérifie l'équation différentielle stochastique

$$\mathrm{d}X[t] = \mu(X[t])\mathrm{d}t + \beta(X[t])\mathrm{d}B_t,$$

où B_t est le mouvement brownien standard en d dimensions et $\mu : \mathcal{X} \to \mathbb{R}^d$ et $\beta : \mathcal{X} \to \mathbb{R}^d \times \mathbb{R}^d$ sont des fonctions continues.

Le processus de diffusion de Langevin à espace d'états $\mathcal{X}$ de dimension d, noté $\{L[t] : t \geqslant 0\}$, est un processus de diffusion avec fonctions de dérive et de diffusion

$$\mu(x) = \frac{\sigma^2}{2}\nabla \log\{\pi(x)\} \qquad \text{et} \qquad \beta(x) = \sigma I_d$$

où ∇ est le gradient, $\sigma > 0$, I_d est la matrice identité en dimension d et π est une densité quelconque à valeur dans $\mathcal{X}$. Ceci signifie que le processus $L(t)$ satisfait l'équation différentielle

$$\mathrm{d}L[t] = \frac{\sigma^2}{2}\nabla \log\{\pi(L[t])\}\mathrm{d}t + \sigma \mathrm{d}B_t.$$

Ici, il s'avère que la densité π est la densité associée à la distribution stationnaire, Π, du processus, c'est-à-dire que $\Pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$. Un processus de diffusion peut, comme une chaîne de Markov, posséder une distribution stationnaire sous certaines conditions (voir [**14**]). En fait, il est possible de démontrer que non seulement $L[t]$ possède Π comme distribution stationnaire, mais également que pour tout $x \in \mathcal{X}$, le processus converge vers cette distribution, c'est-à-dire

$$\lim_{t\to\infty} P_L^t(x, A) = \Pi(A), \qquad \forall\, A \subseteq \mathcal{X},$$

où $P_L^t(x, A) = P(L[t] \in A \,|\, L[0] = x)$, $t \geq 0$ (voir [**21**], théorème 2.1). Ceci trace donc un parallèle avec le théorème 1.2.10 pour les chaînes de Markov.

Étant donné que l'objectif ici est de générer une chaîne de Markov, il faut pouvoir discrétiser ce processus de diffusion. Néanmoins, il est intéressant d'observer que le processus de diffusion de Langevin, duquel est inspirée la chaîne de Markov créée par le MALA, converge vers Π. Il existe plusieurs façons de discrétiser un processus de diffusion, mais dans ce cas-ci, la discrétisation utilisée est la suivante : pour $n \in \mathbb{N}$, on pose

$$X[n+1] - X[n] = \frac{\sigma^2}{2} \nabla \log\{\pi(X[n])\} + \sigma Z[n+1], \tag{1.3.7}$$

où les variables aléatoires $Z[n] \sim \mathcal{N}_d(\mathbf{0}, I_d)$ sont indépendantes, avec $\mathbf{0} = (0, \ldots, 0) \in \mathcal{X}$. Ces dernières jouent le rôle du mouvement brownien dans le cas discret. La question maintenant est de savoir si la discrétisation du processus de Langevin possède les mêmes propriétés intéressantes en ce qui concerne la convergence du processus. Or, malheureusement, lorsque la chaîne de Markov est générée selon (1.3.7), c'est-à-dire

$$X[n+1] \,|\, X[n] = x \sim \mathcal{N}\left(x + \frac{\sigma^2}{2} \nabla \log\{\pi(x)\}, \sigma^2 I_d\right),$$

la chaîne peut ne pas converger vers la bonne distribution cible. À titre d'exemple, si $\pi(x)$ est une $\mathcal{N}(0,1)$ sur $\mathbb{R}$ et que $\sigma^2 = 2$, on a

$$\log\{\pi(x)\} = \frac{-x^2}{2} - \frac{\log(2\pi)}{2} \Rightarrow \frac{\partial}{\partial x} \log\{\pi(x)\} = -x,$$

et donc

$$X[n+1] \,|\, X[n] = x \sim \mathcal{N}(0, 2).$$

Ainsi, dans ce cas, il y a « convergence » immédiate, mais pas vers la distribution désirée. D'autres problèmes surviennent dans des cas particuliers. C'est pour ces raisons que le MALA comporte une deuxième étape qui permet d'obtenir une chaîne réversible par rapport à Π. Celle-ci est l'étape d'acceptation/rejet telle qu'introduite en (1.3.1). Lorsque l'état de la chaîne au temps n est $x \in \mathcal{X}$, les étapes du MALA sont les suivantes :

(1) Générer $y \sim q(x, \cdot)$, où

$$q(x,y) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{1}{2\sigma^2} \left\| y - x - \frac{\sigma^2}{2} \nabla \log\{\pi(x)\} \right\|^2\right\}. \tag{1.3.8}$$

(2) Accepter ce candidat avec probabilité

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

(3) Si le candidat est accepté, poser $X[n+1] = y$; sinon, poser $X[n+1] = x$.

Il est à noter que généralement $q(x, y) \neq q(y, x)$. Puisque le MALA n'est en fait qu'un MH avec une distribution instrumentale $Q(x, \cdot)$ particulière, les conditions du théorème 1.2.10 sont respectées et la chaîne converge bien vers la distribution cible Π. L'inconvénient du MALA est que le calcul du gradient est généralement coûteux computationnellement. Malgré tout, la performance du MALA est de loin supérieure à celle du RWM lorsque d augmente (voir [**18**]).

Pour le MALA, le seul paramètre à ajuster est σ. Ainsi, tout comme pour le RWM, le choix de ce paramètre doit être effectué judicieusement. De trop grandes ou trop petites valeurs mènent à une exploration fastidieuse de l'espace d'états. À nouveau, pour une distribution cible avec composantes indépendantes et identiquement distribuées, il a été démontré dans [**18**] que le choix de ℓ optimal pour une variance de la forme $\sigma_d^2 = \ell/d^{1/3}$ est celui permettant d'obtenir un taux d'acceptation asymptotique de 0,574. Ainsi, le MALA est optimal pour un taux d'acceptation plus grand que le 0,234 du RWM mentionné précédemment. De plus, la variance optimale du RWM est $\mathcal{O}(d^{-1})$ tandis que celle du MALA est $\mathcal{O}(d^{-1/3})$; les pas effectués par le MALA seront donc généralement plus grands que ceux du RWM. Ces deux aspects impliquent donc que le MALA aura tendance à explorer $\mathcal{X}$ beaucoup plus rapidement que le RWM traditionnel.

Lorsque les composantes de la distribution cible ne sont pas indépendantes et ont des variances hétérogènes, il faut utiliser une distribution instrumentale différente de (1.3.8). Autrement, il faudrait spécifier une valeur de σ trop conservatrice, c'est-à-dire pouvant accommoder les différentes variances, afin de ne pas avoir un taux d'acceptation trop faible, et donc une exploration inefficace de $\mathcal{X}$. Pour résoudre ce problème, il est possible d'utiliser le MALA avec une matrice pré-conditionnée, M, dans la discrétisation du processus, soit

$$X[n+1] - X[n] = \frac{\sigma^2}{2} M \nabla \log\{\pi(X[n])\} + \sigma M^{1/2} Z[n+1].$$

Certains choix de M mènent à des algorithmes plus efficaces que d'autres. Dans [**5**], Girolami et Calderhead ont démontré qu'utiliser la géométrie inhérente à l'espace d'états en posant

$M = \mathcal{I}^{-1}(X)$, soit l'inverse de l'information de Fisher, menait à un algorithme très performant. Cela est particulièrement le cas pour des distributions cibles en grande dimension avec des composantes fortement corrélées. Dans ce cas, l'utilisation de $M = \mathcal{I}^{-1}(X)$ résulte en un algorithme ayant une vitesse de convergence beaucoup plus grande que le MALA, ce dernier étant très peu efficace dans de telles situations. En pratique, inverser l'information de Fisher à chaque itération peut s'avérer coûteux computationnellement. Ainsi, en fonction de la complexité de la distribution cible, il ne sera pas toujours judicieux d'utiliser cet algorithme. L'utilisation du MALA ou même du RWM dans certaines situations pourrait être recommandée, même si leur vitesse de convergence est plus lente. Il y a donc un équilibre à trouver entre la vitesse de l'exploration de $\mathcal{X}$ et le coût computationnel requis.

1.3.3. Algorithme Multiple-Try-Metropolis

L'algorithme RWM est efficace dans plusieurs situations, mais possède certains défauts. Par exemple, il est difficile pour ce dernier de faire de grands sauts et d'aller explorer des régions de l'espace éloignées de la valeur de la chaîne au temps présent. L'algorithme *Multiple-try-Metropolis* (MTM), présenté originellement par Liu, Liang et Wong [**8**], apporte une solution à ce problème. Cet algorithme possède le même squelette que l'algorithme MH traditionnel, mais il génère plusieurs candidats lors d'une même itération. Un candidat parmi ceux-ci est ensuite sélectionné et, comme c'est le cas dans l'algorithme MH, ce candidat est accepté ou rejeté avec une certaine probabilité. Le fait de générer plusieurs candidats à chaque itération augmente la chance de choisir des candidats plus éloignés et permet d'explorer plus rapidement l'espace d'états. Tout comme pour le MH, l'algorithme nécessite une densité instrumentale, $q(x, \cdot)$. De plus, il faut une fonction de poids, $w : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, utilisée pour pondérer les candidats générés. Lorsqu'on écrit $w(x, y)$, on calcule le poids du candidat y alors que x est connu. On suppose que $w(x, y) > 0$ lorsque $\pi(y) > 0$. Soit $X[n] = x$, l'état de la chaîne au temps n. Au temps $n + 1$, l'algorithme va comme suit :

(1) Pour $i = 1, \ldots, N$, générer $y^{(i)} \sim q(x, \cdot)$ de telle sorte que les $y^{(i)}$ sont indépendants.

(2) Calculer les poids $w(x, y^{(i)})$, $i = 1, \ldots, N$, puis les normaliser pour obtenir $\overline{w}_i$, c'est-à-dire

$$\overline{w}_i = \frac{w(x, y^{(i)})}{\sum_{j=1}^{N} w(x, y^{(j)})}.$$

(3) Choisir $y = y^{(k)} \in \{y^{(1)}, \ldots, y^{(N)}\}$ proportionnellement aux poids $\overline{w}_i$, $i = 1, \ldots, N$ et poser $W_y = \overline{w}_k$.

(4) Générer $x_*^{(i)} \sim q(y, \cdot)$, $i = 1, \ldots, k-1, k+1, \ldots, N$ et poser $x_*^{(k)} = x$. Ces points sont appelés les points de référence.

(5) Calculer les poids $w(y, x_*^{(i)})$, $i = 1, \ldots, N$ puis calculer

$$W_x = \frac{w(y, x)}{w(y, x) + \sum\limits_{\substack{i=1 \\ i \neq k}}^{N} w(y, x_*^{(i)})}.$$

(6) Accepter le candidat y avec probabilité

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y) q(y, x) W_x}{\pi(x) q(x, y) W_y}\right\}. \tag{1.3.9}$$

(7) Si le candidat est accepté, poser $X[n+1] = y$; sinon, poser $X[n+1] = x$.

Générer plusieurs candidats permet d'explorer plus rapidement $\mathcal{X}$, mais augmente aussi le coût computationnel par rapport au MH. De plus, il faut également générer des points de référence à l'étape 4. Ceux-ci n'aident pas à explorer l'espace d'états, mais sont nécessaires afin de rendre la chaîne réversible, tel que mentionné dans la remarque suivante.

Remarque 1.3.4. Le ratio en (1.3.9) permet d'obtenir la propriété de réversibilité. Supposons que $x \neq y$. En dénotant $P(x, \mathrm{d}y)$ le noyau de transition de la chaîne créée par le MTM ainsi que I, l'indice du candidat choisi, on a que

$$\begin{aligned} \Pi(\mathrm{d}x) P(x, \mathrm{d}y) &= \Pi(\mathrm{d}x) \mathbb{P}\left[\bigcup_{i=1}^{N} \{(Y^{(i)} \in \mathrm{d}y) \cap (I = i)\} \,|\, x\right] \\ &= \Pi(\mathrm{d}x) \sum_{i=1}^{N} \mathbb{P}\left[(Y^{(i)} \in \mathrm{d}y) \cap (I = i) \,|\, x\right]. \end{aligned} \tag{1.3.10}$$

En montrant que chaque élément de la somme en (1.3.10) est réversible, on montre que la réversibilité est respectée. Soit $k \in \{1, \ldots, N\}$. On note $\mathbb{P}\left[(Y^{(k)} \in \mathrm{d}y) \cap (I = k) \,|\, x\right] = p_k$. Selon cette probabilité, pour se déplacer de x à y, on désire que ce soit le k-ième candidat qui soit choisi. Ainsi, il faut que la densité instrumentale propose le candidat y et que celui-ci soit choisi puis accepté, mais il faut également prendre en compte tous les autres candidats

possibles pouvant être générés. De ce fait, on a

$$\begin{aligned}
\Pi(\mathrm{d}x)p_k &= \pi(x)\left(\int\dots\int q(x,y)W_y \min\left\{1, \frac{\pi(y)q(y,x)W_x}{\pi(x)q(x,y)W_y}\right\}\prod_{\substack{i=1\\i\neq k}}^{N} Q(x,\mathrm{d}y^{(i)})Q(y,\mathrm{d}x_*^{(i)})\right)\mathrm{d}x\mathrm{d}y \\
&= \left(\int\dots\int \min\left\{\pi(x)q(x,y)W_y, \pi(y)q(y,x)W_x\right\}\prod_{\substack{i=1\\i\neq k}}^{N} Q(x,\mathrm{d}y^{(i)})Q(y,\mathrm{d}x_*^{(i)})\right)\mathrm{d}x\mathrm{d}y \\
&= \left(\int\dots\int \min\left\{\frac{\pi(x)q(x,y)w(x,y)}{w(x,y)+\sum_{\substack{i=1\\i\neq k}}^{N} w(x,y^{(i)})}, \frac{\pi(y)q(y,x)w(y,x)}{w(y,x)+\sum_{\substack{i=1\\i\neq k}}^{N} w(y,x_*^{(i)})}\right\}\times\right. \\
&\qquad \left.\prod_{\substack{i=1\\i\neq k}}^{N} Q(x,\mathrm{d}y^{(i)})Q(y,\mathrm{d}x_*^{(i)})\right)\mathrm{d}x\mathrm{d}y,
\end{aligned}$$

ce qui est symétrique en x et y, impliquant que (1.3.10) l'est également.

Les N candidats n'ont pas à être générés à partir de la même distribution. Certaines généralisations de l'algorithme utilisent des distributions différentes pour chaque candidat, comme c'est le cas dans [**9**]. L'algorithme MTM laisse également une grande liberté par rapport au choix de la fonction de poids. En fait, la seule contrainte appliquée à celle-ci est qu'elle soit positive. Cela laisse place à beaucoup de choix disponibles. Tel que mentionné dans [**8**] puis [**2**], l'impact de la fonction de poids sur la performance du MTM a été peu étudié et le plus souvent, des choix généraux sont privilégiés. Dans la pratique, les choix du tableau 1.1 ont été proposés.

Tableau 1.1. Différentes fonctions de poids utilisées dans le MTM

$w(x,y)$
$\pi(y)$
$\pi(y)q(y,x)$
$\frac{\pi(y)}{q(x,y)}$
$\pi(y)\lvert x-y\rvert^{\alpha}$

Les notions et algorithmes présentés dans ce chapitre forment la base de la théorie sur laquelle les méthodes MCMC reposent. Bien entendu, seulement la surface de ce vaste domaine

a été effleurée. Néanmoins, les notions introduites ici devraient suffire à la compréhension des résultats présentés dans les prochains chapitres.

Chapitre 2

Distribution instrumentale en équilibre

Le principe d'une distribution instrumentale dite informative est d'y incorporer de l'information à propos de la cible Π afin d'améliorer l'échantillonnage. Le MALA est un exemple de distribution instrumentale informative. L'objectif de ce chapitre est de présenter une nouvelle classe de distributions instrumentales informatives et de détailler leurs propriétés. De plus l'impact de ces distributions, lorsqu'utilisées de paire avec les algorithmes présentés au chapitre 1 (RWM, MALA, MTM), est analysé. Les concepts présentés dans ce chapitre proviennent en grande partie de Zanella [**24**], où ce dernier introduit cette nouvelle classe de distributions dans un contexte où l'espace d'états est discret. Nous reprenons ici plusieurs idées de cet article et poursuivons la réflexion dans un contexte où l'espace d'états est continu.

2.1. Une nouvelle distribution instrumentale informative

Pour débuter ce chapitre, un cas spécifique est présenté afin de fournir un peu d'intuition sur ce nouveau type de distribution instrumentale. Supposons que l'on désire échantillonner d'une distribution, Π, dont la densité associée, π, est bornée. Dans le contexte du RWM, on possède une distribution instrumentale symétrique, $Q_\sigma(x, \cdot)$, qui échantillonne les candidats de manière aveugle autour de $x \in \mathcal{X}$, l'état actuel de la chaîne générée. L'indice σ indique que la distribution possède un paramètre d'échelle, σ. Comme c'est le cas dans la pratique, on suppose ici, et dans les chapitres ultérieurs, que $\mathcal{X} = \mathbb{R}^d$. Un choix raisonnable de Q_σ est donc la distribution instrumentale associée à la densité (1.3.4), c'est-à-dire la densité normale centrée en x où le paramètre d'échelle est l'écart-type, σ. De façon générale, on dénote la densité associée à Q_σ par q_σ. Étant donné que la distribution Q_σ est symétrique, cela signifie

que $q_\sigma(x,y) = q_\sigma(y,x)$ pour tout $x, y \in \mathcal{X}$. Quelques fois, lorsque la densité q_σ est centrée en x, elle peut être exprimée sous une forme utile lorsqu'un paramètre d'échelle est utilisé.

Remarque 2.1.1. Soit $q_\sigma(x,y)$, une densité centrée en x symétrique par rapport à x et y pour tout $x, y \in \mathcal{X}$ et où σ est un paramètre d'échelle. Pour certaines densités q_σ, il est alors possible d'écrire

$$q_\sigma(x,y) = \frac{1}{\sigma^d} r\left(\frac{y-x}{\sigma}\right), \tag{2.1.1}$$

où r est une densité bornée de paramètre d'échelle unitaire telle que $r(-z) = r(z)$. Cela est vrai, entre autres, pour les lois normale, Student et uniforme continue. Ce n'est pas vrai en général pour toute densité centrée en x possédant un paramètre d'échelle σ.

Dorénavant, lorsqu'on utilisera une densité instrumentale symétrique, il s'agira d'une densité pouvant s'écrire sous la forme (2.1.1). Dans ce chapitre, on s'intéressera aux cas où $\sigma \downarrow 0$ et $\sigma \uparrow \infty$ dans un contexte bien particulier. Les propositions suivantes seront donc utiles pour la suite des choses.

Proposition 2.1.2. *Soit Q_σ, une distribution instrumentale symétrique avec densité associée, q_σ. Si $f : \mathcal{X} \to [0,\infty)$ est une fonction bornée, alors, pour tout $x \in \mathcal{X}$, on a*

$$\int_{\mathcal{X}} f(y) Q_\sigma(x, \mathrm{d}y) \xrightarrow{\sigma \downarrow 0} \int_{\mathcal{X}} f(y) \delta_x(\mathrm{d}y),$$

où $\delta_x(\cdot)$ est la mesure de Dirac centrée en x. On dira dans ce cas que $Q_\sigma(x,\cdot)$ converge faiblement vers $\delta_x(\cdot)$.

Démonstration. Il suffit de démontrer qu'il y a convergence vers $f(x)$ puisque $\int_{\mathcal{X}} f(y)\delta_x(\mathrm{d}y) = f(x)$. La remarque 2.1.1 permet d'écrire

$$\int_{\mathcal{X}} f(y) Q_\sigma(x, \mathrm{d}y) = \int_{\mathcal{X}} f(y) q_\sigma(x,y) \mathrm{d}y = \int_{\mathcal{X}} \frac{f(y)}{\sigma^d} r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y.$$

Un changement de variable $u = (y-x)/\sigma$ donne

$$\int_{\mathcal{X}} \frac{f(y)}{\sigma^d} r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y = \int_{\mathcal{X}} f(u\sigma + x) r(u) \mathrm{d}u.$$

La fonction f est bornée. Ainsi, pour une constante $M < \infty$, on a $f(u\sigma+x)r(u) \leqslant Mr(u)$, qui est intégrable puisque la fonction r est une densité. Le théorème de la convergence dominée (voir annexe A.1) permet d'écrire

$$\int_{\mathcal{X}} f(u\sigma + x) r(u) \mathrm{d}u \xrightarrow{\sigma \downarrow 0} \int_{\mathcal{X}} f(x) r(u) \mathrm{d}u = f(x).$$

□

Proposition 2.1.3. *Soit Q_σ, une distribution instrumentale symétrique avec densité associée, q_σ. Soit $\{\mu_\sigma\}_{\sigma>0}$, la suite de mesures définies par $\mu_\sigma(x, A) = \sigma^d Q_\sigma(x, A)$, pour tout $A \subseteq \mathcal{X}$ mesurable et pour tout $x \in \mathcal{X}$. Alors, si $f : \mathcal{X} \to [0, \infty)$ est intégrable par rapport à la mesure de Lebesgue, on a, pour tout $x \in \mathcal{X}$:*

$$\int_{\mathcal{X}} f(y)\mu_\sigma(x, \mathrm{d}y) \xrightarrow{\sigma\uparrow\infty} \int_{\mathcal{X}} f(y)r(\mathbf{0})\mathrm{d}y,$$

où $\mathbf{0} = (0, \ldots, 0)^\top \in \mathcal{X}$.

Démonstration. Selon la remarque 2.1.1, il est possible d'écrire

$$\int_{\mathcal{X}} f(y)\mu_\sigma(x, \mathrm{d}y) = \int_{\mathcal{X}} f(y)\sigma^d q_\sigma(x, y)\mathrm{d}y = \int_{\mathcal{X}} f(y)r\left(\frac{y-x}{\sigma}\right)\mathrm{d}y.$$

La densité r est bornée et la fonction f est intégrable. Par le théorème de la convergence dominée, il est possible d'écrire

$$\int_{\mathcal{X}} f(y)r\left(\frac{y-x}{\sigma}\right)\mathrm{d}y \xrightarrow{\sigma\uparrow\infty} \int_{\mathcal{X}} f(y)r(\mathbf{0})\mathrm{d}y,$$

avec $r(\mathbf{0}) < \infty$ puisque la densité r est bornée. □

Ainsi, lorsque $\sigma \downarrow 0$, la distribution instrumentale Q_σ converge faiblement vers la mesure de Dirac ; tandis que lorsque $\sigma \uparrow \infty$, la mesure μ_σ converge vers une mesure proportionnelle à la mesure de Lebesgue. Ces deux limites seront grandement utiles dans ce chapitre et le chapitre suivant.

Rappelons que la distribution Q_σ échantillonne les candidats de manière aveugle. Pour amener cette distribution à échantillonner plus souvent vers des endroits de plus grande densité π, il est possible de modifier Q_σ en posant

$$Q_{\pi,\sigma}(x, \mathrm{d}y) = \frac{\pi(y)Q_\sigma(x, \mathrm{d}y)}{Z_\sigma(x)}. \tag{2.1.2}$$

Ici, $Z_\sigma(x)$ est une constante de normalisation telle que $Z_\sigma(x) = \int_{\mathcal{X}} \pi(z)Q_\sigma(x, \mathrm{d}z)$. La densité instrumentale $q_\sigma(x, y)$ est symétrique en x et y. Cela implique donc que pour tout $x, y \in \mathcal{X}$, on a $Q_\sigma(x, \mathrm{d}y)\mathrm{d}x = q_\sigma(x, y)\mathrm{d}y\mathrm{d}x = q_\sigma(y, x)\mathrm{d}x\mathrm{d}y = Q_\sigma(y, \mathrm{d}x)\mathrm{d}y$. Ainsi, il est possible d'écrire

$$\frac{Z_\sigma(x)Q_{\pi,\sigma}(x, \mathrm{d}y)\mathrm{d}x}{\pi(y)} = Q_\sigma(x, \mathrm{d}y)\mathrm{d}x = Q_\sigma(y, \mathrm{d}x)\mathrm{d}y = \frac{Z_\sigma(y)Q_{\pi,\sigma}(y, \mathrm{d}x)\mathrm{d}y}{\pi(x)},$$

ce qui implique que

$$\pi(x)Z_\sigma(x)Q_{\pi,\sigma}(x,\mathrm{d}y)\mathrm{d}x = \pi(y)Z_\sigma(y)Q_{\pi,\sigma}(y,\mathrm{d}x)\mathrm{d}y. \tag{2.1.3}$$

Supposons, pour l'instant, qu'il soit possible de tirer un échantillon de $Q_{\pi,\sigma}$. Ainsi, la chaîne générée par un MH qui utiliserait la distribution instrumentale $Q_{\pi,\sigma}$ sans étape d'acceptation/rejet serait réversible par rapport à une distribution $\Pi_\sigma(\mathrm{d}x) = \Pi(\mathrm{d}x)Z_\sigma(x)$. Ceci implique que seule la constante de normalisation, $Z_\sigma(x)$, empêche la chaîne d'être directement réversible par rapport à Π. En temps normal, la probabilité d'acceptation (1.3.1) à l'étape d'acceptation/rejet permet de rendre la chaîne générée par Q_σ réversible par rapport à Π.

Examinons maintenant ce qui se produit dans les deux cas extrêmes, c'est-à-dire lorsque $\sigma \uparrow \infty$ et lorsque $\sigma \downarrow 0$. Dans le premier cas, en multipliant par $\sigma^d/r(\mathbf{0})$ les deux côtés de l'égalité (2.1.3), nous pouvons affirmer que $Q_{\pi,\sigma}$ est réversible par rapport à

$$\Pi^*_\sigma(\mathrm{d}x) = \Pi(\mathrm{d}x)\frac{\sigma^d Z_\sigma(x)}{r(\mathbf{0})} = \Pi(\mathrm{d}x)\int_{\mathcal{X}} \frac{\pi(z)}{r(\mathbf{0})}\sigma^d Q_\sigma(x,\mathrm{d}z) = \Pi(\mathrm{d}x)\int_{\mathcal{X}} \frac{\pi(z)}{r(\mathbf{0})}\mu_\sigma(x,\mathrm{d}z).$$

En utilisant la proposition 2.1.3, le fait que la densité r est bornée et l'intégrabilité de π, la limite lorsque $\sigma \to \infty$ satisfait

$$\lim_{\sigma\to\infty}\int_{\mathcal{X}} \frac{\pi(z)}{r(\mathbf{0})}\mu_\sigma(x,\mathrm{d}z) = \int_{\mathcal{X}} \frac{\pi(z)}{r(\mathbf{0})}r(\mathbf{0})\mathrm{d}z = 1.$$

Il s'en suit donc que $\lim_{\sigma\to\infty}\Pi^*_\sigma(\mathrm{d}x) = \pi(x)\mathrm{d}x = \Pi(\mathrm{d}x)$. Pour de grandes valeurs de σ, la chaîne générée par $Q_{\pi,\sigma}$ sera donc approximativement réversible par rapport à Π, et ce, avant même l'étape d'acceptation/rejet. Ceci est une propriété très intéressante, puisque cela signifie que la probabilité d'acceptation (1.3.1) n'aura qu'une très faible correction à faire pour rendre la chaîne réversible par rapport à Π. Ceci implique que le taux d'acceptation des candidats y sera plus élevé dans ce cas. En effet, si l'on note $q_{\pi,\sigma}$ la densité associée à la distribution $Q_{\pi,\sigma}$, alors pour de grandes valeurs de σ et $\forall\, x, y \in \mathcal{X}$, nous aurons $\pi(y)q_{\pi,\sigma}(y,x) \approx \pi(x)q_{\pi,\sigma}(x,y)$, et, par conséquent, la probabilité d'acceptation (1.3.1) sera proche de 1. Cela favorise l'exploration de l'espace d'états et donc la rapidité de la convergence.

Dans le deuxième cas, en utilisant la proposition 2.1.2 et le fait que le densité π est bornée, il est possible d'écrire

$$\lim_{\sigma\to 0} Z_\sigma(x) = \lim_{\sigma\to 0} \int_{\mathcal{X}} \pi(z) Q_\sigma(x, \mathrm{d}z) = \int_{\mathcal{X}} \pi(z)\delta_x(\mathrm{d}z) = \pi(x).$$

Par conséquent, $\lim_{\sigma\to 0} \Pi_\sigma(\mathrm{d}x) = \pi(x)^2\mathrm{d}x$ et $Q_{\pi,\sigma}$ est réversible par rapport à $\pi(x)^2\mathrm{d}x$ lorsque $\sigma \to 0$. Puisque $\pi(x)^2\mathrm{d}x$ est potentiellement très différent de $\pi(x)\mathrm{d}x$, il va sans dire que la chaîne n'est pas réversible par rapport à Π. Il est, par contre, possible de modifier $Q_{\pi,\sigma}$ afin que cette propriété soit respectée. Par exemple, on peut poser

$$Q_{\sqrt{\pi},\sigma}(x, \mathrm{d}y) = \frac{\sqrt{\pi(y)} Q_\sigma(x, \mathrm{d}y)}{Z_\sigma(x)}.$$

Dans ce cas, par le même raisonnement qu'auparavant, on a que la chaîne générée par $Q_{\sqrt{\pi},\sigma}$ est réversible par rapport à $\sqrt{\pi(x)} Z_\sigma(x) \to \pi(x)$ lorsque $\sigma \downarrow 0$. Ainsi, il est possible de modifier les distributions Q_σ en ajoutant de l'information à propos de π afin d'obtenir des propriétés intéressantes. Il reste à étudier la possibilité d'échantillonner d'une telle distribution en pratique. Ce problème sera traité à la section 2.3, mais tout d'abord, nous définissons les nouvelles distributions instrumentales de manière plus formelle.

2.2. Distribution instrumentale en équilibre

Tel qu'il a été vu à la section précédente, il est possible de modifier une distribution instrumentale non-informative afin de la rendre informative. Pour ce faire, il suffit de multiplier cette distribution non-informative par une fonction de π. Cette idée permet de définir une nouvelle classe de distributions instrumentales de la manière suivante.

Définition 2.2.1. Soit X, une chaîne de Markov à espace d'états $\mathcal{X}$ et soit Π, sa distribution stationnaire avec densité bornée, π. Pour une chaîne à l'état $x \in \mathcal{X}$, soit $Q_\sigma(x, \cdot)$ une distribution instrumentale symétrique centrée en x paramétrisée par un paramètre d'échelle σ. Soit $g : \mathcal{X} \times \mathcal{X} \to [0, \infty)$, une fonction continue bornée. La distribution $Q_{g,\sigma}(x, \cdot)$ est appelée une *distribution instrumentale biaisée* si pour tout $y \in \mathcal{X}$, on a

$$Q_{g,\sigma}(x, \mathrm{d}y) = \frac{g(x, y)\, Q_\sigma(x, \mathrm{d}y)}{Z_\sigma(x)}, \tag{2.2.1}$$

où la fonction $Z_\sigma(x)$ est une constante de normalisation telle que

$$Z_\sigma(x) = \int_{\mathcal{X}} g(x, z)\, Q_\sigma(x, \mathrm{d}z). \tag{2.2.2}$$

La constante de normalisation est assurément bornée puisque la fonction g est bornée. En effet, cela implique que pour une constante $M < \infty$ et pour tout $x, z \in \mathcal{X}$, la constante de normalisation est bornée par $\int_{\mathcal{X}} g(x,z) Q_\sigma(x, \mathrm{d}z) \leqslant \int_{\mathcal{X}} M Q_\sigma(x, \mathrm{d}z) = M$. La fonction g est générique, mais les cas intéressants seront ceux où celle-ci sera fonction de π. La fonction g permet alors de biaiser la distribution instrumentale non-informative Q_σ en ajoutant de l'information sur π. Ainsi, tout dépendant du choix de g, les distributions instrumentales biaisées peuvent devenir informatives. Par exemple, avec le choix $g(x,y) = \pi(y)/\pi(x)$, on obtient $Q_{g,\sigma}(x, \mathrm{d}y) \propto \pi(y) Q_\sigma(x, \mathrm{d}y)$, qui est la distribution $Q_{\pi,\sigma}$ de la section 2.1. Il se peut également que la distribution ne soit pas informative, comme lorsque $g(x,y) = 1$. Ce cas nous ramène au point de départ puisque $Q_{g,\sigma} = Q_\sigma$. Notons que pour la suite des choses, seules les fonctions g ne dépendant pas de σ seront considérées.

Dans un contexte pratique, il ne suffit pas de définir de nouvelles distributions instrumentales ; il faut bien sûr pouvoir tirer un échantillon de ces distributions biaisées. Évidemment, lorsque la fonction g dépend de $\pi(y)$, il n'est pas possible d'échantillonner à partir de $Q_{g,\sigma}$ directement puisque l'on cherche justement un algorithme pour réussir à échantillonner efficacement de π. Ce problème sera traité à la section 2.3. La prochaine difficulté concerne le choix de g. Tel qu'il a été vu à la section 2.1, certains choix de g semblent conduire à des distributions ayant des propriétés plus intéressantes que d'autres. Pour discriminer entre les différentes fonctions possibles, la propriété suivante est définie.

Définition 2.2.2. (Definition 1 dans Zanella [**24**]) Soit X, une chaîne de Markov ayant comme famille de noyaux de transition $\{Q_\sigma\}_{\sigma>0}$. On dit que $\{Q_\sigma\}_{\sigma>0}$ est en *équilibre local* par rapport à une distribution Π si pour chaque Q_σ, la chaîne X est réversible par rapport à une distribution Π_σ convergente en distribution vers Π lorsque $\sigma \downarrow 0$. La famille est dite en *équilibre global* si Π_σ converge en distribution vers Π lorsque $\sigma \uparrow \infty$.

La définition 1 dans Zanella [**24**] ne contient pas l'équilibre global. Bien que le cas global soit brièvement mentionné dans l'article, celui-ci ne s'intéresse qu'au cas local en général. Dans l'objectif de développer les idées présentées par ce dernier, nous considérons le cas global plus en détails dans ce chapitre. Dans la section 2.1, la distribution $Q_{\pi,\sigma}$ est en équilibre global, tandis que $Q_{\sqrt{\pi},\sigma}$ est en équilibre local. Tel que mentionné dans cet exemple, l'utilisation de distributions instrumentales en équilibre local ou global dans le

cadre du MH comporte des avantages. En effet, sous le bon régime, soit de petites valeurs de σ dans le cas local et de grandes valeurs de σ dans le cas global, le taux d'acceptation du MH sera élevé. Cela favorise l'exploration de l'espace d'états et mène à un algorithme plus efficace. Ainsi, la prochaine étape consiste à déterminer quelles fonctions g permettent d'obtenir les propriétés d'équilibre local et global.

En pratique, quelques précisions s'imposent relativement à l'utilisation de distributions en équilibre. Le paramètre d'échelle, σ, détermine l'amplitude des pas entre les éléments de la chaîne générée. Il a été mentionné précédemment que des valeurs trop petites de σ résultent en un algorithme inefficace. Or, la propriété d'équilibre local ne se manifeste justement que pour de petites valeurs de σ. Ces deux aspects peuvent sembler paradoxaux, mais rappelons qu'il faut également tenir compte de la dimension d de la distribution cible. En effet, plus la dimension augmente, plus la valeur de σ doit diminuer sous peine d'obtenir un taux d'acceptation minuscule. De ce fait, les distributions instrumentales en équilibre local devraient être efficaces en grande dimension. De la même manière, intuitivement, les distributions en équilibre global devraient être efficaces en petite dimension. Cet aspect sera exploré plus en détails à la section 2.4 et au chapitre 3.

2.2.1. Équilibre local

Afin de déterminer quelles fonctions g permettent d'obtenir la propriété d'équilibre local, l'intérêt est porté à une sous-classe des distributions instrumentales biaisées.

Définition 2.2.3. Soit $Q_{g,\sigma}$ une distribution instrumentale biaisée. Celle-ci est appelée une *distribution instrumentale ponctuelle* si g est de la forme

$$g(x, y) = h\left(\frac{\pi(y)}{\pi(x)}\right),$$

où $h : [0, \infty) \to [0, \infty)$ est une fonction continue bornée par une fonction linéaire, c'est-à-dire que $\exists\, a, b > 0$ tels que $h(t) \leq a + bt$, et ce, pour tout $t \geq 0$.

La condition sur la fonction h permet d'éviter des problèmes d'intégration. Dorénavant, une distribution instrumentale ponctuelle sera notée directement $Q_{h,\sigma}$, pour faire référence à la fonction h. Parmi ce sous-groupe de fonctions, certaines permettent d'obtenir la propriété d'équilibre local. Afin de pouvoir trouver quelles sont ces fonctions, il convient d'énoncer le

lemme technique suivant. Celui-ci permet de vérifier que si la fonction h est bornée linéairement, la limite de l'intégrale $\int_{\mathcal{X}} \pi(x) Z_\sigma(x) \mathrm{d}x$ est bien définie lorsque $\sigma \downarrow 0$.

Lemme 2.2.4. (Lemma 1 dans Zanella [**24**]) *Soit une fonction continue $h : [0, \infty) \to [0, \infty)$, avec $h(1) = 1$ et $h(t) \leq a + bt$ pour $a, b \geq 0$ et $\forall\, t \geq 0$. Soient une distribution, Π, associée à une densité bornée $\pi : \mathcal{X} \to [0, \infty)$, une distribution instrumentale symétrique, $Q_\sigma(x, \cdot)$, centrée en x, ainsi que la fonction $Z_\sigma(x)$ définie en* (2.2.2)*. Alors, il est possible d'écrire*

$$\lim_{\sigma \to 0} \int_{\mathcal{X}} \pi(x) Z_\sigma(x) \mathrm{d}x = 1.$$

La démonstration de ce lemme se retrouve en annexe A.1. Avec l'aide de celui-ci, il est maintenant possible de présenter le théorème identifiant les fonctions permettant d'obtenir la propriété d'équilibre local, tel que démontré dans Zanella [**24**]. La deuxième partie de la démonstration présentée ci-dessous diffère de celle de Zanella [**24**] étant donné que ce dernier se concentre sur les distributions à espace d'états discret alors que l'on considère les distributions à espace d'états continu.

Théorème 2.2.5. (Theorem 1 dans Zanella [**24**]) *Soit $Q_{h,\sigma}$, une distribution instrumentale ponctuelle. Celle-ci est en équilibre local par rapport à une distribution Π avec support $\mathcal{X}$ si et seulement si*

$$h\left(\frac{\pi(y)}{\pi(x)}\right) = \frac{\pi(y)}{\pi(x)} h\left(\frac{\pi(x)}{\pi(y)}\right) \quad \forall\, x, y \in \mathcal{X}. \tag{2.2.3}$$

Démonstration. ($\Leftarrow$) On suppose que la condition (2.2.3) est respectée et, sans perte de généralité, que $h(1) = 1$. Par la définition d'une distribution instrumentale ponctuelle et en utilisant la symétrie de Q_σ, il est possible d'écrire

$$\frac{Z_\sigma(x) Q_{h,\sigma}(x, \mathrm{d}y)\mathrm{d}x}{h\left(\frac{\pi(y)}{\pi(x)}\right)} = Q_\sigma(x, \mathrm{d}y)\mathrm{d}x = Q_\sigma(y, \mathrm{d}x)\mathrm{d}y = \frac{Z_\sigma(y) Q_{h,\sigma}(y, \mathrm{d}x)dy}{h\left(\frac{\pi(x)}{\pi(y)}\right)}.$$

En utilisant la condition (2.2.3), ceci implique que

$$\pi(x) Z_\sigma(x) Q_{h,\sigma}(x, \mathrm{d}y)\mathrm{d}x = \pi(y) Z_\sigma(y) Q_{h,\sigma}(y, \mathrm{d}x)\mathrm{d}y,$$

où

$$Z_\sigma(x) = \int_{\mathcal{X}} h\left(\frac{\pi(z)}{\pi(x)}\right) Q_\sigma(x, \mathrm{d}z).$$

Ainsi, $Q_{h,\sigma}$ est réversible par rapport à la distribution $\pi(x) Z_\sigma(x)\mathrm{d}x / \int_{\mathcal{X}} \pi(z) Z_\sigma(z)\mathrm{d}z$. L'intérêt est porté ici à la distribution limite pour $\sigma \downarrow 0$. Par le lemme 2.2.4, la limite du

dénominateur est bien définie et vaut 1. Pour le numérateur, il est possible d'utiliser la proposition 2.1.2 afin de calculer la limite de $Z_\sigma(x)$. En effet, la fonction h est bornée puisque la densité π l'est et h est continue. De plus, étant donné que la distribution Q_σ est symétrique, la proposition 2.1.2 permet d'écrire

$$\lim_{\sigma\to 0} Z_\sigma(x) = \lim_{\sigma\to 0} \int_{\mathcal{X}} h\left(\frac{\pi(z)}{\pi(x)}\right) Q_\sigma(x, \mathrm{d}z) = \int_{\mathcal{X}} h\left(\frac{\pi(z)}{\pi(x)}\right) \delta_x(\mathrm{d}z) = h(1) = 1,$$

et ce, pour tout $x \in \mathcal{X}$. Il suit que la densité $\pi(x)Z_\sigma(x)/\int_{\mathcal{X}} \pi(z)Z_\sigma(z)\mathrm{d}z$ converge ponctuellement vers $\pi(x)$, pour tout $x \in \mathcal{X}$. Par le lemme de Scheffé (voir annexe A.1), il y a donc également convergence en distribution vers $\pi(x)dx = \Pi(\mathrm{d}x)$.

($\Rightarrow$) On suppose que $Q_{h,\sigma}$ est en équilibre local par rapport à Π. Tel que vu à la section 2.1, il est possible d'écrire

$$h\left(\frac{\pi(x)}{\pi(y)}\right) Z_\sigma(x) Q_{h,\sigma}(x, \mathrm{d}y) = h\left(\frac{\pi(y)}{\pi(x)}\right) Z_\sigma(y) Q_{h,\sigma}(y, \mathrm{d}x).$$

En notant Q_h^* la distribution limite de $Q_{h,\sigma}$ lorsque $\sigma \downarrow 0$ et en prenant la limite des deux côtés, on obtient

$$h\left(\frac{\pi(x)}{\pi(y)}\right) h(1) Q_h^*(x, \mathrm{d}y) = h\left(\frac{\pi(y)}{\pi(x)}\right) h(1) Q_h^*(y, \mathrm{d}x).$$

De plus, la propriété d'équilibre locale étant respectée, il faut que

$$h\left(\frac{\pi(x)}{\pi(y)}\right) = \nu\pi(x) \qquad \text{et} \qquad h\left(\frac{\pi(y)}{\pi(x)}\right) = \nu\pi(y),$$

où ν est une constante de normalisation. Ainsi,

$$h\left(\frac{\pi(y)}{\pi(x)}\right) = \frac{\pi(y)}{\pi(x)} h\left(\frac{\pi(x)}{\pi(y)}\right).$$

□

Plusieurs fonctions respectent la condition (2.2.3). Le tableau 2.1 présente quelques exemples de fonctions, de même que les distributions instrumentales ponctuelles qui y sont associées. Il est intéressant de noter que la fonction $h(t) = \min\{1, t\}$ correspond à la probabilité d'acceptation de l'algorithme RWM. Plus généralement, toutes les fonctions du tableau 2.1 respectent la condition (1.3.5) et sont donc toutes des probabilités d'acceptation valides si elles sont bornées par 1. Ce lien est particulièrement intéressant puisqu'il existe parmi les différentes probabilités d'acceptation une probabilité optimale, soit la forme (1.3.1), utilisée dans l'algorithme Metropolis-Hastings. La question se pose donc de savoir s'il est possible

Tableau 2.1. Différentes fonctions h et distributions instrumentales ponctuelles associées pour $t = \pi(y)/\pi(x)$.

$h(t)$	$Q_{h,\sigma}(x, dy) \propto$
$\sqrt{t}$	$\sqrt{\pi(y)}Q_\sigma(x, dy)$
$\frac{t}{1+t}$	$\frac{\pi(y)}{\pi(y)+\pi(x)}Q_\sigma(x, dy)$
$\min\{1, t\}$	$\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} Q_\sigma(x, dy)$
$\max\{1, t\}$	$\max\left\{1, \frac{\pi(y)}{\pi(x)}\right\} Q_\sigma(x, dy)$

de trouver une fonction h qui résulte en un algorithme optimal. Malheureusement, il n'a pas été possible de trouver une telle fonction. Par contre, en pratique, le choix de cette fonction n'a pas beaucoup d'impact tel qu'il sera vu plus loin.

Jusqu'à présent, les distributions biaisées ont été créées en utilisant une distribution instrumentale de base Q_σ qui était symétrique. D'ailleurs, Zanella [**24**] ne se concentre que sur celles-ci. Afin de généraliser l'utilisation des distributions instrumentales en équilibre local, il est possible d'utiliser une fonction g différente afin de pouvoir obtenir cette propriété sans devoir passer par l'utilisation d'une distribution symétrique. Ici, une distribution non symétrique est telle que $q_\sigma(x, y) \neq q_\sigma(y, x)$. Par exemple, le MALA utilise une distribution non symétrique, puisque la densité normale utilisée n'est pas centrée en x. C'est dans ce contexte que la proposition suivante peut être utile.

Proposition 2.2.6. *Soit $Q_{g,\sigma}$ une distribution instrumentale biaisée telle que Q_σ est non symétrique. Alors, cette distribution est en équilibre local par rapport à Π si*

$$g(x, y) = h\left(\frac{\pi(y)}{\pi(x)}\right) s(x, y),$$

où la fonction h respecte les conditions d'une distribution ponctuelle ainsi que la condition (2.2.3) *et $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est telle que $s(x, y)q_\sigma(x, y)$ est symétrique pour tout $x, y \in \mathcal{X}$.*

Démonstration. En utilisant la fonction g, la distribution $Q_{g,\sigma}$ devient

$$Q_{g,\sigma}(x, dy) = \frac{h\left(\frac{\pi(y)}{\pi(x)}\right) s(x, y)Q_\sigma(x, dy)}{Z_\sigma(x)} = \frac{h\left(\frac{\pi(y)}{\pi(x)}\right) Q^*_\sigma(x, dy)}{Z_\sigma(x)},$$

où $Q^*_\sigma(x, dy)$ est maintenant symétrique en x, y et ce pour tout $x, y \in \mathcal{X}$. Puisque Q^*_σ est symétrique, $Q_{g,\sigma}$ respecte la définition d'une distribution instrumentale ponctuelle. De plus, la fonction h respecte la condition (2.2.3) et la propriété d'équilibre local suit du théorème 2.2.5. □

2.2.2. Équilibre global

Nous nous concentrons maintenant sur le régime global. Tout comme pour le cas local, on désire identifier les fonctions g permettant d'obtenir la propriété d'équilibre global. Tel qu'il a été vu à la section 2.1, la fonction $g(x, y) = \pi(y)$ permet d'obtenir cette propriété. Ce cas est le seul cas global considéré dans Zanella [**24**]. Dans l'objectif de développer les idées présentées par celui-ci, le théorème suivant permet d'identifier quels sont les autres cas possibles. La démonstration est donc semblable à celle du théorème 2.2.5, sauf que l'on s'intéresse ici à la limite lorsque $\sigma \uparrow \infty$, soit le régime global. Pour obtenir l'équilibre local, nous nous sommes intéressés à une sous-classe des fonctions g (voir Définition 2.2.3). De manière similaire, pour la recherche de l'équilibre global, on ne s'intéressera uniquement qu'aux fonctions où $\int_{\mathcal{X}} g(x, z)\mathrm{d}z < \infty$ pour tout $x \in \mathcal{X}$.

Théorème 2.2.7. *Soit $Q_{g,\sigma}$ une distribution instrumentale biaisée où $\int_{\mathcal{X}} g(x, z)\mathrm{d}z < \infty$ pour tout $x \in \mathcal{X}$. Celle-ci est en équilibre global par rapport à une distribution Π avec support $\mathcal{X}$ si et seulement si*

$$g(x, y) \propto \pi(y), \quad \forall\, x, y \in \mathcal{X}.$$

Démonstration. ($\Leftarrow$) On suppose que $g(x, y) \propto \pi(y)$. En utilisant la définition 2.2.1 et la symétrie de Q_σ, on observe que

$$\frac{Z_\sigma(x)Q_{g,\sigma}(x, \mathrm{d}y)\mathrm{d}x}{g(x, y)} = Q_\sigma(x, \mathrm{d}y)\mathrm{d}x = Q_\sigma(y, \mathrm{d}x)\mathrm{d}y = \frac{Z_\sigma(y)Q_{g,\sigma}(y, \mathrm{d}x)\mathrm{d}y}{g(y, x)}.$$

La fonction $g(x, y)$ est proportionnelle à $\pi(y)$, ce qui implique que

$$\pi(x)Z_\sigma(x)Q_{g,\sigma}(x, \mathrm{d}y)\mathrm{d}x = \pi(y)Z_\sigma(y)Q_{g,\sigma}(y, \mathrm{d}x)\mathrm{d}y.$$

L'intérêt est porté à la limite lorsque $\sigma \uparrow \infty$. Afin de pouvoir utiliser la proposition 2.1.3, il est nécessaire d'inclure un facteur σ^d de chaque côté de l'égalité précédente. Ceci mène à

une fonction $Z_\sigma(x)$ modifiée ; on note

$$Z^*_\sigma(x) = \sigma^d Z_\sigma(x) = \int_{\mathcal{X}} g(x,z)\sigma^d Q_\sigma(x,\mathrm{d}z) = \int_{\mathcal{X}} g(x,z)\mu_\sigma(x,\mathrm{d}z),$$

où μ_σ est la mesure définie dans la proposition 2.1.3. La distribution $Q_{g,\sigma}$ est alors réversible par rapport à la distribution

$$\frac{\pi(x)Z_\sigma(x)\mathrm{d}x}{\int_{\mathcal{X}} \pi(z)Z_\sigma(z)\mathrm{d}z} = \frac{\sigma^d\pi(x)Z_\sigma(x)\mathrm{d}x}{\sigma^d\int_{\mathcal{X}} \pi(z)Z_\sigma(z)\mathrm{d}z} = \frac{\pi(x)Z^*_\sigma(x)\mathrm{d}x}{\int_{\mathcal{X}} \pi(z)Z^*_\sigma(z)\mathrm{d}z}. \tag{2.2.4}$$

On désire montrer que cette distribution converge vers Π lorsque $\sigma \uparrow \infty$. On considère le numérateur en premier. Puisque la fonction $g(x,z)$ est proportionnelle à $\pi(z)$, elle est intégrable par rapport à z et la proposition 2.1.3 permet alors d'écrire, pour une constante $0 < c < \infty$,

$$\lim_{\sigma\to\infty} Z^*_\sigma(x) = \lim_{\sigma\to\infty}\int_{\mathcal{X}} g(x,z)\mu_\sigma(x,\mathrm{d}z) = \lim_{\sigma\to\infty}\int_{\mathcal{X}} c\pi(z)\mu_\sigma(x,\mathrm{d}z) = \int_{\mathcal{X}} c\pi(z)r(\mathbf{0})\mathrm{d}z = cr(\mathbf{0}).$$

Le numérateur de (2.2.4) converge donc ponctuellement vers $cr(\mathbf{0})\pi(x)$, où la densité r est celle introduite dans la remarque 2.1.1. Pour déterminer la limite du dénominateur, il est possible d'utiliser le théorème de la convergence dominée (voir annexe A.1). En utilisant le fait que μ_σ dépend de la densité r, qui est bornée par une constante $M < \infty$, on a

$$Z^*_\sigma(x) = \int_{z\in\mathcal{X}} g(x,z)\mu_\sigma(x,\mathrm{d}z) = \int_{z\in\mathcal{X}} c\pi(z)r\left(\frac{z-x}{\sigma}\right)\mathrm{d}z \leqslant \int_{z\in\mathcal{X}} c\pi(z)M\mathrm{d}z = cM.$$

De ce fait, l'intégrande du dénominateur en (2.2.4) est borné par $\pi(x)Z^*_\sigma(x) \leqslant \pi(x)cM$, qui est intégrable et indépendant de σ. Puisque $Z^*_\sigma(x) \to cr(\mathbf{0})$ à mesure que $\sigma \uparrow \infty$, il est possible d'écrire, par le théorème de la convergence dominée,

$$\lim_{\sigma\to\infty}\int_{\mathcal{X}} \pi(z)Z^*_\sigma(z)\mathrm{d}z = \int_{\mathcal{X}} \pi(z)cr(\mathbf{0})\mathrm{d}z = cr(\mathbf{0}).$$

On a donc que la densité associée à la distribution (2.2.4) converge ponctuellement vers $\pi(x)$. Selon le lemme de Scheffé, la distribution (2.2.4) converge alors vers $\pi(x)\mathrm{d}x = \Pi(\mathrm{d}x)$.

($\Rightarrow$) On suppose que $Q_{g,\sigma}$ est en équilibre global par rapport à Π. Comme auparavant, il est possible d'écrire

$$g(y,x)Z^*_\sigma(x)Q_{g,\sigma}(x,\mathrm{d}y)\mathrm{d}x = g(x,y)Z^*_\sigma(y)Q_{g,\sigma}(y,\mathrm{d}x)\mathrm{d}y. \tag{2.2.5}$$

Étant donné que la fonction $g(x,z)$ est intégrable par rapport à z pour tout $x \in \mathcal{X}$, on a que $Z^*_\sigma(x) \to \int_{\mathcal{X}} g(x,z)r(\mathbf{0})\mathrm{d}z$ lorsque $\sigma \uparrow \infty$ par la proposition 2.1.3. En notant Q^*_g la

distribution limite de $Q_{g,\sigma}$ et en prenant la limite des deux côtés de (2.2.5) lorsque $\sigma \uparrow \infty$, on obtient

$$\left(g(y,x)\int_{\mathcal{X}} g(x,z)r(\mathbf{0})dz\right)Q_g^*(x,\mathrm{d}y)\mathrm{d}x = \left(g(x,y)\int_{\mathcal{X}} g(y,z)r(\mathbf{0})\mathrm{d}z\right)Q_g^*(y,\mathrm{d}x)\mathrm{d}y.$$

L'intégrale $\int_{\mathcal{X}} g(x,z)r(\mathbf{0})\mathrm{d}z = k(x)$ n'est fonction que de x. De plus, la propriété d'équilibre globale implique, pour une constante ν, que

$$g(y,x)k(x) = \nu\pi(x) \Rightarrow g(y,x) = \frac{\nu\pi(x)}{k(x)},$$

où $k(x) \neq 0$, puisque dans le cas contraire, la propriété d'équilibre global ne serait pas respectée. Ainsi, $g(y,x)$ n'est fonction que de x, ce qui signifie que $k(x) = \int_{\mathcal{X}} g(x,z)\mathrm{d}z$ est une fonction constante. On a directement que $g(y,x) \propto \pi(x)$ et ce, pour tout $x, y \in \mathcal{X}$. □

Tout comme pour le cas local, il est possible d'obtenir des distributions en équilibre global à l'aide de distributions Q_σ non-symétriques, tel qu'énoncé dans la proposition suivante.

Proposition 2.2.8. *Soit $Q_{g,\sigma}$, une distribution instrumentale biaisée telle que Q_σ est non-symétrique et soit $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ telle que $s(x,y)q_\sigma(x,y)$ est symétrique pour tout $x, y \in \mathcal{X}$. Alors, cette distribution est en équilibre global par rapport à Π si*

$$g(x,y) \propto \pi(y)s(x,y).$$

Démonstration. En utilisant la fonction g, la distribution $Q_{g,\sigma}$ devient, pour une constante $c > 0$,

$$Q_{g,\sigma}(x,\mathrm{d}y) = \frac{c\pi(y)s(x,y)Q_\sigma(x,\mathrm{d}y)}{Z_\sigma(x)} = \frac{c\pi(y)Q_\sigma^*(x,\mathrm{d}y)}{Z_\sigma(x)},$$

où $Q_\sigma^*(x,\mathrm{d}y)$ est maintenant symétrique en x, y et ce, pour tout $x, y \in \mathcal{X}$. Puisque Q_σ^* est symétrique, $Q_{g,\sigma}$ respecte la définition d'une distribution instrumentale biaisée et la propriété d'équilibre global suit du théorème 2.2.7. □

2.3. Utilisation des distributions en équilibre en pratique

Les distributions biaisées contiennent de l'information sur π grâce à la fonction g. Or, pour pouvoir les utiliser en pratique, il faut pouvoir échantillonner à partir de telles distributions. Que celles-ci soient en équilibre local ou global, elles dépendent de $\pi(y)$ et il n'est donc pas possible d'échantillonner directement de ces distributions. Pour pallier ce problème,

une approximation de $\pi(y)$ est calculée. En utilisant une approximation de Taylor du premier ordre, il est possible de retomber sur des distributions qui peuvent être échantillonnées directement. Cela permet donc d'utiliser les distributions en équilibre en pratique.

2.3.1. Approximations des distributions en équilibre local

Des fonctions permettant d'obtenir l'équilibre local ont été présentées au tableau 2.1. L'une de celles-ci est mentionnée dans la section 5 de Zanella [**24**], où ce dernier utilise comme exemple une distribution en équilibre local utilisant la fonction $h(t) = \sqrt{t}$ et mentionne qu'il est possible de l'approximer par le MALA, l'algorithme présenté à la section 1.3.2. Afin de pousser cette idée plus loin, on détermine vers quels algorithmes peuvent mener des distributions instrumentales en équilibre local utilisant les autres fonctions du tableau 2.1.

Proposition 2.3.1. *Soit $Q_{h,\sigma}$, une distribution en équilibre local qui biaise $Q_\sigma(x,\cdot)$, la distribution normale associée à la densité* (1.3.4)*. Alors, si $h(t) = \sqrt{t}$ ou $h(t) = t/(1+t)$, $Q_{h,\sigma}$ peut être approximée par la distribution associée à la densité du MALA* (1.3.8)*.*

La démonstration est présentée à l'annexe A.1. Ainsi, les fonctions $h(t) = \sqrt{t}$ et $h(t) = t/1 + t$ mènent, en pratique, au même algorithme. Cela rend caduc la question du choix de h si l'on considère uniquement ces deux fonctions. La connexion au MALA est particulièrement intéressante puisque cela signifie que la distribution instrumentale utilisée dans le MALA est un cas particulier des distributions instrumentales en équilibre local. De ce fait, ceci confirme l'efficacité de cet algorithme en grande dimension.

Concernant les deux autres fonctions présentées dans le tableau 2.1, soient $h(t) = \min\{1,t\}$ et $h(t) = \max\{1,t\}$, les résultats sont moins intéressants en ce qui concerne le côté pratique. En effet, en considérant ici la fonction $h(t) = \min\{1,t\}$ et en utilisant une approximation de Taylor du premier ordre comme dans la démonstration de la proposition 2.3.1, on observe

$$\begin{aligned} h\left(\frac{\pi(y)}{\pi(x)}\right) &= \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} \\ &= \exp\{\min\{0, \log(\pi(y)/\pi(x))\}\} \\ &= \exp\left\{\min\left\{0, \log(\pi(y)) - \log(\pi(x))\right\}\right\} \\ &\approx \exp\left\{\min\left\{0, \nabla \log(\pi(x))(y-x)\right\}\right\}. \end{aligned}$$

La fonction $\min\{0, \log(\pi(y)/\pi(x))\} = 0$ lorsque $\pi(y) \geq \pi(x)$. Ainsi, avec cette approximation et en utilisant la densité normale (1.3.4) pour q_σ, la distribution $Q_{h,\sigma}$ peut être approximée par

$$\begin{aligned}
Q_{h,\sigma}(x, \mathrm{d}y) &\propto h\left(\frac{\pi(y)}{\pi(x)}\right) q_\sigma(x, y)\mathrm{d}y \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x) + \min\left\{0, \log(\pi(y)/\pi(x))\right\}\right\}\mathrm{d}y \\
&\approx \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)\right\}\mathbb{1}_{(\pi(y)\geq\pi(x))}\mathrm{d}y \\
&\quad + \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x) + \nabla\log(\pi(x))(y-x)\right\}\mathbb{1}_{(\pi(y)<\pi(x))}\mathrm{d}y \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left\|y-x\right\|^2\right\}\mathbb{1}_{(\pi(y)\geq\pi(x))}\mathrm{d}y \\
&\quad + \exp\left\{-\frac{1}{2\sigma^2}\left\|y-x-\sigma^2\nabla\log(\pi(x))\right\|^2\right\}\mathbb{1}_{(\pi(y)<\pi(x))}\mathrm{d}y \\
&\propto w_1\mathbb{1}_{(\pi(y)\leq\pi(x))}Q_1(x, \mathrm{d}y) + w_2\mathbb{1}_{(\pi(y)>\pi(x))}Q_2(x, \mathrm{d}y), \qquad (2.3.1)
\end{aligned}$$

où $Q_1(x, \cdot)$ est la distribution instrumentale associée à la densité normale (1.3.4) et $Q_2(x, \cdot)$ est une distribution instrumentale informative prenant en compte le gradient de la cible. Ici $w_1, w_2 \in \mathbb{R}$ sont des poids quelconques. La distribution (2.3.1) peut être vue comme une distribution de mélange entre une distribution aveugle et une distribution informative qui utilise de l'information sur la cible π d'une manière très similaire au MALA (1.3.8). En effet, pour des distributions quelconques, $F_1(x), F_2(x), ..., F_n(x)$, une distribution de mélange s'écrit comme étant

$$F(x) = \sum_{i=1}^{n} w_i F_i(x),$$

avec les poids $w_1, ..., w_n$ tels que $w_i \geq 0$ et $\sum_{i=1}^n w_i = 1$. Pour générer un candidat d'une distribution de mélange, on choisit une distribution parmi $F_1, \ldots, F_n$ proportionnellement aux poids $w_1, \ldots, w_n$ et l'on génère ensuite le candidat grâce à la distribution choisie.

Or, dans la situation présente, il n'est pas possible de générer un candidat directement de la distribution (2.3.1). En effet, celle-ci dépend encore de $\pi(y)$ à travers les fonctions indicatrices. Ainsi, même en utilisant une approximation de Taylor, le problème reste entier. Si l'on utilise le maximum au lieu du minimum pour la fonction h, le problème est le même. En effet, dans cette situation, il suffit simplement de changer la direction des inégalités dans

les fonctions indicatrices. Malgré tout, il est intéressant d'observer vers quelles distributions en pratique l'utilisation d'une distribution en équilibre local peut mener. L'utilisation du minimum et du maximum ne permettant pas d'obtenir des distributions instrumentales en équilibre utilisables en pratique, la question du choix optimal de h devient bel et bien caduque. Il s'avère que parmi les fonctions du tableau 2.1, seules les deux premières permettent d'obtenir des distributions instrumentales pouvant être échantillonnées et toutes deux mènent au MALA.

Il convient de mentionner que d'utiliser une approximation de Taylor du premier ordre pour la fonction $\pi(y)$ n'est pas la seule manière d'obtenir une distribution instrumentale en équilibre local en pratique. En effet, il serait possible d'utiliser une approximation d'ordre plus grand ou bien un autre type d'approximation. De plus, l'utilisation d'une distribution Q_σ autre que normale mènerait à différents algorithmes. Il serait même possible d'utiliser une densité non-symétrique en utilisant la forme de la fonction g trouvée à la proposition 2.2.6. Ces approches n'ont pas été considérées ici, mais elles témoignent de la diversité de la classe des distributions en équilibre local.

2.3.2. Approximation des distributions en équilibre global

La classe des fonctions permettant d'obtenir la propriété d'équilibre global est beaucoup plus restreinte que son penchant local. De ce fait, toutes les fonctions g permettant d'obtenir l'équilibre global mènent au même algorithme en pratique.

Proposition 2.3.2. *Soit $Q_{g,\sigma}$, une distribution en équilibre global qui biaise la distribution normale $Q_\sigma(x,\cdot)$ associée à la densité* (1.3.4)*. Alors, $Q_{g,\sigma}$ peut être approximée par la distribution associée à la densité*

$$q_{g,\sigma}(x,y) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{1}{2\sigma^2}\left\|y - x - \sigma^2 \nabla \log\{\pi(x)\}\right\|^2\right\}. \tag{2.3.2}$$

À nouveau, la démonstration, très similaire à celle de la proposition 2.3.1, est présentée en annexe A.1. Cette distribution est très semblable à celle utilisée dans le MALA, sauf qu'il n'y a pas de facteur $1/2$ devant le terme contenant le gradient. Cela peut sembler n'être qu'un changement très mineur mais, tel qu'il sera vu plus loin à l'aide de diverses simulations, cela peut améliorer grandement l'efficacité par rapport au MALA dans certains cas. Le fait que les distributions en équilibre local et global ne diffèrent que par un facteur de $1/2$ devant le

gradient pointe vers une nouvelle distribution instrumentale consistant en une interpolation entre les cas local et global. En effet, en posant $\gamma \in [1, 2]$, la distribution associée à la densité

$$q_\sigma(x, y) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{1}{2\sigma^2}\left\|y - x - \frac{\gamma\sigma^2}{2}\nabla \log\{\pi(x)\}\right\|^2\right\}, \tag{2.3.3}$$

respecte cette idée. Le facteur γ est appelé le degré d'interpolation. En fonction de la valeur de γ, la distribution sera plus près d'être en équilibre local ou global. On note ici que si $\gamma = 0$, on retombe sur l'algorithme RWM traditionnel, tandis que si $\gamma = 1$, il s'agit de la densité instrumentale du MALA. L'ajout d'un paramètre γ permet une plus grande versatilité. En effet, auparavant, la taille des pas de l'algorithme n'était déterminée que par la grandeur de la variance instrumentale, σ^2. Avec la distribution instrumentale associée à la densité (2.3.3), les paramètres γ et σ^2 peuvent maintenant être ajustés séparément. Tout comme il est d'intérêt de trouver une valeur de σ^2 optimale, il faut choisir une valeur de γ appropriée. Cette question est considérée au prochain chapitre.

2.4. Performance asymptotique du régime global

Les approximations présentées à la section précédente rendent possible, en pratique, l'utilisation des distributions en équilibre. Celles en équilibre local mènent au MALA, tandis que celles en équilibre global mènent à une légère modification du MALA. Il a été argumenté précédemment que le régime local est pertinent en grande dimension puisque la variance instrumentale tend vers 0 dans cette situation. Intuitivement, les distributions en équilibre local devraient donc être efficaces en grande dimension. À l'opposé, les distributions en équilibre global ne devraient pas être efficaces dans ce cas puisque celles-ci sont pertinentes lorsque $\sigma^2 \to \infty$. Afin de pouvoir comparer l'efficacité asymptotique des régimes local et global, l'accent est mis sur la mise à l'échelle.

Comme mentionné auparavant, la variance instrumentale est une fonction de d et prend la forme $\sigma_d^2 \propto d^{-\beta}$. Le paramètre β est appelé la mise à l'échelle. Ce dernier régule la vitesse à laquelle la variance doit diminuer en fonction de d. Une petite valeur de ce paramètre est préférable puisque dans ce cas, la variance diminuera moins rapidement en fonction de d et l'algorithme fera de plus grands pas à chaque itération. En fait, la mise à l'échelle est directement reliée à la vitesse d'exploration de l'algorithme. En effet, lorsque la chaîne atteint

la stationnarité, le nombre de pas nécessaires pour explorer l'espace d'états est $\mathcal{O}(d^\beta)$, ce qui confirme qu'une petite valeur de β est préférable. Par contre, si la valeur de β est trop petite, la variance instrumentale sera trop grande par rapport à la dimension et le taux d'acceptation risque de s'effondrer vers 0 à mesure que $d \uparrow \infty$. Ainsi, tel que mentionné dans [**1**], pour une distribution instrumentale particulière, le bon choix de β sera

$$\beta_0 = \min_{\beta_c \geq 0} \left\{ \beta_c : \lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] > 0 \quad \forall \beta \in [\beta_c, \infty) \right\}. \tag{2.4.1}$$

Notons qu'en écrivant $\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)]$, on désire souligner que la dimension de X et Y tend vers l'infini. Ici, on suppose que la stationnarité est atteinte et donc que l'espérance est prise par rapport à la distribution conjointe de X (distribué selon π) et Y (choisi selon la distribution instrumentale). Ainsi, la variance instrumentale sera la plus grande possible en fonction de la dimension tout en s'assurant que le taux d'acceptation moyen reste supérieur à 0 pour tout d. Rappelons que pour des densités cibles iid, la mise à l'échelle du MALA est $\beta_0 = 1/3$ tandis que celle du RWM est $\beta_0 = 1$. L'utilisation de distributions en équilibre local est donc asymptotiquement plus efficace que le RWM puisqu'en grandes dimensions, le MALA peut effectuer des pas d'ordre plus grand que le RWM. On désire déterminer quelle est la mise à l'échelle β_0 requise lorsque l'on utilise une valeur de $\gamma > 1$ dans la densité (2.3.3). Une démonstration du calcul de la mise à l'échelle du RWM et du MALA a été présentée dans Beskos & Stuart [**1**] (voir Theorem 1). Nous nous inspirons de ces travaux en les adaptant pour la distribution instrumentale (2.3.3) utilisant un paramètre $\gamma > 1$. De plus, afin de pouvoir comparer ce résultat avec ceux déjà obtenus pour le MALA et le RWM, on considère seulement une densité cible iid. Le lemme suivant, présenté dans Beskos & Stuart[**1**], est utile pour la démonstration du théorème qui suit.

Lemme 2.4.1. (Lemma 1 dans Beskos & Stuart [**1**]) *Soit $T \in \mathbb{R}$, une variable aléatoire. Alors, en définissant $x \wedge y := \min\{x, y\}$, on a*

(i) Pour tout $c > 0$,

$$\mathbb{E}[1 \wedge e^T] \geqslant e^{-c} \left(1 - \frac{\mathbb{E}[|T|]}{c} \right). \tag{2.4.2}$$

(ii) Si $\mathbb{E}[T] < 0$, alors

$$\mathbb{E}[1 \wedge e^T] \leqslant e^{\mathbb{E}[T]/2} + \frac{2\mathbb{E}[|T - \mathbb{E}[T]|]}{(-\mathbb{E}[T])}. \tag{2.4.3}$$

Démonstration. Pour le premier résultat, on a

$$\mathbb{E}[1 \wedge e^T] \geqslant \mathbb{E}\left[\left(1 \wedge e^T\right) \mathbb{1}\left\{|T| \leq c\right\}\right] \geqslant e^{-c} P(|T| \leqslant c).$$

Par l'inégalité de Markov, il suit que

$$P(|T| \leqslant c) = (1 - P(|T| > c)) \geqslant \left(1 - \frac{\mathbb{E}[|T|]}{c}\right),$$

ce qui donne le résultat escompté. Pour le deuxième résultat, on observe que

$$\begin{aligned}
\mathbb{E}[1 \wedge e^T] &= \mathbb{E}\left[(1 \wedge e^T)\mathbb{1}\left\{|T - \mathbb{E}[T]| \leqslant \frac{-\mathbb{E}[T]}{2}\right\}\right] + \mathbb{E}\left[(1 \wedge e^T)\mathbb{1}\left\{|T - \mathbb{E}[T]| > \frac{-\mathbb{E}[T]}{2}\right\}\right] \\
&\leqslant \mathbb{E}\left[(1 \wedge e^T)\mathbb{1}\left\{T \leqslant \mathbb{E}[T] - \frac{\mathbb{E}[T]}{2}\right\}\right] + P\left(|T - \mathbb{E}[T]| > \frac{-\mathbb{E}[T]}{2}\right) \\
&\leqslant e^{\mathbb{E}[T]/2} + \frac{2\mathbb{E}[|T - \mathbb{E}[T]|]}{-\mathbb{E}[T]},
\end{aligned}$$

en utilisant à nouveau l'inégalité de Markov pour la dernière inégalité. □

Théorème 2.4.2. *Soit une chaîne de Markov à espace d'états $\mathcal{X}$ générée par un algorithme MH avec densité instrumentale q_σ (2.3.3) utilisant $\gamma \in (1, 2]$. Soit une densité cible π avec composantes iid de telle sorte que $\pi(x) = \prod_{i=1}^{d} f(x_i) = \prod_{i=1}^{d} \exp(l(x_i))$ où $l(x) = \log(f(x))$. On suppose que la chaîne a atteint la stationnarité. De plus, les conditions suivantes sont apposées sur f :*

(1) tous les moments de f sont bornés ;

(2) l fait partie de la classe C^∞ ;

(3) l et toutes ses dérivées sont bornées par un polynôme, c'est-à-dire

$$|l(x)|, |l^{(i)}(x)| \leqslant M(x), \quad i \geqslant 1,$$

où $M(x)$ est un polynôme positif.

Alors, $\beta_0 = 1$.

Démonstration. Afin de garder cette démonstration succincte, certains calculs de routine se retrouvent dans l'annexe A.1, notamment les développements de Taylor. Soit X, l'élément actuel de la chaîne à l'état x. On a que $X \sim \pi$ puisque la chaîne a atteint la stationnarité.

Le candidat y est accepté avec probabilité $\alpha(x,y) = 1 \wedge e^{R_d}$. Étant donné que les densités π et q_σ ont des composantes iid, R_d peut être exprimé sous la forme

$$R_d = \log\left(\frac{\pi(y)q_\sigma(y,x)}{\pi(x)q_\sigma(x,y)}\right) = \sum_{i=1}^{d} \log\left(\frac{f(y_i)q_\sigma^*(y_i,x_i)}{f(x_i)q_\sigma^*(x_i,y_i)}\right),$$

où q_σ^* est la version unidimensionnelle de la densité instrumentale (2.3.3). Plus précisément, la composante i du candidat Y est générée selon

$$Y_i = x_i + \frac{\gamma\sigma_d^2}{2} l'(x_i) + \sigma_d Z_i,$$

avec $Z_i \sim \mathcal{N}(0,1)$ indépendant de $X_i \sim f$, $i = 1,\ldots,d$ et $\gamma \in (1,2]$. Puisque Y_i dépend de σ_d, on peut considérer R_d comme fonction de σ_d. Grâce à l'analyse du développement de Taylor de R_d, il sera possible de déterminer pour quelles valeurs de la mise l'échelle β la limite du taux d'acceptation moyen $\mathbb{E}[\alpha(X,Y)]$ est nulle ou supérieure à 0.

Cas 1 : $\sigma_d^2 \propto d^{-\beta}$ où $\beta \geqslant 1$.

Ici, on désire démontrer que $\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] > 0$. Selon le lemme 2.4.1 (i), cela survient si $\lim_{d\to\infty} \mathbb{E}[|R_d|] < \infty$. Pour vérifier cela, on calcule le développement de Taylor du deuxième ordre de R_d centré en $\sigma_d = 0$. Ce développement donne

$$R_d = \mathcal{A}_{1,d} + \mathcal{A}_{2,d} + U_d, \tag{2.4.4}$$

avec

$$\mathcal{A}_{1,d} = \sigma_d R_d'(0) = \sigma_d \sum_{i=1}^{d} C_{1,i}, \qquad C_{1,i} = l'(x_i)Z_i(1-\gamma); \tag{2.4.5}$$

$$\mathcal{A}_{2,d} = \frac{\sigma_d^2}{2!} R_d''(0) = \frac{\sigma_d^2}{2!} \sum_{i=1}^{d} C_{2,i}, \qquad C_{2,i} = (1-\gamma)[l''(x_i)Z_i^2 + l'(x_i)^2\gamma]; \tag{2.4.6}$$

$$U_d = \frac{\sigma_d^3}{3!} R'''(\sigma^*) = \frac{\sigma_d^3}{3!} \sum_{i=1}^{d} U_{i,d}(x_i, Z_i, \sigma_i^*), \tag{2.4.7}$$

où $\sigma_i^* \in [0,\sigma_d]$, $i = 1,\ldots,d$. Notons que les termes $C_{1,i}$ et $C_{2,i}$ sont des polynômes qui dépendent de Z_i et des dérivées de l. Par les conditions (3) et (1) sur f et en raison de l'indépendance entre Z_i et X_i, leurs moments sont bornés. Comme $|R_d| \leqslant |\mathcal{A}_{1,d}| + |\mathcal{A}_{2,d}| + |U_d|$, il suffit de vérifier que la limite de l'espérance de chacun de ces termes est finie.

Débutons par $|U_d|$. Les termes $U_{i,d}(x_i, Z_i, \sigma_i^*)$, $i = 1, \ldots, d$ sont des polynômes qui sont fonction de Z_i, des dérivées de l et de puissances positives de σ_i^* (voir l'annexe A.1 pour l'expression détaillée). Cela signifie, en utilisant la condition (3) sur f, qu'il existe des polynômes positifs M_1, M_2, M_3 (voir [**1**]) tels que

$$|U_{i,d}(x_i, Z_i, \sigma_i^*)| \leqslant M_1(x_i)M_2(Z_i)M_3(\sigma_i^*).$$

En utilisant le fait que X_i et Z_i sont indépendants et par la condition (1), il s'avère que $\mathbb{E}[M_1(X_i)M_2(Z_i)] = \mathbb{E}[M_1(X_i)]\mathbb{E}[M_2(Z_i)] < \infty$ pour tout i. De plus, comme $\beta > 0$, il existe un $\varepsilon > 0$ tel que $\sigma_i^* < \sigma_d < \varepsilon$ pour tout i, donc $M_3(\sigma_i^*)$ est aussi borné par une constante. Ainsi, $\mathbb{E}[|U_{i,d}(X_i, Z_i, \sigma_i^*)|] < K < \infty$ pour une constante K indépendante de i et d. Puisque $\beta \geqslant 1$, la limite du terme résiduel vaut

$$\lim_{d\to\infty} \mathbb{E}[|U_d|] \leqslant \lim_{d\to\infty} \frac{\sigma_d^3}{3!} dK = 0.$$

En ce qui concerne $|\mathcal{A}_{1,d}|$, l'inégalité de Jensen permet d'écrire

$$\mathbb{E}[|\mathcal{A}_{1,d}|] = \mathbb{E}\left[\sqrt{\mathcal{A}_{1,d}^2}\right] \leqslant \mathbb{E}[\mathcal{A}_{1,d}^2]^{1/2} = \mathbb{E}\left[\left(\sigma_d \sum_{i=1}^{d} C_{1,i}\right)^2\right]^{1/2}.$$

Il est possible de simplifier ce dernier terme en utilisant le fait que les $C_{1,i}$, $i = 1, \ldots, d$ sont iid, car ils dépendent de X_i et de Z_i. De plus, on observe que $\mathbb{E}[C_{1,i}] = 0$. Ainsi, il est permis d'écrire

$$\mathbb{E}[|\mathcal{A}_{1,d}|] \leqslant \sigma_d \left(\sum_{i=1}^{d} \mathbb{E}[C_{1,i}^2] + \sum_{i=1}^{d}\sum_{j\neq i} \mathbb{E}[C_{1,i}C_{1,j}]\right)^{1/2} = \sigma_d\sqrt{d}\mathbb{E}[C_{1,1}^2]^{1/2}.$$

Puisque tous les moments de $C_{1,1}$ sont bornés et que $\beta \geqslant 1$, la limite de ce terme sera bornée.

Pour $|\mathcal{A}_{2,d}|$, l'inégalité du triangle puis l'inégalité de Jensen permettent d'écrire

$$\mathbb{E}[|\mathcal{A}_{2,d}|] \leqslant \frac{d\sigma_d^2}{2!}\mathbb{E}[|C_{2,1}|] \leqslant \frac{d\sigma_d^2}{2!}\mathbb{E}[C_{2,1}^2]^{1/2},$$

et la limite de $\mathbb{E}[|\mathcal{A}_{2,d}|]$ sera également bornée. On en déduit que $\lim_{d\to\infty} \mathbb{E}[|R_d|] < \infty$. Par le lemme 2.4.1 (i), on conclut que

$$\lim_{d\to\infty} \mathbb{E}[\alpha(X, Y)] > 0.$$

Cas 2 : $\sigma_d^2 \propto d^{-\beta}$ où $\beta \in (0,1)$.

On cherche à démontrer que $\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] = 0$. Selon le lemme 2.4.1 (ii), cela survient si

$$\mathbb{E}[R_d] \xrightarrow{d\uparrow\infty} -\infty \qquad \text{et} \qquad \frac{\mathbb{E}[|R_d - \mathbb{E}[R_d]|]}{-\mathbb{E}[R_d]} \xrightarrow{d\uparrow\infty} 0. \tag{2.4.8}$$

On se concentre d'abord sur $\mathbb{E}[R_d]$. On utilise une expansion de Taylor d'ordre m de R_d centrée en $\sigma_d = 0$, où $m \in \mathbb{N}$ est tel que $(m+1)\beta > 2$. Ceci donne

$$R_d = \sum_{j=1}^{m} \mathcal{A}_{j,d} + U_d^*,$$

où les termes sont les suivants

$$\mathcal{A}_{j,d} = \frac{\sigma_d^j}{j!} R_d^{(j)}(0) = \frac{\sigma_d^j}{j!} \sum_{i=1}^{d} C_{j,i}, \qquad U_d^* = \frac{\sigma_d^{m+1}}{(m+1)!} R_d^{(m+1)}(\sigma^*) = \frac{\sigma_d^{m+1}}{(m+1)!} \sum_{i=1}^{d} U_{i,d}^*(x_i, Z_i, \sigma_i^*),$$

avec $\sigma_i^* \in [0, \sigma_d]$. Les termes $C_{1,i}$ et $C_{2,i}$ sont identiques à (2.4.5) et (2.4.6), soit ceux obtenus lors du cas 1.

Tout comme pour le cas 1, notre attention est d'abord portée sur le terme résiduel U_d^*. En appliquant les mêmes arguments qu'au cas 1, il est possible de borner le terme résiduel $\mathbb{E}[|U_{i,d}^*(X_i, Z_i, \sigma_i^*)|]$ par une constante K_0 indépendante de d et de i. Le choix de m implique alors que

$$\lim_{d\to\infty} \mathbb{E}[U_d^*] \leqslant \lim_{d\to\infty} |\mathbb{E}[U_d^*]| \leqslant \lim_{d\to\infty} \mathbb{E}[|U_d^*|] \leqslant \lim_{d\to\infty} \frac{\sigma^{m+1}}{(m+1)!} d K_0 = 0$$

et $\mathbb{E}[U_d^*]$ est $\mathcal{O}(d^{1-\frac{(m+1)\beta}{2}})$. Pour les autres termes, puisque $\mathbb{E}[C_{1,1}] = 0$, on a que

$$\mathbb{E}[R_d] = \sum_{j=1}^{m} \mathbb{E}[\mathcal{A}_{j,d}] + \mathbb{E}[U_d^*] = \sum_{j=1}^{m} \frac{d\sigma_d^j}{j!} \mathbb{E}[C_{j,1}] + E[U_d^*] = \sum_{j=2}^{m} \frac{d\sigma_d^j}{j!} \mathbb{E}[C_{j,1}] + E[U_d^*].$$

Les termes $C_{j,1}$, $j = 1, \ldots, m$ sont tous des polynômes qui dépendent de Z_1 et des dérivées de l. De ce fait, de manière similaire à ce qui a été argumenté pour le cas 1, en raison de la condition (3), tous les moments de $C_{j,1}$ sont bornés. Le terme dominant de $\mathbb{E}[R_d]$ est donc le premier terme non nul du développement de Taylor, soit $\mathbb{E}[\mathcal{A}_{2,d}]$. Cette espérance est négative. En effet, comme $\mathbb{E}[l''(X)] = -\mathbb{E}[l'(X)^2]$ (voir lemme 7.3.11 de [**3**]) et en utilisant

la valeur de $C_{2,1}$ trouvée précédemment en (2.4.6), il est possible d'écrire

$$\begin{aligned}\mathbb{E}[C_{2,1}] &= (1-\gamma)\mathbb{E}[l''(X_1)Z_1^2 + l'(X_1)^2\gamma] \\ &= (1-\gamma)(\mathbb{E}[l''(X_1)] + \mathbb{E}[l'(X_1)^2]\gamma) \\ &= -(1-\gamma)^2\mathbb{E}[l'(X_1)^2] < 0.\end{aligned}$$

On en déduit que $\mathbb{E}[\mathcal{A}_{2,d}]$ est négative. Puisqu'il s'agit du terme dominant dans le développement de Taylor de R_d, cela signifie que $E[R_d] \to -\infty$ au rythme de $d^{1-\beta}$.

En ce qui concerne le ratio $\mathbb{E}[|R_d - \mathbb{E}[R_d]|]/-\mathbb{E}[R_d]$, par l'inégalité du triangle et l'inégalité de Jensen, il est possible de borner le numérateur par

$$\begin{aligned}\mathbb{E}[|R_d - \mathbb{E}[R_d]|] &\leqslant \sum_{j=1}^{m} \mathbb{E}[|\mathcal{A}_{j,d} - \mathbb{E}[\mathcal{A}_{j,d}]|] + \mathbb{E}[|U_d^* - \mathbb{E}[U_d^*]|] \\ &\leqslant \sum_{j=1}^{m} \mathbb{V}(\mathcal{A}_{j,d})^{1/2} + \mathbb{E}[|U_d^* - \mathbb{E}[U_d^*]| \\ &= \sum_{j=1}^{m} \mathbb{V}\left(\frac{\sigma_d^j}{j!}\sum_{i=1}^{d} C_{j,i}\right)^{1/2} + \mathbb{E}[|U_d^* - \mathbb{E}[U_d^*]| \\ &= \sum_{j=1}^{m} \frac{\sigma_d^j}{j!}\sqrt{d}\mathbb{V}(C_{j,1})^{1/2} + \mathbb{E}[|U_d^* - \mathbb{E}[U_d^*]|. \qquad (2.4.9)\end{aligned}$$

Le premier terme de cette somme étant dominant, le numérateur de (2.4.8) sera $\mathcal{O}(d^{\frac{1-\beta}{2}})$ tandis que le dénominateur sera $\mathcal{O}(d^{1-\beta})$, ce qui signifie que le ratio sera $\mathcal{O}(d^{\frac{\beta-1}{2}})$ et va tendre vers 0 lorsque $d \uparrow \infty$ puisque $\beta < 1$. Par le lemme 2.4.1 (ii), on conclut que

$$\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] = 0.$$

Puisque la valeur minimale de β permettant d'obtenir un taux d'acceptation asymptotique non nul est 1, on conclut que $\beta_0 = 1$. □

Ainsi, en grande dimension, un algorithme utilisant la distribution instrumentale associée à la densité (2.3.3) avec $\gamma > 1$ devra faire des pas de grandeur $\mathcal{O}(d^{-1})$ tandis que le MALA ($\gamma = 1$) pourra faire des pas de grandeur $\mathcal{O}(d^{-1/3})$. Dans le cas contraire, si l'on utilise une valeur de $\gamma > 1$ et des pas plus grands que $\mathcal{O}(d^{-1})$, le taux d'acceptation sera très faible. Ainsi, il est préférable d'utiliser le MALA ($\gamma = 1$) en grande dimension. Notons que d'utiliser une valeur de $\gamma > 1$ est aussi efficace que le RWM dans le cas asymptotique puisque les deux

algorithmes ont la même mise à l'échelle, $\beta = 1$. Le théorème 2.4.2 confirme donc l'intuition mentionnée précédemment voulant que les distributions en équilibre local soient plus efficaces que celles en équilibre global asymptotiquement. Il reste à évaluer la performance en faible dimension des distributions en équilibre global par rapport au cas local. Cet aspect sera exploré au prochain chapitre à l'aide d'un critère d'efficacité différent de la mise à l'échelle.

2.5. MTM en équilibre local

Le principe des distributions biaisées est d'incorporer de l'information sur Π dans les distributions instrumentales non-informatives à l'aide de la fonction g. Pourtant, il existe déjà un algorithme effectuant indirectement une telle étape, soit le MTM présenté au chapitre 1. Dans celui-ci, au lieu de générer un seul candidat par une distribution Q_σ quelconque comme lors du RWM, plusieurs candidats sont générés à l'aide de Q_σ lors d'une même itération. Par contre, un seul est sélectionné grâce à une fonction de poids. Celle-ci indique, pour chaque candidat proposé, si celui-ci a plus ou moins de chances d'être sélectionné par rapport aux autres candidats. Habituellement, cette fonction dépend de π. De ce fait, le MTM utilise l'information présente dans la fonction de poids afin non pas de générer les candidats comme pour une distribution biaisée, mais de bien les sélectionner. Ainsi, l'ajout d'information est réalisé après avoir proposé les candidats et ne fait donc pas partie de la distribution instrumentale. Toutefois, tel que mentionné dans la conclusion de Zanella [**24**], la distribution instrumentale du MTM ressemble beaucoup à une distribution instrumentale biaisée puisqu'il s'agit d'une fonction de poids générique multipliée par une distribution instrumentale Q_σ quelconque (voir 2.5.1). Les deux méthodes comportent donc des similarités et le lien entre celles-ci est donc exploré plus en détails ici.

Comme certaines fonctions permettent d'obtenir la propriété d'équilibre local pour des distributions biaisées, il serait intéressant de déterminer s'il est possible d'obtenir la même propriété avec un choix particulier de la fonction de poids dans le MTM. Tel que mentionné auparavant, la seule contrainte associée à cette fonction est qu'elle soit positive, ce qui laisse une grande marge de manœuvre. De plus, en pratique, il n'existe aucune ligne directrice permettant de faire un choix éclairé d'une fonction de poids. Ainsi, une fonction de poids permettant d'obtenir la propriété d'équilibre local aurait un avantage sur d'autres.

Malheureusement, le problème n'est pas aussi simple qu'il n'y paraît. En premier lieu, une condition nécessaire à la propriété d'équilibre locale sur la fonction de poids est présentée. Un exemple illustrera par la suite pourquoi cette condition n'est pas suffisante dans le cadre du MTM.

Tout d'abord, si les N candidats du MTM sont générés selon une distribution symétrique Q_σ et que I est l'indice du candidat choisi, alors la distribution instrumentale du MTM est

$$\begin{aligned} Q_{w,\sigma}(x, \mathrm{d}y) &= \mathbb{P}\left(\bigcup_{j=1}^{N} \{(Y^{(j)} \in \mathrm{d}y) \cap (I = j)\} \,|\, x\right) \\ &= N\mathbb{P}\left(\{(Y^{(N)} \in \mathrm{d}y) \cap (I = N)\} \,|\, x\right) \\ &= \frac{NQ_\sigma(x, \mathrm{d}y)w(x, y)}{Z_\sigma(x)} \int \ldots \int \frac{\prod_{j=1}^{N-1} Q_\sigma(x, \mathrm{d}y^{(j)})}{w(x, y) + \sum_{j=1}^{N-1} w(x, y^{(j)})}, \end{aligned} \tag{2.5.1}$$

où, sans perte de généralité, on considère N fois le cas où le dernier candidat est choisi. Notons que les points de référence $x_*^{(j)}$ n'interviennent pas ici puisqu'on ne considère que la distribution instrumentale, mais pas l'étape d'acceptation où ceux-ci surviennent. La constante de normalisation, $Z_\sigma(x)$, est

$$Z_\sigma(x) = \int_{\mathcal{X}} Nw(x, z) \left(\int \ldots \int \frac{\prod_{j=1}^{N-1} Q_\sigma(x, \mathrm{d}y^{(j)})}{w(x, z) + \sum_{j=1}^{N-1} w(x, y^{(j)})} \right) Q_\sigma(x, \mathrm{d}z). \tag{2.5.2}$$

Ici, la fonction de poids joue en quelque sorte le rôle de la fonction g dans la forme (2.2.1). Par contre, en raison de la présence des intégrales, le théorème 2.2.5 ne s'applique pas directement. Ceci étant, il est maintenant possible d'énoncer la condition nécessaire à l'obtention de l'équilibre local. Par souci de simplicité, seules les fonctions de poids indépendantes de σ sont ici considérées.

Proposition 2.5.1. *Soit $Q_{w,\sigma}$, la distribution instrumentale du MTM* (2.5.1)*. Si cette distribution instrumentale est en équilibre local par rapport à une distribution Π, alors pour tous $x, y \in \mathcal{X}$ tels que $x \neq y$ et $w(x, y) > 0$, cela implique que*

$$w(x, y) = \frac{\pi(y)(N-1)w(y, y)w(y, x)}{\pi(x)[w(y, x) + (N-1)w(x, x)] - \pi(y)w(y, x)}. \tag{2.5.3}$$

Démonstration. Cette démonstration est similaire à celle du théorème 2.2.5. En partant de la distribution instrumentale (2.5.1) et en utilisant la symétrie de Q_σ, il est possible

d'écrire

$$\frac{Z_\sigma(x)Q_{w,\sigma}(x,\mathrm{d}y)\mathrm{d}x}{Nw(x,y)\int\ldots\int\frac{\prod_{j=1}^{N-1}Q_\sigma(x,\mathrm{d}y^{(j)})}{w(x,y)+\sum_{j=1}^{N-1}w(x,y^{(j)})}}=\frac{Z_\sigma(y)Q_{w,\sigma}(y,\mathrm{d}x)\mathrm{d}y}{Nw(y,x)\int\ldots\int\frac{\prod_{j=1}^{N-1}Q_\sigma(y,\mathrm{d}x^{(j)})}{w(y,x)+\sum_{j=1}^{N-1}w(y,x^{(j)})}}$$

$$Nw(y,x)Z_\sigma(x)\eta_\sigma(y,x)Q_{w,\sigma}(x,\mathrm{d}y)\mathrm{d}x=Nw(x,y)Z_\sigma(y)\eta_\sigma(x,y)Q_{w,\sigma}(y,\mathrm{d}x)\mathrm{d}y, \tag{2.5.4}$$

où l'on note

$$\eta_\sigma(x,y)=\int\ldots\int\frac{\prod_{j=1}^{N-1}Q_\sigma(x,\mathrm{d}y^{(j)})}{w(x,y)+\sum_{j=1}^{N-1}w(x,y^{(j)})}. \tag{2.5.5}$$

Par hypothèse, $Q_{w,\sigma}$ est en équilibre local. Cela signifie qu'elle est réversible par rapport à une distribution Π_σ telle que $\Pi_\sigma(\mathrm{d}x)Q_{w,\sigma}(x,\mathrm{d}y)=\Pi_\sigma(\mathrm{d}y)Q_{w,\sigma}(y,\mathrm{d}x)$. En considérant l'expression en (2.5.4), cela signifie que

$$\Pi_\sigma(\mathrm{d}x)=\pi_\sigma(x)\mathrm{d}x\propto Nw(y,x)Z_\sigma(x)\eta_\sigma(y,x)\mathrm{d}x, \tag{2.5.6}$$

avec proportionnalité par rapport à une constante de normalisation. Pour que la propriété d'équilibre local soit respectée, il est nécessaire que la densité π_σ converge vers π lorsque $\sigma\downarrow 0$. Il faut donc que la limite de (2.5.6) lorsque $\sigma\downarrow 0$ soit proportionnelle à $\pi(x)$. Afin de déterminer les conditions à apposer sur w pour que ce soit le cas, le calcul de la limite de (2.5.6) lorsque $\sigma\downarrow 0$ est détaillé.

Il est possible d'utiliser la proposition 2.1.2 afin de déterminer la limite de la fonction $\eta_\sigma(y,x)$. En effet, le dénominateur de l'intégrande en (2.5.5) est une fonction bornée par

$$\frac{1}{w(y,x)+\sum_{j=1}^{N-1}w(y,x^{(j)})}\leqslant\frac{1}{w(y,x)}<\infty,$$

puisque $w(y,x)>0$ par hypothèse. Puisque la mesure de probabilité Q_σ converge faiblement vers la mesure de Dirac, en utilisant la proposition 2.1.2, il suit que

$$\lim_{\sigma\to 0}\eta_\sigma(y,x)=\int\cdots\int\frac{\prod_{j=1}^{N-1}\delta_y(\mathrm{d}x^{(j)})}{w(y,x)+\sum_{j=1}^{N-1}w(y,x^{(j)})}=\frac{1}{w(y,x)+(N-1)w(y,y)}.$$

En ce qui concerne la limite de la fonction $Z_\sigma(x)$ en (2.5.2), il est possible de déterminer celle-ci grâce au lemme A.1.1 qui est une version un peu plus générale de la proposition 2.1.2 et dont la démonstration se retrouve à l'annexe A.1. Pour utiliser ce lemme, il faut s'assurer

que la partie de l'intégrande de $Z_\sigma(x)$ ne contenant pas $Q_\sigma(x, \mathrm{d}z)$ est bornée uniformément par une constante, et ce, pour tout σ. Cela est bien le cas puisque

$$Nw(x,z)\eta_\sigma(x,z) = Nw(x,z)\int ... \int \frac{\prod_{j=1}^{N-1} Q_\sigma(x, \mathrm{d}y^{(j)})}{w(x,z) + \sum_{j=1}^{N-1} w(x, y^{(j)})} \leqslant \frac{Nw(x,z)}{w(x,z)} = N.$$

En utilisant le lemme A.1.1, le calcul de la limite donne donc

$$\lim_{\sigma\to 0} Z_\sigma(x) = \lim_{\sigma\to 0} \int_{\mathcal{X}} Nw(x,z)\eta_\sigma(x,z)Q_\sigma(x,\mathrm{d}z) = \int_{\mathcal{X}} \frac{Nw(x,z)\delta_x(\mathrm{d}z)}{w(x,z) + (N-1)w(x,x)} = 1.$$

Ainsi, la limite de (2.5.6) est

$$\lim_{\sigma\to 0} Z_\sigma(x)w(y,x)\eta_\sigma(y,x) = \frac{Nw(y,x)}{w(y,x) + (N-1)w(y,y)}.$$

Pour que la propriété d'équilibre local soit respectée, il faut que cette limite soit proportionnelle à $\pi(x)$ de telle sorte que

$$\frac{Nw(y,x)}{w(y,x) + (N-1)w(y,y)} = z\pi(x) \qquad \text{et} \qquad \frac{Nw(x,y)}{w(x,y) + (N-1)w(x,x)} = z\pi(y),$$

où z est une constante de normalisation. En prenant le ratio de ces deux quantités, il suit que

$$\begin{aligned}
\frac{w(x,y)[w(y,x) + (N-1)w(x,x)]}{w(y,x)[w(x,y) + (N-1)w(y,y)]} &= \frac{\pi(y)}{\pi(x)} \\
w(x,y)\left\{1 + \frac{(N-1)w(x,x)}{w(y,x)}\right\} &= \frac{\pi(y)}{\pi(x)}w(x,y) + \frac{\pi(y)}{\pi(x)}(N-1)w(y,y) \\
w(x,y)\left\{1 + \frac{(N-1)w(x,x)}{w(y,x)} - \frac{\pi(y)}{\pi(x)}\right\} &= \frac{\pi(y)}{\pi(x)}(N-1)w(y,y),
\end{aligned}$$

ce qui mène à

$$w(x,y) = \frac{\pi(y)(N-1)w(y,y)w(y,x)}{\pi(x)[w(y,x) + (N-1)w(x,x)] - \pi(y)w(y,x)}.$$

□

Cette condition est beaucoup plus complexe que la condition d'équilibre local (2.2.3) dans le cas traditionnel. Un exemple de fonction de poids respectant cette condition est

$$w^*(x,y) = \begin{cases} \frac{\pi(y)}{\pi(x)}, & \text{si } y \neq x, \\ \frac{1}{N-1}, & \text{si } y = x. \end{cases}$$

Selon cette fonction de poids, un candidat sera préféré si sa densité cible est grande par rapport à la densité du point actuel sauf dans le cas particulier où $y = x$. Tel que mentionné

auparavant, la condition (2.5.3) n'est pas suffisante à l'obtention de la propriété d'équilibre local. Un exemple avec la fonction de poids $w^*(x,y)$ permet de bien saisir pourquoi. En remplaçant $w^*(x,y)$ par sa valeur à la première étape de la démonstration de la proposition 2.5.1 et en supposant que $x \neq y$, il est possible d'écrire

$$\frac{Z_\sigma(x)Q_{w,\sigma}(x,\mathrm{d}y)\mathrm{d}x}{\pi(y)\int\ldots\int\frac{\prod_{j=1}^{N-1}Q_\sigma(x,\mathrm{d}y^{(j)})}{\pi(y)+\sum_{j=1}^{N-1}\pi(y^{(j)})}} = \frac{Z_\sigma(y)Q_{w,\sigma}(y,\mathrm{d}x)\mathrm{d}y}{\pi(x)\int\ldots\int\frac{\prod_{j=1}^{N-1}Q_\sigma(y,\mathrm{d}x^{(j)})}{\pi(x)+\sum_{j=1}^{N-1}\pi(x^{(j)})}}$$

$$\pi(x)Z_\sigma(x)\kappa_\sigma(y,x)Q_{w,\sigma}(x,\mathrm{d}y)\mathrm{d}x = \pi(y)Z_\sigma(y)\kappa_\sigma(x,y)Q_{w,\sigma}(y,\mathrm{d}x)\mathrm{d}y,$$

où l'on note

$$\kappa_\sigma(x,y) = \int\ldots\int\frac{\prod_{j=1}^{N-1}Q_\sigma(x,\mathrm{d}y^{(j)})}{\pi(y)+\sum_{j=1}^{N-1}\pi(y^{(j)})}, \tag{2.5.7}$$

qui est distincte de la fonction η_σ mentionnée précédemment puisque le dénominateur n'est pas la somme des fonctions de poids en raison de simplifications. Tel que mentionné auparavant, pour que la propriété d'équilibre local soit respectée, il faut que $Q_{w,\sigma}(x,\mathrm{d}y)$ soit réversible par rapport à une distribution Π_σ. Ici, cette distribution est

$$\Pi_\sigma(dx) \propto \pi(x)Z_\sigma(x)\kappa_\sigma(y,x).$$

Or, le terme $\kappa_\sigma(y,x)$ dépend de x et de y. Pour s'assurer que la fonction $\Pi_\sigma(dx)$ ne dépende que de x, il faudrait pouvoir définir une condition semblable à (2.2.3) qui ferait en sorte que le ratio $\kappa_\sigma(y,x)/\kappa_\sigma(x,y)$ soit égal à $\kappa^*_\sigma(x)/\kappa^*_\sigma(y)$. De cette manière, nous aurions $\Pi_\sigma(dx) \propto \pi(x)Z_\sigma(x)\kappa^*_\sigma(x)$, qui ne dépendrait que de x. Or, trouver une telle condition est loin d'être aisé. L'intégrale (2.5.7) est difficilement calculable en pratique. De plus, celle-ci dépend à la fois de la densité cible π ainsi que de la distribution Q_σ utilisée. Cela signifie que la condition ne serait valide que pour la densité cible considérée. De ce fait, il serait possible d'utiliser une fonction de poids permettant d'obtenir l'équilibre local, mais cette fonction serait liée à la densité cible étudiée. Pour cette raison, cette avenue n'a pas été explorée davantage. Enfin, parmi les fonctions de poids du tableau 1.1 qui ne dépendent pas de σ, on note qu'aucune ne respecte la condition (2.5.3).

L'objectif de ce chapitre était de présenter une nouvelle classe de distributions informatives. Deux régimes ont été considérés pour ces distributions, le cas local et le cas global. Étant donné que le cas local est équivalent en pratique à utiliser le MALA, ce cas ne sera

pas exploré plus en détails. En effet, cet algorithme a été étudié abondamment et ses performances ont été largement documentées (voir [**18**]). Nous concentrons notre attention sur l'utilisation de la densité (2.3.3), soit l'interpolation entre le régime local et global. L'étude de cette distribution est intéressante en faible dimension étant donné que le régime local ne survient que lorsque $\sigma^2 \to 0$. De plus, selon nos connaissances, il n'y a rien de comparable dans la littérature. De ce fait, pour la suite des choses, nous désirons étudier la performance de cette densité instrumentale.

Chapitre 3

Choix du paramètre γ

Dans le contexte de l'utilisation du MH avec densité instrumentale normale, il n'y avait qu'un seul paramètre dont la valeur devait être prédéterminée par l'utilisateur de l'algorithme, soit la valeur de la variance instrumentale σ^2. Dans le cadre de l'utilisation de la densité instrumentale (2.3.3), il y a maintenant deux paramètres à déterminer : la variance instrumentale, σ^2, et le nouveau paramètre, γ. L'objectif de ce chapitre est de proposer une méthode de sélection du paramètre γ, afin d'obtenir un algorithme qui soit le plus efficace possible. L'idée est donc similaire à ce qui a déjà été discuté pour le MH et le choix de la variance instrumentale à la section 1.3.1, mais cette fois-ci, elle est appliquée au paramètre γ.

Tout d'abord, la notion d'efficacité doit être définie à l'aide d'un critère de performance spécifique. L'intérêt est ensuite porté au choix de $\gamma \in [1, 2]$. Le choix optimal de ce paramètre dépendra de la dimension du problème. En effet, tel qu'il a été vu au chapitre 2, lorsque $d \uparrow \infty$ (et par le fait même $\sigma^2 \downarrow 0$) il est préférable d'utiliser une valeur de $\gamma = 1$, soit celle correspondant au régime local. En ce qui concerne les deux autres régimes qui surviennent lorsque $d < \infty$, c'est-à-dire les régimes global où $\sigma \uparrow \infty$ et intermédiaire où $0 < \sigma < \infty$, deux questions se posent. On désire d'abord identifier, en pratique, quelles valeurs de γ doivent être utilisées dans ces deux régimes. On désire également établir les frontières des différents régimes. En d'autres mots, on désire identifier en pratique quelles dimensions d correspondent au régime global, au régime intermédiaire et enfin au régime local. Pour terminer, la question du choix optimal de σ est brièvement abordée.

Tout au long de ce chapitre, la mise à l'échelle β (2.4.1) utilisée pour le paramètre σ sera celle du MALA ($\beta = 1/3$), et ce, indépendamment de la valeur de γ considérée. La mise à l'échelle fait en sorte qu'il n'y ait pas de dégénérescence du taux d'acceptation lorsque d croît. Puisque la valeur de γ doit se rapprocher de 1 (le MALA) lorsque $d \to \infty$, le choix d'une mise à l'échelle de $\beta = 1/3$ mènera à une performance optimale en grande dimension. Au chapitre 2, il avait été mentionné que la mise à l'échelle appropriée pour $\gamma > 1$ correspondait à $\beta = 1$. Il peut donc sembler ambigu de combiner une mise à l'échelle $\beta = 1/3$ avec un paramètre $\gamma > 1$. Cela n'a toutefois pas d'impact dans le cas présent, puisque des valeurs $\gamma > 1$ ne sont utilisées qu'en dimension finie. Dans un tel contexte, que l'on considère une variance de la forme $\sigma^2 = \ell/d^{1/3}$ ou $\sigma^2 = \ell/d$, il est toujours possible d'ajuster la valeur de ℓ de sorte à obtenir une variance optimale, puisque la dimension est finie.

3.1. Choix d'un critère d'efficacité

Il existe plusieurs critères d'efficacité permettant de déterminer si une méthode MCMC fonctionne de manière appropriée. Pour une fonction $h : \mathcal{X} \to \mathbb{R}$, lorsqu'on désire estimer $\mathbb{E}_\pi[h(X)]$, un critère naturel est la variance asymptotique en (1.3.6) présentée au chapitre 1. Minimiser cette quantité permet d'augmenter la précision de notre estimation. Malheureusement, cela signifie qu'il faut pour cela utiliser une fonction d'intérêt h spécifique ; par souci de généralité, un autre critère est ici considéré. Tel que mentionné auparavant, lorsqu'on utilise une méthode MCMC, il est préférable que les états consécutifs de la chaîne soient le plus loin possible les uns des autres afin que l'exploration de l'espace d'états soit la plus rapide possible. Le critère suivant cherche à mesurer cet aspect.

Définition 3.1.1. Soient X et X', deux états consécutifs d'une chaîne de Markov à espace d'états $\mathcal{X}$ ayant atteint la stationnarité. La *distance de saut quadratique moyenne* (DSM) est définie par

$$DSM = \mathbb{E}\left[\|X' - X\|^2\right] = \mathbb{E}\left[\sum_{i=1}^{d}(X_i' - X_i)^2\right], \tag{3.1.1}$$

où l'espérance est prise selon la loi conjointe de X et X' et où $\|\cdot\|$ est la norme euclidienne.

On note que lorsque la DSM est calculée la stationnarité est supposée atteinte, ce qui signifie que la densité marginale de X et de X' est π. Afin d'obtenir un algorithme qui soit le plus efficace possible, la DSM devra être maximisée. Ce critère incorpore à la fois le

taux d'acceptation, au travers du calcul de l'espérance, ainsi que la grandeur des sauts à chaque itération. Notons que si une proposition n'est pas acceptée, $\|X' - X\|^2 = 0$ et cette proposition ne contribuera pas à la DSM. Ainsi, pour avoir une grande DSM, il faut que le taux d'acceptation soit assez élevé sans que la grandeur des sauts n'en devienne pour autant trop petite, ce qui survient lorsque la variance instrumentale est faible. En pratique, pour calculer la DSM, il suffit de calculer $\sum_{i=1}^{d}(X_i' - X_i)^2$ pour chaque paire d'états consécutifs de la chaîne, puis de prendre la moyenne sur toute la longueur de celle-ci en évitant la partie tronquée (« burn-in »). Cette quantité est donc toujours disponible et simple à calculer.

L'utilisation de ce critère comporte différents avantages. Tout d'abord, il est possible de relier la DSM d'une unique composante avec l'autocorrélation de délai 1. En notant que $\mathbb{V}(X_i') = \mathbb{V}(X_i)$ et que $\mathbb{E}[X_i' - X_i] = 0$, il est possible d'écrire

$$\mathbb{E}[(X_i' - X_i)^2] = \mathbb{V}(X_i' - X_i) = 2\mathbb{V}(X_i)(1 - \mathrm{Corr}(X_i, X_i')).$$

Ainsi, maximiser la DSM pour une certaine composante implique que l'autocorrélation de délai 1 est minimisée. La DSM devient donc

$$DSM = 2\mathbb{V}(X_1)\left(d - \sum_{i=1}^{d}\mathrm{Corr}(X_i, X_i')\right),$$

et la maximiser revient à minimiser la somme des autocorrélations de délai 1 de toutes les composantes. Minimiser l'autocorrélation est pertinent, puisqu'il est possible de relier cette quantité à la variance de notre estimateur. En effet, si l'on désire estimer $\mathbb{E}_\pi[h(X)]$ par $\sum_{i=1}^{n} h(X[i])/n$, alors la variance de cet estimateur est proportionnelle à l'autocorrélation intégrée définie par (voir [**19**])

$$\tau_h = 1 + 2\sum_{i=1}^{\infty}\mathrm{Corr}(h(X[0]), h(X[i])). \tag{3.1.2}$$

Si la DSM est maximisée, l'autocorrélation de délai 1 sera minimisée et cela contribuera donc à réduire la variance de notre estimateur.

Le deuxième avantage important de ce critère est qu'il s'agit d'un critère pouvant être utilisé dans un régime non-asymptotique. Tel que mentionné dans [**19**], il est possible d'approximer la chaîne générée par un MH par un processus de diffusion lorsque $d \uparrow \infty$. Les auteurs démontrent alors qu'asymptotiquement, toute mesure d'efficacité d'un algorithme

est équivalente à la vitesse de ce processus de diffusion. En fait, différents choix de critère mèneront tous à la même conclusion. Or, dans le cas présent, il est important de définir un tel critère, puisque l'intérêt est porté au choix optimal de γ lorsque $d < \infty$. La principale alternative à la DSM sous cet aspect est l'autocorrélation intégrée (3.1.2) qui peut également être utilisée lorsque $d < \infty$. Un problème relié à ce critère est que même pour de faibles délais, l'autocorrélation théorique peut différer grandement de celle observée. C'est le cas, par exemple, si l'on considère une chaîne unidimensionnelle estimant une distribution uniforme sur $[-2, -1] \cup [1, 2]$. En pratique, l'algorithme n'explorera que la moitié de tout l'espace d'états et, s'il est efficace, l'autocorrélation diminuera rapidement en fonction du délai. Par contre, l'autocorrélation théorique à la stationnarité sera élevée même pour de grands délais puisqu'elle sera calculée en tenant compte de tout l'espace d'états et non seulement de la moitié de celui-ci. Ainsi, la DSM semble être un meilleur choix dans ce contexte.

3.2. Choix de γ en régime global

Il a été mentionné à la section 2.2 que le régime global survenait en faible dimension, soit lorsque $\sigma^2 \to \infty$. Sous ce régime, les algorithmes traditionnels ont tendance à être inefficaces. Par exemple, dans [**22**], l'auteur montre que la DSM tend vers 0 à mesure que $\sigma^2 \uparrow \infty$ lorsqu'on utilise le RWM sur une densité cible symétrique et unimodale. Dans cette section, on désire montrer que ce n'est pas le cas lorsqu'on utilise une densité en équilibre global, et donc que l'utilisation de $\gamma = 2$ mène à un algorithme efficace en faible dimension. Avant toute chose, il faut simplifier l'expression de la DSM afin de déterminer sa limite.

La DSM est une espérance calculée par rapport à la loi conjointe de deux éléments consécutifs d'une chaîne de Markov. Dans le cas présent, on considère que la chaîne est générée selon un algorithme de type MH, c'est-à-dire avec une étape de proposition et une étape acceptation/rejet. Soient X et Y ces deux éléments consécutifs et x et y les états associés à ceux-ci. Afin de calculer la DSM, il faut d'abord déterminer la loi conjointe de X et Y. Dans le cadre du MH, la probabilité de transition de la chaîne de l'état x à l'état y est

$$P(x, \mathrm{d}y) = Q_\sigma(x, \mathrm{d}y)\alpha(x, y) + \delta_x(\mathrm{d}y) \int_{z \in \mathcal{X}} (1 - \alpha(x, z)) Q_\sigma(x, \mathrm{d}z);$$

la loi conjointe de X et Y est $\Pi(\mathrm{d}x)P(x, \mathrm{d}y) = \pi(x)P(x, \mathrm{d}y)\mathrm{d}x$. Ici, $Q_\sigma(x, \cdot)$ est la distribution instrumentale de l'algorithme qui dépend d'un paramètre d'échelle σ, tandis que $\alpha(x, y)$ est

la probabilité d'acceptation présentée en (1.3.1). Afin de simplifier l'expression de la DSM en dimension finie, l'espace d'états $\mathcal{X}$ sera séparé en régions distinctes parallèlement à ce qui a été fait par Sherlock [**22**]. En raison de la forme de la probabilité d'acceptation $\alpha(x, y)$, il n'y a que quatre régions distinctes de $\mathcal{X}$ dans lesquelles le prochain élément de la chaîne (y) peut se retrouver. Ces régions sont :

- **la région d'identité** $R_{id}(x) := \{x\}$, qui est de mesure nulle sous $Q_\sigma(x, \cdot)$;
- **la région d'égalité** $R_{e,\sigma}(x) := \left\{y \in \mathcal{X} : y \notin R_{id}(x), \frac{\pi(y)q_\sigma(y,x)}{\pi(x)q_\sigma(x,y)} = 1\right\}$;
- **la région d'acceptation** $R_{a,\sigma}(x) := \left\{y \in \mathcal{X} : \frac{\pi(y)q_\sigma(y,x)}{\pi(x)q_\sigma(x,y)} > 1\right\}$, soit le restant de la région où la proposition est acceptée automatiquement ;
- **la région de rejet** $R_{r,\sigma}(x) := \{y \in \mathcal{X} : \alpha(x, y) < 1\}$, soit la région où la proposition n'est pas automatiquement acceptée.

On note que ces régions dépendent toutes de la position actuelle de la chaîne (x). Il est possible d'ignorer la région d'identité lors du calcul de la DSM puisque $\|y - x\|^2 = 0$ lorsque $y = x$. De ce fait, la densité conjointe de deux éléments successifs de la chaîne dans les trois autres régions sera $A(\mathrm{d}x, \mathrm{d}y) = \pi(x)q_\sigma(x, y)\alpha(x, y)\mathrm{d}y\mathrm{d}x$. Il est possible de simplifier la DSM en utilisant le fait qu'il y a interchangeabilité entre les régions $R_{a,\sigma}(\cdot)$ et $R_{r,\sigma}(\cdot)$. Plus précisément,

$$y \in R_{a,\sigma}(x) \Leftrightarrow x \in R_{r,\sigma}(y).$$

En effet, si un candidat y est généré dans une région où le ratio $\pi(y)q_\sigma(y, x)/\pi(x)q_\sigma(x, y) > 1$, la probabilité de rejet sera nulle, alors que pour le saut dans le sens inverse (de y à x), cette probabilité sera non nulle puisque le ratio sera inférieur à 1. Le lemme 3.2.1, présenté dans Sherlock [**22**], utilise cette information et réduit l'expression de la DSM à une forme plus simple.

Lemme 3.2.1. (Lemma 1 et Corollary 1 dans Sherlock [**22**]) *Soient X et Y deux éléments consécutifs d'une chaîne de Markov à espace d'états $\mathcal{X}$ générée par un algorithme de type MH avec densité cible π, distribution instrumentale $Q_\sigma(x, \cdot)$ et densité conjointe $A(\mathrm{d}x, \mathrm{d}y)$. Lorsque la stationnarité de la chaîne est atteinte, l'égalité suivante est satisfaite*

$$\int_{x\in\mathcal{X}} \int_{y\in R_{a,\sigma}(x)} \|y - x\|^2 A(\mathrm{d}x, \mathrm{d}y) = \int_{x\in\mathcal{X}} \int_{y\in R_{r,\sigma}(x)} \|y - x\|^2 A(\mathrm{d}x, \mathrm{d}y),$$

et la DSM peut s'exprimer sous la forme suivante :

$$DSM = \int_{x\in\mathcal{X}} \left[\int_{y\in R_{e,\sigma}(x)} \|y-x\|^2 \, Q_\sigma(x, \mathrm{d}y) + 2 \int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 \, Q_\sigma(x, \mathrm{d}y) \right] \pi(x)\mathrm{d}x.$$

Démonstration. La première partie découle de la réversibilité de la chaîne générée ainsi que de l'interchangeabilité des régions $R_{a,\sigma}(\cdot)$ et $R_{r,\sigma}(\cdot)$. En effet, il est possible d'écrire

$$\begin{aligned}
\int_{x\in\mathcal{X}} \int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 \, A(\mathrm{d}x, \mathrm{d}y) &= \int_{x\in\mathcal{X}} \int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 \, \pi(x) q_\sigma(x,y)\alpha(x,y)\mathrm{d}y\mathrm{d}x \\
&= \int_{x\in\mathcal{X}} \int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 \, \pi(y) q_\sigma(y,x)\alpha(y,x)\mathrm{d}y\mathrm{d}x \\
&= \int_{y\in\mathcal{X}} \int_{x\in R_{r,\sigma}(y)} \|y-x\|^2 \, \pi(y) q_\sigma(y,x)\alpha(y,x)\mathrm{d}x\mathrm{d}y \\
&= \int_{y\in\mathcal{X}} \int_{x\in R_{r,\sigma}(y)} \|y-x\|^2 \, A(\mathrm{d}y, \mathrm{d}x) \\
&= \int_{x\in\mathcal{X}} \int_{y\in R_{r,\sigma}(x)} \|y-x\|^2 \, A(\mathrm{d}x, \mathrm{d}y).
\end{aligned}$$

En ignorant la région $R_{id}(x)$, puisque $\|y-x\|^2 = 0$ lorsque $y \in R_{id}(x)$, il est alors possible d'exprimer la DSM en (3.1.1) sous la forme

$$\begin{aligned}
DSM &= \int_{x\in\mathcal{X}} \left[\int_{y\in R_{e,\sigma}(x)} \|y-x\|^2 \, A(\mathrm{d}x, \mathrm{d}y) + \int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 \, A(\mathrm{d}x, \mathrm{d}y) \right. \\
&\qquad \left. + \int_{y\in R_{r,\sigma}(x)} \|y-x\|^2 \, A(\mathrm{d}x, \mathrm{d}y) \right] \\
&= \int_{x\in\mathcal{X}} \left[\int_{y\in R_{e,\sigma}(x)} \|y-x\|^2 \, A(\mathrm{d}x, \mathrm{d}y) + 2 \int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 \, A(\mathrm{d}x, \mathrm{d}y) \right] \\
&= \int_{x\in\mathcal{X}} \left[\int_{y\in R_{e,\sigma}(x)} \|y-x\|^2 \, Q_\sigma(x, \mathrm{d}y) + 2 \int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 \, Q_\sigma(x, \mathrm{d}y) \right] \pi(x)\mathrm{d}x,
\end{aligned}$$

où l'on utilise le fait que pour tout $x \in \mathcal{X}$, on a $\alpha(x,y) = 1$ si $y \in \{R_{e,\sigma}(x) \cup R_{a,\sigma}(x)\}$. □

Les régions $R_{a,\sigma}(x)$ et $R_{e,\sigma}(x)$ dépendent du paramètre d'échelle σ à travers la densité instrumentale. Ainsi, afin de déterminer la limite de la DSM lorsque $\sigma^2 \uparrow \infty$, il est nécessaire de déterminer la limite de ces régions. Ceci est exprimé dans le lemme 3.2.2 pour les distributions en équilibre local et global.

Lemme 3.2.2. *Soit X, une chaîne de Markov à espace d'états $\mathcal{X}$ générée par un algorithme MH avec densité cible π bornée telle que $\sqrt{\pi}$ est intégrable. Soit $Q_{g,\sigma}$, la distribution instrumentale biaisée où $g(x,y) = (\pi(y))^{\gamma/2}$. Soit de plus*

$$A_\sigma(x) := \{R_{e,\sigma}(x) \cup R_{a,\sigma}(x)\} = \left\{ y \in \mathcal{X}\backslash\{x\} : \frac{\pi(y) q_{g,\sigma}(y,x)}{\pi(x) q_{g,\sigma}(x,y)} \geqslant 1 \right\}. \tag{3.2.1}$$

Alors, pour tout $x \in \mathcal{X}$, si $\gamma = 1$, la limite de cet ensemble lorsque $\sigma \to \infty$ vaut

$$\lim_{\sigma\to\infty} A_\sigma(x) = \{y \in \mathcal{X}\backslash\{x\} : \pi(y) \geqslant \pi(x)\},$$

tandis que si $\gamma = 2$, elle vaut

$$\lim_{\sigma\to\infty} A_\sigma(x) = \mathcal{X}\backslash\{x\}.$$

Démonstration. Selon la définition 2.2.1, une densité instrumentale biaisée est de la forme

$$q_{g,\sigma}(x,y) = \frac{(\pi(y))^{\frac{\gamma}{2}} q_\sigma(x,y)}{Z_\sigma(x)},$$

avec

$$Z_\sigma(x) = \int_{\mathcal{X}} (\pi(z))^{\frac{\gamma}{2}} Q_\sigma(x, \mathrm{d}z),$$

où, pour tout $x \in \mathcal{X}$, $q_\sigma(x,\cdot)$ est une densité symétrique pouvant s'écrire sous la forme (2.1.1). Étant donnée la forme des régions $R_{e,\sigma}(x)$ et $R_{a,\sigma}(x)$, on s'intéresse au ratio

$$\frac{\pi(y) q_{g,\sigma}(y,x)}{\pi(x) q_{g,\sigma}(x,y)} = \frac{\pi(y)(\pi(x))^{\gamma/2} q_\sigma(y,x) Z_\sigma(x)}{\pi(x)(\pi(y))^{\gamma/2} q_\sigma(x,y) Z_\sigma(y)} = \left(\frac{\pi(y)}{\pi(x)}\right)^{1-\frac{\gamma}{2}} \frac{Z_\sigma(x)}{Z_\sigma(y)}. \tag{3.2.2}$$

On désire calculer la limite du ratio (3.2.2) lorsque $\sigma \to \infty$. Pour ce faire, il est possible d'utiliser la proposition 2.1.3. En effet, puisque $\sqrt{\pi}$ est intégrable, la fonction $(\pi(x))^{\gamma/2}$ est intégrable pour $\gamma = 1$ et $\gamma = 2$. Il est donc possible d'écrire

$$\sigma^d Z_\sigma(x) = \int_{\mathcal{X}} (\pi(z))^{\frac{\gamma}{2}} \sigma^d Q_\sigma(x, \mathrm{d}z) = \int_{\mathcal{X}} (\pi(z))^{\frac{\gamma}{2}} \mu_\sigma(x, \mathrm{d}z) \xrightarrow{\sigma\uparrow\infty} \int_{\mathcal{X}} (\pi(z))^{\frac{\gamma}{2}} r(\mathbf{0}) \mathrm{d}z,$$

où $\mathbf{0} = (0, \ldots, 0)^\top \in \mathcal{X}$. De ce fait, la limite du ratio (3.2.2) devient

$$\left(\frac{\pi(y)}{\pi(x)}\right)^{1-\frac{\gamma}{2}} \frac{\sigma^d Z_\sigma(x)}{\sigma^d Z_\sigma(y)} \xrightarrow{\sigma\uparrow\infty} \left(\frac{\pi(y)}{\pi(x)}\right)^{1-\frac{\gamma}{2}} \frac{\int_{\mathcal{X}} (\pi(z))^{\gamma/2} r(\mathbf{0}) \mathrm{d}z}{\int_{\mathcal{X}} (\pi(z))^{\gamma/2} r(\mathbf{0}) \mathrm{d}z} = \left(\frac{\pi(y)}{\pi(x)}\right)^{1-\frac{\gamma}{2}}.$$

Donc, si $\gamma = 1$, on a

$$\lim_{\sigma\to\infty} A_\sigma(x) = \left\{ y \in \mathcal{X}\backslash\{x\} : \left(\frac{\pi(y)}{\pi(x)}\right)^{\frac{1}{2}} \geqslant 1 \right\} = \{y \in \mathcal{X}\backslash\{x\} : \pi(y) \geqslant \pi(x)\},$$

tandis que si $\gamma = 2$, on a

$$\left(\frac{\pi(y)}{\pi(x)}\right)^{1-\frac{\gamma}{2}} = 1, \quad \forall\, y \in \mathcal{X},$$

ce qui implique que $\lim_{\sigma\to\infty} A_\sigma(x) = \mathcal{X}\backslash\{x\}$. □

Ainsi, lorsque $\gamma = 2$, $\lim_{\sigma\to\infty} A_\sigma(x) = \mathcal{X}\backslash\{x\}$. Cela n'est pas surprenant puisqu'il s'agit de la principale raison de l'utilisation d'une densité en équilibre global. En effet, lorsque $\sigma \to \infty$ et qu'une telle densité instrumentale est utilisée, le taux d'acceptation vaut 1 et tous les candidats sont acceptés. On remarque également que cela ne survient pas dans le cas local puisque la région d'acceptation ne consiste qu'en une partie de l'espace d'états $\mathcal{X}$. Avec l'aide des lemmes 3.2.1 et 3.2.2, il est possible de déterminer la limite de la DSM lorsque $\sigma \to \infty$ et qu'on utilise une distribution en équilibre global.

Proposition 3.2.3. *Soient X et Y, deux éléments consécutifs d'une chaîne de Markov à espace d'états $\mathcal{X}$ générée par un algorithme MH. Soit π, la densité cible bornée telle que $\mathbb{E}_\pi[\|X\|^2] < \infty$. Si on utilise la distribution instrumentale en équilibre global $Q_{\pi,\sigma}$ afin de biaiser la distribution symétrique Q_σ dont la densité associée q_σ respecte la condition*

$$\pi(y) > 0 \Rightarrow q_\sigma(x, y) > 0 \quad \forall\, x \in \mathcal{X}, \tag{3.2.3}$$

alors

$$\lim_{\sigma\to\infty} DSM > 0.$$

Démonstration. Sans perte de généralité, on suppose que $\sigma \geqslant c$ pour une constante $0 < c < \infty$. Cela permet d'éviter des problèmes d'intégration lorsque $\sigma \downarrow 0$ qui n'est, de toute façon, pas la limite d'intérêt ici. Ceci étant dit, on désire borner inférieurement la limite de la DSM par une expression strictement supérieure à 0. Par le lemme 3.2.1, il est possible d'écrire

$$\begin{aligned} DSM &= \mathbb{E}[\|Y - X\|^2] \\ &= \int_{x\in\mathcal{X}} \pi(x) \left[\int_{y\in R_{e,\sigma}(x)} \|y - x\|^2 \, Q_{\pi,\sigma}(x, \mathrm{d}y) + 2 \int_{y\in R_{a,\sigma}(x)} \|y - x\|^2 \, Q_{\pi,\sigma}(x, \mathrm{d}y) \right] \mathrm{d}x \\ &\geqslant \int_{x\in\mathcal{X}} \pi(x) \int_{y\in A_\sigma(x)} \|y - x\|^2 \, Q_{\pi,\sigma}(x, \mathrm{d}y)\mathrm{d}x, \end{aligned}$$

où l'ensemble $A_\sigma(x)$ est tel que défini en (3.2.1). Afin d'éviter certains problèmes d'intégration, on ne considère que les points $x \in B(\mathbf{0}, a)$, la boule de rayon $0 < a < \infty$ centrée en $\mathbf{0} = (0, \ldots, 0)^T \in \mathcal{X}$. Alors, il s'avère que

$$\begin{aligned} DSM &\geqslant \int_{x \in B(\mathbf{0},a)} \frac{\pi(x)}{Z_\sigma(x)} \int_{y \in A_\sigma(x)} \|y - x\|^2 \pi(y) Q_\sigma(x, \mathrm{d}y)\mathrm{d}x \\ &= \int_{x \in B(\mathbf{0},a)} \frac{\pi(x)}{Z^*_\sigma(x)} \int_{y \in A_\sigma(x)} \|y - x\|^2 \pi(y) \mu_\sigma(x, \mathrm{d}y)\mathrm{d}x, \end{aligned} \tag{3.2.4}$$

où $\mu_\sigma(x, \mathrm{d}y) = \sigma^d Q_\sigma(x, \mathrm{d}y)$ et $Z^*_\sigma(x) = \int_{\mathcal{X}} \pi(z)\mu_\sigma(x, \mathrm{d}z)$.

Pour calculer la limite de la DSM, on désire faire entrer la limite dans les deux intégrales. Pour faire entrer la limite dans la première, on désire utiliser le théorème de la convergence dominée. Pour ce faire, il est nécessaire de borner supérieurement la fonction

$$F_\sigma(x) = \frac{\pi(x)}{Z^*_\sigma(x)} \int_{y \in A_\sigma(x)} \|y - x\|^2 \pi(y) \mu_\sigma(x, \mathrm{d}y) \tag{3.2.5}$$

par une fonction intégrable indépendante de σ, et ce, pour tout $\sigma \geqslant c$. On se concentre d'abord sur le numérateur. Comme vu dans la proposition 2.1.3, on a que $\mu_\sigma(x, \mathrm{d}y) = r((y - x)/\sigma)\mathrm{d}y$ et que $\sup_{z \in \mathcal{X}} r(z) = M < \infty$. De ce fait, le numérateur de $F_\sigma(x)$ est borné par

$$\begin{aligned} \pi(x) \int_{y \in A_\sigma(x)} \|y - x\|^2 \pi(y) \mu_\sigma(x, \mathrm{d}y) &\leqslant \pi(x) \int_{y \in A_\sigma(x)} \|y - x\|^2 \pi(y) M \mathrm{d}y \\ &\leqslant M\pi(x) \int_{y \in \mathcal{X}} \|y - x\|^2 \pi(y) \mathrm{d}y \\ &\leqslant M\pi(x) \int_{y \in \mathcal{X}} (\|y\|^2 + \|x\|^2 + 2\|y\| \|x\|)\pi(y)\mathrm{d}y. \end{aligned}$$

Comme $M_2 = \mathbb{E}_\pi[\|X\|^2] < \infty$, on a également que $M_1 = \mathbb{E}_\pi[\|X\|] < \infty$ par l'inégalité de Jensen. En utilisant cette notation, il suit que l'expression précédente est bornée par

$$M\pi(x) \int_{y \in \mathcal{X}} (\|y\|^2 + \|x\|^2 + 2\|y\| \|x\|)\pi(y)\mathrm{d}y \leqslant M\pi(x)(M_2 + \|x\|^2 + 2M_1 \|x\|).$$

Cette dernière expression est également intégrable sur $\mathcal{X}$ puisque M_2 est borné.

On montre maintenant que le dénominateur, soit la fonction $Z^*_\sigma(x)$, est bornée inférieurement par une constante pour tout $\sigma \geqslant c$. Pour un certain $t > 0$, soit l'ensemble

$$C(t) = \{x \in \mathcal{X} : \pi(x) \geqslant t\},$$

où t est tel que $\mu(C(t)) > 0$, avec μ, la mesure de Lebesgue. Alors, on peut écrire

$$Z_\sigma^*(x) = \int_{\mathcal{X}} \pi(z)\mu_\sigma(x, \mathrm{d}z) \geqslant \int_{C(t)} \pi(z)\mu_\sigma(x, \mathrm{d}z) \geqslant t \int_{C(t)} r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z.$$

Il est clair que l'ensemble $C(t)$ est de mesure finie, car dans le cas contraire, la densité π ne pourrait pas intégrer à 1. En effet, si l'on suppose que ce n'est pas le cas, cela mène à une contradiction puisque

$$1 = \int_{\mathcal{X}} \pi(z)\mathrm{d}z = \int_{C(t)} \pi(z)\mathrm{d}z + \int_{\mathcal{X}\setminus C(t)} \pi(z)\mathrm{d}z \geqslant t \int_{C(t)} \mathrm{d}z = t\mu(C(t)).$$

Comme la densité r est bornée, le théorème de la convergence bornée implique que

$$\int_{C(t)} r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z \xrightarrow{\sigma\uparrow\infty} \int_{C(t)} r(\mathbf{0})\mathrm{d}z = r(\mathbf{0})\mu(C(t)),$$

où l'on a bien $\lim_{\sigma\to\infty} r((z-x)/\sigma) = r(\mathbf{0})$ pour tout $x \in B(\mathbf{0}, a)$ puisque $a < \infty$. Il suit que pour $0 < \varepsilon < r(\mathbf{0})$, il existe un σ_0 tel que pour tout $\sigma \geqslant \sigma_0$, on a

$$\left| r\left(\frac{z-x}{\sigma}\right) - r(\mathbf{0}) \right| < \varepsilon.$$

Par conséquent, pour $\sigma \geqslant \sigma_0$, il s'avère que

$$Z_\sigma^*(x) \geqslant t \int_{C(t)} r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z \geqslant t \int_{C(t)} (r(\mathbf{0}) - \varepsilon)\mathrm{d}z = t\mu(C(t))(r(\mathbf{0}) - \varepsilon) > 0.$$

Lorsque $c \leqslant \sigma < \sigma_0$, étant donnée la condition (3.2.3) apposée sur la densité instrumentale q_σ, on a nécessairement

$$Z_\sigma^*(x) = \int_{\mathcal{X}} \pi(z)\sigma^d Q_\sigma(x, \mathrm{d}z) \geqslant c^d \int_{\mathcal{X}} \pi(z) q_\sigma(x, z)\mathrm{d}z > 0.$$

Cela signifie que $m = \inf_{c\leqslant\sigma<\sigma_0} Z_\sigma^*(x) > 0$. Donc, pour tout $\sigma \geqslant c$, on a

$$Z_\sigma^*(x) \geqslant \min\{m, t\mu(C(t))(r(\mathbf{0}) - \varepsilon)\}.$$

Ainsi, la fonction en (3.2.5) est bornée par

$$F_\sigma(x) \leqslant \frac{M\pi(x)(M_2 + \|x\|^2 + 2M_1 \|x\|)}{\min\{m, t(r(\mathbf{0})\mu(C(t)) - \varepsilon)\}},$$

qui est bien indépendante de σ et intégrable sur $B(\mathbf{0}, a)$.

En utilisant le théorème de la convergence dominée, il est alors possible de faire passer la limite à l'intérieur de la première intégrale de (3.2.4), ce qui donne

$$\lim_{\sigma\to\infty} DSM \geqslant \int_{B(\mathbf{0},a)} \lim_{\sigma\to\infty} \frac{\pi(x)}{Z_\sigma^*(x)} \int_{y\in A_\sigma(x)} \|y - x\|^2 \pi(y)\mu_\sigma(x, \mathrm{d}y)\mathrm{d}x. \tag{3.2.6}$$

On peut calculer séparément les limites de $Z_\sigma^*(x)$ et du reste de l'expression (3.2.6) étant donné que ces deux limites existent. En effet, par la proposition 2.1.3, on peut écrire

$$Z_\sigma^*(x) = \int_{\mathcal{X}} \pi(z)\mu_\sigma(x, \mathrm{d}z) \xrightarrow{\sigma\uparrow\infty} \int_{\mathcal{X}} \pi(z) r(\mathbf{0}) \mathrm{d}z = r(\mathbf{0}).$$

Pour faire entrer la limite dans la deuxième intégrale de l'expression (3.2.6), on utilise le théorème de la convergence dominée. En effet, on a

$$\begin{aligned} \int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y)\mu_\sigma(x, \mathrm{d}y) &= \int_{y\in\mathcal{X}} \|y-x\|^2 \pi(y) \mathbb{1}_{y\in A_\sigma(x)} \mu_\sigma(x, \mathrm{d}y) \\ &\leqslant \int_{y\in\mathcal{X}} \|y-x\|^2 \pi(y) M \mathrm{d}y, \end{aligned} \tag{3.2.7}$$

et la fonction $\|y-x\|^2 \pi(y)$ est indépendante de σ et intégrable, tel que mentionné plus haut. Selon le lemme 3.2.2, la limite de l'ensemble $A_\sigma(x)$ lorsque $\sigma \uparrow \infty$ est $\mathcal{X}\backslash\{x\}$, donc la limite de l'expression (3.2.7) devient

$$\int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y)\mu_\sigma(x, \mathrm{d}y) \xrightarrow{\sigma\uparrow\infty} \int_{y\in\mathcal{X}\backslash\{x\}} \|y-x\|^2 \pi(y) r(\mathbf{0}) \mathrm{d}y.$$

En combinant la limite de $Z_\sigma^*(x)$ et celle de (3.2.7), on obtient

$$\lim_{\sigma\to\infty} DSM \geqslant \int_{B(\mathbf{0},a)} \int_{y\in\{\mathcal{X}\backslash\{x\}\}} \|y-x\|^2 \pi(x)\pi(y) \mathrm{d}y \mathrm{d}x > 0.$$

□

En analysant la démonstration de la proposition 3.2.3, il est possible de saisir pourquoi un algorithme tel que le RWM est inefficace lorsque $\sigma^2 \uparrow \infty$. À mesure que σ devient grand, la densité instrumentale q_σ s'affaisse et les candidats générés sont éloignés des endroits de haute densité cible. En raison de la forme de la probabilité d'acceptation (1.3.1), cela résulte en un faible taux d'acceptation des candidats. Lorsqu'une distribution en équilibre global est utilisée, tous les candidats sont acceptés automatiquement, et ce, peu importe où ceux-ci sont générés dans le paysage de la densité cible. La forme de la distribution instrumentale en équilibre global fait en sorte que la chaîne reste réversible par rapport à Π, malgré le taux d'acceptation de 1.

La proposition 3.2.3 n'appose que peu d'hypothèses sur la densité cible ; la densité instrumentale en équilibre global semble donc plus efficace en faible dimension que le RWM. Dans ce chapitre, on désire déterminer la valeur de γ optimale selon chaque régime. Or, la proposition 3.2.3 ne permet pas de discriminer entre une valeur de $\gamma = 2$ et $\gamma = 1$ en faible

dimension. En d'autres mots, cette proposition ne garantit pas l'optimalité de l'équilibre global comparativement au cas local en faible dimension. Le lemme 3.2.2 montre que tous les candidats ne sont pas acceptés lorsque $\gamma = 1$ contrairement au cas où $\gamma = 2$, mais cela ne signifie pas que la DSM est nécessairement inférieure. Il n'a malheureusement pas été possible de démontrer théoriquement ce dernier point, mais les expériences empiriques présentées à la fin de ce chapitre semblent valider celui-ci. Dans tous les cas, en faible dimension, il est plus efficace, en termes de DSM, d'utiliser une valeur de $\gamma = 2$ par rapport à $\gamma = 1$. Dorénavant, nous supposerons donc que c'est le cas.

3.3. Choix de γ en régime intermédiaire

3.3.1. γ comme fonction de la dimension

Jusqu'à présent, nous nous sommes concentrés sur les régimes local et global. Théoriquement, le régime global survient lorsque $\sigma \uparrow \infty$. D'un point de vue pratique, cela n'est jamais réellement le cas, mais il est possible de s'en rapprocher lorsque la dimension du problème est faible. Cela est dû au fait que la variance instrumentale diminue à mesure que la dimension du problème augmente. Plus précisément, $\sigma_d^2 \propto d^{-\beta}$, où β est la mise à l'échelle. Tel que mentionné dans la section précédente et en considérant les simulations présentées plus loin, il faut utiliser une valeur de $\gamma = 2$ en petite dimension. Par opposition, le régime local survient en grande dimension lorsque $\sigma \downarrow 0$; tel qu'il a été vu au chapitre 2, il faut utiliser le MALA ($\gamma = 1$) dans cette situation.

Par conséquent, dans le régime intermédiaire, soit lorsque $0 < \sigma < \infty$, la seule contrainte à respecter est que le paramètre γ se rapproche de 1 à mesure que d augmente. Deux raisons justifient cette contrainte. Premièrement, cela permet d'obtenir la valeur de γ optimale en grande dimension. Deuxièmement, rappelons que nous avons pris la décision d'utiliser une mise à l'échelle $\beta = 1/3$ indépendamment du paramètre γ utilisé. Il a été argumenté au début du chapitre que cela n'avait pas d'impact en dimension finie. Or, cela n'est pas le cas dans un régime asymptotique si l'on utilise un paramètre $\gamma > 1$. En effet, la mise à l'échelle lorsqu'on considère $\gamma > 1$ est de $\beta = 1$, tel qu'il a été démontré dans le théorème 2.4.2. Ceci implique que d'utiliser une mise à l'échelle $\beta = 1/3 < 1$, en combinaison avec une valeur $\gamma > 1$, mène à un taux d'acceptation asymptotiquement nul. En utilisant une valeur de $\gamma = 1$

en grande dimension, ce scénario est évité, puisque la mise à l'échelle du MALA ($\gamma = 1$) est $\beta = 1/3$. Ainsi, tout comme pour la variance instrumentale, une mise à l'échelle est apposée au paramètre γ afin que celui-ci diminue lorsque la dimension augmente. Le paramètre γ prend donc la valeur 2 en faible dimension et converge vers 1 lorsque $d \uparrow \infty$. Ce nouveau paramètre γ_d, variant en fonction de la dimension, est défini par

$$\gamma_d = 1 + \frac{1}{d^{\lambda_d}}, \tag{3.3.1}$$

où $\lambda_d : \mathbb{N} \to \mathbb{R}$ est une fonction positive qui dépend de la dimension d.

Cette fonction remplit le même rôle que la mise à l'échelle β du paramètre σ^2 : elle régule la vitesse à laquelle le paramètre γ tend vers 1 en grande dimension. Contrairement au paramètre β qui est constant, la fonction λ_d peut varier selon la dimension. Cela permet d'obtenir une plus grande flexibilité et, tel qu'il sera vu plus loin, s'accorde mieux avec les résultats obtenus par simulations. Bien que nous ayons défini γ_d de sorte à décroître lorsque d augmente, nous devons maintenant nous assurer que cette diminution soit assez rapide. Prenons l'exemple trivial où $\lambda_d = 0$ pour tout d, c'est-à-dire où il n'y a pas de décroissance. Même s'il s'agit d'une fonction λ_d admissible, il est clair que $\lim_{d\to\infty} \mathbb{E}[\alpha(x, y)] = 0$ dans ce cas. Ainsi, seules certaines fonctions λ_d feront en sorte que le taux d'acceptation asymptotique ne converge pas vers 0 lorsque $d \to \infty$. Le théorème suivant identifie ces fonctions. La démonstration suit les mêmes étapes que celle du théorème 2.4.2 (et donc du théorème 1 de Beskos & Stuart [**1**]), mais comporte plus d'étapes. La raison est que l'on doit considérer une expansion de Taylor de plus grand ordre étant donné que la mise à l'échelle vaut $\beta = 1/3$, alors qu'elle n'était que de 1 lors de la démonstration du théorème 2.4.2.

Théorème 3.3.1. *Soit une chaîne de Markov à espace d'états $\mathcal{X}$ générée par un algorithme MH avec densité instrumentale q_σ (2.3.3) utilisant un paramètre γ_d du type (3.3.1) et une mise à l'échelle $\beta = 1/3$. On suppose que cette chaîne a atteint la stationnarité et que π, la densité cible, possède des composantes iid et est telle que $\pi(x) = \prod_{i=1}^{d} f(x_i) = \prod_{i=1}^{d} \exp\{l(x_i)\}$, où $l(x) = \log f(x)$. Les conditions suivantes sont apposées sur f :*

(1) tous les moments de f sont bornés ;

(2) $l(x)$ fait partie de la classe C^8 ;

(3) l et ses 8 premières dérivées sont bornées par un polynôme, c'est-à-dire

$$|l(x)|, |l^{(i)}(x)| \leqslant M(x), \qquad 1 \leqslant i \leqslant 8,$$

où $M(\cdot)$ est un polynôme.

Alors $\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] > 0$ *si et seulement si* $\lim_{d\to\infty} \lambda_d \geqslant 1/3$.

Démonstration. Afin de garder cette démonstration succincte, certains calculs de routine se retrouvent à la section A.2 de l'annexe, notamment les développements de Taylor de même que le code MATHEMATICA permettant d'obtenir ceux-ci. On utilisera également les lemmes A.2.1 et A.2.2 dont les démonstrations se retrouvent dans la même section. Soit X, l'élément actuel de la chaîne à l'état x. Le candidat y est accepté avec probabilité $\alpha(x,y) = \min\{1, e^{R_d}\}$. Puisque les composantes de π et de q_σ sont indépendantes, R_d vaut

$$R_d = \log\left(\frac{\pi(y)q_\sigma(y,x)}{\pi(x)q_\sigma(x,y)}\right) = \sum_{i=1}^{d} \log\left(\frac{f(y_i)q_\sigma^*(y_i,x_i)}{f(x_i)q_\sigma^*(x_i,y_i)}\right),$$

où q_σ^* est la version unidimensionnelle de la densité instrumentale (2.3.3). La composante i du candidat Y proposé par cette densité est

$$Y_i = x_i + \frac{\gamma_d \sigma_d^2}{2} l'(x_i) + \sigma_d Z_i, \quad i = 1, \ldots, d,$$

avec $Z_i \sim \mathcal{N}(0,1)$ indépendant de $X_i \sim f$, puisque la stationnarité de la chaîne est atteinte. De plus, les paramètres γ_d et σ_d dépendent de la dimension d. Tout comme pour la démonstration du théorème 2.4.2, on considère ici un développement de Taylor de R_d centré en $\sigma_d = 0$. Celui-ci comporte plus de termes que le développement considéré dans la démonstration du théorème 2.4.2 étant donné que la mise à l'échelle de σ_d^2 est maintenant de $1/3$, et non de 1. On notera donc $\sigma_d^2 = \ell/d^{1/3}$ où $0 < \ell < \infty$.

($\Rightarrow$) Selon le lemme 2.4.1 (i), si $\lim_{d\to\infty} \mathbb{E}[|R_d|] < \infty$, alors $\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] > 0$. Un développement de Taylor d'ordre 6 de R_d évaluée en $\sigma_d = 0$ donne

$$R_d = \sum_{i=1}^{6} \mathcal{A}_{i,d} + U_d,$$

où pour $i = 1, \ldots, 6$, on a

$$\mathcal{A}_{i,d} = \frac{\sigma_d^i}{i!} R^{(i)}(0) = \frac{\sigma_d^i}{i!} \sum_{j=1}^{d} C_{i,j} \qquad \text{et} \qquad U_d = \frac{\sigma_d^7}{7!} R^{(7)}(\sigma^*) = \frac{\sigma_d^7}{7!} \sum_{j=1}^{d} U_{j,d}(x_j, Z_j, \sigma_j^*),$$

avec $\sigma_j^* \in [0, \sigma_d]$, $j = 1, \ldots, d$. Les valeurs de $C_{i,j}$, $i = 1, \ldots, 6$, $j = 1, \ldots, d$ sont détaillées dans l'annexe A.2. Étant donné que $|R_d| \leqslant \sum_{i=1}^{6} |\mathcal{A}_{i,d}| + |U_d|$, on considère la limite de l'espérance de chaque terme de la somme afin de vérifier qu'elles sont bien bornées. Considérons d'abord le terme résiduel U_d. Les termes $U_{j,d}(x_j, Z_j, \sigma_j^*)$, $j = 1, \ldots, d$ sont des polynômes qui sont fonction de Z_j, des dérivées de l et de puissances positives de σ_j^*. Ainsi, par la condition (3), il existe des polynômes M_1, M_2, M_3, tels que

$$|U_{j,d}(x_j, Z_j, \sigma_j^*)| \leqslant M_1(x_j) M_2(Z_j) M_3(\sigma_j^*), \qquad j = 1, \ldots, d.$$

Tous les moments de f et de la loi normale sont bornés et puisque X_j et Z_j sont indépendants, cela implique que $\mathbb{E}[M_1(X_j)M_2(Z_j)] < \infty$, et ce, pour tout j. De plus, puisque $\sigma_j^* \leqslant \sigma_d \leqslant \ell$, cela implique que $M_3(\sigma_j^*) \leqslant K < \infty$, où K est une constante indépendante de j et de d. Il est alors possible de trouver une borne K_0 indépendante de d pour $\mathbb{E}[|U_{j,d}(x_j, Z_j, \sigma_j^*)|]$ et donc

$$\lim_{d\to\infty} \mathbb{E}[|U_d|] \leqslant \lim_{d\to\infty} \frac{d\sigma_d^7}{7!} K_0 = 0,$$

puisque $\sigma_d^2 \propto d^{-1/3}$. Notons également que $\mathbb{E}[|U_d|]$ est $\mathcal{O}(d^{-1/6})$ dans cette situation. On considère maintenant les termes $\mathcal{A}_{i,d}$, $i = 1, \ldots, 5$. Par l'inégalité de Jensen, on a

$$\mathbb{E}[|\mathcal{A}_{i,d}|] = \mathbb{E}\left[\sqrt{\mathcal{A}_{i,d}^2}\right] \leqslant \mathbb{E}[\mathcal{A}_{i,d}^2]^{1/2} = \frac{\sigma_d^i}{i!} \mathbb{E}\left[\left(\sum_{j=1}^{d} C_{i,j}\right)^2\right]^{1/2}. \tag{3.3.2}$$

Les variables $\{C_{i,j}\}$, $j = 1, \ldots, d$ sont des variables iid puisqu'elles dépendent des Z_j et X_j qui sont iid. Ainsi, (3.3.2) devient

$$\begin{aligned}
\mathbb{E}[|\mathcal{A}_{i,d}|] &\leqslant \frac{\sigma_d^i}{i!} \left\{ d\mathbb{E}[C_{i,1}^2] + \sum_{j=1}^{d} \sum_{k\neq j} \mathbb{E}[C_{i,j} C_{i,k}] \right\}^{1/2} \\
&= \frac{\sigma_d^i}{i!} \left\{ d\mathbb{E}[C_{i,1}^2] + d(d-1)\mathbb{E}[C_{i,1}]^2 \right\}^{1/2} \\
&\leqslant \frac{\sigma_d^i}{i!} \left\{ d\mathbb{E}[C_{i,1}^2] + d^2\mathbb{E}[C_{i,1}]^2 \right\}^{1/2} \\
&\leqslant \frac{\sigma_d^i}{i!} \left\{ \sqrt{d}\mathbb{E}[C_{i,1}^2]^{1/2} + d|\mathbb{E}[C_{i,1}]| \right\}.
\end{aligned} \tag{3.3.3}$$

Le lemme A.2.1 stipule que

$$\mathbb{E}[C_{1,1}] = \mathbb{E}[C_{3,1}] = \mathbb{E}[C_{5,1}] = 0.$$

Ainsi, pour $i = 1, 3, 5$, il est possible de borner (3.3.3) par

$$\mathbb{E}[|A_{i,d}|] \leqslant \frac{\sqrt{d}\sigma_d^i}{i!}\mathbb{E}[C_{i,1}^2]^{1/2} = d^{\frac{1}{2}-\frac{i}{6}}\frac{\ell^{\frac{i}{2}}}{i!}\mathbb{E}[C_{i,1}^2]^{1/2}.$$

Selon le lemme A.2.2, on a $\lim\limits_{d\to\infty}\mathbb{E}[C_{i,1}^2] < \infty$ pour $i = 1, \ldots, 6$. Ainsi, pour $i = 3, 5$, il suit directement que $\lim\limits_{d\to\infty}\mathbb{E}[|\mathcal{A}_{i,d}|] < \infty$. Pour $i = 1$, on utilise le lemme A.2.2 qui affirme que $\mathbb{E}[C_{1,1}^2] = (\gamma_d - 1)^2 K_1$, où $K_1 < \infty$ est indépendant de d. En utilisant cela et le fait que $\sigma_d = \sqrt{\ell}/d^{1/6}$, il s'avère que

$$\mathbb{E}[|\mathcal{A}_{1,d}|] \leqslant \sqrt{d}\sigma_d\mathbb{E}[C_{1,1}^2]^{1/2} = \sqrt{d}\sigma_d(\gamma_d - 1)K_1^{1/2} = d^{\frac{1}{3}-\lambda_d}\sqrt{\ell K_1}.$$

Ainsi,

$$\lim_{d\to\infty}\mathbb{E}[|\mathcal{A}_{1,d}|] \leqslant \begin{cases} \sqrt{\ell K_1} < \infty & \text{si} \quad \lim\limits_{d\to\infty}\lambda_d = 1/3, \\ 0 & \text{si} \quad \lim\limits_{d\to\infty}\lambda_d > 1/3. \end{cases}$$

Lorsque $i = 2$, selon le lemme A.2.2, $\mathbb{E}[C_{2,1}^2] = (\gamma_d - 1)^2 K_{2,d}$, où $\lim\limits_{d\to\infty} K_{2,d} = L_2 < \infty$. En utilisant cela, la limite du premier terme de (3.3.3) devient

$$\frac{\sigma_d^2\sqrt{d}}{2!}\mathbb{E}[C_{2,1}^2]^{1/2} = \frac{\sigma_d^2\sqrt{d}}{2!}(\gamma_d - 1)\sqrt{K_{2,d}} = d^{\frac{1}{6}-\lambda_d}\frac{\ell}{2!}\sqrt{K_{2,d}} \xrightarrow{d\uparrow\infty} 0,$$

puisque $\lim\limits_{d\to\infty}\lambda_d \geqslant 1/3$. Pour le second terme de (3.3.3), on utilise le lemme A.2.1. Celui-ci indique que $\mathbb{E}[C_{2,1}] = -(\gamma_d - 1)^2\mathbb{E}[l'(X_1)^2]$. De ce fait, on a

$$\frac{\sigma_d^2}{2!}d|\mathbb{E}[C_{2,1}]| \leqslant \frac{\sigma_d^2}{2!}d(\gamma_d - 1)^2\mathbb{E}[l'(X_1)^2] = d^{\frac{2}{3}-2\lambda_d}\frac{\ell}{2!}\mathbb{E}[l'(X_1)^2].$$

Ainsi, par les conditions (3) et (1), on a $\mathbb{E}[l'(X_1)^2] < \infty$ de sorte que

$$\lim_{d\to\infty}\mathbb{E}[|\mathcal{A}_{2,d}|] \leqslant \begin{cases} \frac{\ell}{2}\mathbb{E}[l'(X_1)^2] < \infty & \text{si} \quad \lim\limits_{d\to\infty}\lambda_d = 1/3, \\ 0 & \text{si} \quad \lim\limits_{d\to\infty}\lambda_d > 1/3. \end{cases}$$

Pour $i = 4$, puisque $\lim\limits_{d\to\infty}\mathbb{E}[C_{4,1}^2] < \infty$ selon le lemme A.2.2, la limite du premier terme de (3.3.3) vaut

$$\frac{\sigma_d^4}{4!}\sqrt{d}\mathbb{E}[C_{4,1}^2]^{1/2} = d^{-\frac{1}{6}}\frac{\ell^2}{4!}\mathbb{E}[C_{4,1}^2]^{1/2} \xrightarrow{d\uparrow\infty} 0.$$

Pour le deuxième terme de (3.3.3), on utilise le lemme A.2.1. Selon celui-ci, pour le terme $C_{4,1}$, on a $\mathbb{E}[C_{4,1}] = (\gamma_d - 1)K_{4,d}$ où $\lim\limits_{d\to\infty} K_{4,d} = L_4 < \infty$. Le deuxième terme de (3.3.3)

devient donc

$$\frac{\sigma_d^4}{4!}d|\mathbb{E}[C_{4,1}]| \leqslant \frac{\sigma_d^4}{4!}d(\gamma_d - 1)|K_{4,d}| = d^{\frac{1}{3}-\lambda_d}\frac{\ell^2}{4!}|K_{4,d}|.$$

Ainsi,

$$\lim_{d\to\infty}\mathbb{E}[|\mathcal{A}_{4,d}|] \leqslant \begin{cases} \frac{\ell^2}{4!}|L_4| < \infty & \text{si} \quad \lim\limits_{d\to\infty}\lambda_d = 1/3, \\ 0 & \text{si} \quad \lim\limits_{d\to\infty}\lambda_d > 1/3. \end{cases}$$

Enfin, pour $|\mathcal{A}_{6,d}|$, l'inégalité du triangle puis celle de Jensen permettent d'écrire

$$\lim_{d\to\infty}\mathbb{E}[|\mathcal{A}_{6,d}|] \leqslant \lim_{d\to\infty}\frac{d\sigma_d^6}{6!}\mathbb{E}[|C_{6,1}|] = \lim_{d\to\infty}\frac{\ell^3}{6!}\mathbb{E}[|C_{6,1}|] \leqslant \lim_{d\to\infty}\frac{\ell^3}{6!}\mathbb{E}[C_{6,1}^2]^{1/2} < \infty,$$

qui est bien borné puisque $\lim\limits_{d\to\infty}\mathbb{E}[C_{6,1}^2] < \infty$ selon le lemme A.2.2. Ainsi, puisque $\lim\limits_{d\to\infty}\mathbb{E}[|R_d|] \leqslant \lim\limits_{d\to\infty}\sum\limits_{i=1}^{6}\mathbb{E}[|\mathcal{A}_{i,d}|] + \mathbb{E}[|U_d|] < \infty$, par le lemme 2.4.1 (i), on conclut que

$$\lim_{d\to\infty}\mathbb{E}[\alpha(X,Y)] > 0.$$

($\Leftarrow$) Le lemme 2.4.1 (ii) implique que $\lim\limits_{d\to\infty}\mathbb{E}[\alpha(X,Y)] = 0$ lorsque

$$\mathbb{E}[R_d] \xrightarrow{d\uparrow\infty} -\infty \qquad \text{et} \qquad \frac{\mathbb{E}[|R_d - \mathbb{E}[R_d]|]}{-\mathbb{E}[R_d]} \xrightarrow{d\uparrow\infty} 0. \tag{3.3.4}$$

Nous désirons montrer que ceci survient lorsque $\lim\limits_{d\to\infty}\lambda_d < 1/3$. Pour la première limite, puisque $\mathbb{E}[C_{1,1}] = 0$, le terme dominant dans le développement de $\mathbb{E}[R_d]$ est $\mathbb{E}[\mathcal{A}_{2,d}]$. De plus, en utilisant le lemme A.2.1 pour $\mathbb{E}[C_{2,1}]$, il s'avère que

$$\mathbb{E}[\mathcal{A}_{2,d}] = \frac{d\sigma_d^2}{2!}\mathbb{E}[C_{2,1}] = \frac{-(\gamma_d-1)^2 d\sigma_d^2}{2!}\mathbb{E}[l'(X_1)^2] = -d^{\frac{2}{3}-2\lambda_d}\frac{\ell}{2!}\mathbb{E}[l'(X_1)^2] < 0.$$

Il suit que $\mathbb{E}[R_d] \to -\infty$ au rythme de $d^{\frac{2}{3}-2\lambda_d}$. En ce qui concerne le ratio en (3.3.4), on désire montrer que le numérateur ne croît pas plus rapidement que le dénominateur. Tel qu'il a été vu lors de la démonstration du théorème 2.4.2 (voir (2.4.9)), il est possible de borner $\mathbb{E}[|R_d - \mathbb{E}[R_d]|]$ par

$$\begin{aligned}
\mathbb{E}[|R_d - \mathbb{E}[R_d]|] &\leqslant \sum_{i=1}^{6}\frac{\sigma_d^i}{i!}\sqrt{d}\mathbb{V}(C_{i,1})^{1/2} + \mathbb{E}[|U_d - \mathbb{E}[U_d]|] \\
&= \sum_{i=1}^{6}\frac{\ell^{\frac{i}{2}}}{i!}d^{\frac{1}{2}-\frac{i}{6}}\mathbb{V}(C_{i,1})^{1/2} + \mathcal{O}(d^{-1/6}) \\
&= \sqrt{\ell}d^{\frac{1}{3}}\mathbb{E}[C_{1,1}^2]^{1/2} + \frac{\ell}{2}d^{\frac{1}{6}}(\mathbb{E}[C_{2,1}^2] - E[C_{2,1}]^2)^{1/2} + \mathcal{O}(1) + \mathcal{O}(d^{-1/6}),
\end{aligned} \tag{3.3.5}$$

où $\mathbb{E}[|U_d - \mathbb{E}[U_d]|]$ est $\mathcal{O}(d^{-1/6})$ tel que vu dans la première partie. De plus, les éléments $i \geqslant 3$ de la somme sont $\mathcal{O}(1)$ puisque selon les lemmes A.2.1 et A.2.2, $\lim\limits_{d\to\infty} \mathbb{E}[C_{i,1}] < \infty$ et $\lim\limits_{d\to\infty} \mathbb{E}[C_{i,1}^2] < \infty$ pour $i = 1, \ldots, 6$. En utilisant ces deux lemmes, il est possible de simplifier l'expression (3.3.5) et d'obtenir

$$\begin{aligned}\mathbb{E}[|R_d - \mathbb{E}[R_d]|] &\leqslant \sqrt{\ell} d^{\frac{1}{3}}(\gamma_d - 1)K_1^{1/2} + \frac{\ell}{2} d^{\frac{1}{6}}((\gamma_d - 1)^2 K_{2,d} - (\gamma_d - 1)^4 \mathbb{E}[l'(X_1)^2]^2)^{1/2} \\ &\quad + \mathcal{O}(1) + \mathcal{O}(d^{-1/6}) \\ &= d^{\frac{1}{3}-\lambda_d}\sqrt{\ell K_1} + d^{\frac{1}{6}-\lambda_d}\frac{\ell}{2}(K_{2,d} - (\gamma_d - 1)^2 \mathbb{E}[l'(X_1)^2]^2)^{1/2} + \mathcal{O}(1) + \mathcal{O}(d^{-1/6}).\end{aligned}$$

Par conséquent, $E[|R_d - \mathbb{E}[R_d]|]$ est $\mathcal{O}(d^{\frac{1}{3}-\lambda_d})$ tandis que $\mathbb{E}[R_d]$ est $\mathcal{O}(d^{\frac{2}{3}-2\lambda_d})$. Le ratio de ces deux quantités est alors $\mathcal{O}(d^{\lambda_d - \frac{1}{3}})$ et va converger vers 0 à mesure que $d \uparrow \infty$ étant donné que $\lim\limits_{d\to\infty} \lambda_d < 1/3$. Ainsi, par le lemme 2.4.1 (ii), on conclut que

$$\lim_{d\to\infty} \mathbb{E}[\alpha(X, Y)] = 0.$$

□

Le théorème 3.3.1 pose les conditions nécessaires permettant d'obtenir un algorithme utilisable en pratique lorsqu'on utilise un facteur γ_d qui décroît avec la dimension. Ainsi, peu importe à quelle vitesse le paramètre γ_d décroît en dimension finie, il faut que la vitesse devienne assez rapide à mesure que la dimension augmente. Remarquons que si la limite de λ_d est $1/3$, la décroissance de γ_d en grande dimension sera la même que celle de σ_d^2. De plus, bien que le théorème nous indique seulement une borne inférieure pour le taux de décroissance asymptotique, un taux de $1/3$ résultera intuitivement en un algorithme plus efficace. En effet, plus la valeur de γ reste grande et plus les pas effectués auront tendance à être grands, améliorant ainsi l'exploration de l'espace d'états. En prenant $1/3$ comme limite de λ_d, il s'agit de la décroissance la moins rapide donnant un taux d'acceptation asymptotique non nul.

Comme la condition sur la fonction déterminant le rythme de décroissance λ_d ne concerne que sa limite, cela laisse beaucoup de choix disponibles. Une fonction pourrait faire décroître γ_d de manière uniforme en étant constante ou bien avoir une décroissance très lente en faible dimension, et ensuite décroitre de façon plus rapide. Ainsi, il est difficile de déterminer parmi ces fonctions très différentes laquelle pourrait être la meilleure. Néanmoins un utilisateur des

MCMC aimerait pouvoir déterminer, pour une densité cible de dimension d donnée, quelle doit être la valeur de γ à utiliser afin d'obtenir un algorithme qui soit le plus efficace possible. On aimerait donc déterminer quelle fonction λ_d maximise la DSM selon la cible utilisée. Or, cette fonction optimale risque de dépendre de la cible elle-même, ce qui complique donc la recherche d'une telle fonction. Des simulations présentées à la section 3.4 permettront d'éclaircir cette question. Par contre, il reste un paramètre dont la valeur optimale n'a pas encore été déterminée, soit le paramètre σ.

3.3.2. Choix de σ

Bien que les sections précédentes nous aient donné une idée de la valeur optimale du paramètre γ, dans la pratique, il reste à déterminer le paramètre σ. Cette question n'a pas été abordée ici. Dans la littérature, des valeurs de σ optimales ont été trouvées presqu'exclusivement pour des régimes asymptotiques. La valeur optimale de σ a rarement été déterminée pour un régime en dimension finie. En pratique, la valeur de σ utilisée est la valeur du régime asymptotique, et ce, peu importe la dimension. Pour déterminer une valeur de σ optimale, il faudrait maximiser la DSM, ce qui n'est pas aisé comme il a été mentionné précédemment. Néanmoins, les applications numériques présentées au prochain chapitre pourraient nous donner une idée de la valeur optimale du paramètre σ.

3.4. Étude de simulations

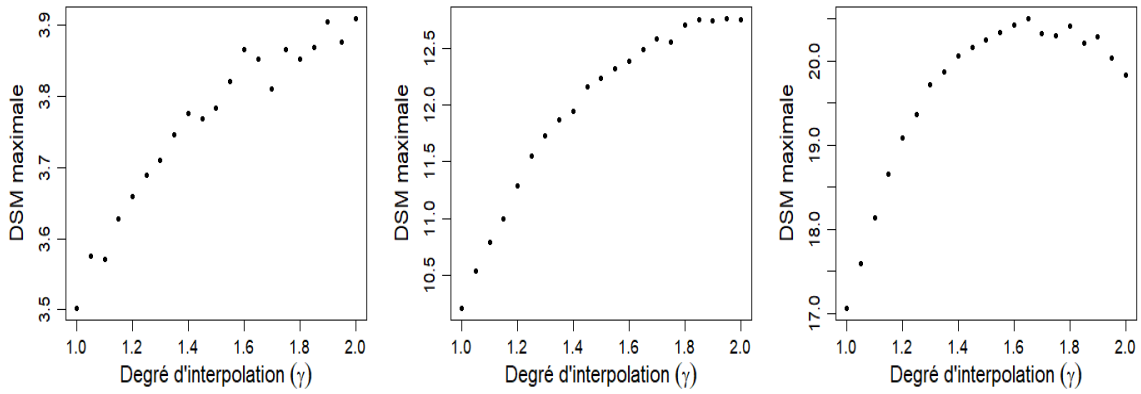
Cette étude a trois objectifs. Tout d'abord, elle sert à illustrer les différents résultats théoriques obtenus jusqu'à présent. Deuxièment, on désire introduire des lignes directrices permettant d'identifier une fonction λ_d optimale. Troisièmement, on veut déterminer quelles dimensions correspondent au régime global, intermédiaire et local en pratique. Tel que vu dans [**18**], lorsque le MALA est utilisé, le régime asymptotique survient assez rapidement, soit autour de $d = 20$. Ceci signifie qu'il est possible d'utiliser la variance asymptotiquement optimale, soit celle de la forme $\sigma^2 = \ell/d^{1/3}$ avec ℓ donnant un taux d'acceptation de 57%, en dimension aussi faible que 20. Les exemples présentés ici cherchent à vérifier si ce régime survient aussi rapidement avec le paramètre γ_d. Dans notre contexte, le régime asymptotique survient lorsque $\gamma = 1$ devient la valeur optimale en termes de DSM.

On désire générer un échantillon de taille 200 000 à l'aide du MH en utilisant la densité instrumentale (2.3.3). On considère deux densités cibles, chacune étudiée selon plusieurs dimensions : $d = 1, 5, 15, 25, 35, 50, 100, 300, 500$. Pour chacune d'entre elles, on désire identifier la valeur de γ qui maximise la DSM. Pour ce faire, pour chaque dimension, on applique l'algorithme en utilisant des valeurs de $\gamma \in \{1; 1{,}05; \ldots; 2\}$ et $\sigma_d^2 \in \{0{,}5/d^{1/3}; 0{,}7/d^{1/3}; \ldots; 6/d^{1/3}\}$. On considère plusieurs valeurs pour le paramètre σ_d^2, puisque la valeur optimale pour ce paramètre est pour l'instant inconnue. De ce fait, pour une valeur de γ donnée, la DSM maximale sera celle obtenue parmi toutes les valeurs de σ_d^2 considérées. En observant l'évolution de la valeur de γ optimale selon la dimension, il sera possible d'identifier les propriétés que possède la fonction λ_d optimale. La première distribution considérée est un mélange de deux normales avec composantes iid, c'est-à-dire

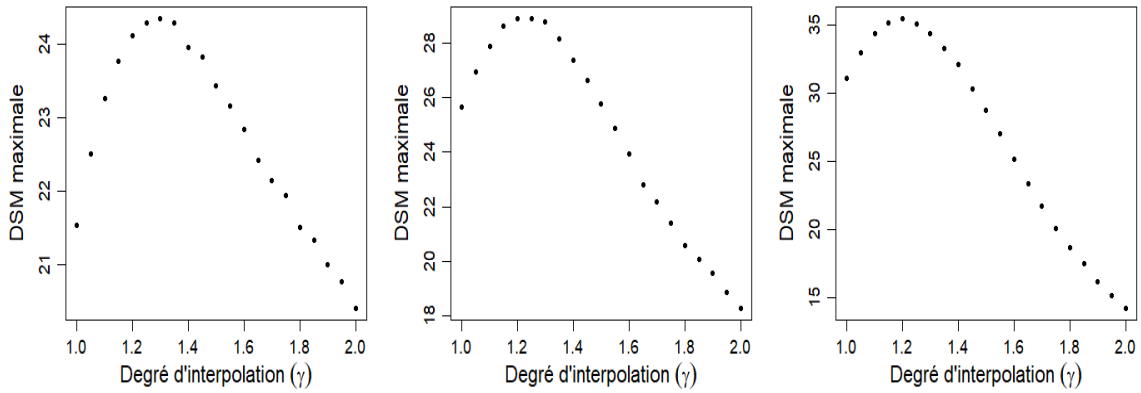
$$X \sim \begin{cases} \mathcal{N}_d(\mu_1, I_d) & \text{avec probabilité } 3/5, \\ \mathcal{N}_d(\mu_2, I_d) & \text{avec probabilité } 2/5, \end{cases} \tag{3.4.1}$$

où $\mu_1 = (1, \ldots, 1)^\top \in \mathbb{R}^d$, $\mu_2 = -\mu_1$ et I_d est la matrice identité de dimension d. Le choix de valeurs pour μ_1 et l'utilisation d'une variance de 1 pour chaque composante fait en sorte que les modes sont assez rapprochés. Cela est important puisqu'en général, les algorithmes du type MH sont peu performants sur des densités ayant des modes isolés (voir [**7**]). Les résultats pour cette première densité cible sont présentés à la figure 3.1.

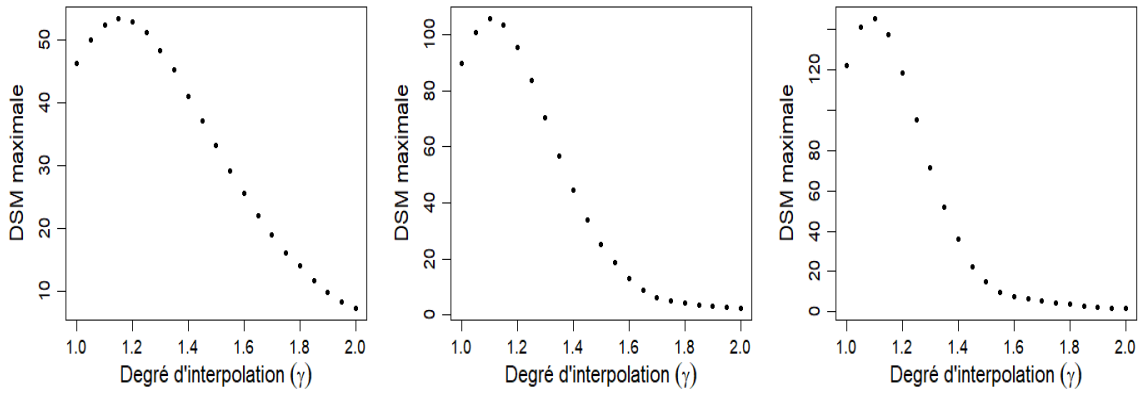
Il est possible de voir que le régime global survient lorsque la dimension du problème est inférieure à 5. En deçà de $d = 5$, il est clair que l'utilisation d'une valeur de $\gamma < 2$ est moins efficace en termes de DSM. Ainsi, même s'il n'a pas été possible de démontrer que l'équilibre global est plus efficace que l'équilibre local en faible dimension, il semble que ce soit empiriquement le cas ici. Le régime intermédiaire, semble débuter après $d = 5$ alors que la valeur de γ optimale est déjà 1,3 lorsque $d = 25$. Passé ce seuil, en plus grande dimension, la décroissance optimale semble être plus faible. En effet, à $d = 100$, le γ optimal vaut 1,15. De ce fait, la décroissance optimale n'est pas uniforme pour toutes les dimensions. Le rythme de cette décroissance est plus rapide en faible dimension et ralentit à mesure que la dimension augmente. Le paramètre λ_d ne semble donc pas être une constante. En ce qui concerne le régime asymptotique, celui-ci survient en très grande dimension. Même lorsque



(a) $d = 1$ **(b)** $d = 5$ **(c)** $d = 15$



(d) $d = 25$ **(e)** $d = 35$ **(f)** $d = 50$



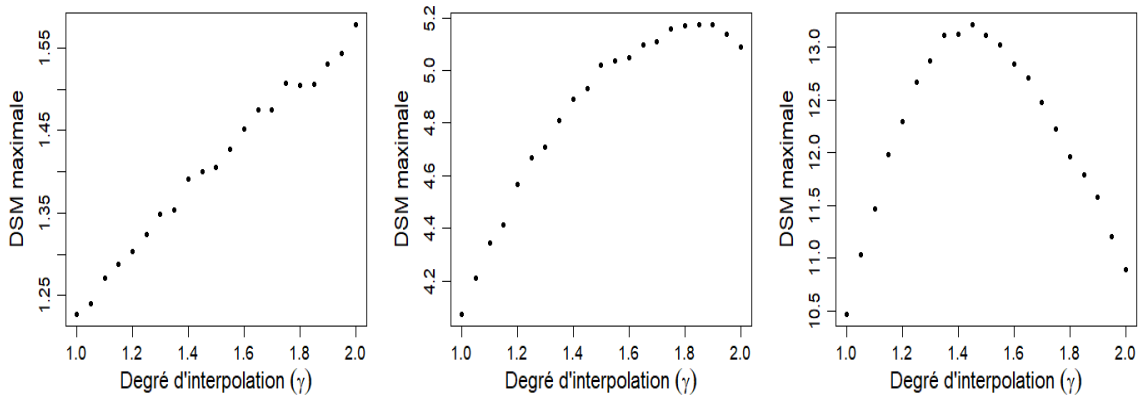
(g) $d = 100$ **(h)** $d = 300$ **(i)** $d = 500$

Figure 3.1. DSM maximale obtenue en fonction du paramètre γ selon plusieurs dimensions sur la densité cible (3.4.1).

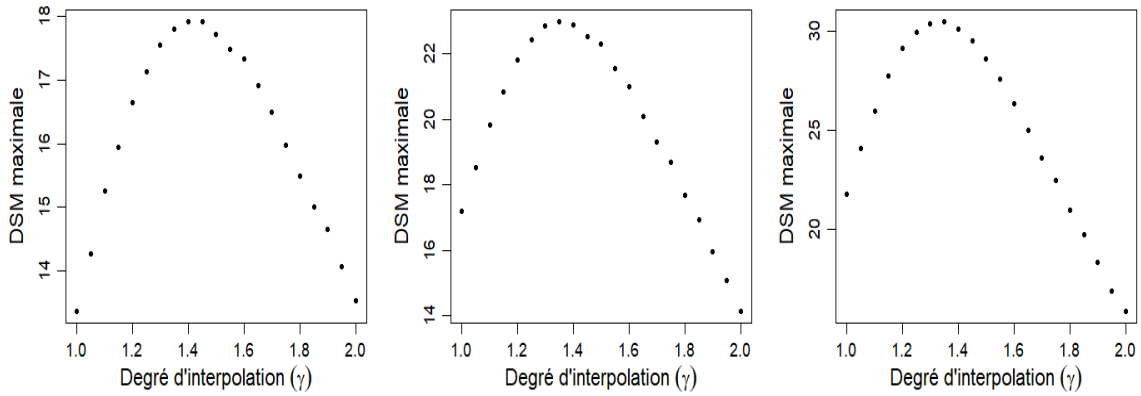
$d = 500$, ce régime n'est pas encore tout à fait atteint puisque la valeur de γ optimale est supérieure à 1. Il s'agit d'une grande différence par rapport à [**18**] où le régime asymptotique survient rapidement, soit autour de $d = 20$. De ce fait, l'utilisation d'une valeur de $\gamma > 1$ est justifiée même pour une dimension relativement grande. De plus, on remarque qu'à mesure que la dimension augmente, les valeurs de γ proches de 2 sont associées à des DSM presque nulles. Cela n'est pas surprenant puisque dans cette situation, le taux d'acceptation tend vers 0, comme vu au chapitre 2 (voir théorème 2.4.2).

L'autre densité cible considérée est une densité $\mathcal{N}_d(\mathbf{0}, A_d)$ où $\mathbf{0} = (0, \ldots, 0)^\top$ et $A_d = \text{diag}(\tau_1, \ldots, \tau_d)$. Ici, pour $i = 1, \ldots, d$, $\tau_i \sim \mathcal{U}(0.5, 2)$. Cela implique que les composantes sont indépendantes, mais ne seront pas identiquement distribuées. On choisit d'utiliser une telle densité au lieu d'une densité normale avec composantes iid puisque si les composantes sont identiquement distribuées, la forme du gradient de la cible fait en sorte que l'on échantillonne directement de la densité cible en prenant $\gamma = 2$.

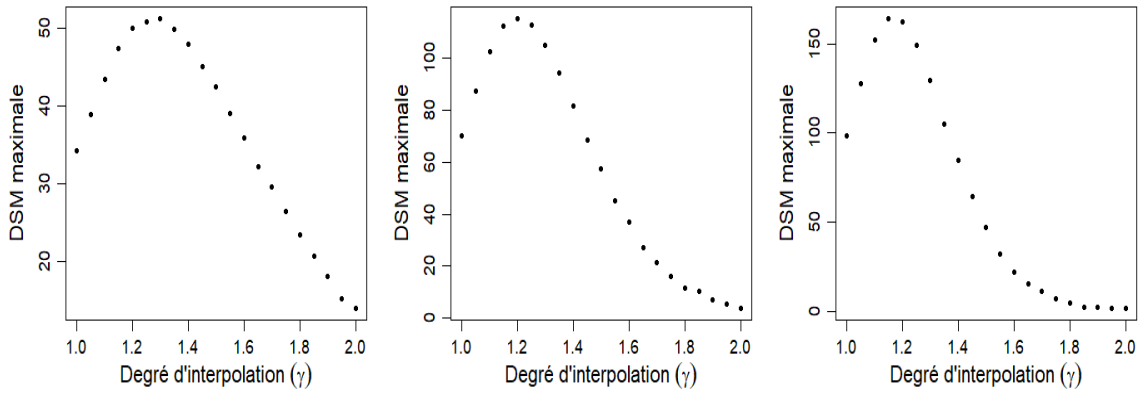
Les résultats associés à cette nouvelle densité cible sont présentés à la figure 3.2. Cette fois-ci, le régime global se termine en deçà de $d = 5$ et il est à nouveau plus efficace d'utiliser une valeur de $\gamma = 2$ en faible dimension. De plus, pour une même dimension, le γ optimal n'est pas le même que pour l'exemple précédent. Par exemple, le γ maximal à $d = 25$ est de 1,4 alors qu'il était de 1,3 pour la densité cible précédente. De ce fait, la décroissance optimale γ_d varie selon la densité cible choisie. Cela complique donc la recherche d'une valeur de γ optimale en dimension finie. On note également qu'au-delà de $d = 15$, tout comme pour le cas précédent, la décroissance devient beaucoup plus lente et à $d = 500$, le γ optimal est de 1,2. Ainsi, le rythme de décroissance semble suivre un schéma similaire à l'exemple précédent, c'est-à-dire qu'il est rapide au départ et lent par la suite. Par contre, les valeurs de γ optimales associées à chaque dimension sont différentes. La fonction λ_d exprimant la décroissance semble donc être la même, mais paramétrisée différemment en fonction de la cible utilisée. De plus, on remarque que le régime asymptotique n'est à nouveau pas tout à fait atteint pour cette cible, même en considérant $d = 500$.



(a) $d = 1$ **(b)** $d = 5$ **(c)** $d = 15$



(d) $d = 25$ **(e)** $d = 35$ **(f)** $d = 50$



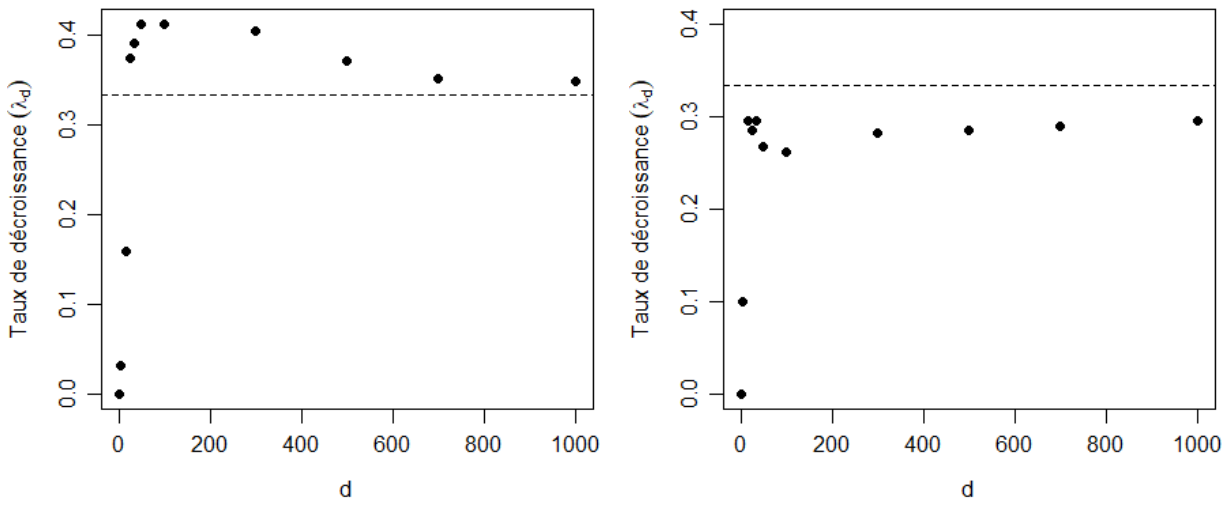
(g) $d = 100$ **(h)** $d = 300$ **(i)** $d = 500$

Figure 3.2. DSM maximale obtenue en fonction du paramètre γ selon plusieurs dimensions sur la densité cible $\mathcal{N}_d(\mathbf{0}, A_d)$.

Pour bien saisir l'évolution de la décroissance de γ_d (3.3.1), la valeur du taux de décroissance, λ_d, a été calculée selon les dimensions considérées. Pour bien visualiser la convergence de ce paramètre, nous avons effectué la même expérience pour $d = 700$ et $d = 1000$. Pour chaque dimension, on calcule la valeur de λ_d associée au paramètre γ_d optimal de la manière suivante :

$$\lambda_d = \frac{-\log(\gamma_d - 1)}{\log(d)}.$$

L'évolution de λ_d est représentée à la figure 3.3. Pour la loi bimodale, le taux de décroissance est supérieur à 1/3 à partir de $d = 25$ ce qui s'accorde avec les conclusions du théorème 3.3.1. De plus, le taux semble tendre vers cette limite à mesure que la dimension augmente. Bien que le taux optimal soit supérieur à 1/3 à partir de $d = 25$, choisir un taux de 1/3 ne diminue pas trop la DSM. Par exemple, à $d = 100$, si l'on choisit un taux de 1/3, la valeur de γ_d obtenue est de 1,21, ce qui n'occasionne pas une grande perte de DSM par rapport à la valeur optimale de 1,15, tel que vu à la figure 3.1. Ainsi, choisir une valeur de 1/3 en grande dimension semble être une approche légitime en pratique. Dans le cas de la normale, la densité ne respecte pas les conditions du théorème 3.3.1 puisque ses composantes ne sont pas identiquement distribuées. Le théorème 3.3.1 ne s'applique pas ici, ce que l'on peut voir



(a) Densité bimodale **(b)** Densité normale

Figure 3.3. Taux de décroissance du paramètre γ optimal en fonction de la dimension. La valeur de 1/3 est en pointillés.

puisque les taux en grande dimension sont inférieurs à 1/3. Malgré tout, ceux-ci ont tout de même tendance à tendre vers une limite en deçà de 1/3. On observe aussi que le rythme de décroissance optimal n'est pas constant pour toutes les dimensions, ce qui justifie l'emploi d'une fonction comme taux de décroissance au lieu d'une constante comme c'est le cas habituellement.

D'un point de vue plus quantitatif, il est également possible de s'intéresser à l'amélioration obtenue en utilisant la valeur de $\gamma > 1$ optimale par opposition au MALA. Cette amélioration est définie par

$$\% \text{ Amélioration} = \frac{DSM_{\gamma\, opt} - DSM_{MALA}}{DSM_{MALA}} \times 100\%. \tag{3.4.2}$$

L'amélioration selon chaque dimension pour les deux cibles considérées est présentée dans le tableau 3.1. On remarque que l'amélioration ne diminue pas nécessairement lorsque la valeur de γ diminue. Ainsi, même si la valeur de γ optimale est proche de 1 en grande

Tableau 3.1. Pourcentage d'amélioration de la DSM par rapport au MALA pour $\gamma > 1$ optimal

	Cible bimodale		Cible normale	
d	γ_{opt}	% Amélioration	γ_{opt}	% Amélioration
1	2	11,39	2	28,66
5	1,95	25,09	1,85	27,10
15	1,65	20,18	1,45	26,33
25	1,3	13,07	1,4	34,20
35	1,25	12,66	1,35	33,63
50	1,2	14,08	1,35	40,13
100	1,15	15,29	1,3	49,71
300	1,1	18,06	1,2	63,88
500	1,1	18,95	1,15	67,06
700	1,1	22,52	1,15	72,88
1000	1,1	25,04	1,15	72,39

dimension, l'amélioration par rapport au MALA est non négligeable. Par exemple, pour la cible normale, l'amélioration est au-delà de 50% à partir de $d = 100$. De plus, l'amélioration est plus grande pour la normale que pour la bimodale. Cela peut s'expliquer par le fait que la valeur de γ optimale pour la cible normale est plus grande pour une même dimension que la cible bimodale. De ce fait, les pas à chaque itération seront plus grands et il y aura un plus grand déplacement que le MALA.

Le tableau 3.1 peut donner une idée de l'amplitude des améliorations obtenues en ajoutant un paramètre γ et en choisissant judicieusement sa valeur. Or, il reste que la recherche de la fonction λ_d optimale est compliquée par le fait que cette fonction dépend de la cible considérée. Dans le prochain chapitre, une étude détaillée de l'efficacité de ce nouvel algorithme est réalisée afin de comparer sa performance à des algorithmes existants. De plus, cette efficacité est analysée dans un contexte où la paramétrisation n'est pas nécessairement optimale, étant donné que celle-ci est difficile à déterminer.

Chapitre 4

Application numérique

Les simulations présentées au chapitre 3 ont pu donner une idée de la performance du MALA en termes de DSM lorsqu'un paramètre $\gamma > 1$ est utilisé. Dans ce chapitre, on désire explorer plus en détails la performance de cet algorithme à travers d'autres critères dans un contexte d'application sur des données réelles. L'objectif est de comparer la performance de cette densité instrumentale par rapport à des algorithmes plus traditionnels comme le RWM et le MALA. Dans la deuxième partie de ce chapitre, on aborde la question du choix du paramètre σ à travers des simulations.

4.1. Régression logistique bayésienne

4.1.1. Contexte

Le problème considéré ici est celui de la régression logistique bayésienne. Seuls les éléments théoriques nécessaires à la compréhension du problème sont ici présentés, mais le lecteur intéressé peut se référer à [**4**] pour plus de détails. Dans ce contexte, on introduit une matrice de design, $\mathbf{X}$, de dimensions $n \times d$, où n est le nombre d'observations et d est la dimension des coefficients de régression $\beta = (\beta_0, \beta_1, \ldots, \beta_{d-1})$, qui sont les paramètres à estimer dans le modèle. La première colonne de $\mathbf{X}$ ne contient que des 1, tandis que les autres colonnes représentent les variables mesurées. La variable réponse est une variable binaire $\mathbf{t} \in \{0,1\}^n$; elle indique habituellement qu'un évènement d'intérêt s'est produit, comme la présence d'une maladie, par exemple. Un des objectifs de la régression logistique est de déterminer

$$p_i = \mathbb{P}(t_i = 1|\beta),$$

qui est la probabilité que l'évènement d'intérêt se réalise pour l'observation i, étant données les variables mesurées pour cette observation. Cette probabilité est modélisée grâce au modèle logit. Pour un individu $i = 1, \ldots, n$, en notant $\mathbf{X_i}$ la i-ème ligne de la matrice $\mathbf{X}$, le modèle indique que

$$p_i = \text{logit}^{-1}(\mathbf{X}_i\beta) = \frac{1}{1 + \text{e}^{-\mathbf{X}_i\beta}}.$$

Pour pouvoir calculer la probabilité d'intérêt, il est alors nécessaire d'estimer les paramètres β. Une approche bayésienne est ici utilisée. Par le théorème de Bayes, la densité *a posteriori* du paramètre β est

$$\pi(\beta|\mathbf{t}) \propto \pi(\beta) f(\mathbf{t}|\beta), \tag{4.1.1}$$

où $\pi(\cdot)$ est la densité *a priori* tandis que f est la vraisemblance. En supposant l'indépendance entre les observations, la vraisemblance vaut

$$f(\mathbf{t}|\beta) = \prod_{i=1}^{n} p_i^{t_i}(1 - p_i)^{1-t_i}.$$

Afin que la densité *a priori* soit la moins informative possible et ne biaise pas les résultats, une densité normale avec grande variance est choisie. Ainsi, à priori, chaque paramètre $\beta_j \sim \mathcal{N}(0, a)$, où $j = 1, \ldots, d$, avec $a = 100$. En raison de la forme de la vraisemblance donnée par le modèle logit et du choix de cette densité *a priori*, il est clair que la densité *a posteriori* (4.1.1) sera complexe. L'inférence bayésienne repose sur l'analyse de la densité *a posteriori*, notamment à travers le calcul de l'espérance *a posteriori* pour l'estimation ponctuelle. De ce fait, il sera donc difficile de travailler directement avec cette densité *a posteriori* afin de faire de l'inférence sur le paramètre β. Tel que mentionné dans l'introduction, c'est ici que les méthodes MCMC peuvent être utiles. En générant un échantillon provenant de la densité *a posteriori*, il sera possible d'avoir une idée de la forme que prend cette densité. Cet échantillon permettra de calculer, entre autres, des estimations ponctuelles des paramètres, de même que des intervalles de crédibilité. L'utilisation des méthodes MCMC présentées dans les chapitres précédents semble donc appropriée au contexte de la régression logistique bayésienne.

Tableau 4.1. Jeux de données analysés

Nom	Variables indépendantes	Observations (n)	Dimension de β (d)	Variable réponse
Pima Indian	7	532	8	Diabète (oui/non)
German Credit	24	1000	25	Crédit (bon/mauvais)
Australian Credit	14	690	15	Crédit (approuvé/refusé)
Heart	13	270	14	Maladie du coeur (oui/non)

4.1.2. Jeux de données, algorithmes et critères de comparaison

Les 4 jeux de données considérés sont ceux étudiés dans [**5**]. Ceux-ci sont présentés dans le tableau 4.1. Pour chaque jeu de données, on désire tirer un échantillon du paramètre β provenant de la densité *a posteriori* (4.1.1) dans le contexte de la régression logistique bayésienne. Le lecteur intéressé peut consulter [**13**] et [**16**] pour plus d'informations sur les variables indépendantes de chaque jeu de données et le contexte dans lequel ces variables ont été mesurées. Ces jeux de données présentent une belle variété de difficultés. La dimension de β varie de $d = 8$ à $d = 25$ et le nombre d'observations, de 270 à 1000. La dimension de β n'étant pas trop élevée pour tous les jeux de données, l'utilisation de la densité instrumentale (2.3.3) avec une valeur de $\gamma > 1$ semble donc justifiée ici. Cette intuition découle également des résultats des simulations présentées dans la dernière section du chapitre précédent, puisque le régime asymptotique survenait en très grande dimension, soit au-delà de 500. Afin d'éviter de potentiels problèmes numériques, les variables ont été normalisées de telle sorte qu'elles sont maintenant de moyenne 0 et d'écart-type 1. Les détails d'implémentation des sept algorithmes considérés sont maintenant présentés ci-dessous. Ceux-ci ont tous été implémentés avec le logiciel R 3.2.4 afin d'assurer une comparaison juste. Ces algorithmes sont :

(1) RWM ;

(2) MALA ;

(3) MALA avec paramètre $\gamma = (1{,}2; 1{,}4; 1{,}6; 1{,}8; 2)$ (densité instrumentale (2.3.3)).

RWM

L'algorithme RWM est choisi puisqu'il s'agit de l'algorithme de base le plus utilisé en pratique. Celui-ci est très simple à implémenter et ne requiert pas un grand coût computationnel. Cela implique que le temps de calcul par itération sera faible. Par contre, cet algorithme n'optimise pas les déplacements effectués, puisque la densité instrumentale utilisée est non-informative et la DSM risque donc d'être plus faible. Tel que mentionné au chapitre 1, afin d'obtenir un taux d'acceptation raisonnable, la variance instrumentale est posée comme étant $\sigma_d^2 = \ell/d$. La constante ℓ est déterminée par essai-erreur en choisissant celle permettant d'obtenir un taux d'acceptation d'environ 25%. Bien que ce critère ne soit valide que lorsque l'on considère un régime asymptotique, celui-ci est utilisé ici, même si la dimension du problème est faible. Cette méthode est souvent utilisée en pratique étant donné qu'il n'existe pas de méthode de sélection de ℓ pour les problèmes en faible dimension.

MALA

Le MALA utilisant une densité instrumentale informative, il devrait explorer plus rapidement l'espace d'états et la DSM devrait donc être supérieure à celle du RWM. Par contre, il est nécessaire de calculer le gradient de la densité *a posteriori* (4.1.1) à chaque itération, ce qui augmente le coût computationnel. Ainsi, il reste à vérifier que ce coût ne soit pas trop grand par rapport aux gains obtenus en ce qui concerne la rapidité de l'exploration. Les critères d'efficacité mesurés prendront en compte cet aspect. Tel que mentionné au chapitre 1, la variance instrumentale est de la forme $\sigma_d^2 = \ell/d^{1/3}$, où la valeur de ℓ est choisie de sorte à obtenir un taux d'acceptation d'environ 57%. À nouveau, ce critère n'est valide qu'asymptotiquement, mais il est ici utilisé en dimension finie.

MALA avec paramètre $\gamma > 1$

On désire comparer la performance du MALA traditionnel avec celle du MALA modifié, c'est-à-dire avec un paramètre $\gamma > 1$. Les simulations effectuées dans la dernière section du chapitre précédent ont montré que le choix optimal de γ selon le DSM dépend de la cible considérée. De ce fait, plusieurs valeurs de γ sont ici étudiées puisqu'il n'est pas possible de déterminer la valeur optimale. En considérant plusieurs valeurs de γ, on désire également vérifier à quel point une valeur éloignée de la valeur optimale de γ peut demeurer efficace face au MALA. En d'autres mots, on désire vérifier si une erreur dans le choix de γ est plus coûteuse que de simplement utiliser le MALA. La valeur de γ étant fixée, il ne reste

qu'à déterminer la valeur de σ. N'ayant pas étudié cet aspect, la valeur de σ est choisie similairement à celle du MALA. Spécifiquement, la variance instrumentale est de la forme $\sigma_d^2 = \ell/d^{1/3}$, où ℓ est choisi de sorte à obtenir un taux d'acceptation d'environ 57%. Dans la deuxième partie de ce chapitre, l'optimalité de ce choix de ℓ sera analysée pour les jeux de données Pima Indian et German Credit.

Lors de la génération de l'échantillon, une période de « burn-in » de 5 000 itérations est nécessaire. Tel que mentionné au chapitre 1, cela implique que ces 5 000 premières itérations ne seront pas considérées dans l'échantillon final. Cela assure que l'échantillon recueilli provient bien de la cible, étant donné que la chaîne générée a eu le temps d'atteindre la stationnarité. De manière similaire à [**5**], les 5 000 itérations suivantes formeront l'échantillon recherché. L'efficacité de chaque algorithme est mesurée à l'aide des critères suivants. Le premier est la distance de saut quadratique moyenne (DSM) telle que présentée au chapitre 3 (voir (3.1.1)). Le deuxième critère est la taille d'échantillon efficace (TEE). Pour chaque composante X_j de la chaîne générée, où $j = 1, \ldots, d$, celle-ci est définie par

$$TEE_j = \frac{N}{1 + 2\sum_{k=1}^{N} \mathrm{Corr}(X_j[0], X_j[k])}, \tag{4.1.2}$$

où N correspond à la taille de l'échantillon généré. Le dénominateur de la TEE correspond à l'autocorrélation intégrée (3.1.2) présentée au chapitre 3. On désire que cette autocorrélation soit minimisée (voir section 3.1), ce qui revient à maximiser la TEE. Tel que vu dans la section 3.1, maximiser la DSM contribue à minimiser l'autocorrélation de délai 1 uniquement. En observant le comportement de la TEE, il est possible de voir à quel point l'autocorrélation est grande pour des délais plus grands. De plus, la TEE peut être interprétée de manière intuitive. En effet, il s'agit de la taille d'un échantillon iid obtenu à partir de la cible qui contiendrait autant d'information que tout l'échantillon généré par la méthode MCMC. Étant donné que la TEE est différente pour chaque composante de la chaîne, afin de résumer l'information, la médiane des TEE est calculée. On utilise ce second critère afin d'aller chercher de l'information complémentaire au sujet de la performance des algorithmes ; cependant, le critère principal considéré est toujours la DSM, tel que justifié à la section 3.1. On mesure également le temps de calcul nécessaire pour l'exécution des algorithmes afin d'avoir une idée de la performance computationnelle. On considère aussi le ratio Temps/DSM afin d'obtenir

un critère combinant à la fois les performances computationnelle et théorique. Enfin, le pourcentage d'amélioration de la DSM par rapport à celui du MALA (3.4.2) est calculé. Tous ces critères sont obtenus en se basant sur la moyenne de 10 exécutions de chaque algorithme.

4.1.3. Résultats

Les résultats pour les quatre jeux de données considérés sont présentés dans les tableaux 4.2, 4.3, 4.4 et 4.5.

Tableau 4.2. Résultats pour le jeu de données Pima Indian ($n = 532, d = 8$)

Algorithme	DSM	Temps (s)	Temps/DSM	TEE	% Amélioration
RWM	0,01746	3,42	196,02	145,98	-78,93
MALA	0,08294	4,69	56,51	605,03	0
MALA ($\gamma = 1,2$)	0,09244	4,66	50,41	669,04	11,43
MALA ($\gamma = 1,4$)	0,09454	4,65	49,16	670,81	13,97
MALA ($\gamma = 1,6$)	0,09229	4,78	51,84	620,55	11,26
MALA ($\gamma = 1,8$)	0,08843	4,72	53,35	562,30	6,60
MALA ($\gamma = 2$)	0,07685	5,07	66,01	428,67	-7,35

Tout d'abord on remarque, sans surprise, que l'algorithme RWM est le plus rapide d'entre tous. Dans la majorité des cas, le MALA est environ 2 fois plus lent. Par contre, la performance du RWM par rapport au MALA en ce qui concerne la DSM et la TEE est très faible. Le ratio Temps/DSM est donc toujours en faveur du MALA. Ainsi, même si le RWM est plus rapide, il est justifié d'utiliser le MALA dans la situation présente. Le temps de calcul pour le MALA avec $\gamma > 1$ est très semblable à celui du MALA. Cela n'est pas surprenant puisqu'il s'agit pratiquement du même algorithme. La valeur optimale de γ, parmi celles considérées, est de 1,4, sauf pour le jeu de données Australian Credit où elle vaut 1,6. Encore une fois, on note que la valeur optimale de γ varie en fonction de la cible utilisée, qui dépend ici du jeu de données considéré. Néanmoins, une valeur d'environ 1,5 semble efficace dans tous les cas considérés.

Tableau 4.3. Résultats pour le jeu de données German Credit ($n = 1000, d = 25$)

Algorithme	DSM	Temps (s)	Temps/DSM	TEE	% Amélioration
RWM	0,00908	7,30	803,97	38,86	-80,66
MALA	0,07970	13,81	173,30	366,85	0
MALA ($\gamma = 1, 2$)	0,09578	13,26	138,46	416,23	20,17
MALA ($\gamma = 1, 4$)	0,09843	13,09	133,03	430,28	23,49
MALA ($\gamma = 1, 6$)	0,09091	13,39	147,29	370,28	14,06
MALA ($\gamma = 1, 8$)	0,01289	14,91	2789,40	273,75	-83,82
MALA ($\gamma = 2$)	0,00534	13,39	2504,49	28,84	-93,29

Tableau 4.4. Résultats pour le jeu de données Australian Credit ($n = 690, d = 15$)

Algorithme	DSM	Temps (s)	Temps/DSM	TEE	% Amélioration
RWM	0,02486	4,40	177,33	96,08	-85,71
MALA	0,17400	6,89	39,63	619,19	0
MALA ($\gamma = 1, 2$)	0,20678	7,46	36,08	755,53	18,83
MALA ($\gamma = 1, 4$)	0,22287	8,88	39,88	786,62	28,08
MALA ($\gamma = 1, 6$)	0,22561	7,93	35,17	721,33	29,66
MALA ($\gamma = 1, 8$)	0,21428	7,94	37,09	636,43	23,15
MALA ($\gamma = 2$)	0,18753	6,81	36,34	524,01	7,77

Choisir une valeur de γ trop grande est plus dramatique dans certains cas que d'autres. Pour le fichier Australian Credit, aucune valeur de γ n'occasionne de perte par rapport au MALA, ce qui peut s'expliquer par le fait que sa valeur de γ optimale est plus élevée que celle des autres jeux de données. Pour les fichiers Pima Indian et Heart, seule la valeur $\gamma = 2$ mène à des algorithmes moins efficaces ; les performances ne sont pas dramatiques, puisque les pertes ne sont que de 7% et 4% respectivement. Pour le jeu de données German Credit, la situation est plus grave. La performance est pire que le MALA lorsque $\gamma = 1, 8$ et $\gamma = 2$. Dans ce dernier cas, la performance de l'algorithme est même inférieure à celle du RWM. Cela peut s'expliquer par le fait que German Credit contient le plus grand nombre de variables parmi tous les jeux de données. En effet, une valeur de σ choisie de sorte à

Tableau 4.5. Résultats pour le jeu de données Heart ($n = 270, d = 14$)

Algorithme	DSM	Temps (s)	Temps/DSM	TEE	% Amélioration
RWM	0,05243	1,88	35,91	93,44	-85,95
MALA	0,37343	3,47	9,30	560,05	0
MALA ($\gamma = 1,2$)	0,43052	3,41	7,92	668,35	15,28
MALA ($\gamma = 1,4$)	0,45313	3,80	8,38	641,03	21,34
MALA ($\gamma = 1,6$)	0,44838	3,37	7,53	616,83	20,07
MALA ($\gamma = 1,8$)	0,41127	3,31	8,05	511,44	10,13
MALA ($\gamma = 2$)	0,35802	3,29	9,21	406,44	-4,12

obtenir un taux d'acceptation d'environ 57%, combiné au fait que γ et d sont grands, mène à une très petite valeur de σ.

Plus précisément, si la valeur de γ est trop grande par rapport à la dimension du problème, le déplacement occasionné par le terme $\gamma\sigma^2\nabla \log(\pi(x))/2$, soit le terme contenant le gradient dans la densité instrumentale (2.3.3), sera très grand. Les candidats se retrouveront dans des endroits de faible densité et le taux d'acceptation sera faible. Afin de réduire l'importance de ce terme et de maintenir un taux d'acceptation d'environ 57%, il est nécessaire de choisir une valeur de σ très petite. Cela résulte en une DSM très faible puisque les déplacements seront petits à chaque itération. Il est donc normal que les résultats soient du même ordre de grandeur que ceux de l'algorithme RWM, puisqu'on ignore en quelque sorte le terme contenant le gradient lorsque σ est très petit.

Par contre, si l'on choisit une valeur de σ telle que le taux d'acceptation est moindre, les résultats sont moins dramatiques pour German Credit. Par exemple, pour un taux d'acceptation d'environ 42% et une valeur de $\gamma = 1,8$, la DSM vaut 0,08293 et est supérieure à celle du MALA. La valeur optimale de σ maximisant la DSM ne semble donc pas être celle menant à un taux d'acceptation de 57% pour toutes les valeurs de γ. D'autres exemples traitant de cet aspect seront présentés plus loin.

De manière générale, il est préférable de rester conservateur et de choisir une valeur de γ relativement petite puisqu'une valeur trop grande peut avoir des conséquences graves dans le contexte où l'on désire conserver un taux d'acceptation élevé. Cela est surtout important lorsque la dimension du problème devient grande, puisque le déplacement occasionné par le paramètre γ aura d'autant plus d'impact dans cette situation. Néanmoins, même pour de petites valeurs de γ, l'amélioration par rapport au MALA est non négligeable et il est donc avantageux de rester conservateur.

En ce qui concerne la TEE, le fait d'utiliser une valeur $\gamma > 1$ mène le plus souvent à une amélioration par rapport au MALA, surtout pour des valeurs de γ proches de la valeur optimale. On remarque cependant que la valeur de la TEE diminue plus rapidement que celle de la DSM lorsque γ s'éloigne de l'optimalité. Par exemple, pour le jeu de données Heart, la DSM lorsque $\gamma = 1,8$ est supérieure à celle du MALA, mais sa TEE est inférieure. Ceci est illustré à l'aide de la fonction d'autocorrélation à la figure 4.1. Dans le graphique du haut, on voit que le MALA a toujours une autocorrélation plus grande par rapport au cas où $\gamma = 1,2$. Dans le graphique du bas, ce n'est plus le cas pour $\gamma = 1,8$, sauf pour le premier délai. Cela n'est pas surprenant, car la DSM pour une valeur de $\gamma = 1,8$ est supérieur à celle du MALA et nous savons que maximiser la DSM minimise l'autocorrélation de délai 1. Par contre, pour les autres délais, l'utilisation de $\gamma = 1,8$ n'est pas optimale en termes d'autocorrélation, malgré une différence minime par rapport au MALA. Bien que l'impact de γ sur la TEE n'ait pas été étudié dans ce mémoire, il est possible d'expliquer intuitivement ce phénomène. Tel que mentionné dans le paragraphe précédent, la valeur de σ choisie diminue à mesure que γ augmente, afin de maintenir un taux d'acceptation d'environ 57%. De plus, une petite valeur de σ résulte le plus souvent en une grande autocorrélation puisque tous les états générés se retrouvent à proximité les uns des autres. En résumé, choisir une valeur de γ trop grande a plus d'impact en termes de TEE qu'en termes de DSM.

4.2. Choix de σ

Tel que mentionné précédemment, choisir une valeur de σ menant à un taux d'acceptation de 57% ne maximise pas la DSM lorsque $\gamma = 2$ pour le jeu de données German Credit. La question se pose donc à savoir si l'utilisation d'un tel choix de σ a eu un impact sur les

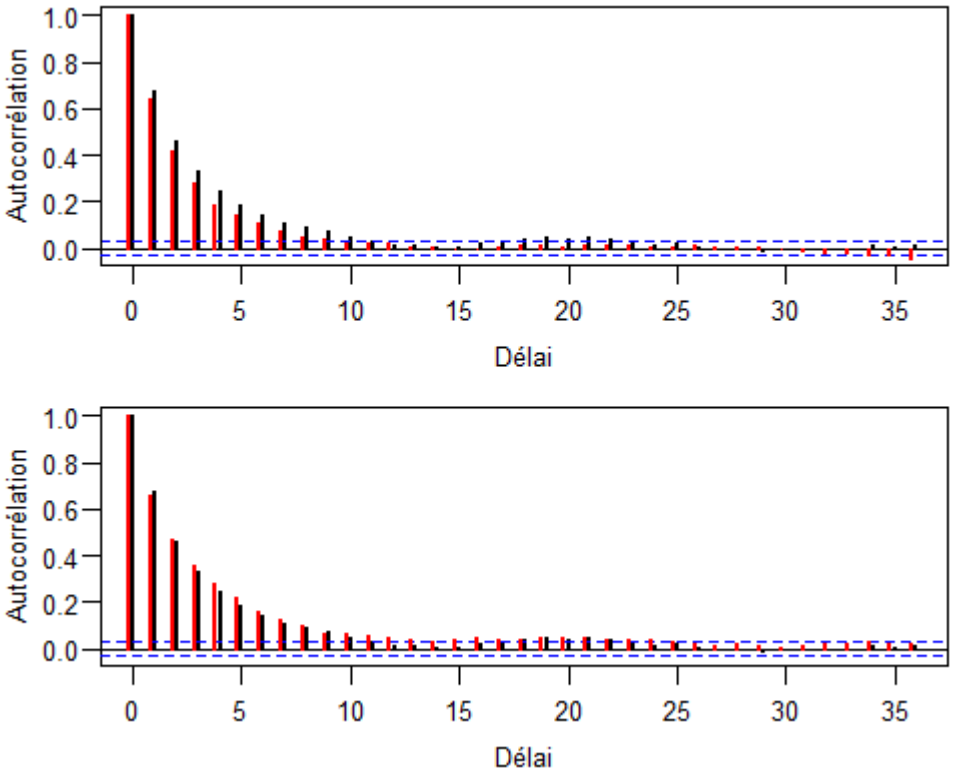
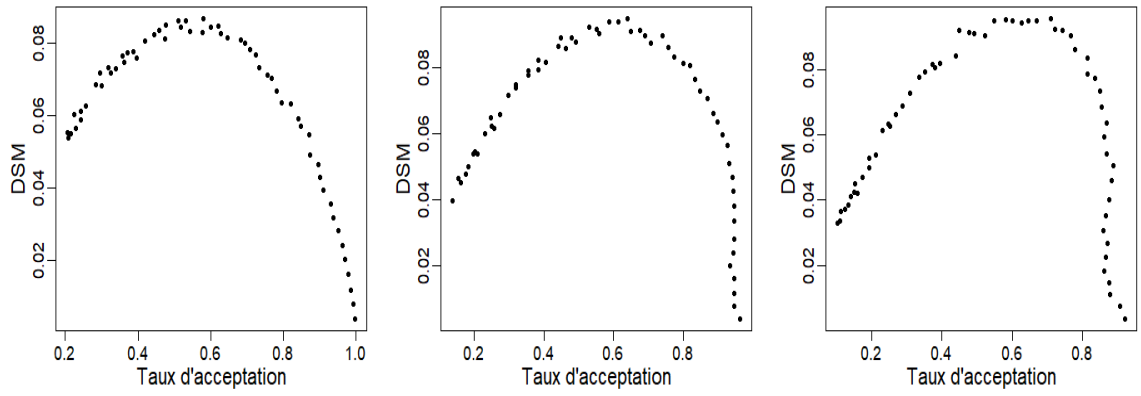


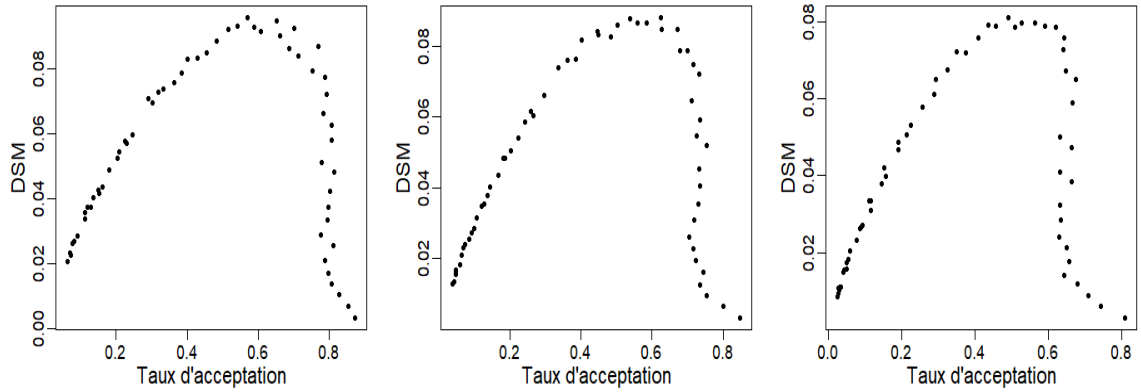
Figure 4.1. Fonction d'autocorrélation de l'échantillon simulé de la première composante du jeu de données Heart. En haut : MALA en noir et MALA ($\gamma = 1,2$) en rouge. En bas : MALA en noir et MALA ($\gamma = 1,8$) en rouge

résultats obtenus dans les tableaux 4.2 à 4.5 lorsque $\gamma > 1$. Les simulations numériques suivantes permettent de répondre en partie à cette question. Pour les jeux de données Pima Indian et German Credit, pour chaque valeur de γ, l'algorithme a été appliqué en utilisant plusieurs valeurs de σ afin de couvrir presque tous les taux d'acceptation possibles. La DSM a été calculée pour chaque valeur de σ afin d'obtenir son évolution en fonction du taux d'acceptation. Aux petits taux d'acceptation correspondent de grandes valeurs de σ et vice-versa.

Pour les données Pima Indian, les résultats sont présentés à la figure 4.2. Pour la valeur de $\gamma = 1$, soit le MALA, la valeur du taux d'acceptation qui maximise la DSM est de 57% comme prévu. Pour les autres valeurs de γ, la valeur maximale ne semble pas toujours être 57%. Pour les valeurs $\gamma = 1,2$ et $\gamma = 1,4$, celle-ci se situe plutôt autour de 62% environ. Par contre, pour les autres valeurs, le taux d'acceptation optimal semble diminuer et lorsque



(a) $\gamma = 1$ **(b)** $\gamma = 1,2$ **(c)** $\gamma = 1,4$



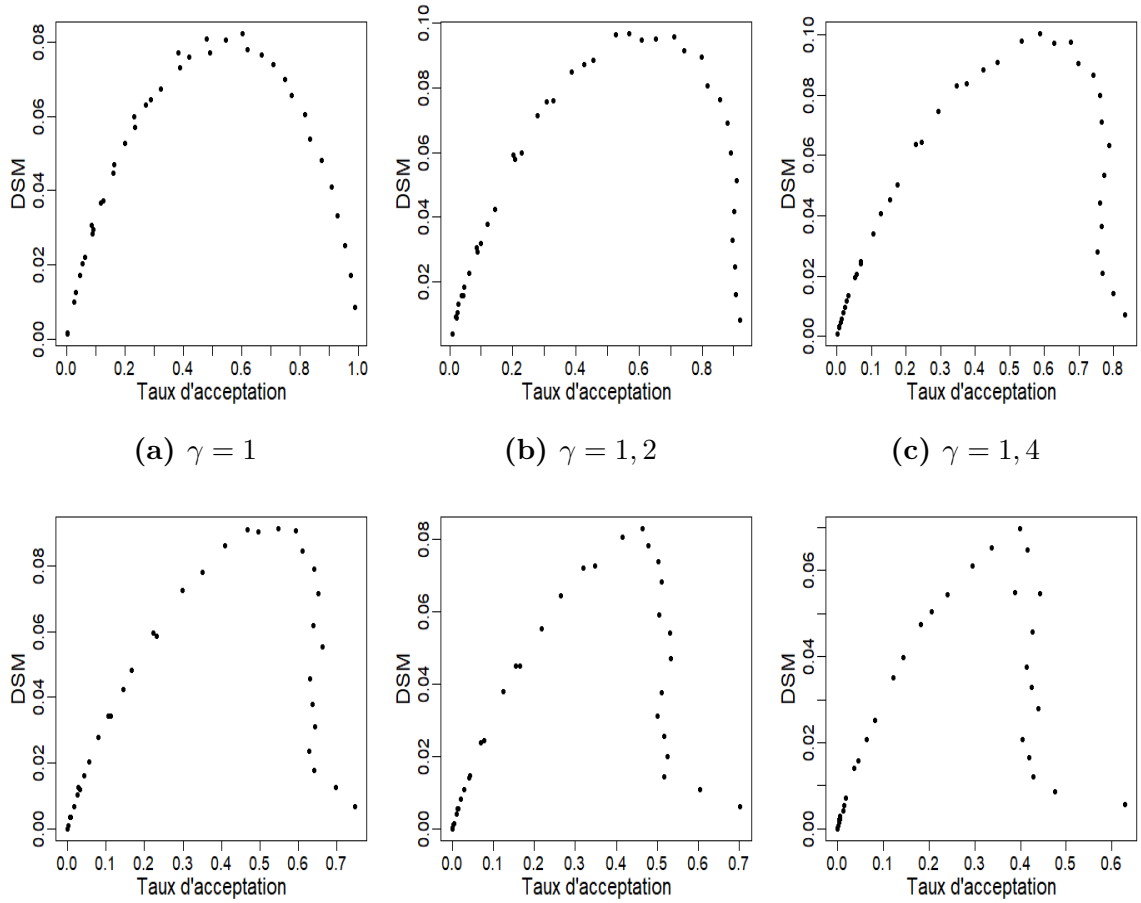
(d) $\gamma = 1,6$ **(e)** $\gamma = 1,8$ **(f)** $\gamma = 2$

Figure 4.2. DSM en fonction du taux d'acceptation selon diverses valeurs de γ pour le jeu de données Pima Indian.

$\gamma = 2$, cette valeur est même en deçà de 57%. On note également que pour chaque valeur de $\gamma > 1$, il y a une suite de valeurs de σ pour laquelle le taux d'acceptation reste constant alors que la DSM diminue. Étant donné que la valeur de γ induit un mouvement latéral à chaque itération à l'aide du terme $\gamma\sigma^2\nabla \log(\pi(x))/2$, il semble donc que de diminuer la valeur de σ n'augmente pas nécessairement le taux d'acceptation pour cette suite de valeurs de σ. L'impact de la diminution de σ sur le taux d'acceptation semble alors contrebalancé par la valeur de γ qui est plus grande. La DSM, quant à elle, diminue puisque la valeur de σ diminue. Pour des valeurs de σ très petites, à partir d'un certain point, le taux d'acceptation recommence à augmenter étant donné que le terme contenant le gradient devient négligeable. Malgré tout cela, il semble que la courbe de la DSM en fonction du taux d'acceptation soit

assez plate au maximum. De ce fait, choisir une valeur de σ donnant un taux d'acceptation autour de 57% n'a pas eu d'impact sur les résultats puisque cette valeur donne le plus souvent une DSM proche de la valeur maximale.

En ce qui concerne le jeu de données German Credit, les résultats sont présentés à la figure 4.3. Le maximum semble bien se trouver autour de 57% pour des valeurs de γ allant de 1 à 1,6. Pour $\gamma = 1,8$, le maximum se situe plutôt autour de 50% tandis que pour $\gamma = 2$, il a diminué à 40%. Ainsi, le taux d'acceptation donnant une valeur maximale de la DSM semble donc diminuer à mesure que γ augmente. Ces deux graphiques permettent de mieux comprendre les résultats obtenus pour le jeu de données German Credit au tableau 4.3. En



(a) $\gamma = 1$ **(b)** $\gamma = 1,2$ **(c)** $\gamma = 1,4$

(d) $\gamma = 1,6$ **(e)** $\gamma = 1,8$ **(f)** $\gamma = 2$

Figure 4.3. DSM en fonction du taux d'acceptation selon diverses valeurs de γ pour le jeu de données German Credit.

effet, lorsque le taux d'acceptation dépasse 50%, la DSM diminue fortement puisque la valeur de σ est très faible, ce qui explique la mauvaise performance obtenue. Il est possible d'obtenir de bons résultats lorsque $\gamma > 1,8$; il suffit de considérer un taux d'acceptation plus faible. Néanmoins, pour des valeurs $\gamma < 1,8$, le taux d'acceptation menant à une DSM optimale demeure aux alentours de 57% et les résultats obtenus pour ces valeurs sont donc valides.

Conclusion

Dans ce mémoire, nous avons étudié une propriété intéressante que peut posséder n'importe quelle densité instrumentale : la propriété d'équilibre. L'étude de l'équilibre local et de l'équilibre global a mené au design d'un algorithme tout à fait nouveau qui généralise le MALA traditionnel grâce à l'ajout du paramètre γ. Cela apporte beaucoup de flexibilité au MALA et permet de l'adapter au problème considéré, selon que celui-ci soit en faible ou grande dimension. Bien que l'idée d'une distribution en équilibre ne soit pas nouvelle, la généralisation du MALA est, selon nos connaissances, inédite.

D'un point de vue pratique, cet algorithme est très simple à implémenter. Pour un utilisateur des MCMC ayant déjà travaillé avec le MALA auparavant, il est très facile de le modifier afin de pouvoir utiliser le nouvel algorithme présenté ici. En ce qui concerne les détails de son implémentation, nous nous sommes concentrés sur le choix optimal du nouveau paramètre γ en fonction de la dimension. Nous avons vu que ce choix dépend de la cible considérée, ce qui empêche de déterminer un choix universel pour toute densité cible. Néanmoins, les simulations et résultats théoriques présentés ont permis de conclure que le paramètre γ doit décroître vers 1 avec un taux de décroissance de $1/3$ dans un régime asymptotique afin d'obtenir un algorithme optimal pour une cible avec composantes iid. De l'autre côté, en faible dimension, même si la valeur optimale est inconnue, un choix de γ plus petit que cette valeur donne tout de même des résultats intéressants. Par contre, un choix trop grand de γ en faible dimension peut mener à une mauvaise performance. En ce qui concerne le choix de σ, cette question n'a pas été abordée en détails, mais les applications numériques présentées laissent croire que de choisir un taux d'acceptation de 57% en combinaison avec la valeur de γ optimale est justifié.

Les résultats obtenus lors des simulations et des applications sur données réelles sont satisfaisants à plusieurs égards. En termes de la DSM et pour une valeur de γ optimale, l'amélioration par rapport au MALA est non négligeable, et ce, sans augmenter pour autant le temps de calcul. De plus, un choix de γ plus faible que la valeur optimale mène également à des résultats intéressants. Dans tous les exemples considérés au chapitre 4, une valeur de γ d'environ 1,5 mène à une amélioration par rapport au MALA. Ainsi, il n'est pas nécessaire de connaître précisément la valeur optimale de γ afin d'obtenir un algorithme performant. De plus, même si l'utilisation de $\gamma > 1$ est théoriquement justifiée en faible dimension uniquement, les simulations montrent qu'il y a amélioration du DSM par rapport au MALA, et ce, jusqu'à 1 000 dimensions. D'ailleurs, cette amélioration peut demeurer significative lorsque la dimension augmente : en grande dimension, même une valeur de γ très proche de 1 peut mener à des améliorations substantielles. Ainsi, ce nouvel algorithme est également efficace sur des problèmes en dimension relativement élevée.

Pour des recherches futures, il serait intéressant d'explorer la relation entre le degré d'interpolation γ et la variance instrumentale σ^2 d'un point de vue théorique. Les applications présentées montrent que le choix d'un σ^2 menant à un taux d'acceptation de 57% ne résulte pas toujours en un algorithme optimal pour tout γ. Par contre, cette règle semble tenir pour la valeur de γ optimale dans les exemples considérés. De plus, il faudrait également démontrer théoriquement que l'équilibre global est effectivement plus efficace en termes de DSM que l'équilibre local en faible dimension. Les simulations effectuées semblent valider ce point, mais une démonstration théorique est nécessaire. Enfin, la valeur de γ optimale pour une dimension précise dépend de la cible considérée, ce qui complique l'implémentation de l'algorithme. Afin d'outrepasser ce problème, il serait peut-être possible de concevoir un algorithme adaptatif qui ajusterait automatiquement la valeur de γ. Pour ce faire, il faudrait déterminer un critère théorique permettant de choisir adéquatement quelle doit être la valeur optimale de γ en fonction des états déjà générés de la chaîne.

Références

[1] Alexandros Beskos et Andrew Stuart : MCMC methods for sampling function space. *In Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, Editors Rolf Jeltsch and Gerhard Wanner*, pages 337–364. European Mathematical Society, 2009.

[2] Roberto Casarin, Radu Craiu et Fabrizio Leisen : Interacting multiple try algorithms with different proposal distributions. *Statistics and Computing*, 23:185–200, 2013.

[3] George Casella et Roger L Berger : *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[4] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari et Donald B Rubin : *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

[5] Mark Girolami et Ben Calderhead : Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73:123–214, 2011.

[6] Wilfred Keith Hastings : Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[7] Michaël Lalancette : Convergence d'un algorithme de type Metropolis pour une distribution cible bimodale. Mémoire de maîtrise, 2017. Université de Montréal.

[8] Jun S Liu, Faming Liang et Wing Hung Wong : The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95:121–134, 2000.

[9] Luca Martino et Jesse Read : On the flexibility of the design of multiple try Metropolis schemes. *Computational Statistics*, 28:2797–2823, 2013.

[10] Kerrie L Mengersen et Richard L Tweedie : Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24:101–121, 1996.

[11] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller et Edward Teller : Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953.

[12] Sean P Meyn et Richard L Tweedie : *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[13] Donald Michie, D. J. Spiegelhalter, C. C. Taylor et John Campbell, éditeurs. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Upper Saddle River, NJ, USA, 1994.

[14] Grigorios A Pavliotis : *Stochastic Processes and Applications : Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, 2014.

[15] Peter H Peskun : Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

[16] Brian D Ripley : *Pattern Recognition and Neural Networks*. Cambridge university press, 1996.

[17] Gareth O Roberts, Andrew Gelman et Walter R Gilks : Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7:110–120, 1997.

[18] Gareth O Roberts et Jeffrey S Rosenthal : Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 60:255–268, 1998.

[19] Gareth O Roberts et Jeffrey S Rosenthal : Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

[20] Gareth O Roberts et Jeffrey S Rosenthal : General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

[21] Gareth O Roberts et Richard L Tweedie : Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2:341–363, 1996.

[22] Chris Sherlock : *Methodology for inference on the Markov modulated Poisson process and theory for optimal scaling of the random walk Metropolis*. Thèse de doctorat, Lancaster University, 2006.

[23] Luke Tierney : Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1728, 1994.

[24] Giacomo Zanella : Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, à venir, 2019.

Annexe A

A.1. Chapitre 2

Théorème (**Convergence dominée**). *Soit* $\{f_n\}_{n\geqslant 0}$ *une suite de fonctions mesurables sur un espace mesurable* $(\mathcal{X}, \mathcal{F}, \mu)$. *On suppose que la suite de fonctions converge ponctuellement vers une fonction* f *lorsque* $n \to \infty$ *et est dominée par une fonction intégrable* g *de telle sorte que*

$$|f_n(x)| \leq g(x),$$

avec $\int_{\mathcal{X}} |g| d\mu < \infty$, *et ce, pour tout* n *et pour tout* $x \in \mathcal{X}$. *Alors* f *est intégrable et*

$$\lim_{n\to\infty} \int_{\mathcal{X}} f_n \mathrm{d}\mu = \int_{\mathcal{X}} f \mathrm{d}\mu.$$

Théorème (**Convergence bornée**). *Soit* $\{f_n\}_{n\geqslant 0}$ *une suite de fonctions mesurables dans sur un espace mesurable* $(\mathcal{X}, \mathcal{F}, \mu)$ *tel que* $\mu(\mathcal{X}) < \infty$. *On suppose que la suite de fonctions converge ponctuellement vers une fonction* f *lorsque* $n \to \infty$ *et est bornée par une constante* K *de telle sorte que*

$$|f_n(x)| \leq K,$$

et ce, pour tout n *et pour tout* $x \in \mathcal{X}$. *Alors* f *est intégrable et*

$$\lim_{n\to\infty} \int_{\mathcal{X}} f_n \mathrm{d}\mu = \int_{\mathcal{X}} f \mathrm{d}\mu.$$

Lemme 2.2.4. (Lemma 1 Zanella [**24**]) *Soit une fonction continue* $h : [0, \infty) \to [0, \infty)$ *avec* $h(1) = 1$ *et* $h(t) \leq a + bt$ *pour* $a, b \geq 0$ *et pour tout* $t \geq 0$. *Soient une distribution* Π *associée à une densité bornée* $\pi : \mathcal{X} \to [0, \infty)$, *une distribution instrumentale symétrique* $Q_\sigma(x, \cdot)$ *centrée en* x *ainsi que la fonction* $Z_\sigma(x)$ *définie en* (2.2.2). *Alors, il est possible d'écrire*

$$\lim_{\sigma \to 0} \int_{\mathcal{X}} \pi(x) Z_\sigma(x) \mathrm{d}x = 1.$$

Démonstration. Pour tout $x \in \mathcal{X}$, il s'avère que $\pi(x)Z_\sigma(x) \to \pi(x)$ lorsque $\sigma \downarrow 0$ (voir théorème 2.2.5). De ce fait, le lemme de Fatou implique que

$$\liminf_{\sigma \to 0} \int_{\mathcal{X}} \pi(x) Z_\sigma(x) \mathrm{d}x \geqslant \int_{\mathcal{X}} \liminf_{\sigma \to 0} \pi(x) Z_\sigma(x) \mathrm{d}x = \int_{\mathcal{X}} \pi(x) \mathrm{d}x = 1.$$

Il faut maintenant montrer que la limite supérieure est bornée par 1. Pour ce faire, on fixe un $\varepsilon > 0$ ainsi qu'un ensemble ouvert $A \subset \mathcal{X}$ tel que $\int_A \mathrm{d}x < \infty$ et $\Pi(A) > 1 - \varepsilon$. Par le théorème de la convergence bornée, il est possible d'écrire

$$\lim_{\sigma \to 0} \int_A \pi(x) Z_\sigma(x) \mathrm{d}x = \int_A \pi(x) \mathrm{d}x = \Pi(A). \tag{A.1.1}$$

Ceci est possible puisque $\int_A \mathrm{d}x < \infty$ et le produit $\pi(x)Z_\sigma(x)$ est borné par une constante. En effet,

$$\begin{aligned} \pi(x) \int_{\mathcal{X}} h\left(\frac{\pi(z)}{\pi(x)}\right) Q_\sigma(x, \mathrm{d}z) &\leq \pi(x) \int_{\mathcal{X}} \left(a + b\frac{\pi(z)}{\pi(x)}\right) Q_\sigma(x, \mathrm{d}z) \\ &= a\pi(x) + b \int_{\mathcal{X}} \pi(z) Q_\sigma(x, \mathrm{d}z) \\ &\leq (a + b) \sup_{x \in \mathcal{X}} \pi(x). \end{aligned}$$

On considère maintenant l'intégrale sur A^c. Puisque $h(t) \leq a + bt$, on a

$$\begin{aligned} \int_{A^c} \pi(x) Z_\sigma(x) \mathrm{d}x &\leq \int_{A^c} \pi(x) \int_{\mathcal{X}} \left(a + b\frac{\pi(y)}{\pi(x)}\right) Q_\sigma(x, \mathrm{d}y) \mathrm{d}x \\ &= a\Pi(A^c) + b \int_{x \in A^c,\, y \in \mathcal{X}} \pi(y) Q_\sigma(x, \mathrm{d}y) \mathrm{d}x. \end{aligned} \tag{A.1.2}$$

En utilisant la symétrie de Q_σ, il est possible de borner cette dernière intégrale comme suit,

$$\begin{aligned} \int_{x \in A^c,\, y \in \mathcal{X}} \pi(y) Q_\sigma(x, \mathrm{d}y) \mathrm{d}x &= \int_{y \in \mathcal{X}} \pi(y) \int_{x \in A^c} Q_\sigma(y, \mathrm{d}x) \mathrm{d}y \\ &= \int_{y \in \mathcal{X}} \pi(y) Q_\sigma(y, A^c) \mathrm{d}y \\ &\leq \Pi(A^c) + \int_{y \in A} \pi(y) Q_\sigma(y, A^c) \mathrm{d}y. \end{aligned} \tag{A.1.3}$$

La densité π est bornée et $Q_\sigma(y, A^c)$ est borné par 1. Par le théorème de la convergence bornée, puisque $\int_A \mathrm{d}z < \infty$, on a que

$$\lim_{\sigma\to 0} \int_{y\in A} \pi(y) Q_\sigma(y, A^c)\mathrm{d}y = \int_{y\in A} \pi(y) \lim_{\sigma\to 0} Q_\sigma(y, A^c)\mathrm{d}y = 0.$$

Cette dernière limite vaut 0, étant donné que A^c est fermé et que $Q_\sigma(y, \cdot)$ converge faiblement vers la mesure $\delta_y(\cdot)$. Par conséquent, lorsque $\sigma \downarrow 0$, la limite $\lim_{\sigma\to 0} Q_\sigma(y, A^c) = 0$ pour tout $y \in A$. En combinant les résultats (A.1.1),(A.1.2) et (A.1.3), on obtient

$$\limsup_{\sigma\to 0} \int_{\mathcal{X}} \pi(x) Z_\sigma(x)\mathrm{d}x \leq \Pi(A) + a\Pi(A^c) + b\Pi(A^c) \leq 1 + (a + b)\varepsilon.$$

Puisque cela est vrai $\forall\, \varepsilon > 0$, il s'en suit que $\limsup_{\sigma\to 0} \int_{\mathcal{X}} \pi(x) Z_\sigma(x)\mathrm{d}x = 1$. Ainsi,

$$\lim_{\sigma\to \mathbf{0}} \int_{\mathcal{X}} \pi(x) Z_\sigma(x)\mathrm{d}x = 1.$$

□

Lemme (**Scheffé**). *Soit $\{f_n\}_{n\geqslant 0}$, une suite de densités de probabilité définies sur l'espace $\mathcal{X}$ par rapport à une mesure μ sur l'espace mesurable $(\mathcal{X}, \mathcal{F})$. Supposons que $\{f_n\}_{n\geqslant 0}$ converge ponctuellement vers une densité f lorsque $n \to \infty$. Alors, si X_n et X ont pour densités respectives f_n et f, X_n converge en distribution vers X lorsque $n \to \infty$.*

Proposition 2.3.1. *Soit $Q_{h,\sigma}$, une distribution en équilibre local qui biaise $Q_\sigma(x, \cdot)$, la distribution normale associée à la densité* (1.3.4)*. Alors, si $h(t) = \sqrt{t}$ ou $h(t) = t/(1 + t)$, $Q_{h,\sigma}$ peut être approximée par la distribution associée à la densité du MALA* (1.3.8)*.*

Démonstration. Tout d'abord, puisque $Q_\sigma(x, \cdot)$ est la distribution normale et en notant $h_1(t) = \sqrt{t}$, on a

$$\begin{aligned} Q_{h_1,\sigma}(x, \mathrm{d}y) &\propto \sqrt{\frac{\pi(y)}{\pi(x)}} q_\sigma(x, y)\mathrm{d}y \\ &\propto \exp\left\{\frac{\log(\pi(y))}{2}\right\} \exp\left\{-\frac{1}{2\sigma^2}(y - x)^\top (y - x)\right\} \mathrm{d}y. \end{aligned} \tag{A.1.4}$$

Une approximation de Taylor du premier ordre évaluée à x de la fonction $\pi(y)$ donne

$$\pi(y) = \exp\{\log(\pi(y))\} \approx \exp\{\log(\pi(x)) + \nabla \log(\pi(x))(y - x)\}.$$

L'expression (A.1.4) peut alors être approximée par

$$\begin{aligned}
Q_{h_1,\sigma}(x,\mathrm{d}y) &\propto \exp\left\{\frac{\log(\pi(y))}{2}\right\}\exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)\right\}\mathrm{d}y \\
&\approx \exp\left\{\frac{\log(\pi(x))}{2}+\frac{\nabla\log(\pi(x))(y-x)}{2}\right\}\exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)\right\}\mathrm{d}y \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)+\frac{\nabla\log(\pi(x))(y-x)}{2}\right\}\mathrm{d}y \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left\{(y-x)^\top(y-x)-\sigma^2\nabla\log(\pi(x))(y-x)\right\}\right\}\mathrm{d}y \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left\|y-x-\frac{\sigma^2}{2}\nabla\log\{\pi(x)\}\right\|^2\right\}\mathrm{d}y,
\end{aligned}$$

qui est bien proportionnel à la distribution instrumentale du MALA. On considère maintenant la fonction $h_2(t) = t/(1+t)$. En notant $k(y) = \pi(x) + \pi(y)$, on a que $\nabla k(y) = \nabla\pi(y)$. De plus, $\nabla\log(\pi(x)) = \nabla(\log\circ\pi)(x) = \nabla\pi(x)/\pi(x)$. Ainsi, l'approximation de Taylor du premier ordre pour la fonction h_2 devient

$$\begin{aligned}
h_2\left(\frac{\pi(y)}{\pi(x)}\right) &= \frac{\pi(y)}{\pi(x)+\pi(y)} \\
&= \exp\Big\{\log(\pi(y)) - \log(\pi(x)+\pi(y))\Big\} \\
&\approx \exp\Big\{\log(\pi(x)) + \nabla\log(\pi(x))(y-x) - \log(2\pi(x)) - \nabla(\log\circ k)(x)(y-x)\Big\} \\
&\propto \exp\left\{\nabla\log(\pi(x))(y-x) - \left(\frac{\nabla k(x)}{\pi(x)+\pi(x)}\right)(y-x)\right\} \\
&= \exp\left\{\nabla\log(\pi(x))(y-x) - \left(\frac{\nabla\pi(x)}{2\pi(x)}\right)(y-x)\right\} \\
&= \exp\left\{\nabla\log(\pi(x))(y-x) - \frac{\nabla\log(\pi(x))(y-x)}{2}\right\} \\
&= \exp\left\{\frac{\nabla\log(\pi(x))(y-x)}{2}\right\}.
\end{aligned}$$

Étant donné que cette approximation est proportionnelle à

$$\exp\left\{\frac{\log(\pi(x))}{2}+\frac{\nabla\log(\pi(x))(y-x)}{2}\right\} \approx h_1\left(\frac{\pi(y)}{\pi(x)}\right),$$

qui est l'approximation de la fonction $h_1(t)$, $Q_{h_2,\sigma}$ pourra aussi approximée par la distribution du MALA. □

Proposition 2.3.2. *Soit $Q_{g,\sigma}$ une distribution en équilibre global qui biaise la distribution normale $Q_\sigma(x,\cdot)$ associée à la densité* (1.3.4)*. Alors $Q_{g,\sigma}$ peut être approximée par la distribution associée à la densité*

$$q_{g,\sigma}(x,y) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{1}{2\sigma^2}\left\|y - x - \sigma^2 \nabla \log\{\pi(x)\}\right\|^2\right\}.$$

Démonstration. La fonction g est telle que $g(y) \propto \pi(y)$, c'est-à-dire que pour une constante $0 < c < \infty$, $g(y) = c\pi(y)$. En utilisant le développement de Taylor présenté à la preuve de la proposition 2.3.1, $Q_{g,\sigma}$ est approximée par

$$\begin{aligned}
Q_{g,\sigma}(x, \mathrm{d}y) &\propto \pi(y) q_\sigma(x,y) \mathrm{d}y \\
&= \exp\left\{ \log(\pi(y))\right\} \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top (y-x)\right\} \mathrm{d}y \\
&\approx \exp\left\{ \log(\pi(x)) + \nabla \log(\pi(x))(y-x)\right\} \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top (y-x)\right\} \mathrm{d}y \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top (y-x) + \nabla \log(\pi(x))(y-x)\right\} \mathrm{d}y \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left\{(y-x)^\top (y-x) - 2\sigma^2 \nabla \log(\pi(x))(y-x)\right\}\right\} \mathrm{d}y \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left\|y - x - \sigma^2 \nabla \log(\pi(x))\right\|^2\right\} \mathrm{d}y.
\end{aligned}$$

□

Développement de Taylor pour la démonstration du théorème 2.4.2

Il est possible de développer la fonction R_d de la manière suivante :

$$\begin{aligned}
R_d &= \sum_{i=1}^d \log\left(\frac{f(y_i) q_\sigma^*(y_i, x_i)}{f(x_i) q_\sigma^*(x_i, y_i)}\right) \\
&= \sum_{i=1}^d \left[l(y_i) - l(x_i) - \frac{1}{2\sigma_d^2}\left\{x_i - y_i - \frac{\sigma_d^2 \gamma}{2} l'(y_i)\right\}^2 + \frac{1}{2\sigma_d^2}\left\{y_i - x_i - \frac{\sigma_d^2\gamma}{2} l'(x_i)\right\}^2\right] \\
&= \sum_{i=1}^d \left[l(y_i) - l(x_i) - \frac{1}{2\sigma_d^2}\left\{(y_i - x_i)^2 + \sigma_d^2 \gamma (y_i - x_i) l'(y_i) + \left(\frac{\sigma_d^2\gamma}{2} l'(y_i)\right)^2\right\}\right. \\
&\quad \left. + \frac{1}{2\sigma_d^2}\left\{(y_i - x_i)^2 - \sigma_d^2 \gamma (y_i - x_i) l'(x_i) + \left(\frac{\sigma_d^2\gamma}{2} l'(x_i)\right)^2\right\}\right] \\
&= \sum_{i=1}^d \left[l(y_i) - l(x_i) - \frac{\gamma}{2}(y_i - x_i)(l'(x_i) + l'(y_i)) + \frac{\sigma_d^2 \gamma^2}{8}\left(l'(x_i)^2 - l'(y_i)^2\right)\right].
\end{aligned}$$

On effectue un développement de Taylor centré en $\sigma_d = 0$. Rappelons que pour $i = 1, \ldots, d$,

$$y_i = x_i + \frac{\sigma_d^2 \gamma}{2} l'(x_i) + \sigma_d Z_i$$

où $Z_i \sim \mathcal{N}(0,1)$ *iid*. Ceci implique que $l(y_i) = l(x_i)$ lorsque $\sigma = 0$. Les trois premières dérivées de R_d sont

$$\begin{aligned}
R_d'(\sigma_d) = \sum_{i=1}^{d} \Bigg[& l'(y_i)[Z_i + \sigma_d \gamma l'(x_i)] - \frac{\gamma}{2}[Z_i + \sigma_d \gamma l'(x_i)][l'(x_i) + l'(y_i)] \\
& - \frac{\gamma}{2}\left[\sigma_d Z_i + \frac{\sigma_d^2 \gamma}{2} l'(x_i)\right][l''(y_i)(Z_i + \sigma_d \gamma l'(x_i))] + \frac{\sigma_d \gamma^2}{4}[l'(x_i)^2 - l'(y_i)^2] \\
& - \frac{\sigma_d^2 \gamma^2}{4}[l'(y_i) l''(y_i)(Z_i + \sigma_d \gamma l'(x_i))]\Bigg]
\end{aligned}$$

$$\begin{aligned}
R_d''(\sigma_d) = \sum_{i=1}^{d} & l''(y_i)[Z_i + \sigma_d \gamma l'(x_i)]^2 + \gamma l'(y_i) l'(x_i) - \frac{\gamma}{2}[\gamma l'(x_i)][l'(x_i) + l'(y_i)] \\
& - \frac{\gamma}{2}\left[Z_i + \sigma_d \gamma l'(x_i)\right]\left[l''(y_i)(Z_i + \sigma_d \gamma l'(x_i))\right] \\
& - \frac{\gamma}{2}\left[\sigma_d Z_i + \frac{\sigma_d^2 \gamma}{2} l'(x_i)\right][l'''(y_i)(Z_i + \sigma_d \gamma l'(x_i))^2 + l''(y_i)(\gamma l'(x_i))] \\
& - \frac{\gamma}{2}[Z_i + \sigma_d \gamma l'(x_i)][l''(y_i)(Z_i + \sigma_d \gamma l'(x_i))] + \frac{\gamma^2}{4}[l'(x_i)^2 - l'(y_i)^2] \\
& - \sigma_d \gamma^2 [l'(y_i) l''(y_i)(Z_i + \sigma_d \gamma l'(x_i))] - \frac{\sigma_d^2 \gamma^2}{4}\Bigg[l''(y_i)^2 (Z_i + \sigma_d \gamma l'(x_i))^2 \\
& + l'(y_i) l'''(y_i)(Z_i + \sigma_d \gamma l'(x_i))^2 + l'(y_i) l''(y_i) \gamma l'(x_i)\Bigg]
\end{aligned}$$

$$
\begin{aligned}
R_d'''(\sigma^*) &= \sum_{i=1}^{d} U_{i,d}(x_i, Z_i, \sigma_i^*) \\
&= \sum_{i=1}^{d} l'''(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^3 + 2\gamma l''(y_i)l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i)) \\
&\quad - \gamma l''(y_i)l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i)) - \gamma^2 l'(x_i)l''(y_i)(Z_i + \sigma_i^* l'(x_i)) \\
&\quad - \frac{\gamma}{2}[Z_i + \sigma_i^*\gamma l'(x_i)][l'''(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^2 + l''(y_i)\gamma l'(x_i)] \\
&\quad - \frac{\gamma}{2}(Z_i + \sigma_i^*\gamma l'(x_i))[l'''(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^2 + \gamma l''(y_i)l'(x_i)] \\
&\quad - \frac{\gamma}{2}[\sigma_i^* Z_i + \frac{(\sigma_i^*)^2\gamma}{2} l'(x_i)][l^{(4)}(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^3 + 2\gamma l'''(y_i)l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i)) \\
&\quad + \gamma l'''(y_i)l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i))] - \frac{\gamma^2}{2} l''(y_i)l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i)) \\
&\quad - \frac{\gamma}{2}[Z_i + \sigma_i^*\gamma l'(x_i)][l'''(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^2 + \gamma l''(y_i)l'(x_i)] \\
&\quad - \frac{\gamma^2}{2} l''(y_i)l'(y_i)(Z_i + \sigma_i^*\gamma l'(x_i)) - \gamma^2[l''(y_i)l'(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))] \\
&\quad - \gamma^2\sigma_i^*[(l''(y_i)(Z_i + \sigma_i^*\gamma l'(x_i)))^2 + (l'''(y_i)l'(y_i)(Z_i + \sigma_i^*\gamma l'(x_i)))^2 \\
&\quad + l''(y_i)l'(y_i)\gamma l'(x_i)] - \frac{\sigma_i^*\gamma^2}{2}[l''(y_i)^2(Z_i + \sigma_i^*\gamma l'(x_i))^2 + l'(y_i)l'''(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^2 \\
&\quad + l'(y_i)l''(y_i)\gamma l'(x_i)] - \frac{(\sigma_i^*)^2\gamma^2}{4}\Bigg[2l'''(y_i)l''(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^3 \\
&\quad + 2\gamma l''(y_i)^2 l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i)) + l^{(4)}(y_i)l'(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^3 \\
&\quad + l'''(y_i)l''(y_i)(Z_i + \sigma_i^*\gamma l'(x_i))^2 + 2\gamma l'''(y_i)l'(y_i)l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i)) \\
&\quad + \gamma l'''(y_i)l'(y_i)l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i)) + \gamma l''(y_i)^2 l'(x_i)(Z_i + \sigma_i^*\gamma l'(x_i))\Bigg].
\end{aligned}
$$

En évaluant les deux premières dérivées à 0, on obtient

$$
\begin{aligned}
R_d'(0) &= \sum_{i=1}^{d}(1-\gamma)l'(x_i)Z_i, \\
R_d''(0) &= \sum_{i=1}^{d}(1-\gamma)[l''(x_i)Z_i^2 + l'(x_i)^2\gamma].
\end{aligned}
$$

Lemme A.1.1. *Soit Q_σ, une distribution instrumentale symétrique avec densité associée q_σ. Soit $\{f_\sigma\}$, une suite de fonctions mesurables sur un espace mesurable $(\mathcal{X}, \mathcal{F}, \mu)$ telles que $f_\sigma \to f$ ponctuellement lorsque $\sigma \downarrow 0$. De plus, on suppose que cette suite est bornée uniformément par une constante K, de telle sorte que*

$$|f_\sigma(x)| \leqslant K,$$

et ce, pour tout σ et pour tout $x \in \mathcal{X}$. Alors, pour tout $x \in \mathcal{X}$,

$$\int_{\mathcal{X}} f_\sigma(z) Q_\sigma(x, \mathrm{d}z) \xrightarrow{\sigma \downarrow 0} \int_{\mathcal{X}} f(z) \delta_x(\mathrm{d}z).$$

Démonstration. Il suffit de démontrer qu'il y a convergence vers $f(x)$ puisque $\int_{\mathcal{X}} f(z)\delta_x(\mathrm{d}z) = f(x)$. La distribution Q_σ étant symétrique, selon la remarque 2.1.1, on a

$$\int_{\mathcal{X}} f_\sigma(z) Q_\sigma(x, \mathrm{d}z) = \int_{\mathcal{X}} f_\sigma(z) q_\sigma(x, z) \mathrm{d}z = \int_{\mathcal{X}} \frac{f_\sigma(z)}{\sigma^d} r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z.$$

Un changement de variables $u = (z - x)/\sigma$ permet d'écrire

$$\int_{\mathcal{X}} \frac{f_\sigma(z)}{\sigma^d} r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z = \int_{\mathcal{X}} f_\sigma(u\sigma + x) r(u) \mathrm{d}u.$$

La suite de fonctions f_σ étant bornée, on a que $f_\sigma(u\sigma + x)r(u) \leqslant Kr(u)$ qui est intégrable, car r est une densité. Par le théorème de la convergence dominée, il suit que

$$\int_{\mathcal{X}} f_\sigma(u\sigma + x) r(u) \mathrm{d}u \xrightarrow{\sigma \downarrow 0} \int_{\mathcal{X}} f(x) r(u) \mathrm{d}u = f(x).$$

□

A.2. Chapitre 3

Expansions de Taylor pour le théorème 3.3.1

Les dérivés complètes $R_d^{(i)}$ ne sont pas fournies explicitement ici. La raison est que celles-ci comportent rapidement un très grand nombre de termes et il ne semble pas pertinent de tout détailler. De plus, seules les valeurs $R_d^{(i)}(0)$ sont utilisées dans la démonstration. Le code MATHEMATICA utilisé pour l'obtention de ces dérivées est tout de même fourni à titre de référence.

```
(*y comme fonction de x*)
y:=x+(s^2/2)*g*l'[x]+s*Z
(*Log du taux d'acceptation*)
R[s]=l[y]-l[x]-(g/2)*(y-x)*(l'[x]+l'[y])+(s^2*g/8)*(l'[x]^2-l'[y]^2)
(*Première dérivée de Taylor*)
D[R[s],s]
D[R[s],{s,1}]/.s->0
(*Seconde dérivée de Taylor*)
D[R[s],{s,2}]
D[R[s],{s,2}]/.s->0
(*Troisième dérivée de Taylor*)
D[R[s],{s,3}]
D[R[s],{s,3}]/.s->0
(*Quatrième dérivée de Taylor*)
D[R[s],{s,4}]
D[R[s],{s,4}]/.s->0
(*Cinquième dérivée de Taylor*)
D[R[s],{s,5}]
D[R[s],{s,5}]/.s->0
(*Sixième dérivée de Taylor*)
D[R[s],{s,6}]
D[R[s],{s,6}]/.s->0
(*Septième dérivée de Taylor*)
D[R[s],{s,7}]
```

En utilisant les résultats fournis par ce code, il est possible de déterminer les valeurs de $C_{i,j}$, $i = 1, \ldots, 6$, $j = 1, \ldots, d$. Ces valeurs sont listées ci-dessous pour $j = 1, \ldots, d$:

$$C_{1,j} = (1 - \gamma_d) Z_j l'(x_j); \tag{A.2.1}$$

$$C_{2,j} = (1 - \gamma_d)[l''(x_j) Z_j^2 + \gamma_d l'(x_j)^2)]; \tag{A.2.2}$$

$$C_{3,j} = \left(\frac{3\gamma_d}{2} - 3\gamma_d^2\right) Z_j l'(x_j) l''(x_j) + \left(1 - \frac{3\gamma_d}{2}\right) Z_j^3 l^{(3)}(x_j); \tag{A.2.3}$$

$$\begin{aligned} C_{4,j} = {} & 3\gamma_d(1 - 3\gamma_d) Z_j^2 l'(x_j) l^{(3)}(x_j) - 3\gamma_d Z_j^2 l''(x_j)^2 - 3\gamma_d^3 l'(x_j)^2 l''(x_j) \\ & + (1 - 2\gamma_d) Z_j^4 l^{(4)}(x_j); \end{aligned} \tag{A.2.4}$$

$$\begin{aligned} C_{5,j} = {} & Z_j^5 \left(1 - \frac{5\gamma_d}{2}\right) l^{(5)}(x_j) - 5\gamma_d Z_j^3 \left\{3\gamma_d^2 l''(x_j) l^{(3)}(x_j) - (1 - 3\gamma_d) l'(x_j) l^{(4)}(x_j)\right\} \\ & - 15\gamma_d^2 Z_j \left\{\frac{\gamma_d}{2} l'(x_j)^2 l^{(3)}(x_j) + l'(x_j) l''(x_j)^2\right\}; \end{aligned} \tag{A.2.5}$$

$$\begin{aligned} C_{6,j} = {} & -\frac{15\gamma_d}{2}\{\gamma_d^2(1 + 3\gamma_d) l'(x_j)^3 l^{(3)}(x_j) + 18\gamma_d Z_j^2 l'(x_j) l''(x_j) l^{(3)}(x_j) \\ & + 3\gamma_d^2 l'(x_j)^2 (l''(x_j)^2 + 4 Z_j^2 l^{(4)}(x_j)) + Z_j^4 (3 l^{(3)}(x_j)^2 + 4 l''(x_j) l^{(4)}(x_j))\} \\ & + \frac{15}{2}(1 - 5\gamma_d)\gamma_d Z_j^4 l'(x_j) l^{(5)}(x_j) + (1 - 3\gamma_d) Z_j^6 l^{(6)}(x_j). \end{aligned} \tag{A.2.6}$$

Lemme A.2.1. *Soient les expressions* (A.2.1) *à* (A.2.6) *associées à* $C_{i,1}, i = 1, \ldots, 6$ *respectivement. On suppose que* $X_1 \in \mathbb{R}$ *est distribué selon une densité* f, $Z_1 \sim \mathcal{N}(0, 1)$ *et* X_1 *et* Z_1 *sont indépendants. De plus, on suppose que tous les moments de* f *sont bornés, que* $l(x) = \log(f(x))$ *est* C^8 *et que* $l(x)$ *et ses 8 premières dérivées sont bornées par un polynôme, c'est-à-dire que*

$$|l(x)|, |l^{(i)}(x)| \leqslant M(x), \quad i \leqslant 8,$$

où $M(x)$ *est un polynôme positif. Alors,* $\lim_{d \to \infty} \mathbb{E}[C_{i,1}] < \infty$, $i = 1, \ldots, 6$ *où, plus particulièrement,*

$$\mathbb{E}[C_{1,1}] = \mathbb{E}[C_{3,1}] = \mathbb{E}[C_{5,1}] = 0.$$

avec

$$\mathbb{E}[C_{2,1}] = -(\gamma_d - 1)^2 \mathbb{E}[l'(X_1)^2] \qquad et \qquad \mathbb{E}[C_{4,1}] = (\gamma_d - 1) K_{4,d},$$

où $\lim_{d \to \infty} K_{4,d} < \infty$.

DÉMONSTRATION. En ce qui concerne $C_{1,1}, C_{3,1}$ et $C_{5,1}$, on remarque que chaque terme de ces polynômes est un multiple d'une puissance impaire de Z_1. En utilisant le fait que X_1 et Z_1 sont indépendants, on obtient directement que

$$\mathbb{E}[C_{1,1}] = \mathbb{E}[C_{3,1}] = \mathbb{E}[C_{5,1}] = 0.$$

Pour $C_{2,1}$, on utilise le fait que $\mathbb{E}[l''(X)] = -\mathbb{E}[l'(X)^2]$ (voir lemme 7.3.11 de [**3**]) ce qui permet d'écrire

$$\begin{aligned}\mathbb{E}[C_{2,1}] &= (1-\gamma_d)\mathbb{E}[l''(X_1)Z_1^2 + \gamma_d l'(X_1)^2] \\ &= (1-\gamma_d)(-\mathbb{E}[l'(X_1)^2] + \gamma_d \mathbb{E}[l'(X_1)^2]) \\ &= -(1-\gamma_d)^2 \mathbb{E}[l'(X_1)^2],\end{aligned}$$

où $\mathbb{E}[l'(X_1)^2] < \infty$, car les dérivées de l sont bornées par un polynôme et tous les moments de f sont bornés. Ceci implique que $\lim_{d\to\infty} \mathbb{E}[C_{2,1}] = 0$.

Pour le terme $C_{4,1}$, on calcule en premier lieu l'espérance par rapport à Z_1 en utilisant le fait que $\mathbb{E}[Z_1^2] = 1$ et $\mathbb{E}[Z_1^4] = 3$, ce qui donne

$$\begin{aligned}\mathbb{E}[C_{4,1}] = 3\gamma_d(1-3\gamma_d)\mathbb{E}[l'(X_1)l^{(3)}(X_1)] - 3\gamma_d\mathbb{E}[l''(X_1)^2] - 3\gamma_d^3\mathbb{E}[l'(X_1)^2 l''(X_1)] \\ + 3(1-2\gamma_d)\mathbb{E}[l^{(4)}(X_1)].\end{aligned} \tag{A.2.7}$$

On simplifie cette expression en démontrant que

$$\begin{aligned}\mathbb{E}[l''(X_1)^2] &= -(\mathbb{E}[l'(X_1)^2 l''(X_1)] + \mathbb{E}[l'(X_1)l^{(3)}(X_1)]) \\ \mathbb{E}[l'(X_1)l^{(3)}(X_1)] &= -\mathbb{E}[l^{(4)}(X_1)].\end{aligned}$$

En effet, en résolvant ces espérances par parties, on obtient

$$\begin{aligned}\mathbb{E}[l''(X_1)^2] &= \int_{-\infty}^{\infty} \exp(l(x))l''(x)^2 \mathrm{d}x \qquad (u = \exp(l(x))l''(x) \quad \mathrm{d}v = l''(x)\mathrm{d}x) \\ &= \exp(l(x))l'(x)l''(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \exp(l(x))l'(x)^2 l''(x)\mathrm{d}x \\ &\quad - \int_{-\infty}^{\infty} \exp(l(x))l'(x)l^{(3)}(x)\mathrm{d}x \\ &= -(\mathbb{E}[l'(X_1)^2 l''(X_1)] + \mathbb{E}[l'(X_1)l^{(3)}(X_1)])\end{aligned}$$

et

$$\begin{aligned}\mathbb{E}[l'(X_1)l^{(3)}(X_1)] &= \int_{-\infty}^{\infty} \exp(l(x))l'(x)l^{(3)}(x)\mathrm{d}x \qquad (u = l^{(3)}(x) \quad \mathrm{d}v = \exp(l(x))l'(x)\mathrm{d}x) \\ &= \exp(l(x))l^{(3)}(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \exp(l(x))l^{(4)}(x)\mathrm{d}x \\ &= \mathbb{E}[l^{(4)}(X_1)],\end{aligned}$$

où $\exp(l(x))l'(x)l''(x)\Big|_{-\infty}^{\infty} = \exp(l(x))l^{(3)}(x)\Big|_{-\infty}^{\infty} = 0$, car les dérivées de l sont bornées par un polynôme et tous les moments de f sont bornés. On réécrit donc (A.2.7) par

$$\begin{aligned}\mathbb{E}[C_{4,1}] &= 3\gamma_d(1-3\gamma_d)\mathbb{E}[l'(X_1)l^{(3)}(X_1)] + 3\gamma_d(\mathbb{E}[l'(X_1)l^{(3)}(X_1)] + \mathbb{E}[l'(X_1)^2l''(X_1)]) \\ &\quad - 3\gamma_d^3\mathbb{E}[l'(X_1)^2l''(X_1)] - 3(1-2\gamma_d)\mathbb{E}[l'(X_1)l^{(3)}(X_1)] \\ &= (-9\gamma_d^2 + 12\gamma_d - 3)\mathbb{E}[l'(X_1)l^{(3)}(X_1)] - 3\gamma_d(\gamma_d^2-1)\mathbb{E}[l'(X_1)^2l''(X_1)] \\ &= -3(3\gamma_d-1)(\gamma_d-1)\mathbb{E}[l'(X_1)l^{(3)}(X_1)] - 3\gamma_d(\gamma_d^2-1)\mathbb{E}[l'(X_1)^2l''(X_1)] \\ &= (\gamma_d-1)\{-3(3\gamma_d-1)\mathbb{E}[l'(X_1)l^{(3)}(X_1)] - 3\gamma_d(\gamma_d+1)\mathbb{E}[l'(X_1)^2l''(X_1)]\} \\ &= (\gamma_d-1)K_{4,d}.\end{aligned}$$

Enfin, la limite de $K_{4,d}$ donne

$$\lim_{d\to\infty} K_{4,d} = -6(\mathbb{E}[l'(X_1)l^{(3)}(X_1)] + \mathbb{E}[l'(X_1)^2l''(X_1)]) < \infty,$$

car les dérivées de l sont bornées par un polynôme et tous les moments de f sont bornés, ce qui implique que $\lim_{d\to\infty} \mathbb{E}[C_{4,1}] = 0$. Pour le dernier terme $C_{6,1}$ on a que

$$\begin{aligned}\lim_{d\to\infty} \mathbb{E}[C_{6,1}] = \lim_{d\to\infty} \Bigg\{ &- \frac{15\gamma_d}{2}\{\gamma_d^2(1+3\gamma_d)\mathbb{E}[l'(X_1)^3l^{(3)}(X_1)] + 18\gamma_d\mathbb{E}[Z_1^2l'(X_1)l''(X_1)l^{(3)}(X_1)] \\ &+ 3\gamma_d^2(\mathbb{E}[l'(X_1)^2l''(X_1)^2] + 4\mathbb{E}[Z_1^2l'(X_1)^2l^{(4)}(X_1)] + \mathbb{E}[Z_1^4(3l^{(3)}(X_1)^2] \\ &+ 4\mathbb{E}[l''(X_1)l^{(4)}(X_1))]\} + \frac{15}{2}(1-5\gamma_d)\gamma_d\mathbb{E}[Z_1^4l'(X_1)l^{(5)}(X_1]) \\ &+ (1-3\gamma_d)\mathbb{E}[Z_1^6l^{(6)}(X_1)]\Bigg\} \\ = &-\frac{15}{2}\{4\mathbb{E}[l'(X_1)^3l^{(3)}(X_1)] + 18\mathbb{E}[Z_1^2l'(X_1)l''(X_1)l^{(3)}(X_1)] \\ &+ 3(\mathbb{E}[l'(X_1)^2l''(X_1)^2] + 4\mathbb{E}[Z_1^2l'(X_1)^2l^{(4)}(X_1)] + \mathbb{E}[Z_1^4(3l^{(3)}(X_1)^2] \\ &+ 4\mathbb{E}[l''(X_1)l^{(4)}(X_1))]\} - 30\mathbb{E}[Z_1^4l'(X_1)l^{(5)}(X_1]) - 2\mathbb{E}[Z_1^6l^{(6)}(X_1)]\end{aligned}$$

qui est bien borné puisque toutes les dérivées de l sont bornées par un polynôme et tous les moments de f sont bornés. □

Lemme A.2.2. *Soient les expressions* (A.2.1) *à* (A.2.6) *associées à* $C_{i,1}, i = 1, \ldots, 6$ *respectivement. On suppose que* $X_1 \in \mathbb{R}$ *est distribué selon une densité* f, $Z_1 \sim \mathcal{N}(0,1)$ *et* X_1 *et* Z_1 *sont indépendants. De plus, on suppose que tous les moments de* f *sont bornés, que* $l(x) = \log(f(x))$ *est* C^8 *et que* $l(x)$ *et ses 8 premières dérivées sont bornées par un polynôme, c'est-à-dire que*

$$|l(x)|, |l^{(i)}(x)| \leqslant M(x), \quad i \leqslant 8,$$

où $M(x)$ *est un polynôme positif. Alors,* $\lim_{d\to\infty} \mathbb{E}[C_{i,1}^2] < \infty$, $i = 1, \ldots, 6$ *où, plus particulièrement,*

$$\mathbb{E}[C_{1,1}^2] = (\gamma_d - 1)^2 K_1, \qquad \mathbb{E}[C_{2,1}^2] = (\gamma_d - 1)^2 K_{2,d},$$

où $K_1 < \infty$ *est indépendant de* d *et où* $\lim_{d\to\infty} K_{2,d} < \infty$.

Démonstration. Pour $C_{1,1} = (1 - \gamma_d) Z_1 l'(x_1)$, on observe que

$$\mathbb{E}[C_{1,1}^2] = (1 - \gamma_d)^2 \mathbb{E}[(Z_1 l'(X_1))^2]$$

et on a alors directement que $K_1 = \mathbb{E}[Z_1^2 l'(X_1)^2]$ qui ne dépend pas de d. De plus, Z_1 et X_1 sont indépendants et $l'(x)$ est bornée par un polynôme. Étant donné que tous les moments de f sont bornés, tout ceci implique que $\mathbb{E}[Z_1^2 l'(X_1)^2] < \infty$ et donc que $\lim_{d\to\infty} \mathbb{E}[C_{1,1}^2] = 0$.
En ce qui concerne le deuxième moment de $C_{2,1}$, on a que

$$\mathbb{E}[C_{2,1}^2] = (1 - \gamma_d)^2 \mathbb{E}[(l''(X_1) Z_1^2 + \gamma_d l'(X_1)^2)^2] = (1 - \gamma_d)^2 K_{2,d}.$$

De plus, il s'avère que

$$\begin{aligned}
\lim_{d\to\infty} K_{2,d} &= \lim_{d\to\infty} \mathbb{E}[l''(X_1)^2 Z_1^4 + 2 l''(X_1) Z_1^2 \gamma_d l'(X_1)^2 + \gamma_d^2 l'(X_1)^4] \\
&= \lim_{d\to\infty} \left\{ \mathbb{E}[l''(X_1)^2 Z_1^4] + \gamma_d \mathbb{E}[2 l''(X_1) Z_1^2 l'(X_1)^2] + \gamma_d^2 \mathbb{E}[l'(X_1)^4] \right\} \\
&= \mathbb{E}[l''(X_1)^2 Z_1^4] + \mathbb{E}[2 l''(X_1) Z_1^2 l'(X_1)^2] + \mathbb{E}[l'(X_1)^4].
\end{aligned}$$

Cette dernière égalité est bornée. En effet, il est possible de borner les deux premières dérivées de l par un polynôme et puisque X_1 et Z_1 sont indépendants et tous les moments de f sont

bornés, le tout est aussi borné. Ceci implique donc que $\lim_{d\to\infty} \mathbb{E}[C_{2,1}^2] = 0$. Pour les autres valeurs $C_{i,1}$, $i = 3, \ldots, 6$, on utilise la même technique pour borner la limite. Pour $C_{3,1}$, on a

$$\begin{aligned}
C_{3,1}^2 &\leqslant |C_{3,1}|^2 \\
&= \left| \left(\frac{3\gamma_d}{2} - 3\gamma_d^2 \right) Z_1 l'(X_1) l''(X_1) + \left(1 - \frac{3\gamma_d}{2} \right) Z_1^3 l^{(3)}(X_1) \right|^2 \\
&\leqslant \left(\left| \frac{3\gamma_d}{2} - 3\gamma_d^2 \right| |Z_1 l'(X_1) l''(X_1)| + \left| 1 - \frac{3\gamma_d}{2} \right| |Z_1^3 l^{(3)}(X_1)| \right)^2 .
\end{aligned}$$

Puisque $|l(x)|, |l^{(i)}(x)| \leqslant M(x)$, $i \leqslant 8$ où $M(\cdot)$ est un polynôme positif, on a que

$$\begin{aligned}
C_{3,1}^2 &\leqslant \left(\left| \frac{3\gamma_d}{2} - 3\gamma_d^2 \right| (1 + Z_1^4)(1 + M(X_1))^2 + \left| 1 - \frac{3\gamma_d}{2} \right| (1 + Z_1^4)(1 + M(X_1))^2 \right)^2 \\
&= (1 + Z_1^4)^2 (1 + M(X_1))^4 \left(\left| \frac{3\gamma_d}{2} - 3\gamma_d^2 \right| + \left| 1 - \frac{3\gamma_d}{2} \right| \right)^2 .
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\lim_{d\to\infty} \mathbb{E}[C_{3,1}^2] &\leqslant \mathbb{E}[(1 + Z_1^4)^2 (1 + M(X_1))^4] \lim_{d\to\infty} \left(\left| \frac{3\gamma_d}{2} - 3\gamma_d^2 \right| + \left| 1 - \frac{3\gamma_d}{2} \right| \right)^2 \\
&= 4\mathbb{E}[(1 + Z_1^4)^2 (1 + M(X_1))^4],
\end{aligned}$$

qui est borné en raison de l'indépendance entre X_1 et Z_1 et du fait que tous les moments de f sont bornés. La même idée est appliquée aux valeurs de $C_{i,1}$, $i = 4, 5, 6$. Pour $C_{4,1}$, on a

$$\begin{aligned}
C_{4,1}^2 &\leqslant |C_{4,1}|^2 \\
&= \Big| 3\gamma_d(1 - 3\gamma_d) Z_1^2 l'(X_1) l^{(3)}(X_1) - 3\gamma_d Z_1^2 l''(X_1)^2 - 3\gamma_d^3 l'(X_1)^2 l''(X_1) \\
&\quad + (1 - 2\gamma_d) Z_1^4 l^{(4)}(X_1) \Big|^2 \\
&\leqslant \Big(3\gamma_d |1 - 3\gamma_d| |Z_1^2 l'(X_1) l^{(3)}(X_1)| + 3\gamma_d |Z_1^2 l''(X_1)^2| + 3\gamma_d^3 |l'(X_1)^2 l''(X_1)| \\
&\quad + |1 - 2\gamma_d| |Z_1^4 l^{(4)}(X_1)| \Big)^2 \\
&\leqslant \Big(3\gamma_d |1 - 3\gamma_d| (1 + Z_1^4)(1 + M(X_1))^3 + 3\gamma_d (1 + Z_1^4)(1 + M(X_1))^3 \\
&\quad + 3\gamma_d^3 (1 + Z_1^4)(1 + M(X_1))^3 + |1 - 2\gamma_d| (1 + Z_1^4)(1 + M(X_1))^3 \Big)^2 \\
&= (1 + Z_1^4)^2 (1 + M(X_1))^6 \left(3\gamma_d |1 - 3\gamma_d| + 3\gamma_d + 3\gamma_d^3 + |1 - 2\gamma_d| \right)^2 .
\end{aligned}$$

Ainsi, la limite du deuxième moment vaut

$$\begin{aligned}\lim_{d\to\infty}\mathbb{E}[C_{4,1}^2] &\leqslant \mathbb{E}[(1+Z_1^4)^2(1+M(X_1))^6]\lim_{d\to\infty}\Big(3\gamma_d|1-3\gamma_d|+3\gamma_d+3\gamma_d^3+|1-2\gamma_d|\Big)^2\\ &= 169\mathbb{E}[(1+Z_1^4)^2(1+M(X_1))^6]\end{aligned}$$

qui est bien borné comme pour $C_{3,1}$. Pour $C_{5,1}$, on a

$$\begin{aligned}C_{5,1}^2 &\leqslant |C_{5,1}|^2\\ &= \bigg|\left(1-\frac{5\gamma_d}{2}\right)Z_1^5 l^{(5)}(X_1) - 15\gamma_d^3 Z_1^3 l''(X_1)l^{(3)}(X_1) - 5\gamma_d(1-3\gamma_d)Z_1^3 l'(X_1)l^{(4)}(X_1)\\ &\quad - 15\frac{\gamma_d^3}{2}Z_1 l'(X_1)^2 l^{(3)}(X_1) - 15\gamma_d^2 Z_1 l'(X_1)l''(X_1)^2\bigg|^2\\ &\leqslant \bigg(\left|1-\frac{5\gamma_d}{2}\right||Z_1^5 l^{(5)}(X_1)| + 15\gamma_d^3|Z_1^3 l''(X_1)l^{(3)}(X_1)| + 5\gamma_d|1-3\gamma_d||Z_1^3 l'(X_1)l^{(4)}(X_1)|\\ &\quad + 15\frac{\gamma_d^3}{2}|Z_1 l'(X_1)^2 l^{(3)}(X_1)| + 15\gamma_d^2|Z_1 l'(X_1)l''(X_1)^2|\bigg)^2\\ &\leqslant \bigg(\left|1-\frac{5\gamma_d}{2}\right|(1+Z_1^6)(1+M(x))^3 + 15\gamma_d^3(1+Z_1^6)(1+M(X_1))^3\\ &\quad + 5\gamma_d|1-3\gamma_d|(1+Z_1^6)(1+M(X_1))^3 + 15\frac{\gamma_d^3}{2}(1+Z_1^6)(1+M(X_1))^3\\ &\quad + 15\gamma_d^2(1+Z_1^6)(1+M(X_1))^3\bigg)^2\\ &= (1+Z_1^6)^2(1+M(X_1))^6\bigg(\left|1-\frac{5\gamma_d}{2}\right| + 15\gamma_d^3 + 5\gamma_d|1-3\gamma_d| + 15\frac{\gamma_d^3}{2} + 15\gamma_d^2\bigg)^2.\end{aligned}$$

Ainsi, étant donné que

$$\lim_{d\to\infty}\bigg(\left|1-\frac{5\gamma_d}{2}\right| + 15\gamma_d^3 + 5\gamma_d|1-3\gamma_d| + 15\frac{\gamma_d^3}{2} + 15\gamma_d^2\bigg)^2 = 2401,$$

la limite de l'espérance est bornée par

$$\lim_{d\to\infty}\mathbb{E}[C_{5,1}^2] \leqslant 2401\mathbb{E}[(1+Z_1^6)^2(1+M(X_1))^6] < \infty.$$

Enfin, pour le dernier terme, on a que

$$
\begin{aligned}
C_{6,1}^2 &= |C_{6,1}|^2 \\
&= \Big| -\frac{15\gamma_d}{2}\{\gamma_d^2(1+3\gamma_d)l'(X_1)^3 l^{(3)}(X_1) + 18\gamma_d Z_1^2 l'(X_1) l''(X_1) l^{(3)}(X_1) \\
&\quad + 3\gamma_d^2 l'(X_1)^2 l''(X_1)^2 + 13\gamma_d^2 Z_1^2 l'(X_1)^2 l^{(4)}(X_1) + 3Z_1^4 l^{(3)}(X_1)^2 + 4Z_1^4 l''(X_1) l^{(4)}(X_1)\} \\
&\quad + \frac{15}{2}(1-5\gamma_d)\gamma_d Z_1^4 l'(X_1) l^{(5)}(X_1) + (1-3\gamma_d) Z_1^6 l^{(6)}(X_1)\Big|^2 \\
&\leqslant \bigg(\frac{15\gamma_d}{2}\{\gamma_d^2(1+3\gamma_d)|l'(X_1)^3 l^{(3)}(X_1)| + 18\gamma_d |Z_1^2 l'(X_1) l''(X_1) l^{(3)}(X_1)| \\
&\quad + 3\gamma_d^2 |l'(X_1)^2 l''(X_1)^2| + 13\gamma_d^2 |Z_1^2 l'(X_1)^2 l^{(4)}(X_1)| + 3|Z_1^4 l^{(3)}(X_1)^2| + 4|Z_1^4 l''(X_1) l^{(4)}(X_1)|\} \\
&\quad + \frac{15}{2}|1-5\gamma_d|\gamma_d |Z_1^4 l'(X_1) l^{(5)}(X_1)| + |1-3\gamma_d||Z_1^6 l^{(6)}(X_1)|\bigg)^2 \\
&\leqslant (1+Z_1^6)^2(1+M(x))^8 \bigg(\frac{15\gamma_d}{2}\{\gamma_d^2(1+3\gamma_d) + 18\gamma_d + 3\gamma_d^2 + 13\gamma_d^2 + 3 + 4\} \\
&\qquad\qquad + \frac{15}{2}|1-5\gamma_d|\gamma_d + |1-3\gamma_d|\bigg)^2.
\end{aligned}
$$

Ainsi, puisque

$$
\lim_{d\to\infty}\bigg(\frac{15\gamma_d}{2}\{\gamma_d^2(1+3\gamma_d) + 18\gamma_d + 3\gamma_d^2 + 13\gamma_d^2 + 3 + 4 + |1-5\gamma_d|\} + |1-3\gamma_d|\bigg)^2 = \frac{546121}{4},
$$

alors

$$
\lim_{d\to\infty}\mathbb{E}[C_{6,1}^2] \leqslant \frac{546121}{4}\mathbb{E}[(1+Z_1^6)^2(1+M(X_1))^8] < \infty,
$$

qui est bien borné puisque Z_1 et X_1 sont indépendants et tous les moments de f sont bornés. □