Université de Montréal

Quelques contributions sur les méthodes de Monte Carlo

par

Yves F. ATCHADÉ

Département de mathématiques et de statistique

Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de

Philosophiæ Doctor (Ph.D.)
en statistique

septembre 2003

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée

Quelques contributions sur les méthodes de Monte Carlo

présentée par

Yves F. ATCHADÉ

a été évaluée par un jury composé des personnes suivantes :

Roch Roy
(président-rapporteur)

François Perron
(directeur de recherche)

Bruno Rémillard
(membre du jury)

Jeffrey S. Rosenthal
(examinateur externe)

Normand Mousseau
(représentant du doyen de la FES)

Thèse acceptée le:
25 septembre 2003

SOMMAIRE

Cette thèse porte sur les méthodes de Monte Carlo pour l'inférence statistique. Nous nous sommes intéressés à divers aspects des méthodes de Monte Carlo. Dans le premier essai, nous développons une méthode de réduction de variance par variables de contrôle. Le cadre est celui des méthodes de Monte Carlo statiques. La réduction de variance par variable de contrôle n'est pas une idée nouvelle. Mais contrairement à l'approche classique, notre méthode fournit un estimateur sans biais. En terme d'erreur quadratique moyenne, nous obtenons un estimateur asymptotiquement équivalent à l'estimateur usuel. Nous appliquons notre idée à l'algorithme de Rejet et à la méthode "Importance Sampling".

Le second essai porte également sur la réduction de variance. Mais le contexte est celui de l'algorithme de Metropolis Indépendant. Nous montrons que l'utilisation de variables de contrôle, très répandue en Monte Carlo statique, peut se développer aussi pour l'algorithme de Metropolis Indépendant. En utilisant des arguments de symétrie et la méthode de Rao-Blackwellisation, nous introduisons également de nouveaux estimateurs statistiquement plus efficaces que l'estimateur usuel.

Dans le troisième essai, nous nous sommes intéressés à l'ergodicité géométrique de l'algorithme de Metropolis-Hastings général. Nous avons développé une condition suffisante (connue par ailleurs comme étant nécessaire) d'ergodicité géométrique pour cet algorithme. Nous prouvons également que le spectre de l'opérateur engendré par l'algorithme de Metropolis Hastings Indépendant est à toute fin pratique, l'ensemble des valeurs prises par la probabilité de rejet de la chaîne. Finallement, nous proposons une version de l'algorithme de Metropolis Hastings qui

a des propriétés de convergence équivalentes à celles d'un algorithme de Metropolis Indépendant et qui explore l'espace des états de la chaîne de façon similaire à l'algorithme de Metropolis Marche Aléatoire Symétrique.

MOTS CLÉS

Statistique, Probabilité, Méthodes de Monte Carlo, Méthodes de Monte Carlo par Chaînes de Markov, Algorithme de Rejet, Algorithme de Metropolis-Hastings, Méthodes de Réduction de Variance, Rao-Blackwellisation, Ergodicité Géométrique, Trou Spectral d'Opérateurs Markoviens Auto-Adjoints.

SUMMARY

This thesis is about Monte Carlo methods for statistical inference. It is built on three independent essays. In the first essay, we propose a variance reduction method based on control variates in the static Monte Carlo framework. The use of control variates for variance reduction is not a new idea. But in this work, we obtain an unbiaised estimator which does not require the estimation of the Fourier coefficients. In terms of its mean quadratic error, our estimator is asymptotically equivalent to the estimator obtained when the Fourier coefficients are known. The idea is apply to the Rejection algorithm and to the Importance Sampling methods.

In our second essay, we develop various variance reduction methods for integrals computation when the algorithm used is the Independent Metropolis-Hastings algorithm. We show that it is possible to use control variates in the context of this algorithm. Moreover, our derived estimator is easy to compute. In another direction and based on the lack of symmetry of the Independence Sampler and the so-called Rao-Blackwellization technique, we introduce new estimators which are more efficient than the usual estimator.

Our third essay is about the geometric ergodicity of the Metropolis-Hastings algorithm. We derive a sufficient condition (also known to be necessary) for the geometric ergodicity of the general Metropolis-Hastings algorithm. Also, we show that the spectrum of the operator of the Independence Sampler is the essential range of the probability of rejection of the algorithm. This result actually solves a conjecture made by Jun S. Liu (1996). Finally, we propose a new Metropolis-Hastings move that combines the properties of the Independence Sampler and the Random Walk Metropolis moves.

Key Words

Statistics, Probability, Monte Carlo Methods, Markov Chain Monte Carlo, Rejection Algorithm, Metropolis-Hastings Algorithm, Variance Reduction Methods, Rao-Blackwellization, Geometric Ergodicity, Spectral Gap for Markovian Self-Adjoint Operators.

DÉDICACE

À ma mère.

REMERCIEMENTS

Cette thèse n'aurait jamais vu le jour sans mon directeur de recherche François Perron. François m'a enseigné la statistique, mais également m'a montré le chemin à prendre pour être un bon chercheur. Sa disponibilité exemplaire, sa grande implication dans nos recherches, la confiance qu'il m'a fait à pouvoir développer mes propres idées et le climat de travail qu'il a instauré autour de lui ont été pour moi de puissants stimulants. Je le remercie pour son soutien indéfectible sur tous mes dossiers accadémiques et pour son soutien financier particulièrement pendant mes premières années difficiles comme étudiant étranger à l'université de Montréal.

Je voudrais remercier Marlène Frigon pour ses conseils, son soutien pour nombre de mes dossiers et pour les magnifiques cours d'analyse fonctionnelle linéaire et nonlinéaire que j'ai eu la chance de suivre avec elle.

Je remercie le Directeur du département Monsieur Yvan Saint-Aubin pour le dynamisme qu'il a apporté au département. Je remercie le professeur Richard Duncan pour sa disponibilité et son aide sur la théorie des opérateurs, les professeurs Martin Goldstein, Jean-François Angers, René Ferland de l'UQAM, Angelo Canti qui était à l'époque à l'Université Concordia, Mario Lefebvre de l'École Polytechnique de Montréal, Martin Bilodeau, Daniel Dufresne pour leur enseignement.

Je voudrais remercier la Faculté des Études Supérieures de l'Université de Montréal, mon directeur de recherche François Perron, l'ISM, le fonds CRSNG et le fonds FCAR pour leur support financier.

Il me fait plaisir de remercier ma mère pour avoir toujours cru en moi. Son dévouement et ses prières constantes sont mon réconfort. Je remercie mon père,

la première femme de mon père, mes frères et soeurs : Djossè, Bernadette, Popo, Edgard, Clarisse, Sylvie, Aimée, Martin, Clovis et Gina ma préférée (ce n'est un secret pour personne...). Particulièrement Clarisse et Aurèle : merci d'avoir toujours été là lorsque j'ai eu besoin de vous. Je remercie aussi toute ma grande famille, particulièrement Elvire, Coco, Adelaide, Edith, Paterne, Patou, tonton Frédé, tonton Willi, Mama-Gouan, mes grand-mères et tous ceux que malheureusement j'oublie.

Je voudrais remercier spécialement Aïcha Sow pour sa gentillesse, pour toutes les (nombreuses) fois où elle m'a remonté le moral lorsque mes travaux piétinaient, et sa patience face à tous les "il faut que je bosse".

Je remercie mes amis de longue date qui ont fait l'aventure québecoise avec moi : Georges Tsafack, Dovonon Prosper, Franck Goussanou, Damien Fousseni, Jacques Ewoudou, Sévérien Nkurunziza. Lâchez pas les gars !

Je voudrais aussi remercier Mario Lavallée, Président de Finlab Inc. pour son support et l'opportunité qu'il m'a donné de travailler avec lui. Je me suis fait de très bons amis à Finlab qui ont contribué à faire de mon séjour à Montréal une très belle expérience : Hugo Sarkisian, Sylvestre Izelimana, Benoit Crispin, Pierre Therrien, Emilie Blouin. Merci à vous tous. Je remercie tous les étudiants du Département de Mathématiques et Statistique de l'Université de Montréal avec une attention particulière à Pierre Lafaye, Alain Desgagné, Alexandre Leblanc, Geneviève Lefebvre, Terence Ngoakou pour les bons conseils prodigués et les bons moments passés ensemble.

Table des matières

INTRODUCTION

0.1. Introduction aux méthodes de Monte Carlo

Les méthodes de Monte Carlo sont une branche des mathématiques qui étudie les expériences sur les nombres aléatoires. L'approche des méthodes de Monte Carlo consiste à utiliser les lois de probabilité des nombres aléatoires avec des expérimentations rigoureuses sur ces nombres pour arriver à résoudre des problèmes mathématiques (qui sont dans bien des cas difficiles à résoudre autrement). Il existe deux contextes assez distincts d'utilisation des méthodes de Monte Carlo. On peut utiliser les méthodes de Monte Carlo dans le cas où on a un système stochastique dynamique trop complexe pour pouvoir faire l'objet d'une analyse rigoureuse. Grâce aux méthodes de Monte Carlo, un tel système peut alors être simulé sur un ordinateur, ce qui en facilite l'étude. C'est une approche très commune dans les sciences des ingénieurs. Dans un tout autre contexte, les méthodes de Monte Carlo prennent la forme d'un outil de calcul numérique. Ici, on est en présence d'un modèle mathématique dont la résolution conduit à des quantités mathématiques difficiles à évaluer telles que des intégrales, des solutions de programmes d'optimisation, des équations linéaires et nonlinéaires en grandes dimensions.

Dans cette thèse, notre contribution se situe principalement dans cette deuxième optique où les méthodes de Monte Carlo sont vues comme un outil de calcul numérique. En tant que telles, ces méthodes ont eu et continuent d'avoir d'importantes applications dans des domaines aussi variés que la biologie (Karplus et Petsko (1990), Leach (1996)), la chimie (Alder et Wainwright (1959)), l'informatique (Kirkpatrick et al. (1983), Ullman (1984)) la physique (Metropolis et al. (1953), Goodman et Sokal (1989)), la finance (Chib et al. (2002)) et bien sûr la statistique et la liste n'est pas exhaustive. La raison principale est que l'analyse numérique classique souffre de ce que certains ont appelé la malédiction de la dimension. C'est-à-dire le fait que l'efficacité de ces méthodes décroit exponentiellement avec la dimension du problème à résoudre. Par contre, les méthodes de Monte Carlo ont cette propriété intéressante que leur performance ne dépend pas de la dimension du problème (à tout le moins en théorie). Avec la complexification croissante des modèles mathématiques utilisés dans les disciplines scientifiques,

il est à prévoir que l'approche Monte Carlo prendra une place de plus en plus importante comme méthode numérique.

A la base des méthodes de Monte Carlo se trouvent les générateurs de nombres aléatoires uniformément distribués sur l'intervalle unité. C'est-à-dire des programmes informatiques capables de générer des nombres aléatoires dans l'intervalle $(0, 1)$. Il est clair qu'un programme informatique ne peut pas générer des nombres aléatoires. En fait, ce que ces programmes essaient de faire, c'est de générer une séquence de nombres qui paraîtra imprédictible à quiconque ne disposant pas du programme générateur. Il existe une vaste littérature sur l'élaboration de tels générateurs. Nous n'en dirons pas plus sur les générateurs de nombres aléatoires et nous référons le lecteur intéressé à ce que D. Knuth en dit dans son impressionnante monographie Knuth (1997). Voir aussi L'Ecuyer (1994) et le site internet du groupe PLab (http ://random.mat.sbg.ac.at). Dans toute la suite, nous supposons avoir à notre disposition un générateur de nombres aléatoires parfait.

A partir d'un générateur de nombres aléatoires, il est possible par transformation d'uniformes de générer des variables aléatoires distribués suivant d'autres distributions de probabilité. Nous présentons brièvement quelques méthodes universelles de cette approche à la section 0.2. Nous parlerons de la méthode d'inversion, la méthode de rejet, la méthode "Importance Sampling". La grande référence sur le sujet est sans nul doute Devroye (1986) et nous prions le lecteur de bien vouloir s'y référer pour plus de détails. A la section 0.3, nous présentons les idées de base sur l'utilisation de nombres aléatoires pour faire du calcul d'intégrales. Et nous discuterons aussi de la méthode de réduction de variance par variables de contrôle. Ces discussions permettront d'apprécier les contributions que nous avons faites dans notre premier essai "Monte Carlo Simulations via Control Variates".

Mais lorsqu'on travaille sur des modèles complexes, il n'est pas rare qu'on obtienne des distributions de probabilité dont il est impossible ou très couteux de simuler directement. Metropolis et al. (1953) ont développé l'idée fondamentale suivante. Supposons qu'on a une chaîne de Markov dont la distribution stationnaire est la distribution d'intérêt. Si la chaîne est ergodique (nous préciserons

le sens de ce terme plus tard), si on la laisse évoluer suffisamment longtemps, ses réalisations pourront être assimilées à des réalisations de la distribution de probabilité d'intérêt. C'est la génèse des méthodes de Monte Carlo par chaînes de Markov. Par la suite, nous utiliserons l'acronyme MCMC qui correspond à Markov Chain Monte Carlo en anglais. Mais il faudra attendre les années 90 avec l'article séminal de Gelfand et Smith (1990) pour que la communauté statistique commence à s'intéresser à grande échelle aux méthodes MCMC. En statistique, les méthodes MCMC se sont révélées un puissant outil de calcul numérique qui a grandement contribué au développement fulgurant que l'analyse statistique bayésienne a connu au cours des années 90. Voir par exemple le livre Robert (2001) pour une présentation de la statistique bayésienne contemporaine et l'article Berger (2000) qui fait une revue de littérature sur l'utilisation de l'approche bayésienne dans les autres disciplines scientifiques. Aujourd'hui, la littérature sur les méthodes MCMC est très vaste. Ceci d'autant plus que l'analyse des algorithmes MCMC est directement connectée à l'étude générale des chaînes de Markov qui est un domaine vaste en soit. Comme références introductives, nous pouvons citer Tierney (1994), Gilks et al. (1996), Roberts et Rosenthal (1998),Casella et Robert (1999), Liu (1999), Liu (2001). Dans la section 0.4, nous donnons une petite introduction aux méthodes MCMC pour préparer le lecteur aux contenus des deux derniers essais.

0.2. Méthodes de simulation de variables aléatoires

A la base de toutes les méthodes de Monte Carlo se trouvent les générateurs de la loi uniforme sur $(0,1)$. Un générateur d'uniformes est un algorithme déterministe qui produit une suite (V_n) d'éléments sur $(0,1)$ telle que toute séquence $(V_1,\ldots,V_n)$ extraite de (V_n) peut être validée comme une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) uniformes sur $(0,1)$ à travers une batterie quelconque de tests (Niederreiter (1992), L'Ecuyer (1994)). Dans tout ce qui suit, nous supposerons avoir à notre disposition un générateur d'uniformes parfait.

Le lemme suivant est fondamental pour simuler des distributions sur $\mathbb{R}$ à partir d'uniformes.

Lemme 0.2.1. *Soit F une fonction de répartition sur $\mathbb{R}$. On définit*

$$F^{-1}(x) := \inf \{t \in \mathbb{R} : \ F(t) \geq x\}$$

Si $U \sim \mathcal{U}(0,1)$, alors $F^{-1}(U)$ est distribué suivant F.

Exemple 0.2.1 (Comment simuler une loi de Cauchy $\mathcal{C}(0,1)$). *Supposons que X suit une loi de cauchy $\mathcal{C}(0,1)$. Alors la fonction de répartition de X s'écrit : $F(x) = \frac{1}{2} + \frac{1}{\pi} Arctan(x)$. Donc $F^{-1}(t) = \tan\left(\pi(t - \frac{1}{2})\right)$. D'où l'algorithme pour simuler une loi de Cauchy :*

- *Générer $U \sim \mathcal{U}(0,1)$.*
- *Faire $X = \tan\left(\pi(U - \frac{1}{2})\right)$.*

Il y a au moins deux problèmes à l'utilisation de ce lemme. Il n'est pas toujours posssible d'avoir une expression exacte de F^{-1} (exemple de la loi normale) et la méthode se généralise mal au cas multidimensionnel. Toutefois, il existe une version multidimensionnelle de cet algorithme utilisant les copules (Nelsen (1999) page 35, Devroye (1986)).

Une alternative très populaire qui fonctionne tout aussi bien dans le cas multidimensionnel est l'algorithme de Rejet (von Neumann (1951)).

Supposons qu'on veut simuler des observations d'une mesure de probabilité π sur un espace mesurable $(\mathcal{X}, \mathcal{F})$. Supposons qu'on dispose d'une autre mesure de probabilité Q qui est plus facile à simuler que π. Supposons aussi que π est absolument continue par rapport à Q et écrivons $\omega(x) = \frac{\pi(dx)}{Q(dx)}$ la densité de Radon-Nikodym de π par rapport à Q. Supposons enfin que Q est assez proche de π dans le sens que $\omega(x) \leq M$ pour tout $x \in \mathcal{X}$ où M est une constante.

Algorithme 0.2.1 (Algorithme de Rejet). *(1) Simuler Z suivant Q, simuler U suivant $\mathcal{U}(0,1)$.*

(2) Si $U \leq \frac{\omega(Z)}{M}$, accepter Z.

Sinon retourner à l'étape 1.

Soit $Z_1, \ldots, Z_{n+t}$ les variables simulées suivant Q par l'algorithme pour pouvoir en accepter n. Notons $X_1, \ldots, X_n$ les variables acceptées et $Y_1, \ldots, Y_t$ les variables rejetées.

Proposition 0.2.1. *(1) Chaque X_i est distribué suivant π.*

(2) Chaque Y_j est distribution suivant $\frac{Q(\cdot)-\rho\pi(\cdot)}{1-\rho}$, avec $\rho = \frac{1}{M}$.

(3) $n+t$ est distribué suivant une loi binomiale négative de paramètres n et ρ.

Exemple 0.2.2 (Simuler une loi normale $N(0,1)$ à partir d'une cauchy $C(0,1)$). *Prenons $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$ et $g(x) = \frac{1}{\pi(1+x^2)}$. Donc on peut prendre $M = \sup_{x\in\mathbb{R}} \left(\frac{f(x)}{g(x)}\right) = \sqrt{2\pi} \exp\left(-\frac{1}{2}\right)$. D'où l'algorithme suivant qui permet de simuler une loi normale à partir d'une loi de Cauchy.*

Algorithme 0.2.2. *(1) Simuler $V \sim \mathcal{U}(0,1)$, simuler $U \sim \mathcal{U}(0,1)$.*

Faire $Y = \tan\left(\pi\left(V - \frac{1}{2}\right)\right)$.

(2) Si $U \leq 2(1+Y^2)\exp(-\frac{1}{2}(Y^2+1))$, accepter Y.

Sinon retourner à l'étape 1.

Un point intéressant de l'algorithme de rejet est qu'il peut être utilisé même si ω est connu seulement à une constante de normalisation près. Mais si π est compliquée, M peut être difficile à trouver. D'autre part, étant donné qu'il faut générer en moyenne M observations Z pour en accepter une, plus M est grand, plus l'algorithme sera inefficace dans le sens qu'il va "gaspiller" les Z.

0.3. Méthodes de Monte Carlo Statique

0.3.1. Intégration par Monte Carlo

Considérons toujours notre espace de probabilité $(\mathcal{X}, \mathcal{F}, \pi)$. Supposons maintenant qu'on voudrait évaluer une intégrale $\pi(f) := \int f(x)\pi(dx)$, où f est une fonction intégrable $f : \mathcal{X} \longrightarrow \mathbb{R}$. Typiquement, $\mathcal{X} = \mathbb{R}^d$. Dans cette section, nous introduisons les méthodes de Monte Carlo pour le calcul intégral basées sur des observations indépendantes, que nous appellerons les méthodes de Monte Carlo statiques. Ces méthodes se basent sur le théorème suivant.

Théorème 0.3.1. *Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées suivant π. Nous avons :*

(1)

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{p.s.} \pi(f).$$

(2) Si de plus $\pi(|f|^2) < \infty$, alors

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \pi(f)\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2(f)),$$

où $\sigma^2(f) = \pi(|f|^2) - (\pi(f))^2$.

Donc pour calculer $\pi(f)$, on va simuler $X_1, \ldots, X_n$ i.i.d. π (par exemple en utilisant l'algorithme Rejet) et utiliser

$$\hat{\pi}_0(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i). \tag{0.3.1}$$

De plus, le point (2) du théorème 0.3.1 nous enseigne que la vitesse à laquelle $\hat{\pi}_0(f)$ converge vers $\pi(f)$ est de l'ordre de $\frac{1}{\sqrt{n}}$, indépendamment de la dimension de l'espace $\mathcal{X}$. En pratique, on utilise le fait que $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \pi(f)\right)$ converge en loi vers une loi normale pour obtenir un intervalle de confiance pour $\pi(f)$, ce qui tient lieu d'estimateur de l'erreur commise en estimant $\pi(f)$ par $\hat{\pi}_0(f)$. Cette méthode d'estimation d'erreur explique pourquoi l'existence d'un théorème limite central est si importante pour l'utilisation des méthodes de Monte Carlo.

On peut ausi utiliser la méthode "Importance Sampling" de Marshall (1956). A cette fin, on note que $\pi(f) = \int f(x)\omega(x)Q(dx)$. Ainsi, on peut simuler $Z_1, \ldots, Z_n$ i.i.d. Q et utiliser

$$\hat{\pi}_1(f) = \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\omega(Z_i). \tag{0.3.2}$$

Exemple 0.3.1 (Calcul de $\int_2^\infty x^5 f(x)dx$ où f est la densité d'une loi de student en simulant d'une Cauchy). *Soit $f(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$ la densité d'une loi de student à $\nu > 1$ degré de liberté. Supposons qu'on veut estimer $\int_2^\infty x^5 f(x)dx$. On peut utiliser la méthode "Importance Sampling" en simulant des observations de la loi de Cauchy $C(0,1)$. Notons $g(x) = \frac{1}{\pi(1+x^2)}$ la densité de la loi de Cauchy $C(0,1)$. On obtient l'algorithme :*

Algorithme 0.3.1. *(1) Simuler $U_1 \ldots, U_n$ i.i.d. $\mathcal{U}(0,1)$.*

(2) Pour $i = 1$ à n, faire $Y_i = \tan\left(\pi\left(U_i - \frac{1}{2}\right)\right)$, faire $X_i = Y_i$ si $Y_i > 2$ et $X_i = 0$ sinon.

(3) Estimer I par :

$$\hat{I} = \frac{1}{n}\sum_{k=1}^{n} X_i^5 f(X_i)/g(X_i).$$

En faisant une petite simulation, le lecteur peut facilement se convaincre que $\hat{I}$ est meilleur que l'estimateur qu'on obtiendrait en simulant directement de $t(\nu)$, l'écart s'accentuant à mesure que ν devient grand.

Pour que le Théorème 0.3.1 (2) soit valide pour $\hat{\pi}_1(f)$, il faut garantir que $\int f^2(x)\omega^2(x)Q(dx) < \infty$. Une condition suffisante pour cela est que $\pi(f^2) < \infty$ et que la densité ω soit bornée. C'est-à-dire que Q ait des queues plus épaisses que π. Voir Geweke (1989) pour d'autres conditions. Lorsque ω est bornée, on a vu qu'on peut aussi utiliser l'algorithme de Rejet (Algorithme 0.2.1). La question se pose donc de savoir laquelle des deux approches est la meilleure entre utiliser (0.3.1) où $X_1, \ldots, X_n$ sont les observations acceptées de l'algorithme de Rejet ou utiliser (0.3.2) avec les $n + t$ observations Z_i de l'algorithme de Rejet nécessaires pour obtenir ces n observations $X_1, \ldots, X_n$. Il n'est pas facile de répondre à une telle question. On peut trouver des éléments de réponse dans Casella and Robert (1998).

Une autre question intéressante est celle de déterminer la meilleure mesure de probabilité Q à prendre pour utiliser la méthode Importance Sampling. La proposition suivante répond à la question et peut être trouvée dans Robert and Casella (1999).

Proposition 0.3.1. *La variance de $\hat{\mu}_1(f)$ est minimisée pour $Q(dx) \propto |f|(x)\pi(dx)$.*

0.3.2. Méthode de la variable de contrôle pour la réduction de variance

Il existe plusieurs méthodes de réduction de variance en simulation Monte Carlo. Pour plus de détails on peut consulter Ripley (1987), Fishman (1996).

Nous présentons ici la méthode de réduction de variance par variable de contrôle parce qu'on s'y est spécifiquement intéressé.

Le contexte est le même qu'à la section 0.2. Supposons qu'on voudrait évaluer l'intégrale $\pi(f)$. Supposons qu'on connait une autre fonction C telle que $\pi(C) = 0$ et $\pi(C^2) < \infty$. Soit $X_1, \ldots, X_n$ i.i.d. π. L'estimateur par variable de contrôle est :

$$\hat{\pi}_{CV}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \beta\frac{1}{n}\sum_{i=1}^{n} C(X_i) \tag{0.3.3}$$

En choisissant $\beta = \frac{\pi(fC)}{\pi(C^2)}$, on peut montrer que $\hat{\pi}_{CV}(f)$ a une variance plus petite ou égale à la variance de l'estimateur (0.3.1). C'est dans ce sens que l'utilisation d'une variable de contrôle constitue une méthode de réduction de variance. C'est une méthode très efficace particulièrement parce qu'elle fait intervenir très peu de calculs supplémentaires par rapport à l'estimateur de base $\hat{\pi}_0(f)$ comparé à d'autres méthodes de réduction de variance. Mais en pratique, très souvent la valeur optimale de β donnée par $\frac{\pi(fC)}{\pi(C^2)}$ n'est pas connue. Dans notre premier essai, "Monte Carlo Simulation via Control Variates", nous proposons un estimateur où la connaissance de β n'est pas nécessaire, dans le sens que notre estimateur estime β. Nous montrons que notre estimateur est sans biais et est asymptotiquement équivalent à $\hat{\pi}_{CV}(f)$.

Une question naturelle est de savoir s'il est aussi possible d'utiliser l'estimateur $\hat{\pi}_{CV}(f)$ lorsque les (X_i) forment une chaîne de Markov avec une expression simple à estimer pour la valeur optimale de β. Nous nous sommes intéressé à la question et nous y donnons une réponse affirmative dans notre essai "Improving on the Independence Metropolis-Hastings Algorithm".

0.4. Méthodes de Monte Carlo par chaînes de Markov

Lorsque la dimension de l'espace devient très grande, et/où que la distribution d'intérêt est compliquée, il est souvent difficile de simuler des observations i.i.d. de cette distribution. Une solution possible est de simuler d'une autre distribution plus simple et d'utiliser l'échantionneur pondéré comme on l'a vu à la section 0.3.1. Une autre possibilité est de simuler des observations dépendantes

approximativement distribuées selon la distribution d'intérêt par les méthodes MCMC.

0.4.1. Introduction aux Chaînes de Markov

Nous commençons par une introduction aux chaînes de Markov. Notre présentation est basée sur Meyn et Tweedie (1993) et nous référons le lecteur à cet ouvrage pour les preuves et des références supplémentaires. Le contexte est le suivant. Soit $(\mathcal{X}, \mathcal{F}, \pi)$ un espace de probabilité où π est la mesure de probabilité d'intérêt.

Définition 0.4.1. *Une chaîne de Markov (homogène) sur $(\mathcal{X}, \mathcal{F})$ est un processus stochastique $(X_n)_{n\geq 0}$ tel que :*

$$\begin{aligned} \Pr(X_n \in A | X_0, \ldots, X_{n-1}) &= \Pr(X_n \in A | X_{n-1}) \\ &= \Pr(X_1 \in A | X_0). \end{aligned}$$

Définissons $P : \mathcal{X} \times \mathcal{F} \longrightarrow [0,1]$ par $P(x, A) := \Pr(X_1 \in A | X_0 = x)$. P s'appelle le noyau de transition de la chaîne. Plus généralement, on appelle noyau de transition toute fonction $P : \mathcal{X} \times \mathcal{F} \longrightarrow [0,1]$ telle que $P(x, \cdot)$ est une mesure de probabilité pour tout $x \in \mathcal{X}$ et telle que $P(\cdot, A)$ est une fonction mesurable pour tout $A \in \mathcal{F}$. A partir d'un noyau de transition P, on peut définir ses itérés qui sont aussi des noyaux de transition de la façon suivante :
$P^0(x, dy) := \delta_x(dy)$, où δ_x est la mesure de Dirac en x et
$P^n(x, dy) := \int P(x, dz) P^{n-1}(z, dy)$ pour $n \geq 1$.
En fait, $P^n(x, \cdot)$ est la distribution de X_n lorsque $X_0 = x$, de façon que partant d'une distribution initiale μ (c'est-à-dire la distribution de X_0) et de P, on peut obtenir toute la loi du processus (X_n). Et réciproquement, partant d'une distribution μ et d'un noyau de transition P, on peut construire une chaîne de Markov qui admet μ comme distribution initiale et P comme noyau de transition. Il s'agit là d'un résultat non trivial qui fait appel à des théorèmes de construction de processus stochastiques tels que le théorème d'existence de Kolmogorov (Billingsley (1993), section 36) ou le théorème de Ionescu-Tulcea (Neveu (1965), proposition

V 1.1.). Voir Meyn et Tweedie (1993), chapitre 3 pour plus de détails et d'autres références.

Définition 0.4.2. *Une chaîne de Markov de noyau de transition P est dite $\phi-$irreductible s'il existe une mesure de probabilité ϕ telle que :*
pour tout $x \in \mathcal{X}$, pour tout $A \in \mathcal{F}$ tel que $\phi(A) > 0$, il existe n_0 tel que $P^{n_0}(x, A) > 0$.

Définition 0.4.3. *Une chaîne de Markov de noyau de transition P est dite apériodique s'il n'existe pas de partition $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_d$ avec $d \geq 2$ telle que $P(x, \mathcal{S}_{i+1}) = 1$ pour tout $x \in \mathcal{S}_i$, $i = 1, \ldots, d-1$ et $P(x, \mathcal{S}_1) = 1$ pour tout $x \in \mathcal{S}_d$.*

Définition 0.4.4. *Une mesure de probabilité π est dite invariante pour le noyau de transition P si :*
$\pi(A) = \pi P(A) := \int \pi(dx) P(x, A)$, *pour tout $A \in \mathcal{F}$.*

Définition 0.4.5. *Une chaîne de Markov de noyau de transition P et de distribution invariante π est dite ergodique si :*
$\|P^n(x, \cdot) - \pi(\cdot)\| \xrightarrow[n\to\infty]{} 0,$
où $\|\mu\| := \sup\{|\mu(A)|,\ A \in \mathcal{F}\} = \frac{1}{2}\sup_{|f|\leq 1}\left|\int f(x)\mu(dx)\right|$ est la norme de la variation totale de μ.

Un résultat important de la théorie des chaînes de Markov dans des espaces généraux affirme qu'une chaîne de Markov $\phi-$irreductible, apériodique et qui admet une distribution invariante est ergodique.

Si une chaîne de Markov (X_n) de noyau de transition P est ergodique, alors pour toute fonction $f : \mathcal{X} \longrightarrow \mathbb{R}$ telle que $\pi(|f|) < \infty$:

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow[n\to\infty]{} \pi(f),$$

$\mathbb{P}_x-$presque surement pour $\pi-$presque tout x, où $\mathbb{P}_x$ est la loi du processus obtenue lorsque la loi initiale est la mesure de Dirac δ_x. Voir Athreya et al. (1996) pour une démonstration complète et détaillée. Ce résultat est souvent appelé le théorème ergodique pour les chaîne de Markov. Il remplace la loi forte des grands nombre sur laquelle se base les méthodes de Monte Carlo statiques (voir le théorème 0.3.1). On voit donc comment procède les méthodes MCMC : simuler

une chaîne de Markov ergodique de distribution invariante π et estimer $\pi(f)$ par $\hat{\pi}_0(f) = \frac{1}{n}\sum_{k=0}^{n-1} f(X_k)$.

Contrairement aux méthodes de Monte Carlo statiques où la condition $\pi(f^2) < \infty$ suffit pour obtenir un théorème limite central pour $\hat{\pi}_0(f)$, dans le cas des méthodes MCMC, cette condition ne suffit plus. L'existence d'un théorème limite central dépend fondamentalement de la vitesse à laquelle la chaîne converge vers sa distribution invariante.

Définition 0.4.6. *Une chaîne de Markov de noyau de transition P et de distribution invariante π est ergodique à la vitesse $r(n)$ (où $(r(n)_{n\geq 0})$ est une suite de nombres réels convergeant vers 0) s'il existe une fonction $V : \mathcal{X} \longrightarrow \mathbb{R}$ telle que :*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq V(x)r(n).$$

Lorsque $r(n) = \rho^n$ pour un certain $\rho < 1$, on dira que P est géométriquement ergodique (Nummelin et Tuominen (1982), Roberts et Rosenthal (1997)). Mais si $r(n) = n^{-q}$ pour un certain $q > 0$, on dira que la convergence de P est polynômiale (Nummelin et Tuominen (1983), Tuominen et Tweedie (1994), Jarner et Roberts (2002)).

Lorsque la vitesse de convergence d'une chaîne de Markov est géométrique, on obtient aisément un théoreme limite central pour $\hat{\pi}_0(f)$.

Théorème 0.4.1 (Chan & Geyer 94). *Si P est géométriquement ergodique, alors pour toute fonction f telle que $\pi(|f|^{2+\epsilon}) < \infty$, il existe $\sigma^2(f) < \infty$ telle que :*

$$\sqrt{n}\,(\hat{\pi}_0(f) - \pi(f)) \xrightarrow[n\to\infty]{\mathcal{L}} N(0, \sigma^2(f)). \tag{0.4.1}$$

Définition 0.4.7. *Une chaîne de Markov de noyau de transition P est dite réversible par rapport à la mesure de probabilité π ($\pi-$reversible) si $\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$ (en tant que mesures sur $\mathcal{X}\times\mathcal{X}$).*

En évaluant les deux membres de l'égalité de réversibilité $\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$ sur $A\times\mathcal{X}$, on obtient $\pi(A) = \int \pi(dy)P(y,A)$ pour tout $A \in \mathcal{X}$ qui dit que π est une distribution invariante pour P.

Théorème 0.4.2 (Roberts & Rosenthal 97). *Si en plus d'être géométriquement ergodique, P est $\pi-$reversible alors (0.4.1) est vrai même si f vérifie seulement $\pi(f^2) < \infty$.*

Le théorème 0.4.2 se base sur le théorème de Kipnis-Varadhan (Kipnis et Varadhan (1986)) qui est à notre connaissance le résultat le plus général qui existe sur les théorèmes limites centraux pour les chaînes réversibles. De façon générale, lorsqu'on perd l'ergodicité géométrique, la classe des fonctions pour lesquelles un théorème limite central est vérifié devient limitée. Par exemple, si la convergence de la chaîne n'est que polynomiale, on peut obtenir un théorème limite central pour les fonctions bornées (où plus généralement pour des fonctions dominées par un certaine fonction V qui intervient dans la vitesse de convergence de la chaîne). Voir Jarner et Roberts (2002) par exemple.

On va finir cette sous-section en faisant le lien entre les chaînes de Markov et les opérateurs linéaires qu'elles induisent. Pour plus de détails, nous référons à Krengel (1985), Baxter et Rosenthal (1995), Roberts et Rosenthal (1997). Notons $L^2(\pi)$ l'espace de Hilbert des fonctions $f : \mathcal{X} \to \mathbb{R}$ telles que $\pi(|f|^2) < \infty$. Un noyau de transition P de distribution invariant π induit un opérateur linéaire qu'on va noter T sur $L^2(\pi)$ par $Tf(x) := \int P(x, dy)f(y)$. T est un opérateur borné qui admet 1 comme valeur propre et qui est auto-adjoint lorsque P est $\pi-$réversible. Notons T_0 la restriction de T sur $L^2_0(\pi) := \{f \in L^2(\pi) : \ \pi(f) = 0\}$. Notons $r(T_0)$ le rayon spectral de T_0 défini par $r(T_0) := \lim_{n\to\infty} \|T_0^n\|^{\frac{1}{n}}$. Lorsque P est $\pi-$réversible, un résultat bien connu (obtenu par Roberts et Rosenthal (1997)) dit que l'ergodicité géométrique telle que définie à la définition 0.4.6 est équivalente à $r(T_0) < 1$. Dans le cas des chaînes non-réversibles, on peut encore se demander si l'ergodicité géométrique est aussi équivalente à $r(T_0) < 1$. A notre connaissance, il s'agit d'un problème ouvert. Si la réponse est oui, alors même sans l'hypothèse de réversibilité, on aurait un théorème limite central pour toutes les fonctions de $L^2(\pi)$, ce qui serait très intéressant en MCMC.

0.4.2. Algorithmes MCMC usuels

La question qui se pose maintenant est de savoir comment simuler en pratique une chaîne de Markov qui admet π comme distribution invariante. Comme nous l'avons vu, on peut se contenter de simuler des chaînes qui sont réversibles par

rapport à π ce qui est plus facile en général. L'algorithme pour ce faire est l'algorithme de Metropolis-Hastings. Comme précédemment, soit $(\mathcal{X}, \mathcal{F}, \pi)$ un espace de probabilité où π est la mesure de probabilité d'intérêt. Soit Q un noyau de transition sur $(\mathcal{X}, \mathcal{F})$ tel que $Q(x, \cdot)$ est absolument continue par rapport à π et nous écrirons $Q(x, dy) = \omega(x, y)\pi(dy)$. Le noyau de transition Q s'appelle communément le noyau de transition instrumental et les observations générées de ce noyau sont les observations proposées. Posons

$$\alpha(x, y) = \begin{cases} Min(1, \frac{\omega(y,x)}{\omega(x,y)}) & \text{si} \quad \omega(x, y) \neq 0 \\ 1 & \text{sinon.} \end{cases}$$

Algorithme 0.4.1 (Algorithme de Metropolis-Hastings). *Démarrons la chaîne en un point quelconque $X_0 = x_0$ et supposons qu'à la $n-$ième itération on a $X_n = x$. Alors, nous allons*

(1) Simuler $Y \sim Q(x, \cdot)$

Simuler $U \sim \mathcal{U}(0, 1)$.

(2) Si $U \leq \alpha(x, Y)$, faire $X_{n+1} = Y$.

Sinon faire $X_{n+1} = x$.

Le noyau de transition de la chaîne générée par cet algorithme peut s'écrire :

$$P(x, dy) = \alpha(x, y)\omega(x, y)\pi(dy) + r(x)\delta_x(dy) \tag{0.4.2}$$

avec $r(x) = 1 - \int \alpha(x, y)\omega(x, y)\pi(dy)$, la probabilité de rejet de l'algorithme.

On voit donc que :

$$\pi(dx)P(x, dy) = \alpha(x, y)\omega(x, y)\pi(dx)\pi(dy) + r(x)\pi(dx)\delta_x(dy).$$

Etant donné que $\alpha(x, y)\omega(x, y)$ est une fonction symétrique en x et y, on a :

$$\int_{\{x \in A\}} \int_{\{y \in B\}} \alpha(x, y)\omega(x, y)\pi(dx)\pi(dy) = \int_{\{x \in A\}} \int_{\{y \in B\}} \alpha(y, x)\omega(y, x)\pi(dy)\pi(dx),$$

pour tout $A, B \in \mathcal{F}$ et

$$\begin{aligned}\int_{\{x\in A\}}\int_{\{y\in B\}} r(x)\pi(dx)\delta_x(dy) &= \int_{\{x\in A\}} r(x)\pi(dx) \\ &= \int_{\{y\in A\}} r(y)\pi(dy) \\ &= \int_{\{y\in A\}}\int_{\{x\in B\}} r(y)\pi(dy)\delta_y(dx),\end{aligned}$$

ce qui montre que $\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$. Donc, l'algorithme de Metropolis-Hastings produit bien une chaîne de Markov qui admet π comme distribution invariante.

L'algorithme de Metropolis-Hastings est très général et contient plusieurs cas particuliers très utilisés en pratique.

(1) *Algorithme de Metropolis-Hastings Indépendant.*
Dans ce cas, $Q(x,\cdot) = Q(\cdot)$. Cet algorithme a été abondamment étudié (Liu (1996), Smith et Tierney (1996), Mergersen et Tweedie (1996)).

(2) *Algorithme de Metropolis-Hastings Marche Aléatoire*
Dans ce cas, $\mathcal{X} = \mathbb{R}^d$ et on suppose que $Q(x,dy) = q(x,y)dy$ où dy désigne la mesure de Lebesgue sur $\mathbb{R}^d$. L'algorithme de Metropolis Marche Aléatoire propose d'utiliser q telle que $q(x,y)$ ne dépende que de $y-x$. La version la plus populaire de cet algorithme est de prendre q telle que $q(x,y)$ ne dépende que de $|y-x|$ où $|\cdot|$ est la distance euclidienne de $\mathbb{R}^d$. On parle alors de Marche Aléatoire Symétrique. Voir Mergersen et Tweedie (1996)), Jarner et Hansen (2000), Jarner et Tweedie (2001).

(3) *Algorithme de Gibbs*
C'est sans doute l'algorithme le plus utilisé en statistique appliquée et particulièrement en statistique bayésienne (Gelfand and Smith (1990), Robert (2001)). Ici aussi, pour simplifier, on va supposer que $\mathcal{X} = \mathbb{R}^d$ et que π admet une densité par rapport à la mesure de Lebesgue que nous noterons aussi $\pi(x)$. Supposons qu'on sait simuler des d distributions conditionnelles de π : $\pi(x_i|x_{(-i)}) = \frac{\pi(x_1,\ldots,x_d)}{\int \pi(x_1,\ldots,x_d)dx_i}$. Ici $x_{(-i)} = (x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d)$.

Avec ces hypothèses et notations, le noyau $Q = Q_i$ de Gibbs a pour densité :

$$q_i(x, y) = \pi(y_i|x_{(-i)})\delta_{\{x_{(-i)}\}}(y_{(-i)}).$$

Concrètement, partant de $x = (x_1, \ldots, x_d)$, cela revient à simuler $Y_i \sim \pi(\cdot|x_{(-i)})$ et à retourner $Y = (x_1, \ldots, x_{i-1}, Y_i, x_{i+1}, \ldots, x_d)$. Dans ce cas, on a :

$$\begin{aligned} \alpha(x, y) &= Min\left(1, \frac{\pi(y_i|x_{(-i)})\pi(x_{(-i)})\pi(x_i|x_{(-i)})}{\pi(x_i|x_{(-i)})\pi(x_{(-i)})\pi(y_i|x_{(-i)})}\right) \\ &= 1. \end{aligned}$$

Donc, tous les Y proposés sont acceptés. Soit $P_i = Q_i$ le noyau de la chaîne générée par l'algorithme que nous venons de décrire. Pour avoir de bonnes propriétés de convergence, il est necéssaire de composer les P_i soit de façon déterministe $P = P_1 \cdots P_d$, soit de façon aléatoire (uniforme) $P = \frac{1}{d}(P_1 + \cdots + P_d)$. L'algorithme de Gibbs a été étudié par exemple dans Roberts et Polson (1994), Schervish et Carlin (1992).

Il y a principalement deux problèmes qui se posent dans l'utilisation des méthodes MCMC. Etant donné que la chaîne (X_n) est généralement démarrée en un point quelconque $x_0 \in \mathcal{X}$, l'estimateur $\hat{\pi}_0(f)$ est biaisé. Pour réduire ce biais, la pratique courante est de supprimer une partie de la chaîne et utiliser $\frac{1}{n}\sum_{i=t+1}^{t+n} f(X_i)$ à la place de $\hat{\pi}_0(f)$. Ce problème est aussi connu sous le nom de "burn-in" problème. L'idée étant de choisir t tel que X_t soit approximativement distribuée suivant π. Ce problème a conduit aux méthodes de diagnostic de convergence pour lesquelles nous réferons le lecteur à Brooks et Roberts (1998), Mengersen et al. (1999) pour une bonne revue de littérature. Une autre approche de solution plus rigoureuse mais plus ardue adoptée par Rosenthal (1995), Roberts et Tweedie (1999), Douc et al. (2002) consiste à rechercher une borne supérieure calculable pour $\|P^n(x, \cdot) - \pi(\cdot)\|$ et utiliser cette borne pour déterminer t.

Le second problème important des méthodes MCMC est celui de la validité d'un théorème limite central afin de permettre une estimation d'erreur de la méthode. Des problèmes reliés et tous aussi importants sont ceux de l'estimation effective de cette erreur et la possibilité de dévélopper des estimateurs de plus

petite erreur. L'approche de solution consiste d'abord à développer des algorithmes qui ont des vitesses de convergence les plus rapides possibles. C'est dans cette direction que se situe nos contributions dans le troisième essai. Dans notre deuxième essai nous proposons plusieurs méthodes de réduction de variance pour l'algorithme de Metropolis Indépendant.

Avant une description plus détaillée du contenu des essais, nous commençons par donner une revue générale des résultats de convergence connus pour les algorithmes MCMC présentés plus haut.

0.4.3. Propriétés de convergence des algorithmes MCMC

On va commencer par la proposition suivante qui montre qu'il n'est pas très difficile de construire un algorithme de Metropolis-Hastings ergodique.

Proposition 0.4.1. *Supposons que* $\omega(x,y) > 0$ *pour tout* $x, y \in \mathcal{X}$ *et qu'il existe un ensemble* $A \in \mathcal{F}$, $\pi(A) > 0$ *et* $\varepsilon > 0$ *tels que :*

$$\omega(x,y) > \varepsilon \text{ pour tout } x, y \in A. \tag{0.4.3}$$

Alors l'algorithme de Metropolis-Hastings est ergodique.

Dans beaucoup de situations pratiques, $\mathcal{X} = \mathbb{R}^d$ et π admet une densité $\pi(x)$ qui est partout positive et bornée sur des ensembles compacts. Dans une telle situation, en prenant un noyau de transition instrumental $Q(x, dy) = q(x,y)dy$ avec q continue et partout positive, on peut prendre A comme n'importe quel sous ensemble compact de $\mathbb{R}^d$ dans la proposition 3.2.1 et l'algorithme de Metropolis-Hastings résultant est ergodique.

En ce qui concerne la vitesse de convergence des algorithmes MCMC, la situation la plus simple est celle de l'algorithme de Metropolis Indépendant. On a le résultat suivant dû à Mergersen et Tweedie (1996).

Théorème 0.4.3. *Notons* P *le noyau de transition de l'algorithme de Metropolis Indépendant de distribution invariant* π *et de distribution instrumentale* Q. *Soit* $\omega(x) = \frac{Q(dx)}{\pi(dx)}$ *la densité de* Q *par rapport à* π. *Posons* $\rho =$ *inf-ess* $\omega(x)$ *(infimum essentiel pris par rapport à* π*).*

(i): *Si $\rho > 0$ alors pour tout $n \geq 1$,*

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{VT} \leq (1-\rho)^n, \tag{0.4.4}$$

pour tout $x \in \mathcal{X}$.

(ii): *Si $\rho = 0$, alors P ne peut pas converger à une vitesse géométrique.*

La condition $\rho > 0$ est équivalente à dire que Q a des queues plus épaisses que π. Mais lorsque $\rho = 0$, en fait il est possible que P converge à une vitesse polynômiale. Voir par exemple Jarner et Roberts (2002). La situation est moins simple pour l'algorithme de Metropolis-Hastings Marche Aléatoire Symétrique (MHMAS). D'abord on a le résultat négatif clair suivant dû à Jarner et Hansen (2000) et Jarner et Tweedie (2001).

Théorème 0.4.4. *Supposons que pour tout $s > 0$, $\int e^{s|x|}\pi(x)dx = \infty$. Alors l'algorithme de MHMAS ne peut pas être géométriquement ergodique.*

Ce résultat dit que lorsque la densité d'intérêt a des queues plus épaisses que les fonctions exponentielles, aucun algorithme de Metropolis-Hastings de type Marche Aléatoire Symétrique ne peut converger à une vitesse géométrique. Malheureusement, même si la densité π a des queues qui décroissent suffisamment rapidement, ce n'est pas automatique que l'algorithme MHMAS converge géométriquement. Certaines conditions de courbure des lignes de niveau de la densité rentrent également en jeu. Voir Roberts et Tweedie (1996), Jarner et Hansen (2000).

De façon plus générale, si $r(x)$, la probabilité de rejet de l'algorithme de Metropolis-Hastings est telle que sup-ess $r(x) = 1$, Roberts et Tweedie (1996) ont montré qu'un tel algorithme ne peut pas converger à une vitesse géométrique. Dans notre troisième essai, nous prouvons une réciproque de ce résultat. Nous montrons que si sup-ess $r(x) < 1$ et si une condition supplémentaire de compacité est satisfaite, alors l'algorithme de Metropolis-Hastings est géométriquement ergodique. A l'aide de ce résultat, nous avons proposé une version de l' algorithme de Metropolis-Hastings appelé Metropolis Restreint avec comme objectif de combiner les propriétés de l'agorithme de Metropolis Indépendant et de l'algorithme de Metropolis Marche Aléatoire. Lorsque tous ses paramètres sont bien définis,

notre algorithme explore $\mathcal{X}$ d'une manière très similaire à l'agorithme MHMAS mais sa convergence géométrique (ou non) peut se vérifier en comparant tout simplement les queues de π et celles des densités instrumentales utilisées comme dans l'algorithme de Metropolis Indépendant.

La convergence géométrique de l'algorithme de Gibbs est peut-être plus difficile en général. On peut montrer que dans le cas de l'algorithme de Gibbs déterministe, le noyau de transition induit admet une densité par rapport à π : $P(x, dy) = k(x, y)\pi(dy)$. Les résultats qui existent sur l'ergodicité géométrique de cet algorithme supposent des conditions qui sont assez difficiles à vérifier en pratique. Par exemple Roberts et Polson (1994) supposent l'équicontinuité de la famille de fonctions $\{k(x, \cdot),\ x \in \mathcal{X}\}$ et Schervish et Carlin (1992) supposent que l'opérateur intégral induit par P est de type Hilbert-Schmidt.

0.5. Description détaillée des essais

0.5.1. Monte Carlo Simulations via Control Variates

Dans cet essai, nous proposons une méthode de réduction de variance qui utilise des variables de contrôle. Supposons qu'on a une variable aléatoire X et nous nous intéressons à $\mu = \mathrm{E}(X)$ avec $\sigma^2 = \mathrm{var}(X) < \infty$. Supposons qu'on dispose d'un autre vecteur aléatoire Y telle que $\mathrm{E}(Y) = 0$ and $\mathrm{cov}(Y) = I$. Notons $\gamma = \mathrm{Var}(Y, X)$ et supposons que $\gamma \neq 0$. Basé sur un échantillon $\{(X_i, Y_i)\}_{i=1}^n$, l'estimateur ordinaire de μ est donné par $\delta_0 = \frac{1}{n}\sum_{i=1}^n X_i$. Mais on peut utiliser l'information disponible donnée par la propriété $\mathrm{E}(Y) = 0$ et proposer $\delta_1 = \frac{1}{n}\sum_{i=1}^n (X_i - \gamma^t Y_i)$. δ_1 est l'estimateur par variables de contrôle de μ. C'est bien connu que δ_0 et δ_1 sont sans biais, $\mathrm{var}(\delta_0) = \sigma^2/n$ et $\mathrm{var}(\delta_1) = (\sigma^2 - \|\gamma\|^2)/n$. Donc si γ est connu, δ_1 est meilleur que δ_0. Lorsque γ n'est pas connu, une possibilité fréquemment utilisée consiste à estimer γ dans l'expression de δ_1. Une telle approche induit un biais dans δ_1 et peut parfois contribuer à augmenter la variance de δ_1. Notre contribution dans ce travail consiste à proposer un nouvel estimateur δ_2 qui est sans biais, asymptotiquement équivalent à δ_0 et ne nécessite pas le calcul de γ. Soit $\hat{\gamma}_{-i} = \frac{1}{n-1}\sum_{\{j:\, j\neq i\}} X_j Y_j$. Nous montrons que δ_2 donné par $\delta_2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\gamma}_{-i}^t Y_i)$ permet d'atteindre ce but.

Par la suite nous appliquons l'idée à l'algorithme de Rejet et à la méthode Importance-Sampling. Dans le cas de l'algorithme Rejet, nous proposons aussi une version Rao-Blackwellisée de l'estimateur ainsi développé.

0.5.2. Improving on the Independent Metropolis-Hastings algorithm

Dans ce travail, nous montrons comment obtenir de meilleurs estimateurs d'intégrales de type $\pi(f) = \int f(x)\,\pi(dx)$ lorsque l'algorithme de Monte Carlo utilisé est l'algorithme de Metropolis-Hastings Indépendant (MHI). Rappelons que dans l'algorithme MHI, lorsque le processus se trouve au point x au temps n, donc pour $X_n = x$, l'algorithme MHI procède en proposant une observation Y distribuée selon Q et U uniformément distribué sur l'intervalle $(0,1)$. Si $U \leq \alpha(x,Y)$, $X_{n+1} = Y$, sinon $X_{n+1} = x$. Alors $\pi(f)$ est estimé en utilisant $\hat{\pi}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. Il arrive très souvent que des informations auxiliaires sont disponibles sur la distribution instrumentale Q : on connait une variable g telle que $\int g\,dQ = 0$ et $\int g^2\,dQ = 1$. Notre première amélioration utilise ces informations sous forme de variables de contrôle évaluées sur les données générées par la loi de probabilité instrumentale de l'algorithme pour proposer une nouvelle classe d'estmateurs de $\pi(f)$, $\hat{\pi}_\beta(f) = \sum_{i=1}^{n}(f(X_i) - \beta g(Y_i))/n$. Nous obtenons une forme explicite de la variance de notre estimateur pour une taille fixe et montrons que pour des β bien choisis, cette variance devient inférieure à celle de l'estimateur usuel $\hat{\pi}(f)$ lorsque la taille de l'échantillon devient grande. La seconde approche que nous avons proposée capitalise sur le manque de symétrie de l'algorithme MHI. En effet, la chaîne $(X_1,\ldots,X_n)$ de l'algorithme MHI tel que décrite ci-dessus est formée en utilisant la séquence de variables aléatoires $((U_1,Y_1),(U_2,Y_2),\ldots,(U_n,Y_n))$. Soit $\mathcal{S}$ l'ensemble des permutations de $\{1,\ldots,n\}$, et soit $s \in \mathcal{S}$. Si on recommence l'algorithme de MHI mais en utilisant plutôt la sequence $((U_{s(1)},Y_{s(1)}),(U_{s(2)},Y_{s(2)}),\ldots,(U_{s(n)},Y_{s(n)}))$, on obtient un estimateur tout aussi valide de $\pi(f)$ que nous noterons $\hat{\pi}_f(s)$. $\hat{\pi}(f)$ et $\hat{\pi}_f(s)$ sont différents en général mais ont la même distribution. Alors $\sum_{s\in\mathcal{O}} \nu(s)\hat{\pi}_f(s)$ est un meilleur estimateur que $\hat{\pi}(f)$ où $\mathcal{O}$ est un sous-groupe de $\mathcal{S}$ et ν une distribution de probabilité

non dégénérée sur $\mathcal{O}$. Nous donnons un exemple d'un tel sous-groupe et fournissons la forme explicite de l'estimateur qu'il induit. Comme troisième méthode de réduction de variance, nous avons développé les versions Rao-Blackwellisées des deux estimateurs présentés ci-dessus. La Rao-Blackwellisation permet en general de faire des réductions de variance importantes mais au prix de calculs prohibitifs. Nous parvenons aussi à déterminer une borne explicite sur la plus grande amélioration que nos méthodes peuvent produire en comparaison avec l'estimateur usuel.

0.5.3. Geometric Ergodicity of the Restricted Metropolis Algorithm

Cet travail contient principalement deux résultats. Le premier concerne l'ergodicité géométrique de l'algorithme de Metropolis-Hastings (MH). Nous avons vu à la section 3.2.1 que le noyau de transition de la chaîne de Markov engendrée par l'algorithme MH peut s'écrire : $P(x,dy) = \alpha(x,y)\omega(x,y)\pi(dy) + r(x)\delta_x(dy)$ où $r(x) = 1 - \int \alpha(x,y)\omega(x,y)\,\pi(dy)$ et $\delta_x(dy)$ est la masse de Dirac en x. Ce noyau de transition engendre un opérateur linéaire sur l'espace $L^2(\pi)$ des fonctions f réelles de carré π–intégrable par $Kf(x) = \int f(y)P(x,dy)$.

Dans le cas particulier de l'algorithme MH, K s'écrit : $Kf(x) = M_r f(x) + Uf(x)$, où M_r est l'opérateur de multiplication par r la probabilité de rejet : $M_r f(x) = r(x)f(x)$, et U est l'opérateur intégrale $Uf(x) = \int \alpha(x,y)\omega(x,y)f(y)\pi(dy)$. On définit le spectre de K par $\sigma(K) = \{\lambda \in \mathbb{R} :\ K - \lambda I \text{ non inversible}\}$ (où I est l'opérateur identité) et le trou spectral de K par $\tau(K) = 1 - \sup\{|\lambda| :\ \lambda \in \sigma(K) \setminus \{1\}\}$. Lorsque $\tau(K) > 0$, on dit que K admet un trou spectral. Si l'opérateur K de l'algorithme MH admet un trou spectral, il est bien connu que l'algorithme a une convergence géométrique vers sa distribution stationnaire, c'est à dire qu'il existe V une fonction et $\rho < 1$ tels que $\|P^n(x,\cdot) - \pi\|_{TV} \leq \rho^n V(x)$. En général, ρ et V ne sont pas connus, autrement dit, cette borne n'est pas calculable. Mais l'existence d'une convergence géométrique a des consequences pratiques importantes : l'existence d'un théorème central limite pour certaines fonctions de la chaîne, l'applicabilité de certaines méthodes de diagnostic de convergence de la chaîne.

Dans ce travail, nous trouvons une condition suffisante sous laquelle l'opérateur K de l'algorithme de Metropolis-Hastings admet un trou spectral. Nous avons montré que si P est ergodique, sup-ess $r(x) < 1$ et si U est compact alors K admet un trou spectral. Dans le cas particulier de l'algorithme de Metropolis-Hastings Indépendant (MHI), le professeur Jun S. Liu a trouvé le spectre de l'opérateur engendrée K dans le cas où l'espace des états est discret (K est donc une matrice) et a conjecturé le résultat dans le cas où l'espace des états de la chaîne est arbitraire. Nous avons montré que cette conjecture est vraie lorsque l'espace des états est un espace euclidien (qui englobe la plupart des cas rencontrés en pratique). Ce qui permet de retrouver des résultats (déja existants) sur les conditions nécessaires et suffisantes d'ergodicité géométrique pour l'algorithme MHI.

Le deuxième résultat important de ce travail est en fait un algorithme. L'idée part d'un théorème de Jarner et Tweedie (2000) qui dit que si les queues de la distribution d'intérêt π sont trop épaisses (fat tail distribution), l'algorithme de Metropolis-Hastings Marche Aléatoire Symétrique (MHMAS) ne peut pas être géométriquement ergodique. Notons que l'algorithme MHMAS est une des versions les plus utilisées en pratique de l'algorithme de Metropolis-Hastings. Nous avons montré qu'en restreignant le comportement de type marche aléatoire de la chaîne sur un ensemble convexe compact, et en choisissant bien les queues de la distribution instrumentale, on arrive à fabriquer un algorithme qui est géométriquement ergodique. En fait, notre algorithme empêche les éléments de la famille de distributions instrumentales $Q(x, \cdot)$ de s'éloigner trop de π. Il fournit donc une alternative intéressante à l'algorithme MHMAS lorsque les queues de la distribution π sont trop épaisses.

Bibliographie

Alder, B., Wainwright, T. E., 1959. Studies in molecular dynamics. J. Chem. Phys. 31, 459–466.

Athreya, K., Doss, D., Sethuraman, J., 1996. On the convergence of the makov chain simulation method. ann. stat. 24 (1), 69–100.

Baxter, J. R., Rosenthal, J. S., 1995. Rate of convergence of everywhere-positive markov chains. Stat. Prob. Lett. 22, 333–338.

Berger, J., 2000. Bayesian analysis : A look at today and thoughts of tomorrow. JASA 95, 1269–1276.

Billingsley, P., 1993. Probability and Measure. Wiley-Interscience, New-York.

Brooks, S. P., Roberts, G., 1998. Assessing convergence of markov chain monte carlo algorithms. Statistics and Computing 8(4), 319–335.

Casella, G., Robert, C. P., 1998. Post-processing accept-reject samples : recycling and rescaling. J. Comput. Graph. Statist. 7 (2), 139–157.

Chib, S., Nardari, F., Shephard, N., 2002. Markov chain monte carlo methods for stochastic volatility models. Journal of Econometrics 108, 281–316.

Devroye, L., 1986. Nonuniform random variate generation. Springer-Verlag, New York.

Douc, R., Moulines, E., Rosenthal, J. S., 2002. Quantitative bounds on convergence of time-inhomogeneous markoc chains. Technical Report .

Fishman, G., 1996. Monte Carlo : Concepts, Algorithms, and Applications. Springer-Verlag, New York.

Gelfand, A. E., Smith, A. F. M., 1990. Sampling-based approaches to calculating marginal densities. J. Am. Stat. Ass. 85 (410), 398–409.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57 (6), 1317–1339.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (Eds.), 1996. Markov chain Monte Carlo in practice. Interdisciplinary Statistics. Chapman & Hall, London.

Goodman, J., Sokal, A. D., 1989. Multigrid monte carlo method. conceptual foundations. Physical Review D 40, 2035–71.

Jarner, S. F., Hansen, E., 2000. Geometric ergodicity of metropolis algorithms. Sto. Proc. Appl. 85, 341–361.

Jarner, S. F., Roberts, G. O., 2002. Polynomial convergence rates of Markov chains. Ann. Appl. Probab. 12 (1), 224–247.

Jarner, S. F., Tweedie, R. L., 2001. Necessary conditions for geometric and polynomial ergodicity of random walk-type markov chains.. MCMC Preprints .

Karplus, M., Petsko, G. A., 1990. Molecular dynamics simulations in biology. Nature 347, 631–639.

Kipnis, C., Varadhan, S. R. S., 1986. Central limit theorem for additive functionals of reversible markov processes and applications to simple exclusions. Comm. Math. Phys. 104, 1–19.

Kirkpatrick, S., D., G. J. C., Vecchi, M. P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Knuth, D. E., 1997. The Art of Computer Programming, Vol. 2. Addison-Wesley, Massachusetts.

Krengel, U., 1985. Ergodic Theorems. Walter de Gruyter, Berlin, New-York.

Leach, A. R., 1996. Molecular modelling : Principles and Applications. Addison Wesley Longman, Singapore.

L'Ecuyer, P., 1994. Uniform random number generation. Ann. Oper. Res. 53, 77–120, simulation and modeling.

Liu, J., 1999. Markov chain monte carlo and related topics. Technical Report, Dept of Statistics, Stanford University .

Liu, J. S., 1996. Metropolized independent sampling with comparaisons to rejection sampling and importance sampling. Statistics and Computing 6, 113–119.

Liu, J. S., 2001. Monte Carlo Strategies in Scientific Computing. Springer Verlag, New-York.

Marshall, A., 1956. The use of multi-stage sampling schemes in monte carlo computations. In : symposium on Monte Carlo methods. Wiley, New York, pp. 123–140.

Mengersen, K. L., Robert, C. P., Guihenneuc-Jouyaux, C., 1999. MCMC convergence diagnostics : a reviewww. In : Bayesian statistics, 6 (Alcoceber, 1998). Oxford Univ. Press, New York, pp. 415–440.

Mengersen, K. L., Tweedie, R. L., 1996. Rates of convergence of the Hastings and Metropolis algorithms. Ann. Statist. 24 (1), 101–121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1091.

Meyn, S. P., Tweedie, R. L., 1993. Markov chains and stochastic stability. Springer-Verlag London Ltd., London.

Nelsen, R. B., 1999. An introduction to copulas. Springer-Verlag, New York.

Neveu, J., 1965. Mathematical Foundations of the Calculus of Probability. Holden-Day.

Niederreiter, H., 1992. Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia.

Nummelin, E., Tuominen, P., 1982. Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory. Stochastic Process. Appl. 12 (2), 187–202.

Nummelin, E., Tuominen, P., 1983. The rate of convergence in Orey's theorem for Harris recurrent Markov chains with applications to renewal theory. Stochastic Process. Appl. 15 (3), 295–311.

Ripley, B. D., 1987. Stochastic simulation. John Wiley & Sons Inc., New York.

Robert, C. P., 2001. The Bayesian choice, 2nd Edition. Springer-Verlag, New York, from decision-theoretic foundations to computational implementation, Translated and revised from the French original by the author.

Robert, C. P., Casella, G., 1999. Monte Carlo statistical methods. Springer-Verlag, New York.

Roberts, G. O., Polson, N. G., 1994. On the geometric convergence of the Gibbs sampler. J. Roy. Statist. Soc. Ser. B 56 (2), 377–384.

Roberts, G. O., Rosenthal, J. S., 1997. Geometric ergodicity and hybrid Markov chains. Electron. Comm. Probab. 2, no. 2, 13–25 (electronic).

Roberts, G. O., Rosenthal, J. S., 1998. Markov-chain Monte Carlo : some practical implications of theoretical results. Canad. J. Statist. 26 (1), 5–31, with discussion by Hemant Ishwaran and Neal Madras and a rejoinder by the authors.

Roberts, G. O., Tweedie, R. L., 1996. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika 83 (1), 95–110.

Roberts, G. O., Tweedie, R. L., 1999. Bounds on regeneration times and convergence rates for Markov chains. Stochastic Process. Appl. 80 (2), 211–229.

Rosenthal, J. S., 1995. Minorization conditions and convergence rates for markov chain monte carlo. JASA 90, 558–566.

Schervish, M. J., Carlin, B. P., 1992. On the convergence of successive substitution sampling. J. Comput. Graph. Statist. 1 (2), 111–127.

Smith, R., Tierney, L., 1996. Exact transition probabilities for the independence metropolis sampler. Research Report, Statistical Laboratory University of Cambrige 16.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22 (4), 1701–1762, with discussion and a rejoinder by the author.

Tuominen, P., Tweedie, R. L., 1994. Subgeometric rates of convergence of f-ergodic Markov chains. Adv. in Appl. Probab. 26 (3), 775–798.

Ullman, J. D., 1984. Computational Aspects of VLSI. Computer Science Press, Rockville.

von Neumann, J., 1951. Various techniques used in connection with random digits. Natl. Bureau of Standards Appl. Math. Ser. 12, 36–38.

Chapitre 1

MONTE CARLO SIMULATIONS VIA CONTROL VARIATES

Abstract.

This paper proposes methods to improve Monte Carlo estimates of a mean μ using a covariate. In the paper, we propose a new unbiased estimator when the correlation with the covariate is unknown. We apply our results to the importance sampling and the accept-reject algorithm. We also work on Rao-Blackwellizations. Numerical results are given.

AMS Classification: 65C40, 60J22, 60J10.

Keywords: IMPORTANCE SAMPLING, ACCEPT-REJECT ALGORITHM, CONTROL VARIATES, RAO-BLACKWELLIZATION.

1.1. INTRODUCTION

This paper is about variance reduction. For an introduction to the subject see Tierney (1994), Chen et al. (2000) and Evans and Swartz (2000). Consider X a random variable and Y a random vector. Let $\mu = \mathrm{E}(X)$, $\sigma^2 = \mathrm{var}(X)$ and $\gamma = \mathrm{cov}(Y, X)$. Suppose that $\gamma \neq 0$ and assume that $\mathrm{E}(Y) = 0$ and $\mathrm{cov}(Y, Y) = I$. Consider $\{(X_i, Y_i)\}_{i=1}^n$ a sample of size n, $n > 1$. We want to estimate μ. The standard estimator δ_0 is given by

$$\delta_0 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

while, δ_1, the estimator which uses the covariate Y, is given by

$$\delta_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \gamma^t Y_i).$$

It is well known that δ_0 and δ_1 are unbiased, $\mathrm{var}(\delta_0) = \sigma^2/n$ and $\mathrm{var}(\delta_1) = (\sigma^2 - \|\gamma\|^2)/n$. If γ is known then it is always better to use δ_1 than δ_0. If γ is unknown then δ_1 is not available. A naïve approach consists of replacing γ by an estimate in the expression for δ_1. In general, this approach will introduce bias and it may also increase the variance. Our aim is to find an estimator δ_2 such that:

$$\begin{aligned} \delta_2 \quad & \text{is unbiased,} \\ n\{\mathrm{var}(\delta_2) - \mathrm{var}(\delta_1)\} \quad & \to 0 \text{ as } n \to \infty. \end{aligned}$$

In Section 2 our estimator δ_2 is introduced. In Section 3 we apply the results of Section 2 on the importance sampling algorithm and the accept-reject algorithm. In the case of the accept-reject algorithm, we also develop a Rao-Blackwellization. This one is different from those introduced by Casella and Robert (1996), Casella and Robert (1998) and it is closer to the one in Perron (1999). This last development is really a key element in this paper. Finally, we give numerical results.

1.2. BASIC RESULT

The main idea in this section is to develop n unbiased estimates of γ. These estimates depend on i, they are called $\hat{\gamma}_{-i}$, and they are such that $\hat{\gamma}_{-i}$ and Y_i are

independent for each value of i, $i = 1, 2, \ldots, n$. More precisely, we set

$$\begin{aligned}\hat{\gamma}_{-i} &= \frac{1}{n-1}\sum_{\{j:\, j\neq i\}} X_j Y_j, \\ \delta_2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\gamma}_{-i}^t Y_i).\end{aligned}$$

Theorem 1.2.1. *The estimator δ_2 is unbiased and*

$$var(\delta_2) = var(\delta_1) + \frac{1}{n(n-1)}\mathrm{tr}\{cov(XY) + cov(XY, Y)cov^t(XY, Y)\}.$$

PROOF. It is straightforward to verify that δ_2 is unbiased. Moreover,

$$\begin{aligned}\mathrm{var}(\delta_2) &= \frac{1}{n^2}\mathrm{var}\{\sum_{i=1}^{n}(X_i - \hat{\gamma}_{-i}^t Y_i)\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{cov}(X_i, \hat{\gamma}_{-j}^t Y_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{cov}(\hat{\gamma}_{-i}^t Y_i, \hat{\gamma}_{-j}^t Y_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) - \frac{2}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\{k:\, k\neq j\}}\mathrm{cov}(X_i, X_k Y_k^t Y_j) \\ &\quad + \frac{1}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{\{k:\, k\neq i\}}\{\sum_{\{\ell:\, \ell\neq i\}}\mathrm{cov}(X_k Y_k^t Y_i, X_\ell Y_\ell^t Y_i) \\ &\quad + \sum_{\{j:\, j\neq i\}}\sum_{\{\ell:\, \ell\neq j\}}\mathrm{cov}(X_k Y_k^t Y_i, X_\ell Y_\ell^t Y_j)\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}[X_i] - \frac{2}{n^2(n-1)}\sum_{j=1}^{n}\sum_{\{k:\, k\neq j\}}\mathrm{cov}(X_j, X_k Y_k^t Y_j) \\ &\quad + \frac{1}{n^2(n-1)^2}\sum_{i=1}^{n}\{\sum_{\{k:\, k\neq i\}}\mathrm{var}(X_k Y_k^t Y_i) + \sum_{\{j:\, j\neq i\}}\mathrm{cov}(X_j Y_j^t Y_i, X_i Y_i^t Y_j)\} \\ &= \frac{1}{n}\sigma^2 - \frac{2}{n}\|\gamma\|^2 + [\frac{1}{n}\|\gamma\|^2 + \frac{1}{n(n-1)}\mathrm{tr}\{\mathrm{cov}(XY)\}] \\ &\quad + \frac{1}{n(n-1)}\mathrm{tr}\{\mathrm{cov}(XY, Y)\mathrm{cov}^t(XY, Y)\} \\ &= \mathrm{var}[\delta_1] + \frac{1}{n(n-1)}\mathrm{tr}\{\mathrm{cov}(XY) + \mathrm{cov}(XY, Y)\mathrm{cov}^t(XY, Y)\}.\end{aligned}$$

□

1.3. APPLICATIONS

Consider a target distribution P and a proposal Q where P is absolutely continuous with respect to Q and $w = dP/dQ$ is the Radon-Nikodým derivative. Let

$$\mu = \int h \, dP.$$

We are interested in the estimation of μ. Usually, Q is a well known distribution and it is easy to find a function g_0 which is a good approximation of h and for which $\mu_{g_0} = \int g_0 \, dQ$ and $\sigma^2_{g_0} = \int (g_0 - \mu_{g_0})^2 \, dQ$ are known. Therefore, the function g which is given by $g = (g_0 - \mu_{g_0})/\sigma_{g_0}$ can be useful in the construction of a covariate. The first application is based on the importance sampling algorithm while the second application is based on the accept-reject algorithm.

In the importance sampling algorithm we start with $Z_1, Z_2, \ldots, Z_n$, an *iid* sample from Q. We set

$$\begin{aligned} X_i &= w(Z_i)h(Z_i), \\ Y_i &= g(Z_i), \end{aligned}$$

for $i = 1, 2, \ldots, n$.

In the accept reject algorithm we assume that w is bounded from above and we find a constant c where c is an upper bound for w. The algorithm starts with $\{(Z_{i,j}, U_{i,j}) : i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, N_i\}$. Here, the random variables $Z_{i,j}$ form an *iid* sample from Q, the random variables $U_{i,j}$ form an *iid* sample from the uniform distribution on $(0, 1)$ and N_i is a stopping time counting the number of trials necessary for the first occurence of the event $cU_{i,j} \leq w(Z_{i,j})$. In other words, $cU_{i,j} > w(Z_{i,j})$ for $j = 1, 2, \ldots, N_i - 1$ while $cU_{i,N_i} \leq w(Z_{i,N_i})$. In this approach, the marginal distribution of $Z_{i,1}$ is Q for all i while the conditional distribution of $Z_{i,j}$ given that $j = N_i$ becomes the target distribution, that is to say P, for all i as well. Therefore, we can set

$$\begin{aligned} X_i &= h(Z_{i,N_i}), \\ Y_i &= g(Z_{i,1}), \end{aligned}$$

for $i = 1, 2, \ldots, n$.

Finally, this algorithm can be improved via a Rao-Blackwellization. In Casella and Robert (1996), Casella and Robert (1998) or Perron (1999), different Rao-Blackwellizations have been proposed. They can be applied directly to our problem. However, even if these approaches offer considerable improvement, they may not be useful in practice because they are enormously time consuming, to say the least. Here, we shall consider a Rao-Blackwellization such that the number of operations in the evaluation of the estimator grows linearly with n. In fact, we shall take conditional expectations given that the event $T = \{(Z_{i,(j)}, N_i)\colon i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, N_i\}$ is fixed where $Z_{i,(1)}, Z_{i,(2)}, \ldots, Z_{i,(N_i)}$ represent the order statistics based on $Z_{i,1}, Z_{i,2}, \ldots, Z_{i,N_i}$. Let

$$X_i^* = \mathrm{E}(X_i|T),\ Y_i^* = \mathrm{E}(Y_i|T),\ \gamma_i^* = \mathrm{E}(X_iY_i|T),\ \text{and}\ \gamma_{-i}^* = \frac{1}{n-1}(\sum_{j=1}^{n}\gamma_j^* - \gamma_i^*)$$

for $i = 1, 2, \ldots, n$. The Rao-Blackwellized version of δ_2 becomes δ_2^* with

$$\delta_2^* = \frac{1}{n}\sum_{i=1}^{n}(X_i^* - \gamma_{-i}^* Y_i^*).$$

If $N_i = 1$ then $X_i^* = X_i$, $Y_i^* = Y_i$ and $\gamma_i^* = X_iY_i$. If $N_i > 1$ and $w(Z_{i,N_i}) = c$ then

$$\begin{aligned} X_i^* &= X_i, \\ Y_i^* &= \bar{g}_i - \frac{1}{N_i - 1}\{g(Z_{iN_i}) - \bar{g}_i\}, \\ \gamma_i^* &= X_i^*\bar{g}_i - \frac{1}{N_i - 1}h(Z_{iN_i})\{g(Z_{iN_i}) - \bar{g}_i\}. \end{aligned}$$

Otherwise,

$$\begin{aligned} X_i^* &= \sum_{j=1}^{N_i}\rho_{ij}h(Z_{ij})/\sum_{j=1}^{N_i}\rho_{ij}, \\ Y_i^* &= \bar{g}_i - \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\rho_{ij}\{g(Z_{ij}) - \bar{g}_i\}/\sum_{j=1}^{N_i}\rho_{ij}, \\ \gamma_i^* &= X_i^*\bar{g}_i - \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\rho_{ij}h(Z_{ij})\{g(Z_{ij}) - \bar{g}_i\}/\sum_{j=1}^{N_i}\rho_{ij}, \end{aligned}$$

with

$$\begin{aligned} \rho_{ij} &= w(Z_{ij})/\{c - w(Z_{ij})\}, \\ \bar{g}_i &= \frac{1}{N_i}\sum_{j=1}^{N_i} g(Z_{ij}). \end{aligned}$$

Example 1.3.1. Computation of the mean μ of a Gamma Distribution. *This example comes from Casella and Robert (1996). The target is a Gamma distribution with mean α and variance α while the proposal is Gamma distribution with mean α and variance $\alpha^2/2$. In fact, the proposal is the distribution of the sum of two independent random variables exponentially distributed with the same mean. We obtain that*

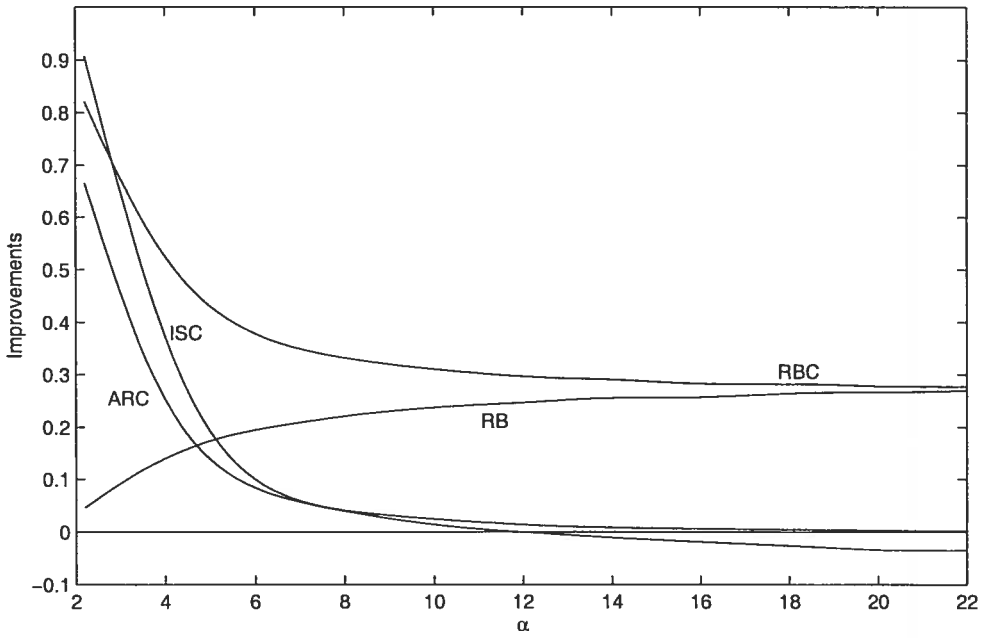
$$w(x) = \left(\frac{\alpha}{2}\right)^2 \frac{1}{\Gamma(\alpha)} x^{\alpha-2} \exp -\{(\alpha-2)x/\alpha\}, \quad \text{for } x > 0$$

and w is bounded from above by $w(\alpha)$ for $\alpha \geq 2$. Clearly, if α is close to 2 from the right then we shall have very few rejections in the accept-reject algorithm. In these situations, the methods using a covariate will be very efficient for reducing the variance. However, if α is much larger than 2, then we shall have many repetitions in the accept-reject algorithm and the improvement provided by the methods using a covariate may be very small. Let us examine numerical results. We consider:

$$h(x) = x \qquad g(x) = \sqrt{2}(x-\alpha)/\alpha$$

so $E\{g(Y_0)\} = 0$, $var\{g(Y_0)\} = 1$ and $\gamma = \sqrt{2}$. The improvement of an estimator δ compared to the standard estimator δ_0 is given by $1 - var(\delta)/Var(\delta_0)$. Our analysis is based on an estimations of these variances using 10 000 replications. The parameter α will vary from 2.2 to 22.0 and the parameter n is fixed at 1000. We consider the improvement provided by a covariate in the importance sampling and in the accept-reject. We have also considered the effect of combining a covariate with a Rao-Blackwellization or simply using a Rao-Blackwellization on its own in the case of the accept-reject algorithm. The results are shown in Graph 1. We are pleased with the improvement provided by the combination of a Rao-Blackwellization and a covariate. The Rao-Blackwellization increases the correlation with the covariate and it gives a better estimate for γ.

Graph1: Improvements of the control variate approach.
ISC: Case of the Importance Sampling method.
ARC: Case of the Accept–Reject algorithm.
RB: Rao–Blackwellization of the AR algorithm.
RBC:Rao–Blackwellization and control variate of the AR algo rithm.



Bibliography

Casella, G., Robert, C. P., 1996. Rao-Blackwellisation of sampling schemes. Biometrika 83 (1), 81–94.

Casella, G., Robert, C. P., 1998. Post-processing accept-reject samples: recycling and rescaling. J. Comput. Graph. Statist. 7 (2), 139–157.

Chen, M.-H., Shao, Q.-M., Ibrahim, J. G., 2000. Monte Carlo Methods in Bayesian Computation. Springer-Verlag, New York.

Evans, M., Swartz, T., 2000. Approximating integrals via Monte Carlo and deterministic methods. Oxford University Press, Oxford.

Perron, F., 1999. Beyond accept-reject sampling. Biometrika 86 (4), 803–813.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22 (4), 1701–1762, with discussion and a rejoinder by the author.

Chapitre 2

IMPROVING ON THE INDEPENDENT METROPOLIS-HASTINGS ALGORITHM

Abstract.

This paper proposes methods to improve Monte Carlo estimates when the Independent Metropolis-Hastings Algorithm (IMHA) is used. Our first approach uses a control variate based on the sample generated by the proposal distribution. We derive the variance of our estimator for fixed sample sizes n and show that, as n tends to infinity, this variance is asymptotically smaller than the one obtained with the IMHA. Our second approach is based on Jensen's inequality. We use a Rao-Blackwellization and exploit the lack of symmetry in the IMHA. An upper bound on the improvements that we can obtain by these methods is derived.

AMS Classification: 65C40, 60J22, 60J10.

Keywords: METROPOLIS-HASTINGS ALGORITHM, CONTROL VARIATES, RAO-BLACKWELLIZATION, SYMMETRY.

2.1. INTRODUCTION

Let Π be a target distribution on $\mathbb{R}^d$ equiped with its Borel sets $\mathcal{B}^d$. In MCMC, we are interested in having a Markov chain (X_n) with stationary distribution Π. There are several ways to achieve this goal. For an introduction to Markov Chain Monte Carlo algorithms, we refer the reader to Robert and Casella (1999) or Liu (2001). The Independent Metropolis-Hastings algorithm (IMHA) is among the most popular algorithms. It is based on a proposal P wich dominates Π. Let $w = d\Pi/dP$ be the Radon Nikodým derivative. The performance of the IMHA is related to w. Here we shall assume that w is essentially bounded, that is $\mathrm{P}[w < M] = 1$ for some $M < \infty$. This condition will imply that our chain is geometrically ergodic. In the construction process, we sample independent and identically distributed random couples (U_i, Y_i), $i = 1, 2, \ldots$. The variables U_i and Y_i are independent with U_i being uniformly distributed on $(0, 1)$ and Y_i being distributed according to P. The initial distribution of the process, namely the distribution of X_0 is not very important. Typically, X_0 is taken to be some arbitrary point of $\mathbb{R}^d$. In the following steps, $X_i = X_{i-1} + (Y_i - X_{i-1})I[U_i w(X_{i-1}) \leq w(Y_i)]$ for $i = 1, 2, \ldots$. For convenience, let the random variable Y_0 have distribution P.

This paper is about variance reduction. For an introduction to the subject see Tierney (1994), Chen et al. (2000) and Evans and Swartz (2000). We are interested in the problem of estimating μ where $\mu = \int h \, d\Pi$. The basic IMHA estimator is given by: $\hat{\mu}_0 = \sum_{i=1}^n h(X_i)/n$. The performance of $\hat{\mu}_0$ as an estimator of μ will depend on Π, w, h. In general, Π and h are fixed by the problem at hand. Finding the best possible choice for w is still a largely unsolved problem. However for w fixed, it is possible to go beyond the IMHA and obtain an improved estimator. Typically, P is a well-known distribution. This means that it is easy to generate random variables from P but it is also easy to select a control variate. Therefore, our first approach uses a control variate g. We propose and study the family of estimators $\hat{\mu}_\beta = \sum_{i=1}^n (h(X_i) - \beta g(Y_i))/n$. Our main result says that $n\{\mathrm{Var}(\hat{\mu}_0) - \mathrm{Var}(\hat{\mu}_\beta)\} \to \beta_0^2 - (\beta - \beta_0)^2$ with $\beta_0 = \mathrm{Cov}(h(X_0), g(X_0))$ as $n \to \infty$. Suprisingly, the expression for β_0 is rather simple in comparison to $\lim_{n\to\infty} n\mathrm{Var}(\hat{\mu}_0)$ and this is a key element in this paper. As a by-product, our

method of proof also provides a very simple way to derive $\mathrm{Var}(\hat{\mu}_0)$ for fixed n, not only for our problem but for any stationary and reversible Markov chain when there is a spectral gap.

Our second approach uses Rao-Blackwellizations and symmetry. Casella and Robert (1996) have shown that one can reduce the variance by a Rao-Blackwellization. Perron (1999) showed that the Rao-Blackwellization suggested by Casella and Robert (1996) in the case of the Accept-Reject Algorithm can be improved. Here we show that the Rao-Blackwellization can be improved in the IMHA as well. We also work on the symmetry of the problem. We can rearrange the estimator by introducing more symmetry and this will reduce its variance. The second major result of this paper shows that if we combine the approach based on a covariate with a Rao-Blackwellization then the variance reduction obtained by the two approaches will add when combined.

2.2. VARIANCE REDUCTION VIA A COVARIATE

Our method of proof is based on the theory of linear operators in $L^2(\Pi)$. Consider the following operators: $Jf(x) = \mathrm{E}[f(X_0)]$ for all x, $Kf(x) = \mathrm{E}[f(X_1)|X_0 = x]$ and $Qf(x) = \mathrm{E}[f(X_1)|Y_1 = x]$. More explicitely, if I represents the identity operator and $f \in L^2(\Pi)$ then, after some simplifications, we obtain for all $x \in R^d$:

$$Jf(x) = \mathrm{E}[f(X_0)]$$

$$(I - K)f(x) = \mathrm{E}[(\frac{1}{\omega(x)} \wedge \frac{1}{\omega(X_0)})(f(x) - f(X_0))]$$

$$Qf(x) = (J + \omega(x)(I - K))f(x),$$

where $a \wedge b = Min(a, b)$.

Remark 2.2.1. *The chain is reversible because* $J(f_1 K f_2) = J(f_2 K f_1)$ *for all bounded functions* f_1, f_2.

Consider now the operator $K_0 = K - J$. Let λ_0 be the norm of K_0. Since w is essentially bounded the operator K has a spectral gap so $\lambda_0 < 1$ and $I - K_0$ is invertible.

Remark 2.2.2. *If* $f_1, f_2 \in L^2(\Pi)$ *then* $Cov[Kf_1(X_0), f_2(X_0)] = Cov[K_0 f_1(X_0), f_2(X_0)]$.

Let g be a covariate. Consider the family of estimators $\hat{\mu}_\beta$ given by

$$\hat{\mu}_\beta = \sum_{i=1}^{n}(h(X_i) - \beta g(Y_i))/n,$$

with $\mathrm{E}[g(Y_0)] = 0$ and $\mathrm{Var}[g(Y_0)] = 1$ and let

$$\beta_0 = \mathrm{Cov}[h(X_0), g(X_0)].$$

Theorem 2.2.1. *Under the hypothesis above, and assuming that* $X_0 \sim \Pi$,

$$\begin{aligned} n\,Var[\hat{\mu}_\beta] &= Cov[(I - K_0)^{-1}(I + K_0)h(X_0), h(X_0)] - \beta_0^2 + (\beta - \beta_0)^2 \\ &\quad -(2/n)\{Cov[h_0(X_0), h(X_0)] - \beta Cov[(I - K_0)h_0(X_0), g(X_0)]\} \end{aligned}$$

with $h_0 = (I - K_0)^{-2}(I - K_0^n)K_0 h$.

PROOF. Since $(I - K_0)$ is invertible, one can verify that

$$\sum_{\ell=0}^{n-1}(n - \ell)K_0^\ell = n(I - K_0)^{-1} - (I - K_0)^{-2}(I - K_0^n)K_0.$$

We know that

$$\mathrm{Var}[\sum_{i=1}^{n}(h(X_i) - \beta g(Y_i))] = \mathrm{Var}[\sum_{i=1}^{n} h(X_i)] - 2\beta\mathrm{Cov}[\sum_{i=1}^{n} h(X_i), \sum_{j=1}^{n} g(Y_j))] + n\beta^2,$$

The variance term is given by

$$\begin{aligned} \mathrm{Var}[\sum_{i=1}^{n} h(X_i)] &= 2\sum_{1\le i\le j\le n}\mathrm{Cov}[h(X_i), h(X_j)] - \sum_{1\le i\le n}\mathrm{Var}[h(X_i)] \\ &= 2\sum_{\ell=0}^{n-1}(n - \ell)\mathrm{Cov}[h(X_{\ell+1}), h(X_1)] - n\mathrm{Var}[h(X_0)] \\ &= 2\mathrm{Cov}[\sum_{\ell=0}^{n-1}(n - \ell)K_0^\ell h(X_0), h(X_0)] - n\mathrm{Var}[h(X_0)] \\ &= n\mathrm{Cov}[(I - K_0)^{-1}(I + K_0)h(X_0), h(X_0)] \\ &\quad -2\mathrm{Cov}[(I - K_0)^{-2}(I - K_0^n)K_0 h(X_0), h(X_0)]. \end{aligned}$$

For $\ell \geq 0$, taking the conditional expectation gives

$$\begin{aligned}
\mathrm{E}[h(X_{\ell+1})|Y_1 = y] &= \mathrm{E}[\mathrm{E}[h(X_{\ell+1})|Y_1 = y, X_1]|Y_1 = y] \\
&= \mathrm{E}[\mathrm{E}[h(X_{\ell+1})|X_1]|Y_1 = y] \\
&= \mathrm{E}[K^\ell h(X_1)|Y_1 = y] \\
&= QK^\ell h(y) \\
&= w(y)(I - K)K^\ell h(y).
\end{aligned}$$

Using this expression we obtain

$$\begin{aligned}
\mathrm{Cov}[\sum_{i=1}^{n} h(X_i), \sum_{j=1}^{n} g(Y_j)] &= \sum_{1 \leq j \leq i \leq n} \mathrm{Cov}[h(X_i), g(Y_j)] \\
&= \sum_{\ell=0}^{n-1} (n - \ell)\mathrm{Cov}[h(X_{\ell+1}), g(Y_1)] \\
&= \sum_{\ell=0}^{n-1} (n - \ell)\mathrm{Cov}[\mathrm{E}[h(X_{\ell+1})|Y_1], g(Y_1)] \\
&= \sum_{\ell=0}^{n-1} (n - \ell)\mathrm{Cov}[(I - K_0)K_0^\ell h(X_0), g(X_0)] \\
&= \mathrm{Cov}[\sum_{\ell=0}^{n-1} (n - \ell)(I - K_0)K_0^\ell h(X_0), g(X_0)] \\
&= n\beta_0 - \mathrm{Cov}[(I - K_0)^{-1}(I - K_0^n)K_0 h(X_0), g(X_0)]
\end{aligned}$$

□

In most pratical situations, the algorithm is not started under the distribution Π. In the next theorem, we show that $\lim_{n\to\infty} n\mathrm{Var}[\hat{\mu}_\beta]$ does not depend on the initial distribution of the chain.

Theorem 2.2.2. *The asymptotic variance $n\,Var[\hat{\mu}_\beta]$ given in Theorem 2.2.1 does not depend on the initial distribution of the chain (X_n).*

PROOF. Note (Y_n) the sequence of independent and identically distributed random variables used in the IMH algorithm to obtain the IMH chain (X_n). Clearly, (X_n, Y_n) is also a Markov chain.

In vertue of Theorem 1.4 of Chen (1999), it is sufficient to show that (X_n, Y_n) is an Harris recurrent Markov chain. This result says that the asymptotic behaviour of an Harris Markov chain is independent from its initial distribution.

We begin by finding the transition kernel R of (X_n, Y_n) as follow:

$$\begin{aligned} R((u,v), A \times B) &:= \Pr\left((X_1, Y_1) \in A \times B | X_0 = u, Y_0 = v\right) \\ &= \int_B \Pr\left(X_1 \in A | X_0 = u, Y_0 = v, Y_1 = y\right) Q(dy) \\ &= \int_B \left[\alpha(u,y) \mathbb{1}_A(y) + (1 - \alpha(u,y)) \mathbb{1}_A(u)\right] Q(dy), \end{aligned}$$

where $\alpha(x,y) = \min\left(1, \frac{\omega(y)}{\omega(x)}\right)$ is the probability of accepting a move to y being in x, and $\omega(x) = \frac{\Pi(x)}{Q(dx)}$ is the Radon-Nikodym density of Π with respect to Q that we assume to be positive everywhere.

For $B \in \mathcal{B}^d$, note $P_{MH,B}$ to denote the Independent Metropolis-Hastings kernel obtained by using the proposal $\mathbb{1}_B(y)Q(dy)/Q(B)$ (if $Q(B) = 0$, take $P_{MH,B}$ to be the null kernel). Then it is easily seen that:

$$R((u,v), A \times B) = Q(B) P_{MH,B}(u, A). \tag{2.2.1}$$

It follows that:

$$\begin{aligned} \int \Pi(du) Q(dv) R((u,v), A \times B) &= Q(B) \int \Pi(du) P_{MH,B}(u, A) \\ &= Q(B) \Pi(A). \end{aligned}$$

This shows that $\Pi \times Q$ is the invariant distribution of R. Next we show that R is $\Pi \times Q-$irreducible.

Fix $(u,v) \in \mathbb{R}^d \times \mathbb{R}^d$ and $A \times B \in \mathcal{B}^d \times \mathcal{B}^d$ with $\Pi(A) > 0$ and $Q(B) > 0$. Since Π is absolutely continuous with respect to Q, $\Pi(A) > 0$ implies that $Q(A) > 0$. We need to show that there exists an integer n_0 such that $R^{n_0}((u,v), A \times B) > 0$. This is easy, since from (2.2.1), we can write: $R((u,v), A \times B) \geq \int_{A \cap B} \alpha(u,y) Q(dy) > 0$ because $Q(A \cap B) > 0$ and $\omega(x) > 0$. Since R is invariant with respect to $\Pi \times Q$ and is $\Pi \times Q-$irreducible, theorem 1 of Tierney (1994) asserts that R is recurrent. From Proposition 3.13 of Nummelin (1984), we deduce that every bounded harmonic function of R (a function h is called harmonic for R if $h(u,v) = Rh(u,v) := \int R((u,v), dxdy) h(x,y)$) is constant (equals to $\Pi \times Q(h)$)

$\Pi \times Q-$almost everywhere. Assume that h is any such bounded harmonic function of R. We shall show that h is actually constant everywhere and we conclude that R is Harris recurrent using theorem 2 of Tierney (1994).

We have: $R((u,v),dxdy) = \alpha(u,y)\delta_y(dx)Q(dy) + (1-\alpha(u,y))Q(dy)\delta_u(dx)$ where δ_x is the Dirac mass on x. Therefore we have:

$$Rh(u,v) = \int \alpha(u,y)h(y,y)Q(dy) + \int h(u,y)(1-\alpha(u,y))Q(dy). \qquad (2.2.2)$$

$\Pi \times Q\left(\{(x,y):\ h(x,y) \neq \Pi \times Q(h)\}\right) = 0$ implies that for all $u \in \mathbb{R}^d$, $Q\left(\{y:\ h(y,y) \neq \Pi \times Q(h)\}\right) = 0$ and $Q\left(\{y:\ h(u,y) \neq \Pi \times Q(h)\}\right) = 0$. We can then ignore these sets in the expression (2.2.2) and take $h(u,v) = \Pi \times Q(h)$ everywhere to get $Rh(u,v) = \Pi \times Q(h)$ and we are finished. □

Remark 2.2.3. *Because (X_n) is a $\Pi-$reversible chain with a spectral gap, a Central Limit Theorem holds for $\hat{\mu}_0(h) := \sum_{i=1}^n h(X_i)/n$ for every $h \in L^2(\Pi)$. That is :*

$$\sqrt{n}\left(\hat{\mu}_0(h) - Jh\right) \overset{w}{\Rightarrow} N(0,\sigma_h^2),$$

where $\sigma_h^2 < \infty$ can be shown to be equal to $\lim_{n\to\infty} n\,Var\left(\hat{\mu}_0(h)\right)$. For more details see Chan and Geyer (1994), Roberts and Rosenthal (1997). In the proof of Theorem 2.2.1, we obtained the following alternative derivation of σ_h^2:

$$\begin{aligned} n\,Var[\hat{\mu}_0(h)] &= Cov[(I-K_0)^{-1}(I+K_0)h(X_0), h(X_0)] \\ &\quad -\frac{2}{n}Cov[(I-K_0)^{-2}(I-K_0^n)K_0h(X_0), h(X_0)] \\ &\underset{n\to\infty}{\longrightarrow} Cov[(I-K_0)^{-1}(I+K_0)h(X_0), h(X_0)], \end{aligned}$$

which is also valid for any $\Pi-$reversible chain with a spectral gap, and for any $h \in L^2(\Pi)$.

Corollary 2.2.1. *In general, the choice $\beta = \beta_0$ is asymptotically optimal. Let $\eta = E[h(Y_0)]$ and $\sigma^2 = Var[h(Y_0)]$. If $g(y) = (h(y)-\eta)/\sigma$ and $\beta \in (0, 2\beta_0)$ then there exists an n_0 which depends on β and β_0 such that $Var[\hat{\mu}_0] > Var[\hat{\mu}_\beta]$ for all $n > n_0$.*

PROOF. From Theorem 2.2.1 we have

$$n\{\mathrm{Var}[\hat{\mu}_0] - \mathrm{Var}[\hat{\mu}_\beta]\} = \beta_0^2 - (\beta-\beta_0)^2 - (2\beta/n)\mathrm{Cov}[(I-K_0)^{-1}(I-K_0^n)K_0h(X_0), g(X_0)]$$

but since

$$\mathrm{Cov}[(I-K_0)^{-1}(I-K_0^n)K_0h(X_0),g(X_0)] \leq \frac{\lambda_0(1+\lambda_0)}{1-\lambda_0}\frac{\mathrm{Var}[h(X_0)]}{\sigma}$$

this implies that

$$n\{\mathrm{Var}[\hat{\mu}_0]-\mathrm{Var}[\hat{\mu}_\beta]\} \to \beta_0^2-(\beta-\beta_0)^2 \text{ as } n\to\infty$$

and it shows that β_0 is the optimal choice. Finally, if $g(y)=(h(y)-\eta)/\sigma$ then $\beta_0>0$ because $\beta_0=\mathrm{Var}[h(X_0)]/\sigma$ and $\lim_{n\to\infty} n\{\mathrm{Var}[\hat{\mu}_0]-\mathrm{Var}[\hat{\mu}_\beta]\}>0$ for any $\beta\in(0,2\beta_0)$. □

Remark 2.2.4. *In practice, β_0 may be unknown. When this is the case β_0 has to be replaced by an estimate. We suggest using*

$$\hat{\beta}_0=\frac{\sum_{i=1}^n w(Y_i)h(Y_i)(g(Y_i)-\bar{g})}{\sum_{i=1}^n g(Y_i)(g(Y_i)-\bar{g})}.$$

Corollary 2.2.2. *The best possible choice of g for reducing the asymptotic variance of $\hat{\mu}_{\beta_0}$ is given by*

$$g(y)=w(y)(h(y)-\mu)/\sqrt{Var[w(Y_0)(h(Y_0)-\mu)]}.$$

For this choice of g we obtain:

$$\hat{\mu}_{\beta_0}=\hat{\mu}_0-(\hat{\mu}_{IS}-\bar{w}\mu)$$

where $\mu=\int h d\Pi$ and $\hat{\mu}_{IS}$ is the importance sampling estimator, that is $\hat{\mu}_{IS}=\sum_{i=1}^n w(Y_i)h(Y_i)/n$.

PROOF. From Theorem 2.2.1 and the proof of Corollary 2.2.1 the asymptotic variance will be minimized if we can maximize β_0^2. We have

$$\begin{aligned}
\beta_0^2 &= \mathrm{Cov}^2[h(X_0),g(X_0)] \\
&= \mathrm{Cov}^2[(h(X_0)-\mu),g(X_0)] \\
&= \mathrm{Cov}^2[w(Y_0)(h(Y_0)-\mu),g(Y_0)] \\
&\leq \mathrm{Var}[w(Y_0)(h(Y_0)-\mu)]
\end{aligned}$$

with equality if $g(y)=w(y)(h(y)-\mu)/\sqrt{\mathrm{Var}[w(Y_0)(h(Y_0)-\mu)]}$. □

2.3. Variance reduction via Rao-Blackwellizations and symmetry

Consider the setting of the (IMHA) given in the introduction. Let $\mathcal{S}$ be the group of permutations on $\{1, 2, \ldots, n\}$. For $s \in \mathcal{S}$, $(s(1), s(2), \ldots, s(n))$ is a permutation of $(1, 2, \ldots, n)$. The Metropolis-Hasting estimator is based on $((U_1, Y_1), (U_2, Y_2), \ldots, (U_n, Y_n))$. For any $s \in \mathcal{S}$ we say that

$\tilde{\mu}_0(s)$ is equal to $\hat{\mu}_0$ evaluated at $((U_{s(1)}, Y_{s(1)}), (U_{s(2)}, Y_{s(2)}), \ldots, (U_{s(n)}, Y_{s(n)}))$.

Remark 2.3.1. *The estimators $\{\tilde{\mu}_0(s) \colon s \in \mathcal{S}\}$ are different from one another but they share the same distribution. For example, Y_1 may appears up to n times in $\tilde{\mu}_0((1, 2, \ldots, n))$ but it can appear at most once in $\tilde{\mu}_0((2, 3, \ldots, 1))$.*

Let π be the distribution on $\mathcal{S}$. Let

$$\pi\tilde{\mu}_0 = \sum_{s \in \mathcal{S}} \pi(s)\tilde{\mu}_0(s).$$

Since all of the $\tilde{\mu}_0(s)$ have the same distribution, if $X_0 \sim \Pi$ we obtain from Jensen's inequality that

$$\mathrm{E}[\pi\tilde{\mu}_0] = \mu \text{ and } \mathrm{Var}[\pi\tilde{\mu}_0] < \mathrm{Var}[\hat{\mu}_0]$$

if π is not a Dirac distribution. A natural thing to do is to consider the uniform distribution on $\mathcal{S}$. However, if we cannot find any simplifications, the algorithm will involve $n!$ evaluations and it will become untractable for large values of n. The simplifications that we have found so far do not help that much. A second approch would be to replace $\pi\tilde{\mu}_0$ by an approximation. For example, it could be a Monte Carlo simulation of fixed size or a numerical approach such as quasi Monte Carlo. A third approach consists in replacing the uniform distribution by another distribution. Let

$$[i] = 1 + (i - 1) \mathrm{mod}_n$$

and

$$\mathcal{S}_0 = \{(k, [k+1], \ldots, [k+n-1]) \colon k = 1, 2, \ldots, n\}.$$

Let

$$\tilde{\mu}_k = \tilde{\mu}_0((k, [k+1], \ldots, [k+n-1])).$$

Consider

$$\begin{aligned} z(i) &= \sum_{j=0}^{n-1}\prod_{k=0}^{j} \mathrm{I}(U_{[i+k]}w(X_0) > w(Y_{[i+k]})) \\ m(i) &= 1+\sum_{j=1}^{n-1}\prod_{k=1}^{j} \mathrm{I}(U_{[i+k]}w(Y_i) > w(Y_{[i+k]})) \\ s_j(i) &= \begin{cases} z([i]) & \text{if} \quad j=1, \\ s_{j-1}(i)+m([i+s_{j-1}(i)]) & \text{if} \quad j>1 \end{cases} \\ \ell(i) &= \max\{j\colon s_j(i)\le n, j\ge 1\}. \end{aligned}$$

We obtain,

$$\tilde{\mu}_k = \frac{1}{n}\{z(k)h(X_0)+\sum_{j=1}^{\ell(k)} m([k+s_j(k)])h(Y_{[k+s_j(k)]})-(s_{\ell(k)+1}(k)-n)h(Y_{[k+s_{\ell(k)}(k)]})\}$$

and if π is the uniform distribution on $\mathcal{S}_0$ then

$$\pi\tilde{\mu}_0 = \frac{1}{n}\sum_{k=1}^{n}\tilde{\mu}_k.$$

This is the symmetrized version of $\hat{\mu}_0$. Casella and Robert (1996) suggest taking the conditional expectation of the Metropolis-Hastings estimator given that $Y_1, Y_2, \ldots, Y_n$ are fixed. Let us call this estimator $\hat{\mu}_0^*$. It is easy to improve $\hat{\mu}_0^*$ by simply considering the conditional expectation given that $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ are fixed, as Perron (1999) did for a different problem. However, in our context, conditioning on the order statistics will involve too many calculations which is unfeasable for large values of n. Here again, the evaluation of $\hat{\mu}_0^*$ on $(Y_1, Y_2, \ldots, Y_n)$ will be different than the evaluation of $\hat{\mu}_0^*$ on $(Y_2, Y_3, \ldots, Y_1)$. Let us say that

$\tilde{\mu}_k^*$ is equal to $\hat{\mu}_0^*$ evaluated at $(Y_k, Y_{[k+1]}, \ldots, Y_{[k+n-1]})$, for example, $\tilde{\mu}_1^* = \hat{\mu}_0^*$.

In general, if we set

$$\begin{aligned}
p^*(i,j) &= 1 \wedge w(Y_j)/w(Y_i) \text{ for } i,j = 1,2,\ldots,n \\
p^*(0,j) &= 1 \wedge w(Y_j)/w(X_0)) \text{ for } j = 1,2,\ldots,n \\
q^*(i,j) &= 1 - p^*(i,j) \\
f^*(i,\ell) &= \prod_{j=1}^{\ell} q^*(i,[i+j]) \text{ with } f^*(i,0) = 1 \\
\varphi^*(k,\ell) &= \prod_{j=1}^{\ell} q^*(0,[k+j-1]) \\
\delta^*(k,1) &= p^*(0,[k]) \text{ and for } \ell > 1 \\
\delta^*(k,\ell) &= \varphi^*(k,\ell-1)p^*(0,k+\ell-1) \\
&\quad + \sum_{j=1}^{\ell-1} \delta^*(k,j) f^*([k+j],\ell-j-1)p^*([k+j],[k+\ell]),
\end{aligned}$$

then

$$\tilde{\mu}_k^* = \frac{1}{n}\sum_{\ell=1}^{n}\{\varphi^*(k,\ell)h(X_0) + [\sum_{j=0}^{n-\ell} f^*([k+\ell],j)]\delta^*(k,\ell)h(Y_{[k+\ell]})\}$$

and if π is the uniform distribution on $\mathcal{S}_0$ then

$$\pi\tilde{\mu}_0^* = \frac{1}{n}\sum_{k=1}^{n}\tilde{\mu}_k^*.$$

Remark 2.3.2. *As mentioned in Casella and Robert (1996) the evaluations of f^* are very time consuming. For $\hat{\mu}_0^*$, $f^*(i,\ell)$ has to be evaluated at $i = 1,2,\ldots,n$, $\ell = 1,2,\ldots,n-i$. For $\pi\tilde{\mu}_0^*$, $f^*(i,\ell)$ has to be evaluated at $i,\ell = 1,2,\ldots,n$. Thus, even if $\pi\tilde{\mu}_0^*$ is an average of n estimators similar to $\hat{\mu}_0^*$ it requires only twice the number of evaluations of f^* needed to compute $\hat{\mu}_0^*$.*

Now we shall see that the reduction in the variance produced by a Rao-Blackwellization is rather limited. In fact, $\mathrm{Var}[\hat{\mu}_0^*]$ is of the order of $O(1/n)$.

Remark 2.3.3. *If we take into account how time consuming is the Rao-Blackwellisation for large values of n, it is better to increase the sample size than to perform a Rao-Blackwellization when we want to reduce the variance in the case where n is large. A better strategy might be to run several parallel chains based on small*

sample sizes with Rao-Blackwellizations instead of running one chain with a large sample size.

Lemma 2.3.1. *Let Z be a statistic based on $(U_1, Y_1), (U_2, Y_2), \ldots, (U_n, Y_n)$ such that the vector of the order statistics $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)})$ is a function of Z and assume that g is a covariate. We obtain that*

$$Cov[E[\sum_{i=1}^{n} h(X_i)|Z], \sum_{j=1}^{n} g(Y_j)] = Cov[\sum_{i=1}^{n} h(X_i), \sum_{j=1}^{n} g(Y_j)].$$

PROOF.

$$\begin{aligned} \mathrm{Cov}[\sum_{i=1}^{n} h(X_i), \sum_{j=1}^{n} g(Y_j)] &= \mathrm{Cov}[\mathrm{E}[\sum_{i=1}^{n} h(X_i)|Z], \sum_{j=1}^{n} g(Y_j)] \\ &\quad + \mathrm{E}[\mathrm{Cov}[(\sum_{i=1}^{n} h(X_i), \sum_{j=1}^{n} g(Y_j))|Z]] \end{aligned}$$

but $\sum_{j=1}^{n} g(Y_j)$ is fixed when Z is fixed. □

Theorem 2.3.1. *Under the conditions of Lemma 2.3.1 we obtain:*

$$\liminf_{n\to\infty} n\, Var[E[\hat{\mu}_0|Z]] \geq Var[w(Y_0)(h(Y_0) - \mu)].$$

PROOF. Assume that $g(y) = w(y)(h(y) - \mu)$ and let $\bar{g} = \sum_{j=1}^{n} g(Y_j)/n$. We obtain:

$$\begin{aligned} n\mathrm{Var}[\mathrm{E}[\hat{\mu}_0|Z]] &= n\mathrm{Var}[\mathrm{E}[\hat{\mu}_0|Z] - \frac{\mathrm{Cov}[\mathrm{E}[\hat{\mu}_0|Z], \bar{g}]}{\mathrm{Var}[\bar{g}]}\bar{g}] + n\frac{(\mathrm{Cov}[\mathrm{E}[\hat{\mu}_0|Z], \bar{g}])^2}{\mathrm{Var}[\bar{g}]} \\ &\geq n\frac{(\mathrm{Cov}[\mathrm{E}[\hat{\mu}_0|Z], \bar{g}])^2}{\mathrm{Var}[\bar{g}]} \\ &= n\frac{(\mathrm{Cov}[\hat{\mu}_0, \bar{g}])^2}{\mathrm{Var}[\bar{g}]} \\ &\to \mathrm{Var}[w(Y_0)(h(Y_0) - \mu)] \text{ as } n \to \infty \end{aligned}$$

where the last equality comes from Lemma 2.3.1 and the limiting result is explained in the proof of Theorem 2.2.1. □

Remark 2.3.4. *Finally, Lemma 2.3.1 tells us that it is possible to combine the result of this section with that of the previous section. It suggests the following estimator:*

$$\pi\hat{\mu}^*_{\beta_0} = \pi\tilde{\mu}^*_0 - \beta_0\bar{g}.$$

Moreover, from Lemma 2.3.1 we see that the improvement of $\pi\hat{\mu}^*_{\beta_0}$ over $\hat{\mu}_0$ is the improvement due to the use of a covariate plus the improvement due to the use of symmetry combined with the Rao-Blackwellization.

Example 2.3.1. Computation of the mean μ of a Gamma Distribution. *This example comes from Casella and Robert (1996). The target is a Gamma distribution with mean α and variance α while the proposal is a Gamma distribution with mean α and variance $\alpha^2/2$. In fact, the proposal is the distribution of the sum of two independent random variables exponentially distributed with the same mean. We obtain that*

$$w(x) = \left(\frac{\alpha}{2}\right)^2 \frac{1}{\Gamma(\alpha)} x^{\alpha-2} \exp -\{(\alpha-2)x/\alpha\}, \quad \text{for } x > 0$$

and that w is bounded from above by $w(\alpha)$ for $\alpha \geq 2$. Clearly, if α is close to 2 from the right than we shall have very few rejections in the IMHA. In these situations, the method using a covariate will be very efficient for reducing the variance. However, if α is much larger than 2 then we shall have many repetitions in the IMHA and the methods developed in section 3 will take advantage of these repetitions. Let us see numerically how it looks like. We consider:

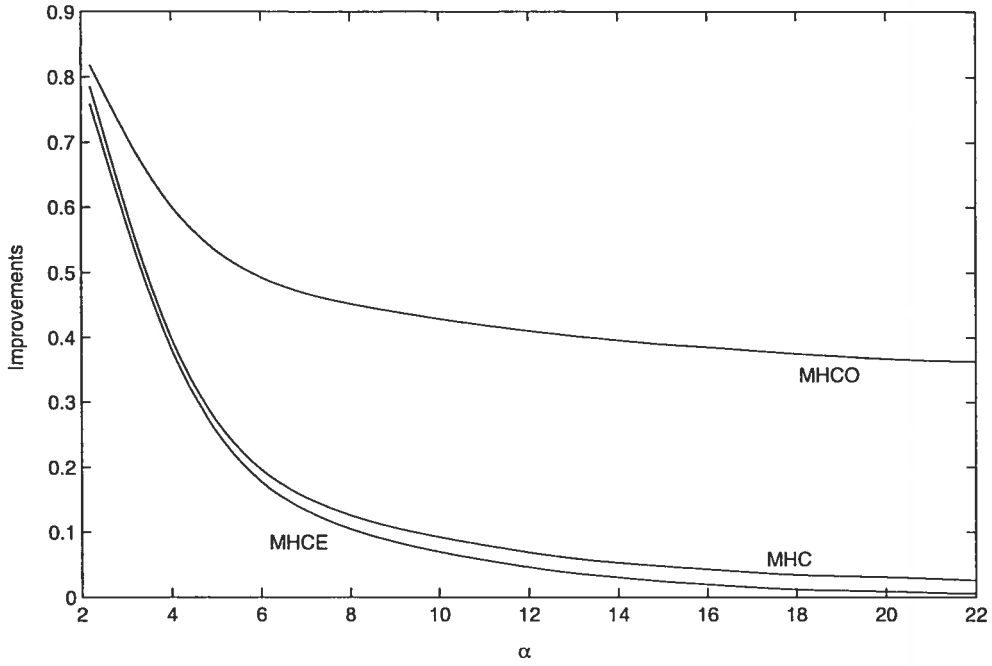
$$h(x) = x \qquad\qquad g(x) = \sqrt{2}(x-\alpha)/\alpha$$

so $E[g(Y_0)] = 0$, $Var[g(Y_0)] = 1$ and $\beta_0 = \sqrt{2}$. The improvement of the estimator $\hat{\mu}$ compared to the IMHA estimator $\hat{\mu}_0$ is given by $1 - Var(\hat{\mu})/Var(\hat{\mu}_0)$. Our analysis will be based on an estimations of these variances based on 10 000 replications. The parameter α will vary from 2.2 to 22.0.

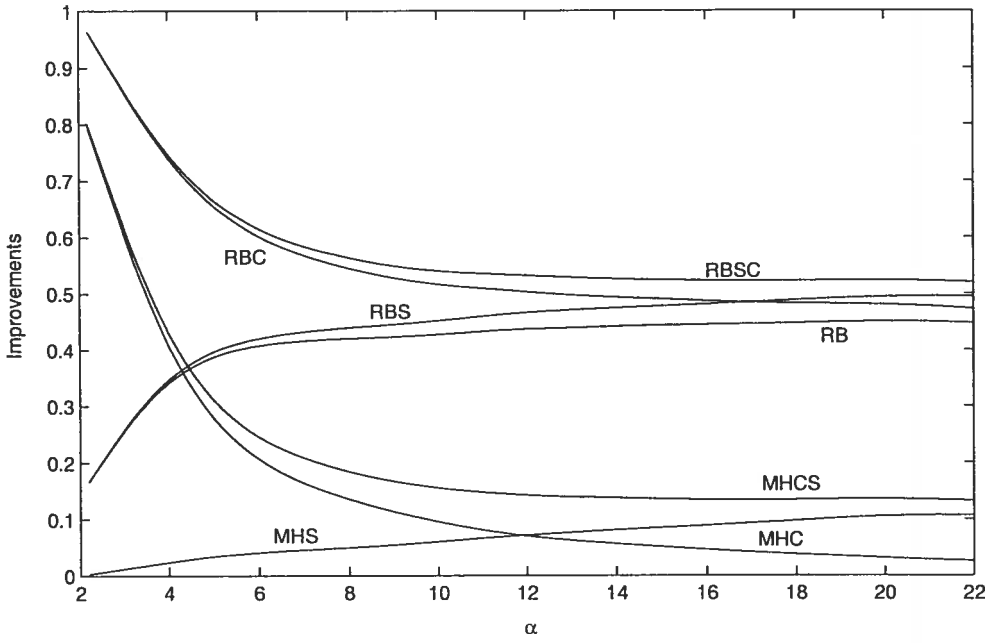
In Graph 1 we consider the case $n = 100$. This graph shows that the covariate approach is very good when the proposal is close to the target and the method given provides a small improvement even though there is a big difference between the porposal and the target. It is rather surprising to see how well the method works when the choice of the covariate is optimized. Maybe, we should work on having a good approximation of the best possible covariate! Finally, the estimation of the best possible choice for β_0 works well. Initially, we had considered the case $n = 1000$ but in this situation it was too difficult to see graphically a difference between the case where β_0 is fixed and the one where it is estimated.

In Graph 2 we study the different estimators proposed in this paper for the case $n = 25$. *The case* $n = 25$ *has been chosen because the Rao-Blackwellization is too time consuming for large values of* n. *We see that the Rao-Blackwellizations help a lot when there are many rejections in the IMHA and, to a lesser extent, this is also true for the symmetric versions. We see also, graphically, that the improvement given by the approach developed in Section 2 is added to the one given by an approach developed in Section 3 when the two approaches are combined. It is surprising that the use of a covariate is still good, considering that the choice of the covariate has been made on the basis of asymptotic considerations. If we take into account evaluation time and algorithm complexity perhaps the use of a symmetric version with a covariate would be a nice approach.*

Graph1: Improvements of the control variate estimators over the classical estimator.
MHC: optimal β.
MHCE: estimated β.
MHCO: optimal β with optimal covariate.



Graph2: Improvements of the new estimators over the classic al estimator.
MHC: control variate. MHS: symmetry. MHCS: control variate and symmetry.
RB: Rao–Blackwellization. RBC: Rao–Blackwellization and con trol variate.
RBSC: Rao–Blackwellization and symmetry and control variate .
RBS: Rao–Blackwellization and symmetry.



Bibliography

Casella, G., Robert, C. P., 1996. Rao-Blackwellisation of sampling schemes. Biometrika 83 (1), 81–94.

Chan, K., Geyer, G., 1994. Discussion paper. Ann. Statist. 22 (4), 1747–1758.

Chen, M.-H., Shao, Q.-M., Ibrahim, J. G., 2000. Monte Carlo Methods in Bayesian Computation. Springer-Verlag, New York.

Chen, X., 1999. Limit theorems for functionals of ergodic markov chains with general state space. Memoirs of AMS 139 (664).

Evans, M., Swartz, T., 2000. Approximating integrals via Monte Carlo and deterministic methods. Oxford University Press, Oxford.

Liu, J. S., 2001. Monte Carlo Strategies in Scientific Computing. Springer Verlag, New-York.

Nummelin, E., 1984. General Irreducible Markov Chains and Nonegative Operators. Cambridge University Press, London.

Perron, F., 1999. Beyond accept-reject sampling. Biometrika 86 (4), 803–813.

Robert, C. P., Casella, G., 1999. Monte Carlo statistical methods. Springer-Verlag, New York.

Roberts, G. O., Rosenthal, J. S., 1997. Geometric ergodicity and hybrid Markov chains. Electron. Comm. Probab. 2, no. 2, 13–25 (electronic).

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22 (4), 1701–1762, with discussion and a rejoinder by the author.

Chapitre 3

GEOMETRIC ERGODICITY OF THE RECTRICTED METROPOLIS ALGORITHM

Abstract.

For an ergodic Metropolis-Hastings algorithm satisfying some compact condition, we show that the induced Markov chain has a spectral gap if and only the probability of rejection is bounded away from unity. Using this result, we propose a modification of the Metropolis algorithm (that we called the Restricted Metropolis algorithm) which is shown to be geometrically ergodic for any density (bounded on compact sets) with tails lighter than the proposal distribution's tails. A simulation example is presented to illustrate. *AMS Classification*: 65C05, 65C40, 60J27, 60J35.

Keywords: METROPOLIS-HASTINGS ALGORITHM, MARKOV CHAINS OPERATOR, GEOMETRIC ERGODICITY

3.1. INTRODUCTION

The Metropolis-Hastings (MH) algorithm initiated by Metropolis et al. (1953) is a very flexible algorithm used to approximately sample from complicated distributions in high dimension spaces.

If π is the probability distribution of interest, such an algorithm generates a Markov chain (X_n) which admits π as its stationary distribution. If the chain is ergodic, then for n sufficiently large, X_n is taken as approximately distributed as π. From a statistical point of view, geometric ergodicity is a desirable property for such algorithms. Specifically, under geometric ergodicity, the Central Limit Theorem for empirical sum of functions of X_n is (theoretically) easy to check as in Roberts and Rosenthal (1997). Geometric ergodicity can also be used to design convergence diagnostic techniques, a big issue in practical use of MCMC methods (see Brooks and Roberts (1998)). There is an interesting discussion about the usefulness of geometric ergodicity of MCMC algorithms in Roberts and Rosenthal (1998).

In Section 1, we show that under a compact condition, an ergodic Metropolis-Hastings chain is geometrically ergodic if the rejection probability of the chain is bounded away from unity. As we shall see, this result can be useful to design new algorithms which are geometrically ergodic. Actually, it is a partial converse of Roberts and Tweedie (1996) who proved that if a Metropolis-Hastings chain is geometrically ergodic, then the rejection probability of the chain is bounded away from unity.

Our techniques of proof differ from those in Roberts and Tweedie (1996) as we mainly use results from Hilbert spaces operators theory. We decompose K_0, the operator induced by the Metropolis-Hastings chain, as the sum of a multiplication operator M_r (multiplication by the rejection probability or the chain denoted r) and an integral operator U. Under a compact condition on U, the Weyl's theorem (Berberian (1970)) states that the "continuous" part of the spectrum of K_0 is the same as the one of M_r which is ess-ran(r), the essential range of r. Then the reversibility and the ergodicity of the chain is used to assure that the "discrete" part of the spectrum of K_0 remains bounded away from unity.

In the particular case of the Independent Metropolis-Hastings algorithm (IMH), the compact condition is always satisfied, but in this case more can be said. In theorem 3.2.2, we show that the spectrum of the IMH chain is exactly the essential range of the rejection probability of the chain. Liu (1996) proved this result when the state space of the chain is finite or discrete (with an additional regularity condition on π). He has also conjectured that the result holds in general spaces. Smith and Tierney (1996) have derived the transition kernel of the successive iterates of the IMH chain in general space. But their results did not clearly solve this conjecture.

In general, it may be difficult to choose a good proposal to use with the MH algorithm. This is certainly one reason why the Metropolis algorithm is still widely used. But as shown by Jarner and Tweedie (2001), exponential or lighter tails is necessary for the Metropolis algorithm to be geometrically ergodic. For heavy tailed distributions, polynomial rate of convergence is possible. In this direction, Jarner and Roberts (2002) have shown that for a symmetric target density with polynomial tails, and for any proposal density with tails that also recede polynomially, the Metropolis algorithm has a polynomial rate of convergence.

In Section 3, as another solution, we propose to modify the Metropolis algorithm by restricting the random walk behaviour on a compact set. This restriction yields a chain that is geometrically ergodic whenever the target density has a lighter tails than the proposal's tails. The idea is as follows. We fix a convex compact set Δ. Ideally, Δ contains all the modes of π. At any $x \in \Delta$, the proposal move is made from a distribution centered at x as in the Metropolis algorithm. But when the chain reaches $x \notin \Delta$, the proposal distribution is centered at $p(x)$ the projection of x on Δ. We show in Theorem 3.3.1 that this algorithm is geometrically ergodic for any density (bounded on compact sets) with tails decaying faster than the proposal's tails.

The main points of the paper are as follows. The geometric ergodicity of the general Metropolis-Hastings algorithm is shown in Theorem 3.2.1 of section 2. The spectrum of the IMH chain is derived in Theorem 3.2.2. The geometric ergodicity of the restricted Metropolis algorithm is shown in Theorem 3.3.1 of

Section 3. A simulation example is also presented to illustrate the benefit of restricting the Metropolis algorithm.

3.2. Geometric Ergodicity of the Metropolis-Hastings Chain

Throughout the section, we fix $(\mathcal{S}, \mathcal{F}, \pi)$ a probability space, where π is our target probability measure. Let $Q(x, \cdot)$ be a transition kernel on $\mathcal{S}$. We shall assume that for all x, $Q(x, \cdot)$ is absolutely continuous with respect to π and we write $Q(x, dy) = \omega(x, y)\pi(dy)$.

Before going further, we recall the Metropolis-Hastings algorithm for simulating a Markov chain which admits π as its invariant distribution.

Algorithm 3.2.1. *The Metropolis-Hastings Algorithm:*

At the ith iteration, $X_i = x$.

Generate $Y_{i+1} = y$ *from* $Q(x, \cdot)$.

set

$$X_{i+1} = \begin{cases} y & \text{with probability } \alpha(x, y) \\ x & \text{with probability } 1 - \alpha(x, y), \end{cases}$$

with

$$\alpha(x, y) = \begin{cases} Min\left(1, \frac{\omega(y,x)}{\omega(x,y)}\right) & if \omega(x, y) \neq 0 \\ 1 & if\ \omega(x, y) = 0. \end{cases}$$

The algorithm leaves a lot of room in choosing Q. When $Q(x, \cdot) = Q(\cdot)$ for all $x \in \mathcal{S}$, the algorithm thus obtained is usually called the Independent Metropolis-Hastings algorithm (IMH). If $\mathcal{S} = \mathbb{R}^d$ and $Q(x, \cdot)$ has a density with respect to the Lebesgue measure with $Q(x, dy) = q(|y - x|)dy$, we obtain the Random Walk Metropolis algorithm (RWMH), where $|x|$ denotes the norm of $x \in \mathcal{S}$.

The algorithm (3.2.1) generates a Markov chain (X_n) with transition kernel

$$P(x, dy) = \alpha(x, y)\omega(x, y)\pi(dy) + r(x)\delta_x(dy), \tag{3.2.1}$$

where

$$r(x) = 1 - \int \alpha(x, y)\omega(x, y)\, \pi(dy) \tag{3.2.2}$$

is the probability of rejection of the chain and $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. It is well known that P is reversible with respect to π, that is:

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx), \tag{3.2.3}$$

as measures on $\mathcal{S} \times \mathcal{S}$. This implies that P admits π as an invariant distribution, $\pi P = \pi$ where

$$\pi P(A) := \int \pi(dx)P(x, A). \tag{3.2.4}$$

Below, we first review some basic concepts from Markov chain theory. For more details, we refer to Meyn and Tweedie (1993). We say that a Markov chain (X_n) with transition kernel P is $\phi-$irreducible if there exists a probability measure ϕ such that

$$\phi(A) > 0 \text{ implies that } \Pr(X_n \in A \text{ for some } n | X_0 = x) > 0 \text{ (for all } x \in \mathcal{S}). \tag{3.2.5}$$

We say that a Markov chain is aperiodic if there does not exist a partition $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_d$ with $d \geq 2$ such that $\Pr(x, \mathcal{S}_{i+1}) = 1$ for all $x \in \mathcal{S}_i$, $i = 1, \ldots, d-1$ and $\Pr(x, \mathcal{S}_1) = 1$ for all $x \in \mathcal{S}_d$.

And a Markov chain (X_n) with transition P is ergodic if

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{var} \longrightarrow_{n \longrightarrow \infty} 0 \tag{3.2.6}$$

for $\pi-$almost every $x \in \mathcal{S}$, where $\|\mu\|_{var} := \frac{1}{2} \sup_{|f| \leq 1} \left| \int \mu(dy) f(y) \right|$ is the total variation norm of a signed measure μ. It is well known that if a ϕ-irreducible Markov chain admits an invariant distribution, and is aperiodic then it is ergodic.

Geometric ergodicity essentially says that the convergence in (3.2.6) takes place at a geometric rate. More precisely, a Markov chain with transition kernel P is geometrically ergodic if there is $\rho < 1$ and a function $V : \mathcal{X} \longrightarrow [1, \infty]$ finite $\pi-$almost everywhere such that:

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{var} \leq V(x)\rho^n \tag{3.2.7}$$

for $\pi-$almost every $x \in \mathcal{S}$.

The transition kernel P of the chain (X_n) induces a linear bounded operator K on $L^2(\pi)$ the space of real-valued square integrable functions defined on $\mathcal{S}$

which is given by:

$$Kf(x) := \int f(y)P(x, dy). \tag{3.2.8}$$

Instead of the operator K defined in (3.2.8), we shall mainly work with K_0, the restriction of K to $L_0^2(\pi) = \{f \in L^2(\pi) : \pi(f) := \int f\, d\pi = 0\}$. It has been shown that (see Roberts and Tweedie (2001) and Roberts and Rosenthal (1997)) for reversible Markov chains, geometric ergodicity is equivalent to:

$$\|K_0 f\| \leq \rho \|f\|, \tag{3.2.9}$$

for some $\rho < 1$.

Define the spectrum of K_0 by $\sigma(K_0) := \{\lambda \in \mathbb{R} : \ K_0 - \lambda I$ is non invertible$\}$ where I is the identity operator of $L_0^2(\pi)$, and write $r(K_0 :) = \sup\{|\lambda| : \ \lambda \in \sigma(K_0)\}$ for the spectral radius of K_0. Then it is well known that (3.2.9) is equivalent to $r(K_0) = \|K_0\| < 1$, where $\|K_0\|$ is the norm of the operator K_0 defined by $\|K_0\| := \sup_{\|f\| \leq 1} \|K_0 f\|$. Whenever $\|K_0\| < 1$, we say that the chain has a spectral gap.

In the study of the spectrum of K_0, we make the distinction between $\sigma_d(K_0)$, the discret spectrum of K_0 and $\sigma_{ess}(K_0) = \sigma(K_0) \setminus \sigma_d(K_0)$ the essential spectrum of K_0. The discret spectrum $\sigma_d(K_0)$ is defined as those λ in $\sigma(K_0)$ which are eigenvalues of K_0 and are isolated points of $\sigma(K_0)$ and such that $\dim \ker (K_0 - \lambda I) < \infty$, where I is the identity operator on $L_0^2(\pi)$. We shall also denote by

$$\text{ess-ran}(r) := \{\lambda \in \mathbb{R} : \pi\{x : |r(x) - \lambda| < \epsilon\} > 0 \text{ for all } \epsilon > 0\}$$

the essential range of r. If ess-inf(r) (respectively ess-sup(r)) is the essential (with respect to π) infimum (respectively essential supremum) of the function r, it is easily seen that ess-ran$(r) \subseteq$ [ess-inf(r), ess-sup(r)] and that both ess-inf(r) and ess-sup(r) belong to ess-ran(r).

The following result from Chan and Geyer (1994) will be useful later.

Proposition 3.2.1. *Suppose that P is an ergodic transition kernel on $(\mathcal{S}, \mathcal{F}, \pi)$ with invariant distribution π. Then K_0 as defined above has no eigenvalue with absolute value 1.*

In the case of the Metropolis-Hastings algorithm, the operator K_0 acts on $f \in L_0^2(\pi)$ as:

$$K_0 f(x) = \int f(y)P(x,dy) \tag{3.2.10}$$

$$= M_r f(x) + Uf(x). \tag{3.2.11}$$

With $M_r f(x) = r(x)f(x)$ and $Uf(x) = \int \alpha(x,y)\omega(x,y)f(y)\pi(dy)$.

In other words, the Metropolis-Hastings operator is a multiplication operator perturbed by an integral operator. The simplest way to study the spectrum of such operator is to assume that U is a compact operator. This will be done through the following condition:

Assumption 3.2.1. *The operator U is a compact $L_0^2(\pi)$ operator.*

The following lemma gives a sufficient condition on the proposal transition kernel Q for this compact condition to hold.

Lemma 3.2.1. *If*

$$\int Q(x,dy)Q(y,dx) < \infty, \tag{3.2.12}$$

then U is an Hilbert-Schmidt operator, thus is compact.

PROOF. It is sufficient to check that if (3.2.12) holds then

$$\int \alpha^2(x,y)\omega^2(x,y)\pi(dx)\pi(dy) < \infty.$$

It follows from $\alpha(x,y)\omega(x,y) = Min\,(\omega(x,y),\omega(y,x))$, that $\alpha^2(x,y)\omega^2(x,y) \leq \omega(x,y)\omega(y,x)$.

Therefore:

$$\alpha^2(x,y)\omega^2(x,y)\pi(dx)\pi(dy) \leq Q(x,dy)Q(y,dx).$$

□

The next theorem is the main result of this section. We have used the Weyl's perturbation theorem (Berberian (1970)) together with some basic Hilbert spaces operators theory to show that K_0 has a spectral gap when ess-sup$(r) < 1$.

Theorem 3.2.1. *Suppose that the compact condition on U holds. Suppose also that the Markov chain generated by the Metropolis-Hastings algorithm (algorithm 3.2.1) with proposal kernel Q is ergodic. Then it is geometrically ergodic if and only if ess-sup$(r) < 1$. The essential supremum being taken with respect to π.*

PROOF. Since U is a compact operator, by the Weyl's theorem (Berberian (1970)), $\sigma_{ess}(K_0) = \sigma_{ess}(M_r)$. The spectrum of the multiplication operator is well known (see Conway (1985) example 2.6 page 271). $\sigma(M_r) = \text{ess-ran}(r)$ and λ is an eigenvalue for M_r if and only if $\pi\{y : r(y) = \lambda\} > 0$ and the indicator function of the set $\{y : r(y) = \lambda\}$ is an associated eigenfunction. Thus

$$\begin{aligned} \sigma_{ess}(M_r) &\subseteq \text{ess-ran}(r), \\ &\subseteq [\text{ess-inf}(r), \text{ess-sup}(r)]. \end{aligned}$$

Since K_0 is self-adjoint, either $\|K_0\| \in \sigma(K_0)$ or $-\|K_0\| \in \sigma(K_0)$ (Halmos (1957) Section 34, Theorem 2) and $\sigma(K_0)$ is bounbed by $\|K_0\|$. Suppose that ess-sup$(r) < 1$. Then if ess-sup$(r) < \|K_0\|$, $\|K_0\| \in \sigma_d(K_0)$ (or $-\|K_0\| \in \sigma_d(K_0)$). But for an ergodic chain, we know from Proposition (3.2.1) that ± 1 cannot be eigenvalues for K_0. Thus $\|K_0\| < 1$.

The necessary part is Proposition 5.1 of Roberts and Tweedie (1996). □

As, we mentioned above, the necessary part of theorem 3.2.1 holds even without the compact condition (3.2.1) (Roberts and Tweedie (1996)).

In practice, it is not very hard to construct an ergodic MH algorithm. The following proposition is adapted from Tierney (1994) and Roberts and Tweedie (1996) and gives simple conditions under which the MH kernel is ergodic. Weaker conditions are possible.

Proposition 3.2.1. *Suppose that there is a set $A \in \mathcal{S}$ with $\pi(A) > 0$ and $\varepsilon > 0$ such that:*

$$\omega(x, y) > 0 \text{ for all } x, y \in \mathcal{S}, \tag{3.2.13}$$

and

$$\omega(x, y) > \varepsilon \text{ for all } x, y \in A. \tag{3.2.14}$$

Then the MH kernel is ergodic.

Condition (3.2.13) is not a restrictive requirement and will be satisfied in many situations. If $\omega(x, y)$ is continuous and satisfies (3.2.13), then (3.2.14) also will be satisfied and we can take A to be any non empty compact subset of $\mathcal{S}$.

In the case of the IMH algorithm, (3.2.12) of Lemma 3.2.1 is always satisfied. Therefore we have the following well known result on the geometric ergodicity

of the IMH algorithm (Tierney (1994), Liu (1996), Smith and Tierney (1996) Mengersen and Tweedie (1996)).

Corollary 3.2.1. *Let r be the probability of rejection of the IMH chain as given by Equation (3.2.2). Then ess-inf$(r) = 0$ and ess-sup$(r) = 1 -$ ess-inf(ω). Therefore the IMH algorithm has a spectral gap if and only if ess-inf$(\omega) > 0$.*

Note that by Proposition 3.2.1, ess-inf$(\omega) > 0$ implies that the IMH algorithm is ergodic. In fact, more can be said about the spectrum of the IMH chain. We shall prove the following:

Theorem 3.2.2. *For the IMH algorithm, $\sigma(K_0) \subseteq$ ess-ran(r). The equality holds if for all $\alpha \in$ ess-ran(r), $\pi\{y : r(y) = \alpha\} = 0$.*

PROOF. For the IMH algorithm, condition (3.2.12) of Lemma 3.2.1 is clearly satisfied. Therefore the operator U in the decomposition (3.2.11) is compact and we have, by the Weyl's perturbation theorem, that $\sigma_{ess}(K_0) = \sigma_{ess}(M_r) \subseteq$ ess-ran(r). Next, we show that for any eigenvalue λ of K_0, $\lambda \in$ ess-ran(r) and conclude that $\sigma(K_0) \subseteq$ ess-ran(r).

First note that for $f \in L_0^2(\pi)$,

$$\begin{aligned} Uf(x) &= \int \alpha(x,y)\omega(y)f(y)\pi(dy) \\ &= \int_{\{y:\,\omega(x)\geq\omega(y)\}} \omega(y)f(y)\pi(dy) + \int_{\{y:\,\omega(x)<\omega(y)\}} \omega(x)f(y)\pi(dy) \\ &= \int_{\{y:\,\omega(y)\leq\omega(x)\}} (\omega(y)-\omega(x))\, f(y)\pi(dy). \end{aligned} \tag{3.2.15}$$

Now, take $\lambda \notin$ ess-ran(r) and suppose that there is a none zero $f_0 \in L_0^2(\pi)$ such that $K_0 f_0(x) = \lambda f_0(x)$. We shall prove that this leads to a contradiction and the result will be proved.

From (3.2.15) and (3.2.11), λ being an eigenvalue of K_0 with eigenfunction f_0 is equivalent to:

$$\int_{\{y:\,\omega(y)\leq\omega(x)\}} \frac{\omega(x)-\omega(y)}{r(x)-\lambda} f_0(y)\pi(dy) = f_0(x). \tag{3.2.16}$$

Consider T the operator $Tf(x) = \int_{\{y:\,\omega(y)\leq\omega(x)\}} \frac{\omega(x)-\omega(y)}{r(x)-\lambda} f(y)\pi(dy)$.

Note $\underline{\omega} =$ ess-inf$(\omega(x))$ and $\kappa =$ ess-inf$\,(|r(x)-\lambda|)$. Since $\lambda \notin$ ess-ran(r), $\kappa > 0$. Because f_0 is $\pi-$integrable and is not identically null, we can find $u > \underline{\omega}$

sufficiently large such that f_0 is not null on $\{x \in \mathcal{S} : \underline{\omega} \leq \omega(x) < u\}$. For any partition $I_n = (u_0 \leq u_1 \leq \ldots \leq u_n)$ of $[\underline{\omega}, u]$, with $u_0 = \underline{\omega}$ and $u_n = u$, we note $D(u_i) := \{x \in \mathcal{S} : u_{i-1} \leq \omega(x) < u_i\}$ and $L_i^2(\pi) := \{h \in L_0^2(\pi) : h(x) = 0 \text{ for } x \notin D(u_i)\}$, $i = 1, \ldots, n$. $L_i^2(\pi)$ is an Hilbert space as a closed subspace of $L_0^2(\pi)$. Let $\chi_{D(u_i)}$ be the indicator function of the set $D(u_i)$ and let M_{D_i} be the multiplication operator defined by $\chi_{D(u_i)}$. Note that $M_{D_i} M_{D_i} = M_{D_i}$. For $h \in L_0^2(\pi)$, we write h_{D_i} for $h\chi_{D(u_i)} = M_{D_i} h$. Note that if T is an operator on $L_0^2(\pi)$, then $M_{D_i} T M_{D_i}$ is an operator on $L_i^2(\pi)$. With these notations, by applying M_{D_1} on both side of (3.2.16) we obtain that $M_{D_1} T f_0 = M_{D_1} f_0$. But:

$$\begin{aligned} M_{D_1} T f_0(x) &= \int_{\{y:\, \omega(y) \leq \omega(x)\}} \frac{\omega(x) - \omega(y)}{r(x) - \lambda} f_0(y) \chi_{D_1}(x) \pi(dy) \\ &= \int_{\{y:\, \omega(y) \leq \omega(x)\}} \frac{\omega(x) - \omega(y)}{r(x) - \lambda} M_{D_1} f_0(y) \chi_{D_1}(x) \pi(dy) \\ &= M_{D_1} T M_{D_1} f_{0,D_1}. \end{aligned}$$

Therefore, (3.2.16) implies that $M_{D_1} T M_{D_1} f_{0,D_1} = f_{0,D_1}$. In the same way, for $2 \leq i \leq n$, (3.2.16) implies that $M_{D_i} T f_0 = M_{D_i} f_0$, and as before, we have:

$$\begin{aligned} M_{D_i} T f_0(x) &= \int_{\{y:\, \omega(y) \leq \omega(x)\}} \frac{\omega(x) - \omega(y)}{r(x) - \lambda} f_0(y) \chi_{D_i}(x) \pi(dy) \\ &= \sum_{k=1}^{i-1} \int_{\{y:\, y \in D(u_k)\}} \frac{\omega(x) - \omega(y)}{r(x) - \lambda} f_0(y) \chi_{D_i}(x) \pi(dy) \\ &\quad + \int_{\{y:\, u_{i-1} < \omega(y) \leq \omega(x)\}} \frac{\omega(x) - \omega(y)}{r(x) - \lambda} f_0(y) \chi_{D_i}(x) \pi(dy) \\ &= M_{D_i} h_i(x) + \int_{\{y:\, \omega(y) \leq \omega(x)\}} \frac{\omega(x) - \omega(y)}{r(x) - \lambda} M_{D_i} f_0(y) \chi_{D_i}(x) \pi(dy) \\ &= M_{D_i} h_i(x) + M_{D_i} T M_{D_i} f_{0,D_i}, \end{aligned}$$

where $h_i(x) = \sum_{k=1}^{i-1} \int_{D(u_k)} \frac{\omega(x) - \omega(y)}{r(x) - \lambda} f_0(y) \pi(dy)$.

This development shows that (3.2.16) implies:

$$\begin{cases} M_{D_1}TM_{D_1}f_{0,D_1} &= f_{0,D_1} \\ M_{D_2}TM_{D_2}f_{0,D_2} &= f_{0,D_2} - M_{D_2}h_2 \\ \quad\vdots & \\ M_{D_n}TM_{D_n}f_{0,D_n} &= f_{0,D_n} - M_{D_n}h_n. \end{cases} \tag{3.2.17}$$

Now we shall prove that (3.2.17) implies that $r\left(M_{D_i}TM_{D_i}\right) \geq 1$ for at least one $i \in \{1,\ldots,n\}$, where $r\left(M_{D_i}TM_{D_i}\right)$ denotes the spectral radius of $M_{D_i}TM_{D_i}$. To see why, assume that $r\left(M_{D_1}TM_{D_1}\right) < 1$. Therefore $M_{D_1}TM_{D_1}f_{0,D_1} = f_{0,D_1}$ implies that $f_{0,D_1} \equiv 0$ which in turn implies that $h_2(x) \equiv 0$. Note that from the second equation of (3.2.17), $h_2 \equiv 0$ implies that $M_{D_2}TM_{D_2}f_{0,D_2} = f_{0,D_2}$. If in addition, $r\left(M_{D_2}TM_{D_2}\right) < 1$, then since we have $M_{D_2}TM_{D_2}f_{0,D_2} = f_{0,D_2}$, we can assert that $f_{0,D_2} \equiv 0$ which together with $f_{0,D1} \equiv 0$ implies that $h_3 \equiv 0$. Continuing this way, we can see that if $r\left(M_{D_i}TM_{D_i}\right) < 1$ for all $i = 1,\ldots,n$ then $f_{0,D_1} = \ldots = f_{0,D_n} \equiv 0$ which contradicts the fact that f_0 is not the null function on $\{x \in \mathcal{S} : \underline{\omega} \leq \omega(x) < u\}$ as chosen.

Until now, what we have proved is that if $\lambda \notin$ ess-ran(r) is an eigenvalue of K_0 with eigenfunction f_0 then for any $u > \underline{\omega}$ taken sufficiently large such that f_0 is not identically zero on $\{x \in \mathcal{S} : \underline{\omega} \leq \omega(x) < u\}$ and any partition $I_n = (u_0 \leq u_1 \leq \ldots \leq u_n)$ of $[\underline{\omega}, u]$, with $u_0 = \underline{\omega}$ and $u_n = u$, there is at least one $i \in \{1,\ldots,n\}$ such that $r\left(M_{D_i}TM_{D_i}\right) \geq 1$. To show the contradiction promised at the begining, we shall simply show that by taking a partition with a sufficiently small increment, we can make all the $r\left(M_{D_i}TM_{D_i}\right) < 1$.

For $g \in L_i^2(\pi)$ with $\|g\| = 1$, and using the Cauchy-Schwartz inequality, we can write:

$$\begin{aligned} \|M_{D_i}TM_{D_i}g\|^2 &= \int_{D(u_i)} \left\{ \int_{\{y:\omega(y)\leq\omega(x)\}} \frac{\omega(x)-\omega(y)}{r(x)-\lambda} g_{D_i}(y)\pi(dy) \right\}^2 \pi(dx) \\ &\leq \left(\frac{u_i - u_{i-1}}{\kappa}\right)^2 \int_{D(u_i)} g^2(y)\pi(dy) \\ &\leq \left(\frac{u_i - u_{i-1}}{\kappa}\right)^2, \end{aligned}$$

and therefore $\|M_{D_i}TM_{D_i}\| \leq (u_i - u_{i-1})/\kappa$. From this, by taking a partition $I_n = (u_0 \leq u_1 \leq \ldots \leq u_n)$ with $\max_{1\leq i\leq n}(u_i - u_{i-1}) < \kappa$, we have for $i = 1, \ldots, n$:

$$\begin{aligned} r\left(M_{D_i}TM_{D_i}\right) &= \lim_{n\to\infty} \|(M_{D_i}TM_{D_i})^n\|^{\frac{1}{n}} \\ &\leq \|M_{D_i}TM_{D_i}\| \\ &\leq \frac{u_i - u_{i-1}}{\kappa} \\ &< 1. \end{aligned}$$

Next, take $\lambda \in$ ess-ran(r) such that $\pi\{y : r(y) = \lambda\} = 0$. Then $\lambda \in \sigma_{ess}(M_r) = \sigma_{ess}(K_0) \subseteq \sigma(K_0)$. This shows that if for all $\lambda \in$ ess-ran(r), $\pi\{y : r(y) = \lambda\} = 0$ then $\sigma(K_0) =$ ess-ran(r).

□

Remark 3.2.1. *(1) A typical case for Theorem 3.2.2 will be that $\mathcal{S} = \mathbb{R}^d$ and π and Q are absolutely continuous with respect to the Lebesgue measure: $\pi(dx) = \pi(x)dx$ and $Q(dx) = q(x)dx$. If in addition, $\pi(x)$ and $Q(x)$ are positive and continuous, then $r(x)$ is also continuous and $\sigma(K_0) = [0, 1 - \text{ess-inf}(\omega)]$.*

(2) One well known consequence of Theorem 3.2.2 is that it is not possible to take advantage of the correlation of the Markov chain to decrease the asymptotic variance of the Monte Carlo estimate. Suppose that (X_n) is a geometrically ergodic Markov chain with stationary distribution π, generated using the IMH algorithm. For any $f \in L_0^2(\pi)$, writting e_f to denote the spectral measure of K_0 on f, $\rho := 1 - \text{ess-inf}(\omega)$ and assuming that $X_0 \sim \pi$, we have:
$\lim_{n\to\infty} \frac{1}{n} E\left(\sum_{k=0}^{n-1} f(X_k)\right)^2 = \int_0^\rho \frac{1+\lambda}{1-\lambda} e_f(d\lambda) \geq \|f\|^2$.

3.3. The Restricted Metropolis Algorithm.

In practice, the RWMH algorithm remains one of the most used MH algorithm. But for many common distributions, the RWMH algorithm will fail to converge at a geometric rate. In fact, as we mentioned in the introduction, exponential or lighter tails is necessary for the RWMH algorithm to be geometrically ergodic (Jarner and Hansen (2000), Jarner and Tweedie (2001)). Here, we propose a MH

algorithm that behaves as the RWMH on a compact set. We show in theorem 3.3.1 that the proposed algorithm is geometrically ergodic whenever the target density's tails decay at least as fast as the proposal density's tails. In this section, we take $\mathcal{S} = \mathbb{R}^d$ and assume that π is absolutely continuous with respect to the Lebesgue measure and we write: $\pi(dx) = \pi(x)dx$. Moreover, we also assume that the function π is positive and bounded on compact sets. Let $Q(x, dy) = q(|y - x|)dy$ be a proposal kernel. We assume that q is positive, continuous and $q(r)$ is nonincreasing for $r > 0$ sufficiently large. Let Δ be a compact, convex and non empty subset of $\mathcal{S}$. We define $p : \mathcal{S} \to \mathcal{S}$ by $p(x) = argmin_{y\in\Delta}\{|y - x|\}$ the projection on Δ. We propose the following Metropolis-Hastings move that we call restricted RWMH:

Algorithm 3.3.1. *(1) Start the chain at any point $x \in \mathcal{S}$: $X_0 = x$.*

(2) Conditional on X_i, generate $Y \sim Q(p(X_i), \cdot)$ and generate $U \sim \mathcal{U}(0, 1)$.

(3) If $U \leq \alpha(X_i, Y)$ then set $X_{i+1} = Y$, otherwise set $X_{i+1} = X_i$, where $\alpha(x, y) = \min\left(1, \frac{\pi(y)q(|x-p(y)|)}{\pi(x)q(|y-p(x)|)}\right)$.

On Δ, this algorithm behaves like the RWMH algorithm with kernel $Q(x, \cdot)$. For $x \notin \Delta$, the proposal is taken from $Q(p(x), \cdot)$ where $p(x)$ is the projection of x on Δ. Typically, if x_0 is any position parameter of π, one can take $\Delta = B(x_0, R)$, the closed ball centered at x_0 with radius $R > 0$ taken sufficiently large for Δ to include all the essential features of π. In this case, $p(x) = x_0 + (\min(1, R/|x - x_0|))(x - x_0)$ is easy to compute.

Clearly, the transition density becomes:

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

where

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

and

$$r(x) = 1 - \int \alpha(x, y)q(|y - p(x)|)dy. \tag{3.3.1}$$

Theorem 3.3.1. *Suppose that:*

$$\limsup_{|x|\to\infty} \frac{\pi(x)}{q(|x|)} < \infty. \tag{3.3.2}$$

Then the Restricted RWMH algorithm (3.3.1) is geometrically ergodic.

PROOF. From the conditions satisfied by π and q, it follows that for any compact set $A \in \mathcal{F}$, there is $\varepsilon > 0$ such that $\omega(x,y) = \frac{q(|y-p(x)|)}{\pi(y)} > \varepsilon$ for all $x, y \in A$. Therefore from Proposition 3.2.1 the restricted RWMH is ergodic.

To finish the proof, it remains to show that the compact condition in Assumption (3.2.1) is satisfied and next prove that r, the rejection probability as given by (3.3.8) is bounded away from 1.

Showing the compact condition amounts to show that
$I := \int dx \int q(|y-p(x)|)q(|x-p(y)|)dy < \infty$.
Take $M > 0$ such that q is nonincreasing on $\{r > M\}$, and write $\tau := \sup\{|z|\,,\ z \in \Delta\} < \infty$. Then for $|y| > M + \tau$ and for $x \in \mathcal{S}$, $|y - p(x)| \geq |y - p(y)| \geq |y| - \tau > M$. This implies that for $x \in \mathcal{S}$, and for $|y| > \tau + M$, $q(|y-p(x)|) \leq q(|y-\tau|)$.

By writing $A := \int_{\{|x|\leq\tau+M\}} dx \int_{\{|y|\leq\tau+M\}} q(|y-p(x)|)q(|x-p(y)|)dy < \infty$, we therefore have:

$$\begin{aligned}
I &= \int_{\{|x|\leq\tau+M\}} dx \int_{\{|y|\leq\tau+M\}} q(|y-p(x)|)q(|x-p(y)|)dy \\
&\quad + \int_{\{|x|\leq\tau+M\}} dx \int_{\{|y|>\tau+M\}} q(|y-p(x)|)q(|x-p(y)|)dy \\
&\quad + \int_{\{|x|>\tau+M\}} dx \int_{\{|y|\leq\tau+M\}} q(|y-p(x)|)q(|x-p(y)|)dy \\
&\quad + \int_{\{|x|>\tau+M\}} dx \int_{\{|y|>\tau+M\}} q(|y-p(x)|)q(|x-p(y)|)dy \\
&\leq A + \int_{\{|x|>\tau+M\}} q(|x-\tau|)dx \left(\int_{\{|x|\leq\tau+M\}} q(|x-p(y)|)dx + \right. \\
&\quad \left. \int_{\{|y|\leq\tau+M\}} q(|y-p(x)|)dy \right) + \left(\int_{\{|y|>\tau+M\}} q(|x-\tau|)dx \right)^2 \\
&< \infty.
\end{aligned}$$

Using condition (3.3.2) and the same arguments as above, we can see that:

$$\inf_{z\in\Delta, x\in\mathcal{S}} \frac{q(|x-z|)}{\pi(x)} = \eta > 0,$$

and

$$\sup_{x\in\mathcal{S}} \frac{q(|x-p(x)|)}{\pi(x)} = \eta_0 < \infty.$$

Therefore we have:

$$\begin{aligned}\alpha(x,y) &= \min\left(1, \frac{\pi(y)q(|x-p(y)|)}{\pi(x)q(|y-p(x)|)}\right)\\ &\geq \min\left(1, \frac{q(|x-p(y)|)/\pi(x)}{q(|y-p(y)|)/\pi(y)}\right)\\ &\geq \frac{\eta}{\eta_0},\end{aligned}$$

for $|y| > \tau + M$. It follows that:

$$\begin{aligned}\int \alpha(x,y)q(|y-p(x)|)dy &\geq \frac{\eta}{\eta_0}\int_{\{|y|>\tau+M\}} q(|y-p(x)|)dy\\ &\geq \frac{\eta^2}{\eta_0}\pi\left(\{|y|>\tau+M\}\right).\end{aligned}$$

This implies that $r(x) \leq 1-\frac{\eta^2}{\eta_0}\pi\left(\{|y|>\tau+M\}\right) < 1$. And the theorem is proved. □

Remark 3.3.1. *With condition (3.3.2) of Theorem 3.3.1, the IMH algorithm with a proposal distribution q is also geometrically ergodic. But depending on the features of the density π, the mixing rate of the IMH chain may be slow. The restricted version of the RWMH algorithm retains the geometric ergodicity of the IMH but will still explore Δ typically like the RWMH algorithm. Actually, the convergence rate of the Restricted RWMH is much more similar to that of the IMH than what theorem (3.3.1) can reveal. In the next theorem, we show that the restricted version of RWMH algorithm is uniformly geometrically ergodic.*

For $x_0 \in \Delta$, write as above $\omega(x_0,x) = \frac{q(|x-x_0|)}{\pi(x)}$ and note:

$$\delta(x_0) := \inf_{y\in\mathbb{R}^d} \omega(x_0,y), \tag{3.3.3}$$

$$\delta := \sup_{x_0\in\Delta} \delta(x_0).$$

Also write:

$$\theta(x_0) := \inf_{x,y\in\mathbb{R}^d} \frac{q(|y-p(x)|)}{q(|y-x_0|)} > 0, \tag{3.3.4}$$

$$\tau(x_0) := \sup_{x,y\in\mathbb{R}^d} \frac{q(|y-p(x)|)}{q(|y-x_0|)} < \infty, \tag{3.3.5}$$

and

$$\theta := \sup_{x_0\in\Delta} \theta(x_0) > 0.$$

Note that on Δ, all the $\delta(x_0)$ are positive together or are all equal to zero together.

Theorem 3.3.2. **(i):** *If $\delta > 0$, then*

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq (1-\theta\delta)^n\,, \tag{3.3.6}$$

for every $x \in \mathbb{R}^d$.

(ii): *Assume that $\delta = 0$. Then P cannot converge at a geometric rate.*

PROOF. **(i):** Because $\delta > 0$, $\delta(x_0) > 0$ for all $x_0 \in \Delta$. Let $x_0 \in \Delta$ be fixed. The transition kernel of the RM algorithm is:

$$P(x,dy) = \alpha(x,y)q(|y-p(x)|)dy + r(x)\mathbb{1}_{\{x\}}(dy), \tag{3.3.7}$$

where

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

and

$$r(x) = 1 - \int \alpha(x,y)q(|y-p(x)|)dy. \tag{3.3.8}$$

If $\alpha(x,y) = 1$, $P(x,dy) \geq \frac{q(|y-p(x)|)}{\pi(y)}\pi(dy) \geq \theta(x_0)\delta(x_0)\pi(dy)$. In the same way, if $\alpha(x,y) = \frac{\pi(y)q(|x-p(y)|)}{\pi(x)q(|y-p(x)|)}$, then $P(x,dy) \geq \frac{q(|x-p(y)|)}{\pi(x)}\pi(dy) \geq \theta(x_0)\delta(x_0)\pi(dy)$.

In any case, we have $P(x,dy) \geq \theta(x_0)\delta(x_0)\pi(dy)$ for $x \in \mathbb{R}^d$, meaning that the whole space is a small set for P. Therefore following from Meyn and Tweedie (1993) theorem 16.1., we have:

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq (1-\theta(x_0)\delta(x_0))^n\,,$$

for all $x \in \mathbb{R}^d$ which implies that:

$$\begin{aligned}\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} &\leq \inf_{\substack{x_0\in\Delta \\ \delta(x_0)>0}} (1-\theta(x_0)\delta(x_0))^n \\ &\leq (1-\theta\delta)^n\,.\end{aligned}$$

(ii): Since $\delta = 0$, $\delta(x_0) = 0$ for all $x_0 \in \Delta$. We shall show that if for some $x_0 \in \Delta$, we have $\delta(x_0) = 0$, then ess-sup $r(x)$, the probability of rejection of the chain is 1. Then proposition 5.1 of Roberts and Tweedie (1996) helps us to conclude that P cannot be geometrically ergodic. We do this very much like Mengersen and Tweedie (1996), theorem 2.1.

We fix $x_0 \in \Delta$ such that $\delta(x_0) = 0$ and we write $D_n = \left\{x \in \mathbb{R}^d : \omega(x_0, x) \leq \frac{1}{n}\right\}$. Since $\delta(x_0) = 0$, $\pi(D_n) > 0$ for all $n \geq 1$. For $x \in \mathbb{R}^d$, we note $A_x = \left\{y \in \mathbb{R}^d : \frac{\pi(y)q(|x-p(y)|)}{\pi(x)q(|y-p(x)|)} \geq 1\right\}$ and $R_x = \left\{y \in \mathbb{R}^d : \frac{\pi(y)q(|x-p(y)|)}{\pi(x)q(|y-p(x)|)} < 1\right\}$. We have:

$$\begin{aligned} r(x) &= \int_{R_x} (1 - \alpha(x,y))q(|y - p(x)|)dy \\ &= 1 - \int_{A_x} q(|y - p(x)|)dy - \int_{R_x} \alpha(x,y)q(|y - p(x)|)dy. \end{aligned} \tag{3.3.9}$$

Take $y \in A_x$, we have: $1 \leq \frac{\pi(y)q(|x-p(y)|)}{\pi(x)q(|y-p(x)|)} \leq \frac{\omega(x_0,x)}{\omega(x_0,y)}\frac{\theta(x_0)}{\tau(x_0)}$. Therefore, for $x \in D_n$ and $y \in A_x$, $\omega(x_0, y) \leq \frac{\theta(x_0)}{\tau(x_0)}\frac{1}{n}$. This implies that

$$\int_{A_x} q(|y - p(x)|)dy \leq \int_{A_x} \tau(x_0)\omega(x_0, y)\pi(dy) \leq \theta(x_0)\frac{1}{n}, \tag{3.3.10}$$

for all $x \in D_n$. In the same way, for $x \in D_n$ and $y \in R_x$, we have:

$$\begin{aligned} \int_{R_x} \alpha(x,y)q(|y - p(x)|)dy &= \int_{R_x} \frac{\pi(y)q(|x - p(y)|)}{\pi(x)}dy \\ &\leq \omega(x_0, x)\tau(x_0) \\ &\leq \tau(x_0)\frac{1}{n}. \end{aligned} \tag{3.3.11}$$

Equations (3.3.10) and (3.3.11) in (3.3.9) gives:

$$r(x) \geq 1 - (\theta(x_0) + \tau(x_0))\frac{1}{n},$$

for all $x \in D_n$, $n \geq 1$. Since $\pi(D_n) > 0$ for all $n \geq 1$, it follows that ess-sup $r(x)$ (with respect to π) is equal to 1.

□

Example 3.3.1. *In this simulation study, we illustrate the benefits of restricting the RWMH in terms of the existence of the Central Limit Theorem. We say that the Central Limit Theorem holds for a function h if there exists $0 < \sigma_h^2 < \infty$ such that:*

$$\sqrt{n}\left(S_n(h) - \pi(h)\right) \overset{w}{\Rightarrow} N(0, \sigma_h^2),$$

where $S_n(h) = \frac{1}{n}\sum_{i=1}^n h(X_i)$. It is well known that for a reversible geometrically ergodic Markov chain, the Central Limit Theorem holds for any $h \in L^2(\pi)$, see Roberts and Rosenthal (1997) for example.

Let $t(\nu)$ be the student distribution on $\mathbb{R}$ with ν degree of freedom. Clearly, from Theorem 2.2 of Jarner and Tweedie (2001), the RWMH with target density $t(\nu)$ is not geometrically ergodic. But using a proposal distributed as $t(r)$ with $r \leq \nu$, we know from Theorem 3.3.1 that the restricted version of the RWMH presented above is geometrically ergodic. Thus the Central Limit Theorem holds for any h with $\pi(h^2) < \infty$.

We use $h(x) = |x|$, $r = 0.5$ and $\nu = 3, 4$. For each of these 4 combinaisons ($\nu = 3, 4$ times ordinary and restricted RWMH), we ran $N = 1,000$ independent chains starting at 0 each with length $n = 1,000,000$. For the restricted versions, we have used $\Delta = B(0, R)$ where R is choosen such that $\Pr(|t(\nu)| < R) = 0.9$. For each of these 4 combinaisons, the $i-th$ run is used to estimate $\pi(|x|)$ by $S^i(|X|) = \frac{1}{n}\sum_{k=1}^{n}\left|X_k^{(i)}\right|$. Figures 1 and 2 show the QQplot versus the normal distribution and the histogram of the normalised empirical sum $S^i(|X|) - \frac{1}{N}\sum_{i=1}^{N} S^i(|X|)$.

For $\nu = 3$, as expected, the Central Limit Theorem seems to hold for $h(x) = |x|$ in the case of the restricted version of the RWMH algorithm, while it clearly fails for the ordinary RWMH. For $\nu = 4$, the central limit theorem seems to hold for both versions. This agrees with Theorem 3.3.1 and theoretical results obtained by Jarner and Tweedie (2001) on the RWMH algorithm.

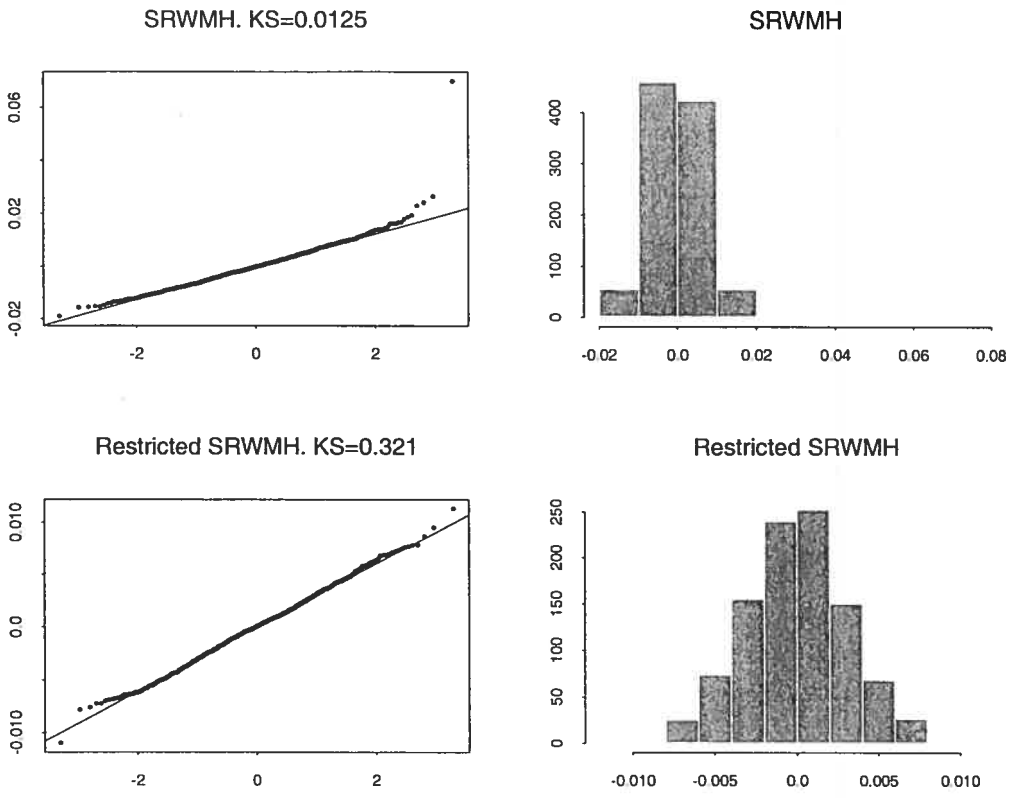


Figure 1: QQ-plots and histograms of 1,000 normalised ergodic averages of the function $h(x) = |x|$. Target density $t(3)$, proposal density $t(0.5)$. The KS value is the p-value of the Kolmogorov-Smirnov test of normality.

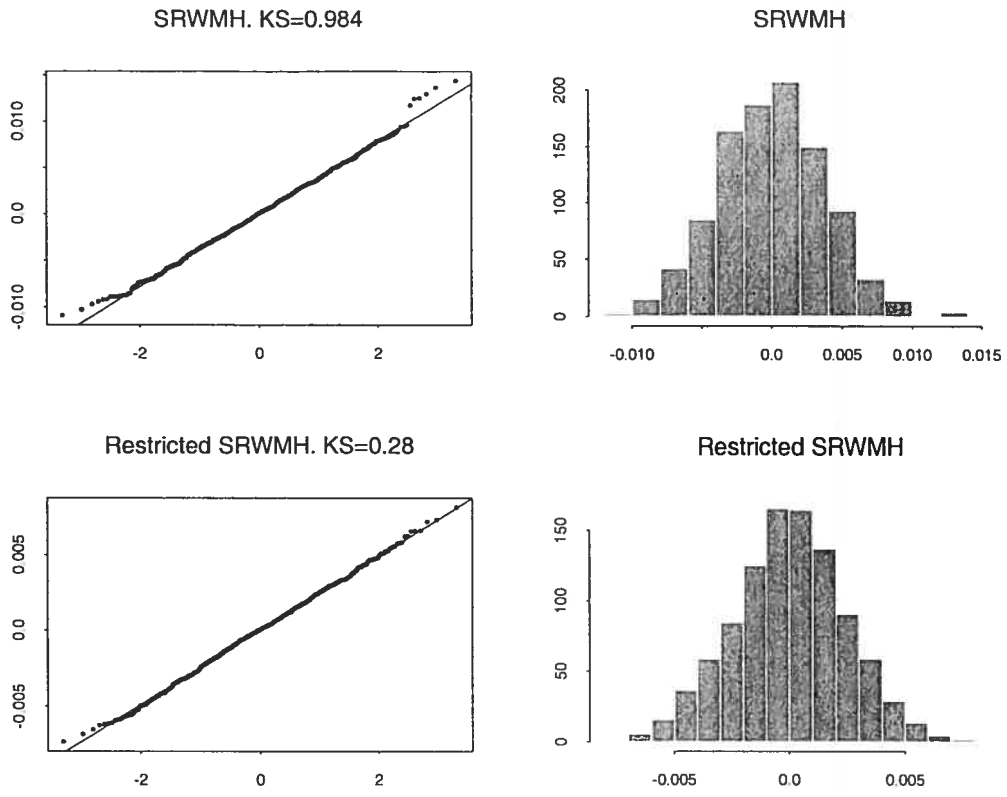


Figure 2: QQ-plots and histograms of 1,000 normalised ergodic averages of the function $h(x) = |x|$. Target density $t(4)$, proposal density $t(0.5)$. The KS value is the p-value of the Kolmogorov-Smirnov test of normality.

From a practical point of view, there are many problems remaining to be solved with this restricted version of the RWMH algorithm. In our sense, the most important is how to choose the set Δ. In many situations there may not be enough information available on π to make a good choice for Δ. One natural possibility is to use some adaptive schemes to update Δ during the simulation. See for example Gilks et al. (1998), Haario et al. (2000). The regenerative framework for adaption proposed by Gilks et al. (1998) seems particularly interesting in this respect as we suspect the restricted RWMH algorithm to have more frequent regenerations than the RWMH algorithm. But more work need to be done in this direction.

Aknowledgements: The authors are grateful to Professor Richard Duncan for helpful discussions on Hilbert spaces operators theory.

1087–1091.

Meyn, S. P., Tweedie, R. L., 1993. Markov chains and stochastic stability. Springer-Verlag London Ltd., London.

Roberts, G. O., Rosenthal, J. S., 1997. Geometric ergodicity and hybrid Markov chains. Electron. Comm. Probab. 2, no. 2, 13–25 (electronic).

Roberts, G. O., Rosenthal, J. S., 1998. Markov-chain Monte Carlo: some practical implications of theoretical results. Canad. J. Statist. 26 (1), 5–31, with discussion by Hemant Ishwaran and Neal Madras and a rejoinder by the authors.

Roberts, G. O., Tweedie, R. L., 1996. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika 83 (1), 95–110.

Roberts, G. O., Tweedie, R. L., 2001. Geometric l^2 and l^1 convergence are equivalent for reversible markov chains. J. Appl. Prob. 38 A, 37–41.

Smith, R., Tierney, L., 1996. Exact transition probabilities for the independence metropolis sampler. Research Report, Statistical Laboratory University of Cambrige 16.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22 (4), 1701–1762, with discussion and a rejoinder by the author.

CONCLUSION

Dans cette thèse, nous avons tenté d'apporter des contributions aux méthodes de réduction de variance des algorithmes de Monte Carlo et à la vitesse de convergence de l'algorithme de Metropolis-Hastings.

Dans le cadre des méthodes de Monte Carlo statiques, nous avons développé un nouvel estimateur par variable de contrôle sans biais qui ne sacrifie pas à la variance. Nous avons également donné un exemple d'utilisation de l'algorithme dans le cas de l'algorithme de Rejet et la méthode "Importance-Sampling".

Il existe peu de résultats rigoureux sur la réduction de variance pour les algorithmes MCMC, bien que plusieurs idées pratiques aient été avancées dans cette direction. La difficulté à manipuler des observations dépendantes et les expressions mathématiques difficiles à utiliser auxquelles on aboutit expliquent principalement cette situation. Nous avons essayé de combler ce vide en proposant deux nouvelles idées de réduction de variance pour l'algorithme de Metropolis-Hastings Indépendant. Nous avons montré que la méthode de réduction de variance par variable de contrôle peut également s'appliquer à cet algorithme tout en restant simple à implémenter. Et nous avons développé un autre estimateur qui utilise le manque de symétrie de l'algorithme par rapport aux observations proposées.

Nos contributions ont ensuite portées sur la vitesse de convergence de l'algorithme de Metropolis-Hastings. Nous avons développé une condition nécessaire et suffisante pour l'ergodicité géometrique de cet algorithme. Notre approche utilise principalement l'analyse fonctionnelle et la théorie des opérateurs auto-adjoints dans les espaces de Hilbert. Nous avons résolu la conjecture de Jun S. Liu sur le spectre de l'operateur associé à l'algorithme de Metropolis-Hastings Indépendant

en montrant que ledit spectre est l'image essentielle de la fonction de probabilité de rejet de la chaîne. Finalement, nous avons proposé un nouvel algorithme de Metropolis que nous avons nommé algorithme de Metropolis Restreint. Nous montrons que notre algorithme combine les aspects intéressants des algorithmes de Metropolis Indépendant et de Metropolis Marche Aléatoire.