Université de Montréal

Convergence d'un algorithme de type Metropolis pour une distribution cible bimodale

par

Michaël Lalancette

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

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

14 juillet 2017

SOMMAIRE

Nous présentons dans ce mémoire un nouvel algorithme de type Metropolis-Hastings dans lequel la distribution instrumentale a été conçue pour l'estimation de distributions cibles bimodales. En fait, cet algorithme peut être vu comme une modification de l'algorithme Metropolis de type marche aléatoire habituel auquel on ajoute quelques incréments de grande envergure à des moments aléatoires à travers la simulation. Le but de ces grands incréments est de quitter le mode de la distribution cible où l'on se trouve et de trouver l'autre mode.

Par la suite, nous présentons puis démontrons un résultat de convergence faible qui nous assure que, lorsque la dimension de la distribution cible croît vers l'infini, la chaîne de Markov engendrée par l'algorithme converge vers un certain processus stochastique qui est continu presque partout. L'idée est similaire à ce qui a été fait par Roberts et al. (1997), mais la technique utilisée pour la démonstration des résultats est basée sur ce qui a été fait par Bédard (2006).

Nous proposons enfin une stratégie pour trouver la paramétrisation optimale de notre nouvel algorithme afin de maximiser la vitesse d'exploration locale des modes d'une distribution cible donnée tout en estimant bien la pondération relative de chaque mode. Tel que dans l'approche traditionnellement utilisée pour ce genre d'analyse, notre stratégie passe par l'optimisation de la vitesse d'exploration du processus limite.

Finalement, nous présentons des exemples numériques d'implémentation de l'algorithme sur certaines distributions cibles, dont une ne respecte pas les conditions du résultat théorique présenté.

Mots-Clés : MCMC, algorithme Metropolis de type marche aléatoire, distribution bimodale, convergence faible, échelonnage optimal.

SUMMARY

In this thesis, we present a new Metropolis-Hastings algorithm whose proposal distribution has been designed to successfully estimate bimodal target distributions. This sampler may be seen as a variant of the usual random walk Metropolis sampler in which we propose large candidate steps at random times. The goal of these large candidate steps is to leave the actual mode of the target distribution in order to find the second one.

We then state and prove a weak convergence result stipulating that if we let the dimension of the target distribution increase to infinity, the Markov chain yielded by the algorithm converges to a certain stochastic process that is almost everywhere continuous. The theoretical result is in the flavour of Roberts et al. (1997), while the method of proof is similar to that found in Bédard (2006).

We propose a strategy for optimally parameterizing our new sampler. This strategy aims at optimizing local exploration of the target modes, while correctly estimating the relative weight of each mode. As is traditionally done in the statistical literature, our approach consists of optimizing the limiting process rather than the finite-dimensional Markov chain.

Finally, we illustrate our method via numerical examples on some target distributions, one of which violates the regularity conditions of the theoretical result.

Keywords : MCMC, random walk Metropolis algorithm, bimodal distribution, weak convergence, optimal scaling.

TABLE DES MATIÈRES

Sommaire iii

Summary v

Liste des figures xi

Remerciements xiii

Introduction 1

Chapitre 1. L'algorithme Metropolis-Hastings de type marche aléatoire 5

1.1. Les méthodes de Monte-Carlo 5

1.2. Chaînes de Markov à espace d'états continu 6

1.3. Construction et implémentation de l'algorithme RWM 10

1.3.1. Définition 10

1.3.2. Justification 11

1.3.3. Premiers exemples 14

1.4. Un nouvel algorithme 18

1.4.1. La distribution cible 19

1.4.2. La distribution instrumentale 19

1.4.3. Retour sur l'exemple 21

Chapitre 2. Optimisation d'algorithmes RWM 23

2.1. Plusieurs critères d'optimalité 23

2.2. Optimisation de l'algorithme RWM classique 25

2.2.1. Le cas iid 26

2.2.2. Le cas presque iid 28

2.3. Convergence de la chaîne engendrée par le nouvel algorithme 28

2.3.1. Légère reformulation de l'algorithme 28

2.3.2. Convergence faible 29

2.4. Optimisation du nouvel algorithme 32

2.4.1. La stratégie proposée 32

2.4.1.1. Le choix du paramètre d'échelle pour les petits pas 32

2.4.1.2. Le choix de la densité instrumentale pour les grands pas 34

2.4.1.3. Le choix de la probabilité de proposer un grand pas 37

2.4.2. Un critère plus objectif 39

Chapitre 3. Démonstration des théorèmes 2.3.1 et 2.3.2 41

3.1. Générateurs 41

3.2. Résultats préliminaires 45

3.3. Le comportement asymptotique lors des petits pas 52

3.4. Le processus limite formé par la première composante de l'algorithme 61

3.5. Le processus limite formé par la seconde composante de l'algorithme 66

Chapitre 4. Simulations 69

4.1. Retour sur l'exemple normal bimodal 70

4.1.1. Choix des paramètres 70

4.1.2. Résultats 73

4.2. Une distribution cible avec structure de dépendance 73

4.2.1. Choix des paramètres 75

4.2.2. Résultats 76

4.2.3. Comparaison avec les résultats obtenus sur le modèle de base 76

4.3. Une distribution cible presque unimodale 78

4.3.1. Choix des paramètres 79

4.3.2. Résulats obtenus avec l'algorithme classique 80

4.3.3. Résultats obtenus avec le nouvel algorithme 80

4.4. Effet de la distance entre les modes 82

Conclusion 85

Bibliographie 87

Annexe A. Lemmes utilisés dans le chapitre 3 ..... A-i

Annexe B. Codes R ..... B-i

B.1. Code pour implémenter le nouvel algorithme ..... B-i

B.2. Code pour l'étude de la distance entre les modes ..... B-vi

LISTE DES FIGURES

1.1 Estimation de la densité d'une $N(15; 9)$ à l'aide de l'algorithme RWM. 15

1.2 Estimation de la densité bimodale correspondant à la distribution décrite en (1.3.3) à l'aide de l'algorithme RWM avec $\sigma = 2{,}5$. 16

1.3 Trace de l'algorithme RWM avec $\sigma = 2{,}5$. 16

1.4 Trace de l'algorithme RWM avec $\sigma = 8$. 17

1.5 Estimation de la densité bimodale correspondant à la distribution décrite en (1.3.3) à l'aide de l'algorithme RWM avec $\sigma = 8$. 18

1.6 Trace du nouvel algorithme. 21

1.7 Estimation de la densité bimodale correspondant à la distribution décrite en (1.3.3) à l'aide du nouvel algorithme. 22

4.1 Taux d'acceptation des petits pas proposés en fonction du paramètre d'échelle l. 71

4.2 Probabilité estimée de changement de mode lorsqu'un grand pas est proposé selon une loi uniforme en fonction du paramètre d'échelle c. 72

4.3 Probabilité estimée de changement de mode lorsqu'un grand pas est proposé selon une loi normale en fonction du paramètre d'échelle σ. 72

4.4 Trace de la chaîne engendrée par la première composante de l'algorithme. . . . 74

4.5 Estimation de la densité cible f_1. 74

4.6 Taux d'acceptation des petits pas proposés en fonction du paramètre d'échelle l. 75

4.7 Trace de la chaîne engendrée par la première composante de l'algorithme. . . . 76

4.8 Estimation de la densité cible f_1. 77

4.9 Densité de la première composante de $\mathbf{X}$. 79

4.10 Trace de la chaîne engendrée par la première composante (algorithme RWM). 80

4.11 Estimation de la densité cible f_1 (algorithme RWM). 81

4.12 Trace de la chaîne engendrée par la première composante (nouvel algorithme). 81

4.13 Estimation de la densité cible f_1 (nouvel algorithme). 82

4.14 Valeur de c_{opt} en fonction de μ_1. 83

4.15 Valeur de p en fonction de μ_1. 84

REMERCIEMENTS

Je tiens en tout premier lieu à remercier ma superviseure, la professeur Mylène Bédard. Le soutien et la guidance qu'elle m'a offerts tout au long de mon parcours de deuxième cycle furent indispensables à ma réussite. Elle a toujours trouvé les mots et le ton pour me motiver et m'encourager, sans toutefois me mettre sous pression. Son calme et sa bonne humeur éternels ont su me réconforter chaque fois que j'entrais dans son bureau avec des résultats que je jugeais insuffisants. De plus, elle m'a accordé sa confiance dès le départ et elle m'a toujours laissé une grande liberté dans l'orientation de notre recherche. J'espère que ces deux années de travail commun ne sont pas les dernières.

J'aimerais aussi adresser mes remerciements à tous mes amis et collègues étudiants qui rendent l'ambiance au Département de Mathématiques et de Statistique si colorée et accueillante. Essayer de tous les nommer ici ne serait qu'une insulte puisque certains noms seraient probablement oubliés. Leur présence et leur écoute m'ont aidé à résoudre ou à oublier les problèmes rencontrés et à toujours quitter l'université avec le sourire aux lèvres.

Merci également au personnel administratif et à l'équipe de soutien informatique du Département qui nous rendent la vie plus facile par leur dévouement et leur professionnalisme. Merci aux professeurs et chargés de cours pour un enseignement passionnant ainsi que pour de nombreuses discussions enrichissantes.

Merci à Philippe, dont j'ai la chance de partager les intérêts de recherche. En étant toujours si enclin à transmettre ses connaissances sur les méthodes MCMC, et peut-être sans s'en rendre compte, il a plusieurs fois influencé le contenu de ce mémoire.

Enfin, je remercie les membres du jury qui ont amélioré la qualité de ce mémoire et qui m'ont surtout fait énormément réfléchir par leur lecture rigoureuse et par leurs nombreux commentaires.

La dernière victime de mes remerciements, mais non la moindre, est Emy. Elle sait supporter, souvent mieux que moi-même, mes propres moments difficiles. Je n'ose pas imaginer ce que je serais devenu sans son support infini. Je lui suis très reconnaissant de m'accompagner dans toutes mes aventures, que ce soit à La Baie, à Montréal, ou à Toronto.

INTRODUCTION

En sciences pures, en génie et particulièrement en statistique, on est souvent amené à approximer numériquement une espérance (qu'on peut écrire sous la forme d'une intégrale). L'analyse numérique nous offre plusieurs outils pour obtenir de telles approximations, mais le temps de calcul requis peut être démesuré lorsque la dimension de la fonction à intégrer est grande. En statistique, des exemples de situations nécessitant l'intégration d'une fonction de grande dimension sont l'étude d'un grand nombre de variables réponses de façon simultanée, ou encore l'inférence sur plusieurs paramètres de façon simultanée en statistique bayésienne. La solution alors envisagée est souvent d'utiliser une méthode de Monte-Carlo, qui se résume à échantillonner directement de la distribution de la variable aléatoire (ou du vecteur aléatoire) dont on cherche à calculer l'espérance. Il suffit par la suite d'estimer cette espérance à l'aide de la moyenne échantillonnale.

Les distributions d'intérêt étant parfois très complexes, il n'est pas toujours possible d'échantillonner directement à partir de ces distributions. Dans de telles situations, nous pouvons contourner le problème en échantillonnant à partir de distributions instrumentales plus simples que la distribution cible, puis en acceptant par la suite les bons candidats et en laissant de côté les moins bons. C'est en fait le principe des méthodes de Monte-Carlo par chaînes de Markov (MCMC). Ces algorithmes soumettent les candidats générés à une étape d'acceptation ou de rejet, qui s'assure que l'échantillon formé pourra être éventuellement considéré comme un échantillon provenant de la distribution d'intérêt. De manière spécifique, les méthodes MCMC sont donc des algorithmes qui permettent de générer une chaîne de Markov qui admet la distribution cible comme distribution stationnaire. Ces méthodes tirent leur origine de [**14**], mais ont été introduites dans la communauté statistique par [**11**]. Un très grand nombre d'algorithmes MCMC ont été proposés à ce jour, et l'un des plus utilisés est sans aucun doute l'algorithme de Metropolis-Hastings de type marche aléatoire (RWM). Cet algorithme est particulièrement simple à utiliser puisqu'il ne requiert que la connaissance de la fonction de densité (ou de probabilité) associée à la distribution cible, à une constante multiplicative près. Il nécessite par la suite de savoir générer une marche

aléatoire sur le support de cette distribution cible ; cet algorithme sera décrit de manière formelle dans le premier chapitre. De plus, nous savons comment optimiser l'exploration de l'espace d'états par l'algorithme RWM pour certaines classes de distributions cibles (voir par exemple [2], [3], [20] et [21]). Par exemple, dans [20], les auteurs considèrent une distribution cible formée de d composantes indépendantes et identiquement distribuées (iid) et obtiennent un résultat théorique permettant de trouver la variance instrumentale optimale de l'algorithme RWM lorsqu'utilisé avec une distribution instrumentale gaussienne. Or, il existe des distributions cibles pour lesquelles cet algorithme, même lorsqu'optimisé, échoue à explorer entièrement l'espace d'états et par conséquent à produire un échantillon ayant les propriétés recherchées. Ceci signifie que certaines régions possédant une probabilité positive demeureront inexplorées et seront considérées comme des régions de probabilité nulle dans l'échantillon généré. Les distributions bimodales, qui émergent fréquemment en statistique bayésienne (voir par exemple [9]), souffrent fréquemment de ce problème.

Dans ce mémoire, nous présentons donc un nouvel algorithme RWM fortement inspiré de l'algorithme RWM classique qui se spécialise dans l'estimation d'une certaine classe de distributions bimodales. Spécifiquement, nous considérons une distribution cible de dimension d dont la première composante possède une densité distincte des $d-1$ autres composantes, qui sont elles-mêmes iid. Ce contexte nous donne la possibilité d'étudier des distributions cibles bimodales de grande dimension, ce qui est nécessaire pour la dérivation subséquente de résultats théoriques d'échelonnage optimal (ces résultats sont généralement obtenus dans un contexte asymptotique où d croît vers l'infini). L'idée derrière la nouvelle méthode consiste à proposer des candidats locaux, qui permettront de bien explorer chacun des modes séparément, et de parfois proposer des candidats globaux, qui visent à étudier la densité cible dans son ensemble, et qui permettront donc de sauter d'un mode à un autre. Cette dernière particularité est cruciale dans l'exploration de densités cibles bimodales, particulièrement lorsqu'il existe une région de très faible densité entre les modes. De par son design, il est facile de déduire que cet algorithme explorera son espace de manière plus efficace que l'algorithme RWM et qu'il offrira un échantillon beaucoup plus représentatif de la distribution cible. Afin de performer de manière optimale, son implémentation requiert cependant l'ajustement de davantage de paramètres, une facette que nous étudierons également dans ce mémoire.

Il importe de remarquer que certaines méthodes ont été proposées afin d'échantillonner à partir de distributions très étendues ou multimodales. Par exemple, [7] propose l'algorithme RAPT, une méthode qui consiste à séparer l'espace en différentes régions et à utiliser des distributions instrumentales distinctes dans chacune de ces régions. Notons que cette méthode est adaptative, ce qui signifie que les distributions instrumentales sont améliorées au fil des itérations en utilisant toute l'information de la chaîne disponible à ce moment. Cet

algorithme a éventuellement été amélioré par [1] pour devenir l'algorithme RAPTOR, dans lequel les frontières des différentes régions sont également mises à jour de manière adaptative. Une autre approche est celle du «tempering», une idée proposée par [13] qui consiste à «réchauffer» la densité cible en aplatissant ses modes afin de permettre un plus grand déplacement de la chaîne puis de la «refroidir» à nouveau pour continuer la simulation. Un autre algorithme à considérer est celui proposé par [12], qui encourage les sauts entre des points où la densité est environ la même. Cela permet des sauts directement d'un mode à l'autre, sans avoir à se soucier des régions de très faible densité qui séparent traditionnellement les modes d'une distribution bimodale et qui bloquent le déplacement de l'algorithme RWM classique. Les méthodes énoncées pour échantillonner de distributions cibles multimodales ont des points communs importants : elles requièrent généralement un temps de calcul plus important qu'un algorithme de type Metropolis-Hastings (pour un même nombre d'itérations) et leur implémentation peut se montrer plus laborieuse. L'algorithme présenté dans ce mémoire se distingue par sa simplicité d'implémentation et par le faible temps de calcul requis.

Dans le chapitre 1, nous présentons des notions de base sur les chaînes de Markov, incluant plusieurs propriétés utiles pour l'introduction de l'algorithme RWM. Nous définissons ensuite ce dernier et justifions qu'en théorie, il estime bien la distribution cible à laquelle nous nous intéressons. Nous présentons ensuite un exemple dans lequel l'algorithme RWM, bien que théoriquement correct, n'arrive pas à estimer en un nombre raisonnable d'itérations la distribution cible considérée. Nous concluons ce chapitre en présentant le nouvel algorithme étudié dans ce mémoire. Dans le chapitre 2, nous présentons le principal résultat sur l'optimisation de l'algorithme RWM et nous établissons qu'il existe un résultat analogue pour notre nouvel algorithme. Or, dans notre cas, nous voyons que le résultat de convergence ne mène pas à une stratégie d'optimisation unique. C'est pourquoi nous présentons une stratégie empirique afin d'optimiser le nouvel algorithme en pratique. Le chapitre 3 est entièrement dédié à la démonstration du résultat de convergence présenté dans le chapitre 2, alors que le chapitre 4 consiste en quelques exemples d'implantation du nouvel algorithme selon la stratégie présentée au chapitre 2.

Chapitre 1

L'ALGORITHME METROPOLIS-HASTINGS DE TYPE MARCHE ALÉATOIRE

1.1. Les méthodes de Monte-Carlo

Comme plusieurs domaines des mathématiques appliquées, la statistique requiert souvent la résolution (numérique) d'intégrales multiples. Non seulement les intégrales peuvent-elles s'écrire, en général, comme des espérances mathématiques, mais celles qui sont rencontrées en statistique (entre autres en inférence bayésienne) se présentent souvent naturellement sous forme d'espérances. La méthode de Monte-Carlo est des plus intuitives pour approximer, de façon probabiliste, une espérance. Par exemple, supposons que la quantité à estimer soit $\mu(h) = \mathbb{E}\left[h(X)\right]$, où X est une variable aléatoire sur un espace échantillonnal $\mathcal{X}$ et $h : \mathcal{X} \mapsto \mathbb{R}$ est une fonction quelconque. L'idée est de générer un échantillon, disons $X_1, ..., X_n$, formé d'éléments indépendants et identiquement distribués (iid) selon la loi de X, puis d'utiliser l'estimateur de Monte-Carlo

$$\hat{\mu}(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i). \tag{1.1.1}$$

Cet estimateur est bien entendu sans biais, pourvu que l'espérance à estimer soit finie, et il est également convergent ; si $h(X) \in L^1$, la loi forte des grands nombres affirme que $\hat{\mu}(h) \to \mu(h)$ presque sûrement. Le principal inconvénient de cette méthode ne se situe donc pas au niveau de la qualité de l'estimation. La grande difficulté est de générer un échantillon selon la loi de X. En effet, cette loi est parfois trop compliquée pour qu'on puisse directement en générer des observations. De plus, on ne possède pas toujours toute l'information sur celle-ci. Une multitude d'algorithmes ont été proposés afin de résoudre ce problème à partir de certaines connaissances spécifiques à propos de cette loi.

Une situation qui se présente particulièrement souvent en statistique bayésienne est la suivante : la densité *a priori* d'un paramètre $\theta \in \mathbb{R}^d$ est une fonction π, alors que la vraisemblance de l'échantillon observé $\mathbf{x}$ est $f(\mathbf{x} \mid \cdot)$. La fonction de densité *a posteriori* de θ est donc l'unique fonction $\pi(\cdot \mid \mathbf{x})$ qui satisfait

$$\pi(\theta \mid \mathbf{x}) \propto \pi(\theta) f(\mathbf{x} \mid \theta)$$

et

$$\int_{\mathbb{R}^d} \pi(\theta \mid \mathbf{x}) \, d\theta = 1.$$

En résumé, on ne connaît que le «coeur» de la densité, c'est-à-dire qu'on connaît l'expression de la fonction de densité à une constante multiplicative près. Comment générer une observation provenant de cette loi ? Est-ce même possible ? L'algorithme de Metropolis-Hastings offre une réponse à cette question. Cet algorithme ne permet cependant pas d'obtenir un échantillon iid, mais plutôt une chaîne de Markov à valeurs dans l'espace $\mathcal{X}$, qui est souvent continu. Nous introduisons donc quelques notions sur les chaînes de Markov à espace d'états continu.

1.2. Chaînes de Markov à espace d'états continu

Dans cette section, nous présentons quelques définitions et propriétés des chaînes de Markov qui mèneront éventuellement à la définition de l'algorithme de Metropolis-Hastings. Les définitions, ainsi que la notation, sont principalement inspirés de [15], mais également de [8], [19], [22] et [23].

Définition 1.2.1. *Soient S un espace, T un ensemble quelconque et $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Un processus stochastique sur T est une fonction $X : \Omega \times T \mapsto S$. S est appelé l'espace d'états.*

Le processus X peut être vu comme une fonction de T dans S dont la valeur au point t est une variable aléatoire, $\forall t \in T$. En particulier, T est souvent dénombrable, dans lequel cas on peut supposer $T = \mathbb{N}$. Le processus X est alors une suite $\{X(n) : n \in \mathbb{N}\}$ de variables aléatoires. Lorsque la réalisation $\omega \in \Omega$ est fixée, tous les $X(n)$ sont réalisés en même temps. Or, on peut voir l'indice n comme étant un temps discret, et donc imaginer que les variables aléatoires sont réalisées l'une après l'autre. Lorque les n premiers termes de la suite ont été réalisés, on peut trouver la loi de $X(n+1)$ conditionnellement à ce que l'on a déjà observé. On est en présence d'une chaîne de Markov si cette loi dépend seulement de la valeur de $X(n)$.

Définition 1.2.2. *Soit $X = \{X(n) : n \in \mathbb{N}\}$ une suite de variables aléatoires définies sur un espace d'états S. X est une chaîne de Markov si, $\forall n$ et $\forall A \subset S$ mesurable,*

$$\mathbb{P}\left(X(n+1) \in A | \, X(1), ..., X(n)\right) = \mathbb{P}\left(X(n+1) \in A | \, X(n)\right).$$

La chaîne est dite à espace d'états continu si S est continu, et donc si les $X(n)$ sont des variables aléatoires continues.

On dit que la chaîne est homogène si, $\forall A \subset S$ mesurable, $\mathbb{P}\left(X(n+1) \in A | \, X(n)\right)$ est indépendant de n. À partir de maintenant, nous supposerons non seulement que les chaînes de Markov sont homogènes, mais également qu'elle sont continues, c'est-à-dire que l'espace d'états S est continu et que les variables aléatoires $X(n)$ sont absolument continues par rapport à une certaine mesure ϕ sur S, puisque ce sera le cas des chaînes de Markov considérées dans ce mémoire.

Définition 1.2.3. *Soit X une chaîne de Markov sur l'espace d'états S. Le noyau de transition de X,*

$$P = \{P(x, \cdot) : x \in S\},$$

est la famille des mesures de probabilité engendrée par la distribution conditionnelle de $X(n+1) \mid X(n)$, c'est-à-dire que $P(x, A) = \mathbb{P}\left(X(n+1) \in A | \, X(n) = x\right)$, où $x \in S$ et $A \subset S$ est ϕ-mesurable. De plus, on note P^m le noyau de transition en m pas, c'est-à-dire la famille des mesures de probabilité engendrée par la distribution conditionnelle de $X(n+m) \mid X(n)$.

Les chaînes qui nous intéresseront seront ϕ-irréductibles, c'est-à-dire qu'à partir d'un point donné, on pourra atteindre n'importe quelle région avec probabilité positive, pourvu que cette région soit de mesure ϕ non nulle. De plus, elles seront apériodiques, c'est-à-dire qu'il n'y aura pas de cycles qui se formeront.

Définition 1.2.4. *Soient X une chaîne de Markov sur l'espace d'états S et P son noyau de transition. X est dite ϕ-irréductible si, $\forall x \in S$ et $\forall A \subset S$ ϕ-mesurable tel que $\phi(A) > 0$, $\exists n \in \mathbb{N}$ tel que*

$$P^n(x, A) > 0.$$

Définition 1.2.5. *Soient X une chaîne de Markov sur l'espace d'états S et P son noyau de transition. X est dite périodique s'il existe un entier $n \geq 2$ et des sous-ensembles mesurables, disjoints et non vides $A_1, ..., A_n \subset S$ tels que*

$$x \in A_n \Rightarrow P(x, A_1) = 1$$

et $\forall i \in \{1, ..., n-1\}$,

$$x \in A_i \Rightarrow P(x, A_{i+1}) = 1.$$

La collection d'ensembles $A_1, ..., A_n$ est alors appelée un n-cycle. Une chaîne qui n'est pas périodique est dite apériodique.

Remarque 1.2.1. *En particulier, si $P(x, \{x\}) > 0$, $\forall x \in S$, alors X est clairement apériodique.*

Nous introduisons maintenant le concept de distribution stationnaire. La définition implique que, si à un certain temps n, $X(n)$ est distribué selon cette distribution, alors $X(n+1)$ l'est également, puis, par induction, $X(n+2), ...$ On dit alors que la chaîne est stationnaire, puisque la distribution de probabilité de ses éléments ne change plus avec le temps.

Définition 1.2.6. *Soient X une chaîne de Markov sur l'espace d'états S et P son noyau de transition. Une distribution Π est dite stationnaire pour X si, $\forall A \subset S$ mesurable,*

$$\Pi(A) = \int_S \Pi(dx)P(x, A). \tag{1.2.1}$$

Finalement, nous introduisons une condition suffisante pour la stationnarité, qui s'avérera plus simple à vérifier que la stationnarité elle-même.

Définition 1.2.7. *Soient X une chaîne de Markov sur l'espace d'états S et P son noyau de transition. X est dite Π-réversible si, $\forall x, y \in S$,*

$$\Pi(dx)P(x, dy) = \Pi(dy)P(y, dx). \tag{1.2.2}$$

Remarque 1.2.2. *En rappelant que $P(x, \cdot)$ est une mesure de probabilité sur S, et donc que $\int_S P(x, dy) = 1$, on remarque qu'en intégrant y sur S puis x sur A de chaque côté de (1.2.2), on retrouve (1.2.1) :*

$$\begin{aligned}
\int_A \int_S \Pi(dx)P(x, dy) &= \int_A \int_S \Pi(dy)P(y, dx) \\
\int_A \Pi(dx) \int_S P(x, dy) &= \int_S \Pi(dy) \int_A P(y, dx) \\
\Pi(A) &= \int_S \Pi(dy)P(y, A).
\end{aligned}$$

Une chaîne Π-réversible possède donc Π comme distribution stationnaire.

Étant donnée une chaîne de Markov avec une certaine distribution stationnaire, il serait sensé de croire que si la chaîne évolue pendant une longue période de temps, alors la distribution des états convergera éventuellement vers Π. Ce n'est pas le cas en toute généralité. Il existe cependant un résultat mentionnant que si une chaîne est Π-irréductible, apériodique, et qu'elle admet la distribution stationnaire Π, alors la chaîne de Markov convergera vers sa distribution stationnaire dans la métrique de la distance en variation totale (voir [**22**],

par exemple). Ceci signifie que sous ces conditions, la distribution stationnaire représente, en quelque sorte, la distribution limite des éléments d'une chaîne de Markov. Cette propriété s'avérera très pratique dans le cadre des algorithmes MCMC.

Rappelons que l'estimateur de Monte-Carlo (1.1.1) est simplement une moyenne échantillonnale. Or, l'échantillon qu'on obtient dans le cas d'un algorithme MCMC est une chaîne de Markov. Si cette chaîne possède une distribution stationnaire Π (on verra que ce sera le cas), on devine qu'elle pourra être utilisée pour estimer une espérance par rapport à la distribution Π. Or, on aimerait s'assurer qu'une moyenne échantillonnale basée sur une telle chaîne, tout comme une moyenne échantillonnale basée sur des réalisations iid, estime bien l'espérance, au sens où elle converge vers celle-ci. On aura besoin, pour cela, d'une dernière définition.

Définition 1.2.8. *Soit X une chaîne de Markov ϕ-irréductible et apériodique sur l'espace d'états S. Supposons également qu'elle possède une distribution stationnaire. X est dite Harris-récurrente si, $\forall A \subset S$ mesurable tel que $\phi(A) > 0$, $\forall x \in A$,*

$$\mathbb{P}\left(X(n) \in A \text{ i.s.} \middle| X(0) = x\right) = 1.$$

Ici, si $\{A_n : n \in \mathbb{N}\}$ est une suite d'événements, l'événement «A_n infiniment souvent» est défini comme

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

Intuitivement, une chaîne Harris-récurrente visite chaque région de probabilité non nulle une infinité de fois. Peu importe le point où on se trouve à un certain moment, on sait donc qu'on revisitera un voisinage de ce point. Nous sommes maintenant prêts à énoncer une version «markovienne» de la loi forte des grands nombres.

Théorème 1.2.1. *Soit X une chaîne de Markov Π-irréductible, apériodique et Harris-récurrente sur l'espace d'états S et P son noyau de transition. Supposons également que Π est sa distribution stationnaire. Finalement, soit h une fonction de S telle que $\int_S |h(x)|\Pi(dx) < \infty$. On a alors*

$$\frac{1}{n}\sum_{i=1}^{n} h(X(i)) \longrightarrow \int_S h(x)\Pi(dx)$$

presque sûrement lorsque $n \to \infty$, et ce, peu importe la distribution du point de départ $X(0)$.

Ce théorème et sa démonstration peuvent être trouvés dans [15], théorème 17.1.7. Ce résultat nous amène à l'idée fondamentale de l'algorithme de Metropolis-Hastings, et du même coup des méthodes MCMC. Si on arrive à construire une chaîne de Markov sur le domaine de Π, la distribution à estimer, qui respecte les conditions du théorème et dont la

distribution stationnaire est elle-même Π, alors l'estimateur de Monte-Carlo $\hat{\mu}(h)$ convergera vers $\mu(h)$ avec probabilité 1.

1.3. Construction et implémentation de l'algorithme RWM

1.3.1. Définition

Soit Π la distribution de probabilité qui nous intéresse, que nous appelons la distribution cible, et π la densité qui y est associée, que nous appelons la densité cible. Supposons que cette distribution est absolument continue par rapport à ϕ, que nous considérons ici être la mesure de Lebesgue sur S. L'algorithme proposé par [14], puis amélioré par [11], vise à construire une chaîne de Markov de distribution stationnaire Π. Afin d'y arriver, on commence par choisir un point de départ $X(0) \in S$ de façon aléatoire ou déterministe, tel que $\pi(X(0)) > 0$. Puis, pour générer $X(n+1)$ à partir de $X(n)$, on génère une «proposition» Y selon une fonction de densité absolument continue par rapport à ϕ qui peut dépendre de l'état actuel, $g(\cdot \mid X(n))$. Nous appellerons g la densité instrumentale. On pose $X(n+1) = Y$ avec probabilité $\alpha(X(n), Y)$ et on dit alors que la proposition est acceptée. Sinon, la proposition est refusée et on pose $X(n+1) = X(n)$. Plusieurs formes sont possibles pour la fonction d'acceptation α, mais [18] a démontré que le meilleur choix était l'une des fonctions relevées par [11], soit

$$\alpha(x, y) = 1 \wedge \frac{\pi(y) g(x \mid y)}{\pi(x) g(y \mid x)},$$

où $a \wedge b = \min\{a, b\}$. Pour que l'algorithme soit non seulement convergent, mais aussi efficace, le choix de la densité instrumentale g doit être judicieux. Pour $S = \mathbb{R}^d$, la distribution instrumentale la plus populaire, qui est également l'une des plus simples à implémenter, est sans doute la loi normale centrée à l'état actuel et formée de composantes indépendantes. Celle-ci satisfait

$$g(\mathbf{y} \mid \mathbf{x}) = (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x})^\top (\mathbf{y} - \mathbf{x})\right\}, \tag{1.3.1}$$

où $\sigma > 0$. L'algorithme qui est alors obtenu est un cas particulier de celui qui est appelé Metropolis marche aléatoire (RWM), puisqu'on peut facilement vérifier que la chaîne de Markov obtenue est une marche aléatoire sur $\mathbb{R}^d$. C'est l'algorithme RWM qui sera étudié dans ce mémoire. D'abord, on remarque que dans ce cas $g(\mathbf{y} \mid \mathbf{x}) = g(\mathbf{x} \mid \mathbf{y})$, ce qui mène à la simplification suivante pour la probabilité d'acceptation :

$$\alpha(\mathbf{x}, \mathbf{y}) = 1 \wedge \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}.$$

Remarquons que dans la construction de cet algorithme, la densité cible n'apparaît qu'à travers le rapport $\pi(\mathbf{y})/\pi(\mathbf{x})$ présent dans la probabilité d'acceptation. Ainsi, la constante de normalisation de cette densité n'a pas besoin d'être connue. Cet algorithme permet donc de résoudre le problème d'inférence bayésienne présenté à la section 1.1.

1.3.2. Justification

Nous allons maintenant vérifier que la chaîne obtenue à partir de cet algorithme respecte les conditions du théorème 1.2.1.

Lemme 1.3.1. *Soient $S = \mathbb{R}^d$ et $\mathbf{X}$ la chaîne de Markov engendrée par l'algorithme RWM avec la distribution instrumentale donnée par (1.3.1). Si Π est absolument continue par rapport à ϕ, la mesure de Lebesgue sur $\mathbb{R}^d$, alors $\mathbf{X}$ est Π-irréductible, apériodique et elle admet Π comme distribution stationnaire.*

Démonstration. Pour la Π-irréductibilité, on pose $\mathbf{x} \in \mathbb{R}^d$ tel que $\pi(\mathbf{x}) > 0$ et A un sous-ensemble mesurable tel que $\Pi(A) > 0$. On a alors

$$\begin{aligned}
P(\mathbf{x}, A) &= \mathbb{P}\left(\mathbf{X}(n+1) \in A \middle| \mathbf{X}(n) = \mathbf{x}\right) \\
&\geq \mathbb{P}\left(\mathbf{Y} \in A, \mathbf{X}(n+1) = \mathbf{Y} \middle| \mathbf{X}(n) = \mathbf{x}\right) \\
&= \mathbb{E}\left[\mathbb{I}_A(\mathbf{Y}) \mathbb{I}\left\{\mathbf{X}(n+1) = \mathbf{Y}\right\} \middle| \mathbf{X}(n) = \mathbf{x}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_A(\mathbf{Y}) \mathbb{I}\left\{\mathbf{X}(n+1) = \mathbf{Y}\right\} \middle| \mathbf{X}(n) = \mathbf{x}, \mathbf{Y}\right] \middle| \mathbf{X}(n) = \mathbf{x}\right] \\
&= \mathbb{E}\left[\mathbb{I}_A(\mathbf{Y}) \mathbb{P}\left(\mathbf{X}(n+1) = \mathbf{Y} \middle| \mathbf{X}(n) = \mathbf{x}, \mathbf{Y}\right) \middle| \mathbf{X}(n) = \mathbf{x}\right] \\
&= \mathbb{E}\left[\mathbb{I}_A(\mathbf{Y}) \alpha(\mathbf{x}, \mathbf{Y}) \middle| \mathbf{X}(n) = \mathbf{x}\right] \\
&= \int_A \alpha(\mathbf{x}, \mathbf{y}) g(\mathbf{y} \mid \mathbf{x}) \, d\mathbf{y} \\
&= \int_A \left(1 \wedge \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right) g(\mathbf{y} \mid \mathbf{x}) \, d\mathbf{y}.
\end{aligned}$$

Posons $A(\mathbf{x}, r) = A \cap B(\mathbf{x}, r)$, où $B(\mathbf{x}, r)$ est la boule de rayon $r > 0$ centrée en $\mathbf{x}$ et où r est choisi assez grand pour que $\Pi(A(\mathbf{x}, r)) > 0$. L'existence d'un tel r nous est assurée par la continuité de la mesure de probabilité, puisque $\lim_{r\to\infty} \Pi(A(\mathbf{x}, r)) = \Pi(\lim_{r\to\infty} A(\mathbf{x}, r)) = \Pi(A) > 0$. De plus, sur $A(\mathbf{x}, r)$, la fonction $g(\cdot \mid \mathbf{x})$ est bornée inférieurement par $(2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} > 0$. On a donc

$$P(\mathbf{x}, A) \geq \int_{A(\mathbf{x},r)} \left(1 \wedge \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right) g(\mathbf{y} \mid \mathbf{x}) \, d\mathbf{y}$$

$$\geq (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \int_{A(\mathbf{x},r)} \left(1 \wedge \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right) d\mathbf{y}. \tag{1.3.2}$$

Il nous faut simplement montrer que l'intégrale dans (1.3.2) est positive, ce qui sera suffisant pour prouver la Π-irréductibilité. Posons $C = \{\mathbf{y} \in A(\mathbf{x}, r) : \pi(\mathbf{y}) \geq \pi(\mathbf{x})\}$. Si $\Pi(C) > 0$, alors on conclut directement que

$$\int_{A(\mathbf{x},r)} \left(1 \wedge \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right) d\mathbf{y} \geq \int_C d\mathbf{y} = \phi(C) > 0,$$

puisque $\Pi(C) > 0$ implique que $\phi(C) > 0$, dû à la continuité absolue de Π par rapport à ϕ, la mesure de Lebesgue sur $\mathbb{R}^d$. Si $\Pi(C) = 0$, alors on a

$$\int_{A(\mathbf{x},r)} \left(1 \wedge \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right) d\mathbf{y} = \int_{A(\mathbf{x},r)} \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\, d\mathbf{y} = \frac{1}{\pi(\mathbf{x})} \int_{A(\mathbf{x},r)} \pi(\mathbf{y})\, d\mathbf{y} = \frac{\Pi(A(\mathbf{x},r))}{\pi(\mathbf{x})} > 0.$$

Pour l'apériodicité, il suffit de noter que, $\forall \mathbf{x}$ tel que la densité $\pi(\mathbf{x}) > 0$, on a

$$\begin{aligned}
P(\mathbf{x}, \{\mathbf{x}\}) &= \mathbb{P}\left(\mathbf{X}(n+1) = \mathbf{x} | \, \mathbf{X}(n) = \mathbf{x}\right) \\
&= \mathbb{E}\left[\mathbb{I}\{\mathbf{X}(n+1) = \mathbf{x}\} | \, \mathbf{X}(n) = \mathbf{x}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{\mathbf{X}(n+1) = \mathbf{x}\} | \, \mathbf{X}(n) = \mathbf{x}, \mathbf{Y}\right] | \, \mathbf{X}(n) = \mathbf{x}\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(\mathbf{X}(n+1) = \mathbf{x} | \, \mathbf{X}(n) = \mathbf{x}, \mathbf{Y}\right) | \, \mathbf{X}(n) = \mathbf{x}\right] \\
&= \mathbb{E}\left[1 - \alpha(\mathbf{x}, \mathbf{Y}) | \, \mathbf{X}(n) = \mathbf{x}\right] \\
&= 1 - \mathbb{E}\left[\alpha(\mathbf{x}, \mathbf{Y}) | \, \mathbf{X}(n) = \mathbf{x}\right] \\
&> 0.
\end{aligned}$$

En effet, on ne peut pas avoir $\mathbb{E}\left[\alpha(\mathbf{x}, \mathbf{Y}) | \, \mathbf{X}(n) = \mathbf{x}\right] = 1$, puisque cela signifierait que $\alpha(\mathbf{x}, \mathbf{Y}) = 1$ presque partout, et donc que $\pi(\mathbf{y}) \geq \pi(\mathbf{x}) > 0$ presque partout, ce qui est absurde puisque π est une densité sur $\mathbb{R}^d$.

Pour la stationnarité de la distribution Π, nous allons simplement montrer que $\mathbf{X}$ est Π-réversible. Il faut donc montrer que, $\forall \mathbf{x} \neq \mathbf{y}$,

$$\Pi(d\mathbf{x}) P(\mathbf{x}, d\mathbf{y}) = \Pi(d\mathbf{y}) P(\mathbf{y}, d\mathbf{x}),$$

car pour $\mathbf{x} = \mathbf{y}$, l'égalité est triviale. Puisque Π est absolument continue par rapport à la mesure de Lebesgue, nous avons $\Pi(d\mathbf{x}) = \pi(\mathbf{x}) d\mathbf{x}$. Pour ce qui est de la densité associée au noyau de transition $P(\mathbf{x}, \cdot)$, il s'agit de $g(\cdot \mid \mathbf{x}) \alpha(\mathbf{x}, \cdot)$. En effet, pour tout ensemble A mesurable tel que $\mathbf{x} \notin A$, on a

$$\int_A g(\mathbf{y} \mid \mathbf{x}) \alpha(\mathbf{x}, \mathbf{y})\, d\mathbf{y} = \mathbb{E}\left[\alpha(\mathbf{x}, \mathbf{Y}) \mathbb{I}_A(\mathbf{Y}) | \, \mathbf{X}(n) = \mathbf{x}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{\mathbf{X}(n+1)=\mathbf{Y}\}\middle|\,\mathbf{X}(n)=\mathbf{x},\mathbf{Y}\right]\mathbb{I}_A(\mathbf{Y})\middle|\,\mathbf{X}(n)=\mathbf{x}\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_A(\mathbf{Y})\mathbb{I}\{\mathbf{X}(n+1)=\mathbf{Y}\}\middle|\,\mathbf{X}(n)=\mathbf{x},\mathbf{Y}\right]\middle|\,\mathbf{X}(n)=\mathbf{x}\right]\\
&= \mathbb{E}\left[\mathbb{I}_A(\mathbf{Y})\mathbb{I}\{\mathbf{X}(n+1)=\mathbf{Y}\}\middle|\,\mathbf{X}(n)=\mathbf{x}\right]\\
&= \mathbb{E}\left[\mathbb{I}_A(\mathbf{X}(n+1))\middle|\,\mathbf{X}(n)=\mathbf{x}\right]\\
&= P(\mathbf{x},A).
\end{aligned}
$$

On a donc $P(\mathbf{x},d\mathbf{y}) = g(\mathbf{y}\mid\mathbf{x})\alpha(\mathbf{x},\mathbf{y})d\mathbf{y}$, $\forall \mathbf{x}\neq\mathbf{y}$. On peut alors conclure que

$$
\begin{aligned}
\Pi(d\mathbf{x})P(\mathbf{x},d\mathbf{y}) &= (\pi(\mathbf{x})d\mathbf{x})(g(\mathbf{y}\mid\mathbf{x})\alpha(\mathbf{x},\mathbf{y})d\mathbf{y})\\
&= \pi(\mathbf{x})g(\mathbf{y}\mid\mathbf{x})\alpha(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}\\
&= \pi(\mathbf{x})g(\mathbf{y}\mid\mathbf{x})\left(1\wedge\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right)d\mathbf{x}d\mathbf{y}\\
&= g(\mathbf{y}\mid\mathbf{x})\left(\pi(\mathbf{x})\wedge\pi(\mathbf{y})\right)d\mathbf{x}d\mathbf{y}.
\end{aligned}
$$

Par symétrie de la densité instrumentale g, on a donc $\Pi(d\mathbf{x})P(\mathbf{x},d\mathbf{y}) = \Pi(d\mathbf{y})P(\mathbf{y},d\mathbf{x})$. La chaîne $\mathbf{X}$ étant Π-réversible, elle admet Π comme distribution stationnaire.

□

Afin de complètement satisfaire les conditions du théorème 1.2.1, il ne reste que la Harris-récurrence de $\mathbf{X}$ à vérifier. Or, par construction de l'algorithme de Metropolis-Hastings, cette propriété est une conséquence de la Π-irréductibilité de $\mathbf{X}$ (voir [**23**] et [**16**]). On peut donc directement appliquer le théorème 1.2.1 et affirmer que l'estimateur de Monte-Carlo basé sur la chaîne obtenue par l'algorithme RWM avec la densité instrumentale donnée par (1.3.1) converge vers l'espérance à estimer lorsque la taille échantillonnale tend vers ∞, et ce, avec probabilité 1. Ce résultat est indépendant du point de départ $\mathbf{X}(0)$ choisi (ou de la loi de celui-ci, s'il est réalisé de façon aléatoire). Or, ce point de départ peut très bien influencer la vitesse de convergence ; on imagine facilement que, par exemple, si $\mathbf{X}(0)$ est choisi très loin du mode de π, les premiers pas de la chaîne, qui serviront principalement à se rapprocher du mode, seront peu représentatifs de la distribution cible. Par contre, une stratégie très utilisée en pratique permet d'éliminer cet effet potentiel du point de départ : le *burn in*. L'idée est de réaliser une chaîne de longueur, disons, n, mais d'utiliser comme échantillon les valeurs $\mathbf{X}(m+1),...,\mathbf{X}(n)$, pour un certain $m<n$. On dit alors qu'on a «brûlé» les m premières observations, puisqu'on suppose que la stationnarité n'avait pas été atteinte. Pour rendre le choix de m le moins arbitraire possible, il existe une panoplie de tests diagnostiques afin de vérifier si on a «atteint» la stationnarité (en réalité, on n'atteint que très rarement la distribution stationnaire, mais on s'en rapproche infiniment). Voir, par exemple, [**6**] pour

une revue de différents tests diagnostiques. Or, cet enjeu ne sera pas abordé ici. Plutôt que de s'intéresser au point de départ, nous allons supposer que $\mathbf{X}(0)$ est distribué selon Π. Par la stationnarité de la distribution Π pour $\mathbf{X}$, cela implique que la distribution marginale de chaque élément de la chaîne est Π.

1.3.3. Premiers exemples

Nous allons maintenant présenter deux exemples d'application de l'algorithme RWM : un premier où l'algorithme estime bien la distribution cible, et un second où il performe beaucoup moins efficacement. Soit $X \sim N(15; 9)$. On souhaite utiliser l'algorithme RWM pour estimer $\mathbb{E}[X^2]$. Bien sûr, on sait que

$$\mathbb{E}\left[X^2\right] = \mathbb{V}\text{ar}(X) + \mathbb{E}[X]^2 = 234.$$

On utilise, comme distribution instrumentale, la loi $N(X; 1)$, où X désigne l'état actuel de la chaîne. Rappelons qu'on n'a pas besoin de l'expression de la densité instrumentale g ; puisque celle-ci est symétrique, elle n'apparaît pas dans la probabilité d'acceptation α. On fixe $X(0) = 15$, le mode de la densité cible, puis on utilise un générateur de la loi $N(0; 1)$ afin de générer les nouvelles observations. La proposition Y est alors acceptée avec probabilité

$$\alpha(X, Y) = 1 \wedge \frac{\pi(Y)}{\pi(X)},$$

où X est l'état actuel de la chaîne et π représente la densité de la loi $N(15; 9)$. Avec $n = 1\,000\,000$ observations, on obtient 233,7348 comme estimation. Or, la qualité de cette estimation dépend de la bonne représentativité de l'échantillon par rapport à la distribution cible. On pourrait donc directement estimer la densité cible en utilisant notre échantillon et, par exemple, la méthode des noyaux. Comme on le voit à la figure 1.1, dans cette situation, l'algorithme est efficace puisqu'il estime très bien la densité, et du même coup l'espérance. La densité réelle, en rouge, et la densité estimée à l'aide de la fonction «density» de R, en noir, sont confondues.

La loi normale univariée étant «lisse», elle est relativement facile à estimer. Par contre, les méthodes MCMC étant particulièrement utilisées pour des distributions multidimensionnelles et complexes, on peut penser à un exemple un peu plus réaliste : un mélange de deux lois normales multivariées. Soit

$$\mathbf{X} \sim \begin{cases} N_{10}(-\mu; 9I_{10}), \text{ avec probabilité } \frac{1}{2} \\ N_{10}(\mu; 9I_{10}), \text{ avec probabilité } \frac{1}{2} \end{cases}, \tag{1.3.3}$$

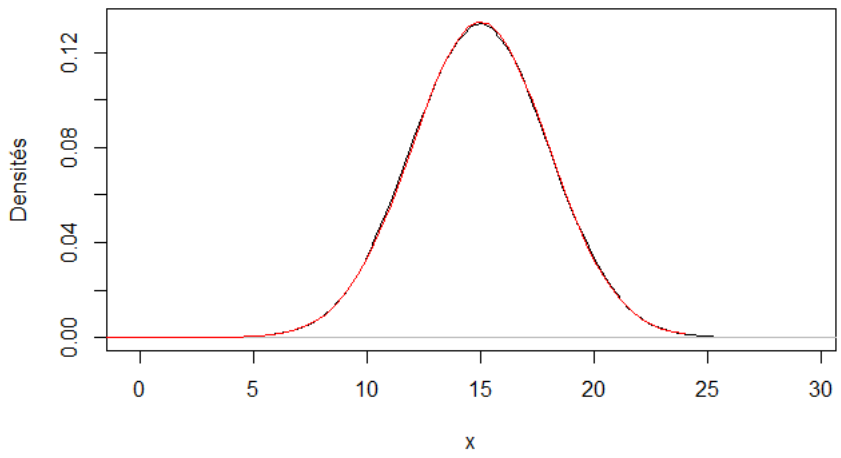

FIGURE 1.1. Estimation de la densité d'une $N(15; 9)$ à l'aide de l'algorithme RWM.

où $\mu = (15, 0, ..., 0)^\top \in \mathbb{R}^{10}$ et I_{10} représente la matrice identité de format 10×10. Cette distribution est dite bimodale, puisqu'on peut facilement démontrer que sa densité est une fonction bimodale (elle admet des modes aux points $\pm\mu$).

Supposons qu'on souhaite utiliser l'algorithme RWM afin de simplement estimer $\mathbb{E}[\mathbf{X}]$. Évidemment, on sait que $\mathbb{E}[\mathbf{X}] = \mathbf{0}$. De façon analogue à l'exemple précédent, on va utiliser la loi $N_{10}(\mathbf{X}; \sigma^2 I_{10})$ comme distribution instrumentale, avec $\sigma = 2{,}5$ (on verra dans le prochain chapitre comment la valeur 2,5 a été choisie). Que devrait être le point de départ de la chaîne ? Afin de ne pas favoriser, *a priori*, l'un des deux modes de la distribution cible, on choisit $\mathbf{X}(0) = 0$. On pose $n = 100\,000$ observations ; cette valeur est relativement petite pour une simulation, mais elle nous permettra de mieux visualiser le comportement de l'algorithme. Pour $i \neq 1$, les estimations qu'on obtient pour $\mathbb{E}[X_i]$ sont proches de 0, mais notre estimation pour $\mathbb{E}[X_1]$ est d'environ 14,96. Cette valeur est très éloignée de 0. Que s'est-il passé ? La réponse se trouve dans l'échantillon qu'on a obtenu. En utilisant ce dernier pour estimer la densité de X_1 (qui est une combinaison convexe de deux densités normales) toujours à l'aide de la fonction «density» de R, on obtient la figure 1.2. La vraie densité est en rouge alors que la densité estimée est en noir.

Le problème est que la chaîne engendrée par l'algorithme RWM a tendance à se diriger là où la densité cible est grande (par la forme de la probabilité d'acceptation). Ainsi, en commençant à $X_1(0) = 0$, elle a choisi «aléatoirement» une direction vers où se diriger. Puis elle a invariablement trouvé l'un des deux modes de la distribution de X_1. Puisque π est grande aux alentours de ce mode, la chaîne a beaucoup de facilité à s'y déplacer. Or, les

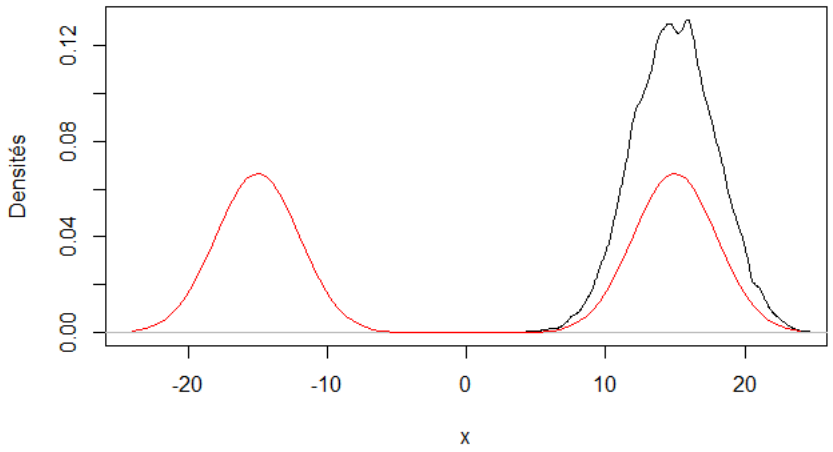

FIGURE 1.2. Estimation de la densité bimodale correspondant à la distribution décrite en (1.3.3) à l'aide de l'algorithme RWM avec $\sigma = 2{,}5$.

pas proposés étant en général plus courts que la distance entre les deux modes (qui est ici 30), la chaîne doit, pour se rendre à l'autre mode, traverser une région de très faible densité. Puisque la probabilité d'acceptation d'un pas vers cette région est infime, les changements de mode seront rares. En fait, en visualisant la trace de la première composante de notre chaîne, représentée à la figure 1.3, on peut voir que celle-ci s'est dirigée vers le mode positif et qu'il n'y a eu aucun changement de mode parmi les n itérations. Notre estimation de $\mathbb{E}[X_1]$

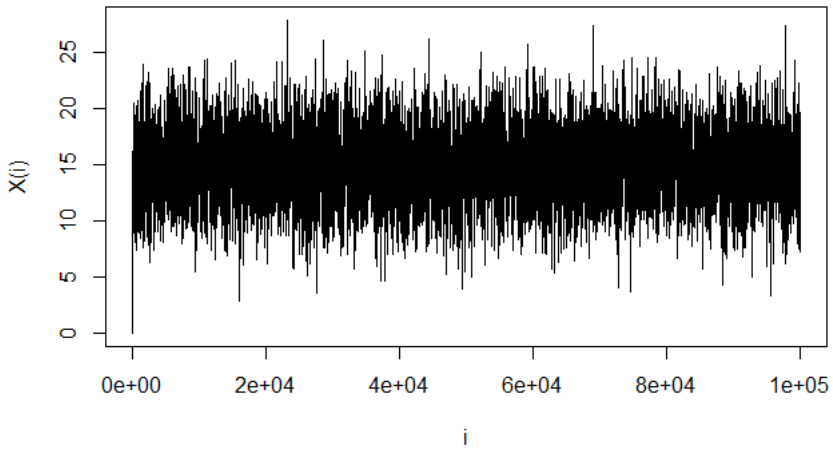

FIGURE 1.3. Trace de l'algorithme RWM avec $\sigma = 2{,}5$.

est donc positive, puisque basée sur seulement la moitié du support de X_1. Au contraire, si

la chaîne s'était dirigée vers le mode négatif au début de la simulation, on aurait observé l'inverse, et notre estimation serait négative.

Rappelons que la distribution instrumentale (1.3.1) dépend d'un paramètre d'échelle σ. Dans cet exemple, on a posé $\sigma = 2{,}5$, mais il est clair qu'une valeur plus grande induirait des pas proposés plus grands, et donc augmenterait les chances de changer de mode. Par exemple, avec $\sigma = 8$, on obtient la chaîne représentée sur la figure 1.4. On voit qu'il s'est produit trois

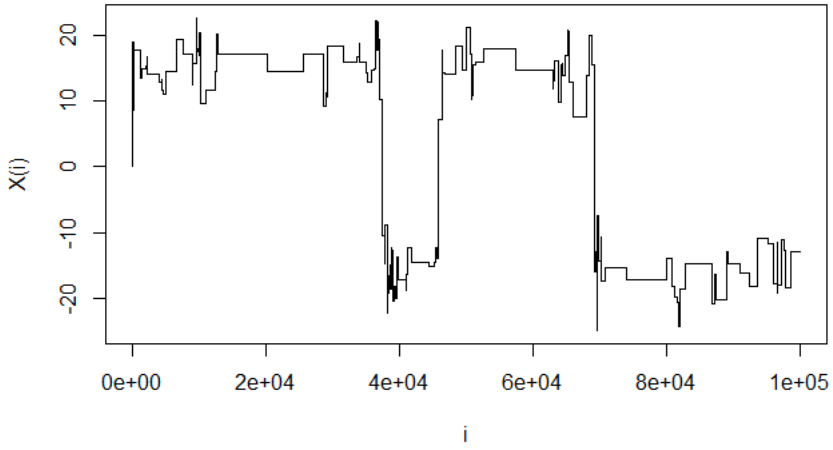

FIGURE 1.4. Trace de l'algorithme RWM avec $\sigma = 8$.

changements de mode. Intuitivement, plus il y a de changements de mode, mieux le temps passé dans chacun des deux modes devrait être distribué. En effet, en observant la densité estimée $\hat{\pi}$ (en noir) à la figure 1.5, on voit qu'elle s'est rapprochée de π (en rouge). Par contre, l'apparition de plateaux sur la trace de la chaîne (figure 1.4) traduit une forte diminution du taux d'acceptation, défini par la proportion des pas qui ont été acceptés parmi les n qui ont été proposés. Malgré qu'il ne soit pas idéal d'avoir un taux d'acceptation trop grand (on verra pourquoi sous peu), on ne veut pas non plus qu'il soit trop petit, puisque cela signifie que la chaîne reste prise à certains états pendant de nombreuses itérations, ce qui se traduit par une mauvaise exploration du support. Or, ce taux est passé d'environ 23% à 0,2% lorsque le paramètre d'échelle est passé de 2,5 à 8. La raison est simple : puisque le paramètre σ est le même pour chacune des 10 composantes, les pas proposés sont en général plus grands, et ce, pour chaque composante. Or, l'acceptation ou le refus d'un pas se fait dans les 10 dimensions en même temps. Une grande valeur de σ est défavorable pour les composantes 2 à 10 ; ces dernières ont alors une plus grande chance de tomber dans une région éloignée où la densité est très faible, ce qui cause un rejet presque assuré. Le sujet du choix optimal de σ sera abordé dans le chapitre suivant. Ce qu'il faut retenir ici est que, lorsque la distribution

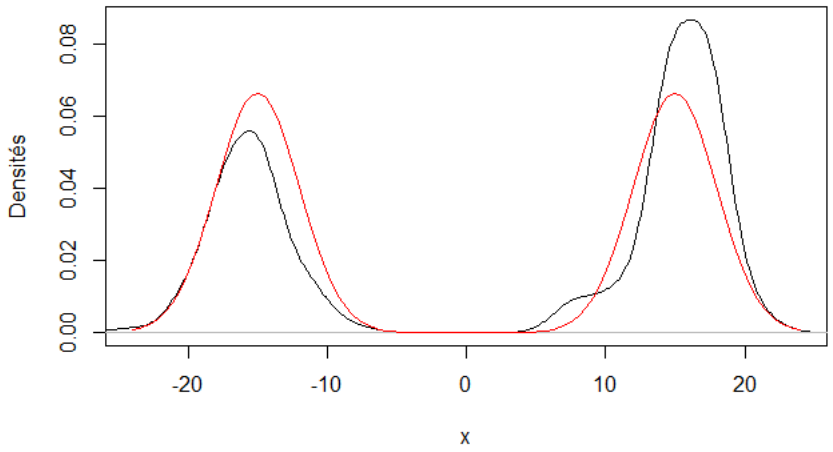

FIGURE 1.5. Estimation de la densité bimodale correspondant à la distribution décrite en (1.3.3) à l'aide de l'algorithme RWM avec $\sigma = 8$.

cible est multidimensionnelle et bimodale (ce serait également vrai pour une distribution k-modale, $k > 2$), l'algorithme RWM avec distribution instrumentale normale semble mener à une impasse ; un choix trop petit du paramètre d'échelle nous empêche de visiter les deux modes de la distribution aussi souvent qu'on le voudrait et un choix trop grand entraîne un taux d'acceptation catastrophique.

Qu'en serait-il si on choisissait un écart-type instrumental pour la première composante X_1 différent de celui pour les autres composantes $X_2, ...X_{10}$? Nous pourrions alors changer de mode de façon plus fluide, sans toutefois pénaliser les composantes 2 à 10. Un inconvénient important de cette approche est que malgré le fait que ces dernières exploreraient bien leur espace, on ne pourrait pas en dire autant de la composante X_1. En effet, le choix d'une grande variance favorable aux changements de mode, et donc à l'exploration globale de la densité cible, n'est en aucun cas optimale pour explorer les caractéristiques locales de chacun des deux modes. Il faut donc penser à une méthode alternative pour échantillonner de ce genre de distribution cible.

1.4. UN NOUVEL ALGORITHME

L'objectif principal de ce mémoire est l'introduction d'un nouvel algorithme de type RWM particulièrement efficace pour estimer les distributions bimodales sur $\mathbb{R}^d$. Comme on l'a vu, l'algorithme RWM est défini comme un algorithme de Metropolis-Hastings dont

la densité instrumentale est symétrique autour de l'état actuel de la chaîne, ce qui cause une simplification de la probabilité d'acceptation α. En gardant en tête cette définition, on souhaite choisir la fonction de densité instrumentale g de façon à ce que $g(\mathbf{x} \mid \mathbf{y}) = g(\mathbf{y} \mid \mathbf{x})$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. De plus, on souhaite que g propose des pas qui ne sont, en général, pas trop grands (afin d'avoir un bon taux d'acceptation) et qu'elle favorise en même temps les changements de mode.

1.4.1. La distribution cible

Dans le but de pouvoir étudier le nouvel algorithme d'un point de vue théorique, nous allons d'abord fixer la forme de la distribution cible. La forme que nous choisissons, bien que peu générale, est celle qui nous permettra d'établir des résultats d'optimalité dans le prochain chapitre. Rappelons toutefois qu'il est possible de généraliser cet algorithme pour une classe plus vaste de distributions cibles. Soit π_d la densité cible, où d est la dimension, et supposons que

$$\pi_d\left(\mathbf{x}^{(d)}\right) = \prod_{i=1}^{d} f_i(x_i).$$

Or, on supposera, par souci de simplicité, que $f_2 = f_3 = ... = f_d \equiv f$, alors que f_1 a une forme particulière. En pratique, f_1 sera bimodale. On peut donc écrire

$$\pi_d\left(\mathbf{x}^{(d)}\right) = f_1(x_1) \prod_{i=2}^{d} f(x_i). \tag{1.4.1}$$

Notons qu'il serait facile de laisser tomber la condition $f_i = f_j$ pour $i, j \neq 1$, quitte à imposer quelques conditions asymptotiques sur la suite $\{f_n : n \in \mathbb{N}\}$, mais cela alourdirait les résultats théoriques des prochains chapitres.

1.4.2. La distribution instrumentale

L'idée de départ est d'utiliser deux stratégies afin de proposer les incréments ; la première consiste à proposer des petits pas dans le but d'explorer localement le support (le mode dans lequel on se trouve à ce moment) alors que l'autre consiste à proposer de grands pas dont le but est de changer de mode. On définit donc la distribution instrumentale de la façon suivante. Supposons que l'état actuel de la chaîne est $\mathbf{X}$. On réalise la variable aléatoire $\delta \sim$ Bernoulli(p), où $p \in (0, 1)$, dont le rôle est de décider quelle sera la nature du pas proposé, $\mathbf{Y}$. Si $\delta = 0$, alors on propose un petit pas, c'est-à-dire

$$\mathbf{Y} \sim N_d\left(\mathbf{X}; \sigma^2 I_d\right),$$

où $\sigma > 0$. Si $\delta = 1$, alors on propose un grand pas. On génère donc une proposition Y_1, pour la première composante, selon une certaine densité $g_1(\cdot \mid X_1)$, définie sur $\mathbb{R}$, qui respecte la condition de symétrie. La fonction g_1 devra être choisie de façon à favoriser des grands pas pour maximiser la probabilité de trouver l'autre mode. Puis, indépendamment de Y_1, on génère

$$\mathbf{Y}^- \sim N_{d-1}\left(\mathbf{X}^-; \sigma^2 I_{d-1}\right),$$

où $\mathbf{X}^-$ désigne le vecteur $\mathbf{X}$ sans sa première composante et $\mathbf{Y}^-$ est défini de façon analogue. Puisque seule la densité f_1 est bimodale, on ne souhaite pas que les pas proposés pour les autres composantes soient grands. C'est ce qui explique que la variance σ^2 ici soit la même que celle utilisée lors des petits pas.

On peut montrer que la chaîne de Markov obtenue en utilisant cet algorithme respecte les conditions du théorème 1.2.1 de façon analogue à ce qui a été fait dans le lemme 1.3.1 où la distribution instrumentale était normale. En effet, chaque proposition $\mathbf{Y}$ est générée selon une loi normale avec probabilité $1 - p > 0$. Cela nous assure l'irréductibilité. Pour l'apériodicité, on a toujours $P(\mathbf{x}, \{\mathbf{x}\}) > 0$, $\forall \mathbf{x}$ tel que $\pi_d(\mathbf{x}) > 0$. La réversibilité, quant à elle, est respectée pour tout algorithme de type RWM. Finalement, la Harris-récurrence découle encore de l'irréductibilité.

Bien sûr, pour que l'algorithme soit efficace, le choix des paramètres p et σ^2 devra être judicieux. Pour ce qui est de g_1, le choix optimal, s'il en existe un, sera probablement très dépendant de la forme de f_1. Ainsi, dépendamment de ce qui est connu sur celle-ci, quelques conditions s'imposeront d'elles-mêmes. Par exemple, on pourrait vouloir que $g_1(\cdot \mid X_1)$ ne soit pas croissante lorsqu'on s'éloigne de X_1, de façon à ce que, si on a surestimé la distance entre les deux modes de f_1, on ne propose pas des pas qui sont systématiquement trop grands. Une autre condition logique est que $g_1(\cdot \mid X_1)$ n'admette pas un mode trop élevé au point X_1, puisque l'objectif ici est de changer de mode, et non de rester dans le même voisinage. Deux choix possibles qui respectent ces conditions et qui seront considérés dans ce mémoire sont la densité de la loi normale centrée à X_1 avec une assez grande variance et la densité d'une loi uniforme sur un intervalle de la forme $(X_1 \pm c)$, avec c suffisamment grand. On verra que la loi uniforme est le meilleur de ces deux choix et qu'en réalité, il s'agit du meilleur choix parmi les densités qui respectent les deux conditions énoncées ci-haut, pourvu qu'on choisisse bien la valeur de c.

1.4.3. Retour sur l'exemple

Reprenons le dernier exemple de la section 1.3.3, où on souhaitait estimer la densité de la loi normale bimodale. On se rappelle que, pour cette distribution, l'algorithme RWM classique nous avait mené dans une impasse ; un petit paramètre d'échelle empêchait les changements de mode alors qu'un paramètre d'échelle suffisamment grand pour nous permettre quelques changements de mode réduisait le taux d'acceptation à moins de 1%. On utilise maintenant la nouvelle méthode, c'est-à-dire l'algorithme RWM avec notre distribution instrumentale particulière, avec $n = 100\,000$ comme précédemment. On pose $p = 0{,}1$ et $\sigma = 2{,}5$, puisqu'on se rappelle que ce choix pour le paramètre d'échelle nous permettait un bon déplacement à l'intérieur des modes. Pour la densité g_1, on choisit celle de la loi uniforme avec $c = 40$, c'est-à-dire $g_1(y \mid x) = \frac{1}{80}\mathbb{I}_{(x \pm 40)}(y)$.

La trace de la chaîne formée par la première composante de l'algorithme, représentée sur la figure 1.6, montre qu'il y a eu plus de changements de mode qu'on ne peut le compter.

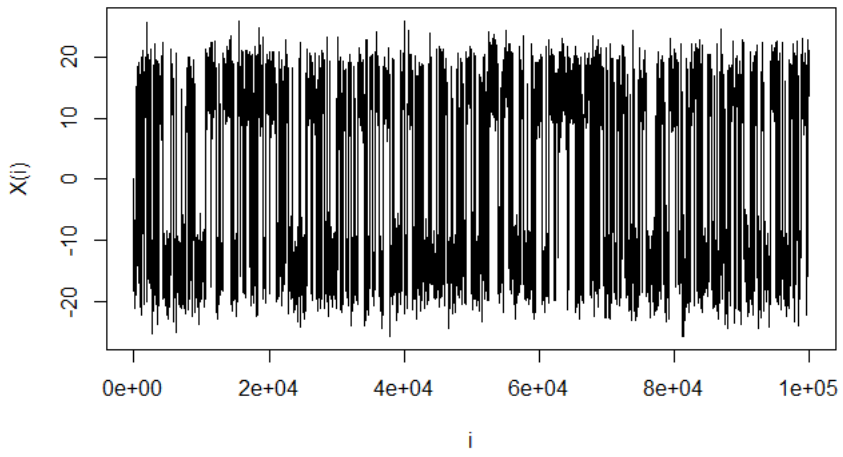

FIGURE 1.6. Trace du nouvel algorithme.

Cela se traduit par la densité estimée représentée en noir sur la figure 1.7, qui accorde une pondération d'environ $\frac{1}{2}$ à chacun des modes et qui ressemble à la densité cible π, en rouge (bien que davantage d'itérations seraient nécessaires pour une estimation plus précise de chacun des modes). De plus, malgré l'envergure des pas, le taux d'acceptation est resté relativement haut aux alentours de 20%.

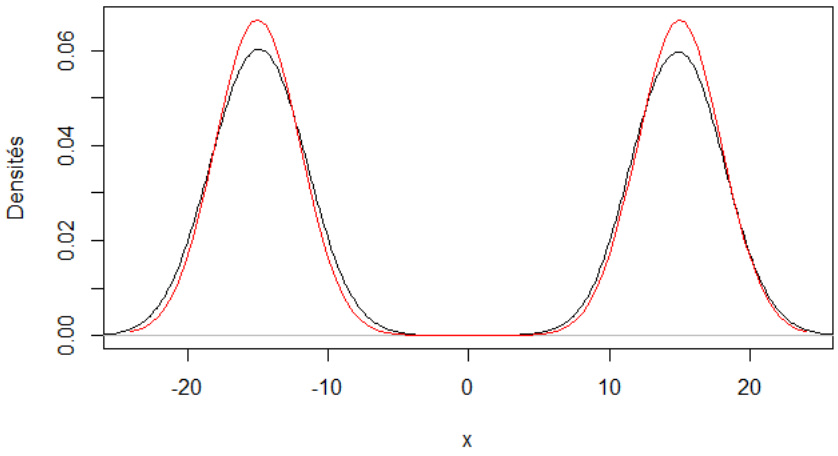

FIGURE 1.7. Estimation de la densité bimodale correspondant à la distribution décrite en (1.3.3) à l'aide du nouvel algorithme.

Chapitre 2

OPTIMISATION D'ALGORITHMES RWM

2.1. Plusieurs critères d'optimalité

En pratique, l'optimisation de la performance d'un algorithme de Metropolis-Hastings est difficilement réalisable et mène souvent à une solution qui n'est pas unique. La qualité d'un algorithme peut être évaluée par plusieurs mesures. En effet, si on se base sur le biais et la variance d'un estimateur spécifique comme mesure de performance, alors il existe un très grand nombre de mesures d'efficacité différentes, dépendant de l'estimateur qui nous intéresse. De plus si on s'intéresse à plus d'un estimateur, alors il serait potentiellement intéressant d'utiliser une mesure qui soit valide en plus grande généralité.

L'un de ces critères plus généraux mesure le «déplacement» de l'algorithme, c'est-à-dire à quel point la chaîne explore rapidement le domaine de la distribution cible. Il en découle le critère de déplacement moyen, qui correspond à maximiser

$$\mathbb{E}\left[\|X(n+1) - X(n)\|\right],$$

où on suppose que $X(n)$ est distribuée selon la distribution cible π_d, $X(n+1)$ est le prochain état de la chaîne généré à partir de $X(n)$ et $\|\cdot\|$ est une certaine norme. La quantité qu'on souhaite maximiser est donc l'amplitude moyenne de chaque incrément, en supposant que l'algorithme ait convergé, qui peut être vue comme une mesure du déplacement de la chaîne sur le domaine.

Une seconde mesure générale est l'autocorrélation de la chaîne. En effet, si les autocorrélations de la chaîne sont petites en valeur absolue, on voit que la variance de l'estimateur de Monte-Carlo (1.1.1), qui n'est qu'une moyenne échantillonnale, se rapproche de la variance

qu'on aurait si on était en présence d'un échantillon iid :

$$\mathbb{V}\mathrm{ar}\left(\hat{\mu}(h)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}\left(h(X(i))\right) + \frac{2}{n^2}\sum_{i<j}\sum\mathbb{C}\mathrm{ov}\left(h(X(i)); h(X(j))\right).$$

On obtient ainsi le critère d'autocorrélation qui consiste à minimiser, pour un n fixe, le second terme de cette variance. Or, en utilisant la version multivariée de la méthode delta et en supposant que l'algorithme ait convergé suffisamment pour que $\mathbb{E}\left[X(i)\right] \approx \mathbb{E}\left[X(j)\right]$, on trouve

$$\mathbb{C}\mathrm{ov}\left(h(X(i)); h(X(j))\right) \approx \left(h'(\mathbb{E}\left[X(i)\right])\right)^2 \mathbb{C}\mathrm{ov}\left(X(i); X(j)\right).$$

Dû à la construction de l'algorithme RWM (et également de plusieurs autres méthodes MCMC), les corrélations de la chaîne de Markov engendrée ne seront traditionnellement pas négatives. Le critère consiste donc à minimiser la somme des covariances $\mathbb{C}\mathrm{ov}\left(X(i); X(j)\right)$ en essayant de les annuler, ce qui est équivalent à minimiser les autocorrélations d'ordre 1 à n de la chaîne. Notons que si on se concentre sur l'autocorrélation d'ordre 1, on retombe sur un cas particulier du premier critère d'optimisation ; il a été démontré dans [**17**] que la minimisation de l'autocorrélation d'ordre 1 est équivalente à la maximisation du déplacement quadratique moyen.

Une dernière mesure d'efficacité plutôt spécifique d'un algorithme est la précision de l'estimation qui en découle. Supposons que l'on s'intéresse uniquement à l'espérance

$$\mu(h) = \mathbb{E}\left[h(X)\right],$$

où X est distribué selon la distribution cible π_d. Définissons la précision de l'estimateur de Monte Carlo par

$$\left(\mathbb{E}\left[\|\hat{\mu}(h) - \mu(h)\|\right]\right)^{-1}, \tag{2.1.1}$$

où $\{X(i) : 1 \leq i \leq n\}$ est l'échantillon obtenu en supposant que $X(0)$ est distribué selon π_d (c'est-à-dire après la période de *burn in*) et $\|\cdot\|$ est une certaine norme. Le critère associé consiste donc à maximiser (2.1.1).

Dans ces trois cas, pour ne nommer que ceux-là, on obtiendrait des solutions bien différentes, qui seraient assurément fonction de la norme choisie. Cela nous prive d'une méthode unique. Voyons un exemple qui montre que l'optimisation en pratique peut, de plus, être difficile à réaliser.

Exemple 2.1.1. *Supposons que l'on souhaite utiliser l'algorithme RWM pour obtenir un échantillon de la distribution $N_2(\mu; \Sigma)$, où $\mu \in \mathbb{R}^2$ et $\Sigma \in \mathbb{R}^{2\times 2}$ est symétrique définie positive. On propose les incréments à l'aide d'une loi $N_2\left(0; \sigma^2 I_2\right)$, où I_2 désigne la matrice identité de format 2×2. On choisit d'utiliser le critère de déplacement quadratique moyen, donc on*

cherche la valeur de σ qui maximisera

$$\mathbb{E}\left[\|\mathbf{X}(n+1)-\mathbf{X}(n)\|_2^2\right],$$

où $\mathbf{X}(n) \sim N_2(\mu;\Sigma)$ et où $\|\cdot\|_2$ désigne la norme euclidienne. La loi conditionnelle de $\mathbf{X}(n+1) \mid \mathbf{X}(n)$ est la suivante :

$$\mathbf{X}(n+1) = \begin{cases} \mathbf{Y}, \text{ avec probabilité } \alpha(\mathbf{X}(n),\mathbf{Y}) \\ \mathbf{X}(n), \text{ avec probabilité } 1-\alpha(\mathbf{X}(n),\mathbf{Y}), \end{cases}$$

où

$$\mathbf{Y} \mid \mathbf{X}(n) \sim N_2\left(\mathbf{X}(n);\sigma^2 I_2\right).$$

Notre mesure d'efficacité devient donc

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{X}(n+1)-\mathbf{X}(n)\|_2^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\mathbf{X}(n+1)-\mathbf{X}(n)\|_2^2 \middle| \mathbf{X}(n),\mathbf{Y}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|\mathbf{Y}-\mathbf{X}(n)\|_2^2 \cdot \mathbb{I}\{\mathbf{X}(n+1)=\mathbf{Y}\} \middle| \mathbf{X}(n),\mathbf{Y}\right]\right] \\
&= \mathbb{E}\left[\|\mathbf{Y}-\mathbf{X}(n)\|_2^2 \cdot \mathbb{P}\left(\mathbf{X}(n+1)=\mathbf{Y} \middle| \mathbf{X}(n),\mathbf{Y}\right)\right] \\
&= \mathbb{E}\left[\|\mathbf{Y}-\mathbf{X}(n)\|_2^2\, \alpha(\mathbf{X}(n),\mathbf{Y})\right],
\end{aligned}$$

avec

$$\alpha(\mathbf{X}(n),\mathbf{Y}) = \left(1 \wedge \exp\left\{\frac{1}{2}\left((\mathbf{X}(n)-\mu)^\top \Sigma^{-1}(\mathbf{X}(n)-\mu) - (\mathbf{Y}-\mu)^\top \Sigma^{-1}(\mathbf{Y}-\mu)\right)\right\}\right).$$

Le problème d'optimisation possède une solution, mais celle-ci n'est pas analytiquement simple puisque la fonction à maximiser en σ est de la forme

$$\begin{aligned}
&\frac{1}{\sigma^2}\int_{\mathbb{R}^4}\left((y_1-x_1)^2+(y_2-x_2)^2\right)\left(1 \wedge \exp\left\{\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\left(\Sigma^{-1}\right)_{ij}((x_i-\mu_i)(x_j-\mu_j)\right.\right. \\
&\left.\left.-(y_i-\mu_i)(y_j-\mu_j))\right\}\right)\exp\left\{-\frac{1}{2}\sum_{i=1}^{2}\left(\sum_{j=1}^{2}\left(\Sigma^{-1}\right)_{ij}(x_i-\mu_i)(x_j-\mu_j)+\frac{(y_i-x_i)^2}{\sigma^2}\right)\right\} d\mathbf{x}\,d\mathbf{y}.
\end{aligned}$$

L'approximation numérique demanderait la résolution d'intégrales multidimensionnelles, ce qui devient très coûteux en terme de temps de calcul lorsque la dimension croît.

2.2. Optimisation de l'algorithme RWM classique

Dans le cas de l'algorithme RWM dont la distribution instrumentale est normale et formée de composantes iid, c'est-à-dire dont les incréments sont générés selon une loi $N_d(0;\sigma^2 I_d)$, l'optimisation est relativement simple, puisque le seul paramètre à choisir est l'écart-type

instrumental σ. Intuitivement, il est logique de croire qu'il existe une valeur optimale dans $(0, \infty)$. En effet, comme on l'a vu dans l'exemple du précédent chapitre, un écart-type trop petit causera de petits incréments et donc une faible exploration du support. De même, un écart-type trop grand génèrera des candidats éloignés du mode de la distribution, c'est-à-dire où la densité cible sera très faible, et qui seront donc rejetés presqu'à coup sûr (on se rappelle que la probabilité d'acceptation fait intervenir le rapport des densités $\pi(y)/\pi(x)$). Malgré cela, comme on l'a vu dans l'exemple précédent, trouver la valeur optimale de σ, en supposant qu'il y en ait une et une seule, n'est pas trivial.

2.2.1. Le cas iid

Pour pallier à ce problème, [20] ont trouvé une solution plutôt élégante. Ils ont considéré l'algorithme RWM avec distribution cible iid

$$\pi_d\left(\mathbf{x}^{(d)}\right) = \prod_{i=1}^{d} f(x_i),$$

où f respecte quelques conditions de régularité, avec distribution instrumentale

$$\mathbf{Y}^{(d)} \mid \mathbf{X}^{(d)} \sim N_d\left(\mathbf{X}^{(d)}; \frac{l^2}{d} I_d\right),$$

où $\mathbf{X}^{(d)}$ désigne l'état actuel de la chaîne, et avec la probabilité d'acceptation usuelle

$$\alpha(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}) = 1 \wedge \frac{\pi_d\left(\mathbf{y}^{(d)}\right)}{\pi_d\left(\mathbf{x}^{(d)}\right)}.$$

La seule modification apportée ici est que la variance instrumentale est maintenant proportionnelle à l'inverse de la dimension d. Ils ont ensuite fait croître d vers l'infini, ce qui a pour effet de diminuer l'envergure des pas, et ils ont accéléré la chaîne par un facteur de d (c'est-à-dire qu'ils écrasent celle-ci de façon à ce qu'elle se déplace d fois plus rapidement). Ils ont démontré que la chaîne de Markov engendrée par l'algorithme converge alors vers un certain processus stochastique. Le sens du mot «converge» sera précisé dans la prochaine section. Le processus limite est ce qu'on appelle une diffusion continue de Langevin. Il s'agit en quelque sorte de la généralisation d'un mouvement brownien.

Définition 2.2.1. *Soit $W : \Omega \times [0, \infty) \mapsto \mathbb{R}$ un mouvement brownien standard. Soient $\mu : \Omega \times ([0, \infty) \times \mathbb{R}) \mapsto \mathbb{R}$ un processus stochastique défini sur le même espace de probabilité que W et σ, v des constantes strictement positives. Le processus stochastique Z, défini sur Ω*

par l'équation différentielle stochastique

$$dZ(t) = \sqrt{v\sigma^2} \cdot dW(t) + \frac{v}{2} \cdot \mu(t, Z(t))dt, \tag{2.2.1}$$

est un processus de diffusion de Langevin. Le processus μ, qui doit satisfaire quelques conditions techniques, est appelé sa fonction de dérive, σ est sa volatilité et v est sa mesure de vitesse.

Or, la fonction de dérive et la volatilité du processus de Langevin obtenu par [20] sont indépendantes du paramètre l. Seule sa mesure de vitesse, $v(l)$, en dépend. Cette mesure de vitesse peut donc être considérée comme l'unique mesure de la performance du processus limite. En effet, on peut montrer que si Z est un processus de diffusion de Langevin de vitesse 1, un processus de diffusion de Langevin U de vitesse $\gamma > 0$ est défini par $U(t) = Z(\gamma t)$. Ainsi, si $\gamma_1 > \gamma_0$, alors toute l'exploration du domaine que fait un processus de Langevin de vitesse γ_0 dans un intervalle de longueur t est effectuée en moyenne dans un intervalle de longueur $\gamma_0 t/\gamma_1 < t$ par un processus de Langevin de vitesse γ_1.

Ainsi, une stratégie logique consiste à maximiser la mesure de vitesse $v(l)$ du processus de diffusion de Langevin obtenu. Cette dernière est donnée par

$$v(l) = 2l^2\Phi\left(-\frac{l\sqrt{B}}{2}\right),$$

où $B = \mathbb{E}\left[((\log f)'(X_i))^2\right]$ est une mesure de rugosité de la densité f. Elle possède un unique maximum au point

$$\hat{l} \approx \frac{2{,}38}{\sqrt{B}}.$$

Il nous suffit donc de choisir $l = \hat{l}$ pour que la vitesse d'exploration de notre algorithme soit optimale. Bien sûr, la constante B est parfois difficile à calculer, dépendant de la complexité de la densité f considérée. Or, [20] ont également démontré que le taux d'acceptation moyen asymptotique est

$$\lim_{d\to\infty} \mathbb{E}\left[\alpha(\mathbf{X}^{(d)}, \mathbf{Y}^{(d)})\right] = 2\Phi\left(-\frac{l\sqrt{B}}{2}\right)$$

qui est une fonction strictement décroissante de l. Ils en concluent alors que le taux d'acceptation asymptotique optimal est

$$2\Phi\left(-\frac{\hat{l}\sqrt{B}}{2}\right) \approx 2\Phi(-1{,}19) \approx 0{,}234,$$

et ce, peu importe la valeur de B. De plus, par monotonicité, aucune valeur de $l \neq \hat{l}$ ne produit le même taux d'acceptation. Il suffit donc de simuler l'algorithme pendant quelques

itérations en essayant différentes valeurs de l et de calculer la proportion des propositions qui sont acceptées, jusqu'à ce que celle-ci se rapproche suffisamment de 0,23. On sait alors qu'on a trouvé le paramètre l optimal ; on est prêt à lancer la vraie simulation. Notons que bien que ce résultat soit valide asymptotiquement, c'est-à-dire pour des distributions cibles de grandes dimensions, il donne généralement de bons résultats en pratique pour des distributions cibles ayant un nombre relativement petit de composantes (10 et plus).

2.2.2. Le cas presque iid

En étudiant la démonstration de ce résultat, on voit facilement qu'il reste vrai si la distribution cible comporte quelques composantes qui ne sont pas distribuées selon la densité f, en autant que ces dernières soient en nombre fini même lorsque $d \to \infty$. Puisqu'elles sont en nombre fini, elles ne devraient avoir aucune importance par rapport aux densités identiques f qui sont, elles, en nombre infini lorsque $d \to \infty$. Par exemple, une distribution cible de la forme

$$\pi_d\left(\mathbf{x}^{(d)}\right) = \prod_{i=1}^{k} f_i(x_i) \prod_{i=k+1}^{d} f(x_i),$$

où $k < d$ est constant par rapport à d, ne pose pas problème si on impose sur les f_i des conditions de régularité semblables à celles imposées sur f. En posant $k = 1$, par exemple, on retrouve la distribution cible donnée par (1.4.1).

2.3. Convergence de la chaîne engendrée par le nouvel algorithme

Nous allons maintenant présenter les deux principaux résultats de ce mémoire, soit les théorèmes 2.3.1 et 2.3.2. Remarquons que ces derniers sont analogues à celui de [**20**].

2.3.1. Légère reformulation de l'algorithme

Supposons que la distribution instrumentale π_d est définie par (1.4.1), où les densités f_i sont des fonctions C^2 telles que $(\log f_i)'$ est continue au sens de Lipschitz et telles que, si X_i est distribué selon f_i, $(\log f_i)'(X_i) \in L^4$ et $\frac{f_i''(X_i)}{f_i(X_i)} \in L^2$, pour $i \in \{1, 2\}$. Autrement dit, nous assumons que les fonctions f_i'' existent et sont continues, et que les espérances

$$\mathbb{E}\left[\left((\log f_i)'(X_i)\right)^4\right]$$

et

$$\mathbb{E}\left[\left(\frac{f_i''(X_i)}{f_i(X_i)}\right)^2\right]$$

sont finies. Il est suffisant de n'imposer des conditions que sur f_1 et f_2 puisqu'on a supposé que $f_2 = f_3 = ... = f_d$. Supposons que la distribution instrumentale est définie comme dans la section 1.4, à l'exception qu'on laisse la variance instrumentale et le paramètre p dépendre de la dimension d, en posant $\sigma^2 = \frac{l^2}{d}$ et $p = 1 \wedge \frac{\beta}{d}$, avec l, β des paramètres strictement positifs.

Remarquons que d'un point de vue pratique, ces modifications n'influencent pas l'algorithme. En effet, dans un contexte donné, on ne s'intéresse pas à l'évolution de d ; si on a $d = 10$, choisir, par exemple, $l^2 = 2$ et $\beta = 0{,}6$ est tout à fait équivalent à choisir $\sigma^2 = 0{,}2$ et $p = 0{,}06$.

2.3.2. Convergence faible

Soit $\mathbf{X}^{(d)}$ la chaîne de Markov engendrée par l'algorithme et $X_j^{(d)}$ la chaîne engendrée par la j-ème composante de l'algorithme. De plus, supposons que le point de départ $\mathbf{X}^{(d)}(0)$ est distribué selon la distribution cible π_d. Comme on l'a vu, les pas effectués par la chaîne $\mathbf{X}^{(d)}$ sont de plus en plus petits lorsque d croît. Afin que sa limite lorsque $d \to \infty$ ne soit pas simplement une fonction constante, on va l'accélérer. Soit $\mathbf{Z}^{(d)}$ le processus défini par

$$\mathbf{Z}^{(d)}(t) = \mathbf{X}^{(d)}(\lfloor dt \rfloor), \tag{2.3.1}$$

où $\lfloor \cdot \rfloor$ désigne la fonction plancher, et définissons sa j-ème composante, $Z_j^{(d)}$, de façon similaire. $\mathbf{Z}^{(d)}$ est la version accélérée de $\mathbf{X}^{(d)}$. Ainsi, lorsque $d \to \infty$, $\mathbf{Z}^{(d)}$ effectue une infinité de pas dans chaque intervalle de temps, mais ceux-ci sont infiniment petits. Il serait donc logique de croire que lorsque $d \to \infty$, le processus accéléré, qui bouge de plus en plus souvent mais qui effectue des pas de plus en plus petits, devrait «se rapprocher» d'un processus continu. Or, si $\mathbf{Z}^{(d)}$ converge, il converge vers un processus infini-dimensionnel. Pour cette raison, on va étudier le comportement asymptotique des composantes $Z_j^{(d)}$ de façon individuelle.

Pour $j \neq 1$, le processus limite est en effet continu : il s'agit du même processus de diffusion de Langevin obtenu par [20] dans le cas iid. Pour $j = 1$, on verra que le processus limite est également une diffusion de Langevin dont le terme de dérive, par contre, est légèrement différent de celui qu'on connaît puisqu'il dépend de la nouvelle densité bimodale f_1. De plus, au processus limite formé par la première composante, c'est-à-dire à la limite de $Z_1^{(d)}$, sont ajoutés des sauts générés par un certain algorithme de Metropolis-Hastings dont

la fréquence des pas proposés correspond à un processus de Poisson de taux β. Ces sauts proviennent en réalité des grands pas qui sont proposés par l'algorithme avec probabilité $\frac{\beta}{d}$ (lorsque d est grand). Le fait d'accélérer l'algorithme annule donc le fait que ces grands pas sont proposés de moins en moins souvent.

Dénotons par $\Longrightarrow$ la convergence faible de processus stochastiques dans la topologie de Skorokhod. Pour une définition rigoureuse, le lecteur est référé à [5] ou encore à [10]. Le théorème 7.8 du chapitre 3 de [10] établit que ce type de convergence implique, entre autres, la convergence en distribution de n'importe quel sous-ensemble fini de termes de chaque processus vers le sous-ensemble correspondant de termes du processus limite. Le théorème suivant établit la limite faible dans la topologie de Skorokhod du processus $Z_1^{(d)}$.

Théorème 2.3.1. *Soit $\mathbf{Z}^{(d)}$ le processus accéléré défini par (2.3.1) et supposons que la distribution cible et la distribution instrumentale sont définies tel que dans la section 2.3.1. Soit Z_{ML} un processus de diffusion de Langevin qui respecte l'équation différentielle stochastique (2.2.1), de volatilité 1, de vitesse $v(l)$, de fonction de dérive donnée par*

$$\mu(t, Z_{ML}(t)) = (\log f_1)'(Z_{ML}(t))$$

et dont le point de départ $Z_{ML}(0)$ est distribué selon la densité f_1. Ici, W est un mouvement Brownien standard,

$$v(l) = 2l^2\Phi\left(-\frac{l\sqrt{B}}{2}\right) \tag{2.3.2}$$

et

$$B = \mathbb{E}\left[((\log f)'(X_2))^2\right].$$

Or, supposons que l'on crée des discontinuités dans le processus Z_{ML} de la façon suivante. Soit N un processus de Poisson de taux β. Supposons qu'à chaque point t où N saute, une variable aléatoire $Y(t)$ est générée selon la densité instrumentale $g_1(\cdot \mid Z_{ML}(t^-))$, où $Z_{ML}(t^-) = \lim_{s\to t^-} Z_{ML}(s)$, et supposons alors qu'avec probabilité $\alpha(l, Z_{ML}(t^-), Y(t))$, Z_{ML} effectue un saut vers $Y(t)$, c'est-à-dire qu'on pose $Z_{ML}(t) = Y(t)$. Sinon, Z_{ML} reste continu (et en fait constant) au point t, c'est-à-dire que $Z_{ML}(t) = Z_{ML}(t^-)$. Ici, la probabilité d'acceptation des sauts est donnée par

$$\alpha(l, x, y) = \Phi\left(\frac{A(x,y) - \frac{l^2}{2}B}{l\sqrt{B}}\right) + e^{A(x,y)}\Phi\left(\frac{-A(x,y) - \frac{l^2}{2}B}{l\sqrt{B}}\right), \tag{2.3.3}$$

où

$$A(x, y) = \log\left(\frac{f_1(y)}{f_1(x)}\right).$$

On a alors

$$Z_1^{(d)} \Longrightarrow Z_{ML}.$$

Remarque 2.3.1. *Le processus limite peut donc être vu comme une diffusion de Langevin combinée à la version continue d'un algorithme de type Metropolis-Hastings qui propose des candidats selon un processus de Poisson. La probabilité d'acceptation $\alpha(l, \cdot, \cdot)$ engendre bel et bien une chaîne de Markov qui respecte les conditions d'irréductibilité, d'apériodicité, de réversibilité et de récurrence, ce qui peut être facilement vérifié en utilisant une approche similaire à celle exposée dans le chapitre précédent.*

Nous présentons maintenant, dans le théorème suivant, la limite faible dans la topologie de Skorokhod des processus $Z_j^{(d)}$, pour $j \geq 2$.

Théorème 2.3.2. *Soit $\mathbf{Z}^{(d)}$ le processus accéléré défini par (2.3.1) et supposons que la distribution cible et la distribution instrumentale sont définies tel que dans la section 2.3.1. Soit Z_L un processus de diffusion de Langevin qui respecte l'équation différentielle stochastique (2.2.1), de volatilité 1, de vitesse $v(l)$, de fonction de dérive donnée par*

$$\mu(t, Z_L(t)) = (\log f)'(Z_L(t))$$

et dont le point de départ $Z_L(0)$ est distribué selon la densité f, où W et v sont définis comme dans le théorème 2.3.1. On a alors, pour $j \neq 1$,

$$Z_j^{(d)} \Longrightarrow Z_L.$$

L'idée de réduire la variance instrumentale par un facteur de d lorsqu'on accélère le processus n'est pas nouvelle. C'est, par exemple, une méthode utilisée pour construire le mouvement Brownien à partir d'une marche aléatoire. Or, on pourrait se demander pourquoi la proportion des pas proposés qui sont grands, p, est elle aussi réduite par un facteur de d. La raison est simple : cela nous permet d'accélérer le processus engendré par la première composante de l'algorithme afin d'en étudier le comportement lors des petits pas. En effet, l'amplitude des grands pas ne diminue pas avec d. Ainsi, si ceux-ci se produisaient une proportion de fois p fixe, lorsqu'on accélérerait le processus, la limite serait composée d'une infinité de grands pas dans chaque intervalle de temps et on perdrait le point de vue utilisé pour l'étude de la première composante. On ne pourrait donc pas accélérer le processus engendré par la première composante de l'algorithme. Or, cela équivaut à perdre toute l'information sur les petits pas, puisque ceux-ci deviennent de plus en plus petits lorsque d croît. De plus, si on laissait p fixe, le processus limite engendré par les $d-1$ autres composantes de l'algorithme serait toujours une diffusion de Langevin, mais dont la mesure de vitesse dépendrait des grands pas, ce qui la rendrait difficile à optimiser (voir [2]). Ajoutons finalement que d'un point de vue intuitif, il est tout à fait raisonnable de réduire la proportion des grands pas par un facteur de d. En effet, plus les composantes iid sont nombreuses

par rapport à la composante bimodale, plus la proportion de petits pas devra être grande afin d'étudier efficacement ces $d-1$ composantes à l'intérieur de chacun des modes.

2.4. Optimisation du nouvel algorithme

En se basant sur les résultats théoriques énoncés dans la section précédente, nous tentons maintenant de déterminer les valeurs des paramètres pour lesquelles le nouvel algorithme sera le plus efficace possible du point de vue de l'exploration de son domaine.

2.4.1. La stratégie proposée

Rappelons que trois choix s'imposent pour implanter cet algorithme : le choix du paramètre d'échelle des petits pas l, celui de la probabilité de proposer un grand pas p (qui peut se faire à travers le choix du paramètre β) et finalement le choix de la densité instrumentale g_1. Ces choix devraient être effectués dans un but d'optimisation de la performance des processus limites. Ceci se fait en deux temps : l'optimisation de l'exploration locale (représentée par la mesure de vitesse du processus limite continu) et la maximisation de la fréquence des changements de mode (représentés par les sauts du processus limite). Or, contrairement au cas iid, nous constaterons qu'il n'existe pas d'unique critère de performance et donc pas d'unique solution à ce problème d'optimisation. Par conséquent, nous proposons, dans cette section, une stratégie pour les choix de l, g_1 et β qui repose en partie sur les processus limites, mais également sur le processus discret.

2.4.1.1. *Le choix du paramètre d'échelle pour les petits pas*

D'abord, comme dans le cas iid, la mesure de vitesse v donnée par (2.3.2) est le seul élément du processus continu qui dépend de l et elle possède un unique maximum au point $\hat{l} \approx \frac{2{,}38}{\sqrt{B}}$. Or, on aimerait également maximiser la probabilité d'acceptation limite des grands pas, $\alpha(l,\cdot,\cdot)$, donnée par (2.3.3), afin de favoriser les changements de mode. Dans cette optique, nous étudions donc le comportement de cette nouvelle probabilité d'acceptation en fonction de l. Il s'avère que, pour des valeurs fixées de x et y, $\alpha(l,x,y)$ est une fonction décroissante de l. En effet, on peut calculer que

$$\frac{d}{dl}\alpha(l,x,y) = \frac{A(x,y)}{l^2\sqrt{B}}\left(e^{A(x,y)}\phi\left(\frac{-A(x,y)-\frac{l^2}{2}B}{l\sqrt{B}}\right) - \phi\left(\frac{A(x,y)-\frac{l^2}{2}B}{l\sqrt{B}}\right)\right)$$

$$-\frac{\sqrt{B}}{2}\left(\phi\left(\frac{A(x,y)-\frac{l^2}{2}B}{l\sqrt{B}}\right)+e^{A(x,y)}\phi\left(\frac{-A(x,y)-\frac{l^2}{2}B}{l\sqrt{B}}\right)\right),$$

où ϕ représente la densité de la loi normale standard. En utilisant le fait que $\phi(t) = (2\pi)^{-\frac{1}{2}}\exp\left\{\frac{-t^2}{2}\right\}$, on remarque que

$$\phi\left(\frac{A(x,y)-\frac{l^2}{2}B}{l\sqrt{B}}\right)=e^{A(x,y)}\phi\left(\frac{-A(x,y)-\frac{l^2}{2}B}{l\sqrt{B}}\right),$$

ce qui nous permet de conclure que

$$\frac{d}{dl}\alpha(l,x,y)=-\sqrt{B}\phi\left(\frac{A(x,y)-\frac{l^2}{2}B}{l\sqrt{B}}\right)<0. \tag{2.4.1}$$

Ainsi, aucune valeur de $l>0$ ne serait optimale en ce qui concerne la probabilité d'acceptation. Cela s'explique par le fait que, même lorsqu'un grand pas est proposé pour la première composante de l'algorithme, l'acceptation dépend largement des pas proposés pour les $d-1$ autres composantes. Plus l est grand, plus ces propositions seront grandes en moyenne, et donc plus elles seront difficiles à accepter. Pour x, y fixés, la plus grande valeur possible pour la probabilité d'acceptation limite du pas de x vers y est

$$\lim_{l\to 0}\alpha(l,x,y)=1\wedge e^{A(x,y)},$$

qui correspond à la probabilité d'acceptation habituelle de l'algorithme RWM pour la densité cible unidimensionnelle f_1. Ce n'est pas surprenant, puisque poser $l=0$ ferait en sorte que les pas proposés pour les composantes 2 à d seraient nuls, donc le rapport $\frac{f(Y_i)}{f(X_i)}$ serait égal à 1 pour tout $i\geq 2$. La probabilité d'acceptation ne dépendrait alors que de la première composante et faire tendre d vers ∞ n'y changerait rien.

Évidemment, fixer $l=0$ forcerait les composantes $2, ..., d$ à demeurer immobiles et détruirait la convergence de la chaîne. L'approche à favoriser dans ce cas consiste à fixer $l=\hat{l}$, le choix optimal pour le déplacement des $d-1$ dernières composantes, et par la suite à tenter de maximiser le taux de changements de mode à travers le choix de g_1. En utilisant $\hat{l}$, la valeur qui maximise la vitesse du processus continu, on obtient comme probabilité d'acceptation limite

$$\alpha(\hat{l},x,y)=\Phi\left(\frac{A(x,y)}{2{,}38}-1{,}19\right)+e^{A(x,y)}\Phi\left(\frac{-A(x,y)}{2{,}38}-1{,}19\right). \tag{2.4.2}$$

On remarque premièrement que, puisque $\hat{l}\propto B^{-\frac{1}{2}}$, ce choix élimine le paramètre B dans l'expression de la probabilité d'acceptation asymptotique, ce qui nous permet de poursuivre cette étude sans se soucier de la forme de la densité f. C'est un avantage considérable. La «perte» en terme de la probabilité d'acceptation limite si on pose $l=\hat{l}$, par rapport à ce qui

se passe lorsque $l \to 0$, peut être représentée par le rapport

$$\begin{aligned}\frac{\alpha(\hat{l},x,y)}{\lim_{l\to 0}\alpha(l,x,y)} &= \Phi\left(\frac{|A(x,y)|}{2{,}38}-1{,}19\right)+e^{|A(x,y)|}\Phi\left(\frac{-|A(x,y)|}{2{,}38}-1{,}19\right) \\ &= r(|A(x,y)|),\end{aligned} \tag{2.4.3}$$

avec

$$r(t) \equiv \Phi\left(\frac{t}{2{,}38}-1{,}19\right)+e^{t}\Phi\left(\frac{-t}{2{,}38}-1{,}19\right).$$

À l'aide d'un calcul similaire à ce qui a été fait pour obtenir (2.4.1), on a

$$\begin{aligned}\frac{d}{dt}r(t) &= \frac{1}{2{,}38}\phi\left(\frac{t}{2{,}38}-1{,}19\right)+e^{t}\Phi\left(\frac{-t}{2{,}38}-1{,}19\right)-\frac{e^{t}}{2{,}38}\phi\left(\frac{-t}{2{,}38}-1{,}19\right) \\ &= e^{t}\Phi\left(\frac{-t}{2{,}38}-1{,}19\right) \\ &> 0.\end{aligned}$$

Le rapport (2.4.3) étant croissant en $|A(x,y)|$, il est donc borné inférieurement par $r(0) = 2\Phi(-1{,}19) \approx 0{,}234$. La stratégie proposée, et celle qui semble la plus élégante, est donc d'utiliser $\hat{l}$. Ainsi, le taux d'acceptation limite des grands pas ne sera jamais plus de quatre fois inférieur au taux qu'on obtiendrait avec une valeur de $l < \hat{l}$. Cette perte est loin d'être catastrophique. Comme dans le cas iid, la quantité $\hat{l}$ dépend de B qui est souvent inconnu. Or, pour la «trouver», il suffit de simuler l'algorithme RWM (ou encore, de façon équivalente, le nouvel algorithme avec $p = 0$) pour un petit nombre d'itérations et de choisir l afin que le taux d'acceptation se rapproche de 0,234.

2.4.1.2. *Le choix de la densité instrumentale pour les grands pas*

Étant donné $l = \hat{l}$, nous souhaitons maintenant optimiser la probabilité d'effectuer un changement de mode à travers le choix de la densité instrumentale g_1. En pratique, le choix de cette densité sera assez arbitraire. En effet, un changement de mode implique premièrement la proposition d'un candidat dans l'autre mode, et par la suite l'acceptation de ce candidat à l'aide de la probabilité $\alpha(\hat{l},\cdot,\cdot)$. Or, l'atteinte de ce but dépend entièrement de la densité cible f_1 et de la définition de changement de mode (que nous n'avons pas spécifiée). Dans ce contexte, nous nous restreignons alors aux choix de g_1 qui respectent les deux conditions énoncées à la section 1.4 afin de guider notre choix. On impose donc que $g_1(\cdot \mid X_1)$ ne soit pas croissante (mais pas nécessairement strictement décroissante) lorsqu'on s'éloigne de l'état actuel X_1, et que le mode de $g_1(\cdot \mid X_1)$ au point X_1 ne soit pas trop élevé. Tel que mentionné dans le chapitre précédent, parmi les densités g_1 symétriques qui respectent la condition de

non-croissance, celle qui maximise la probabilité de proposer un changement de mode est la densité de la loi uniforme centrée à l'état actuel X_1, pourvu que le choix du paramètre d'échelle c soit bon.

En général, les modes d'une distribution bimodale sont entourés d'une région où la densité est suffisamment haute. C'est dans cette région que l'on veut proposer les grands pas afin de changer de mode. Supposons que les régions de haute densité entourant les modes de la densité f_1 sont $S_1 = (a_1; b_1)$ et $S_2 = (a_2; b_2)$, avec $a_1 < b_1 < a_2 < b_2$. Supposons que l'état actuel de la première composante de la chaîne est $X_1 = x$ et supposons $x \in S_1$, mais notons que le raisonnement qui suit est possible pour $x \in S_2$. On souhaite proposer un changement de mode, c'est-à-dire une valeur de $Y_1 \in S_2$. La probabilité que ceci se produise est alors

$$\mathbb{P}\left(Y_1 \in S_2 | \, X_1 = x\right) = \int_{a_2}^{b_2} g_1(y \mid x) \, dy. \tag{2.4.4}$$

Proposition 2.4.1. *Parmi toutes les densités g_1 symétriques en x qui respectent la condition de non-croissance ci-haut, celle qui maximise la probabilité (2.4.4) est celle de la loi uniforme sur l'intervalle $[x \pm c]$, avec $c = b_2 - x$.*

Démonstration. Soit g_1 une densité symétrique en x qui respecte la condition de non-croissance. Cette condition implique que, sur l'intervalle $(x; \infty)$, la fonction $g_1(\cdot \mid x)$ n'est pas croissante. Ainsi,

$$\frac{1}{b_2 - a_2} \int_{a_2}^{b_2} g_1(y \mid x) \, dy \leq \frac{1}{b_2 - a_2} \int_{a_2}^{b_2} g_1(a_2 \mid x) \, dy = g_1(a_2 \mid x). \tag{2.4.5}$$

De même,

$$\frac{1}{a_2 - x} \int_{x}^{a_2} g_1(y \mid x) \, dy \geq \frac{1}{a_2 - x} \int_{x}^{a_2} g_1(a_2 \mid x) \, dy = g_1(a_2 \mid x). \tag{2.4.6}$$

En combinant (2.4.5) et (2.4.6), nous obtenons

$$\frac{1}{b_2 - a_2} \int_{a_2}^{b_2} g_1(y \mid x) \, dy \leq \frac{1}{a_2 - x} \int_{x}^{a_2} g_1(y \mid x) \, dy,$$

qui est équivalent à

$$\int_{a_2}^{b_2} g_1(y \mid x) \, dy \leq \frac{b_2 - a_2}{a_2 - x} \int_{x}^{a_2} g_1(y \mid x) \, dy. \tag{2.4.7}$$

Remarquons maintenant que, par symétrie de la densité $g_1(\cdot \mid x)$ autour de x, on doit avoir

$$\int_{x}^{a_2} g_1(y \mid x) \, dy + \int_{a_2}^{b_2} g_1(y \mid x) \, dy + \int_{b_2}^{\infty} g_1(y \mid x) \, dy = \frac{1}{2}.$$

En utilisant ceci, (2.4.7) devient

$$\int_{a_2}^{b_2} g_1(y \mid x)\ dy \leq \frac{b_2 - a_2}{a_2 - x}\left(\frac{1}{2} - \int_{a_2}^{b_2} g_1(y \mid x)\ dy - \int_{b_2}^{\infty} g_1(y \mid x)\ dy\right).$$

On remarque que le terme $\int_{a_2}^{b_2} g_1(y \mid x)\ dy$ se répète dans cette inégalité. En l'isolant, on trouve

$$\left(1 + \frac{b_2 - a_2}{a_2 - x}\right)\int_{a_2}^{b_2} g_1(y \mid x)\ dy \leq \frac{b_2 - a_2}{a_2 - x}\left(\frac{1}{2} - \int_{b_2}^{\infty} g_1(y \mid x)\ dy\right) \leq \frac{1}{2}\frac{b_2 - a_2}{a_2 - x}.$$

On peut donc finalement borner la probabilité qui nous intéresse :

$$\begin{aligned}\int_{a_2}^{b_2} g_1(y \mid x)\ dy &\leq \frac{1}{2}\left(1 + \frac{b_2 - a_2}{a_2 - x}\right)^{-1}\frac{b_2 - a_2}{a_2 - x} \\ &= \frac{1}{2}\left(\frac{b_2 - x}{a_2 - x}\right)^{-1}\frac{b_2 - a_2}{a_2 - x} \\ &= \frac{1}{2}\frac{b_2 - a_2}{b_2 - x}. \end{aligned} \tag{2.4.8}$$

Or, remarquons que si g_1 est la densité de la loi uniforme sur l'intervalle $[x \pm (b_2 - x)]$, alors on a, pour $y \in (a_2; b_2)$,

$$g_1(y \mid x) = \frac{1}{2(b_2 - x)}$$

et la borne (2.4.8) est donc atteinte.

□

Remarquons d'abord que la borne supérieure pour la probabilité de proposer un changement de mode à partir de $x \in S_1$ vers S_2, donnée par (2.4.8), ne dépend de la densité cible f_1 qu'à travers le rapport $\lambda_{x,2}(f_1) = \frac{b_2 - a_2}{b_2 - x}$. Ce rapport est semblable à

$$\lambda_{1,2}(f_1) = \frac{b_2 - a_2}{\frac{a_2 + b_2}{2} - \frac{a_1 + b_1}{2}},$$

qui représente en quelque sorte le rapport entre la largeur du deuxième mode, que l'on «vise», et la distance entre les deux modes. On peut définir, de façon similaire, $\lambda_{2,1}(f_1)$ comme le rapport entre la largeur du premier mode et la distance entre les deux modes. Outre les valeurs de $\lambda_{1,2}$ et $\lambda_{2,1}$, la forme exacte de f_1 n'a pas une grande influence sur la performance de l'algorithme.

La signification de ce résultat est que, même lorsqu'on choisit la meilleure famille de densités instrumentales pour les grands pas (la famille des densités uniformes), la performance de l'algorithme, en terme de la fréquence des changements de mode, dépend entièrement de la largeur des modes et de la distance qui les sépare. En effet, plus les modes sont étroits, plus la probabilité de les atteindre avec un pas distribué selon une loi uniforme est petite. De même, plus ils sont éloignés, plus la valeur de c qu'on utilise doit être grande, ce qui augmente nos chances de proposer un pas «dans le vide», c'est-à-dire dans la région de faible densité qui sépare les modes.

Bien sûr, la valeur optimale $c = b_2 - x$ dépend de l'état actuel de la chaîne ainsi que de la valeur b_2 ; notons que cette dernière est généralement difficile à définir. De plus, cette analyse ne tient pas compte de la probabilité d'acceptation ; elle repose plutôt sur la supposition que, dans les intervalles S_1 et S_2, la densité f_1 est suffisamment élevée pour que les pas dans ces régions soient facilement acceptés. Ce qu'il faut retenir est simplement qu'il devrait exister une valeur pour le paramètre d'échelle de g_1, c, pour laquelle la densité instrumentale uniforme nous donne une probabilité de changer de mode qui est environ optimale. De plus, cette valeur devrait être semblable à la distance qui sépare les deux modes, qui n'est pas nécessairement connue *a priori*. Nous allons donc proposer une méthode numérique, tout comme dans le cas de l, afin d'approximer la meilleure valeur de c.

Notons que, lors des simulations, nous allons comparer les densités instrumentales uniforme et normales. Nos résultats seront en faveur de la densité uniforme.

2.4.1.3. *Le choix de la probabilité de proposer un grand pas*

Pour ce qui est du paramètre β, on a vu dans le théorème 2.3.1 qu'il n'intervient dans le processus limite qu'en tant que taux du processus de Poisson qui génère les sauts. La maximisation du déplacement quadratique moyen de la première composante à travers le processus limite consisterait alors à choisir β le plus grand possible, ce qui signifierait, en pratique, $p = 1$. Or, en dimension finie, cette valeur de p n'est évidemment pas optimale, ce qui signifie que ce critère de déplacement quadratique moyen n'est pas idéal dans le contexte présent. Pour le voir, il suffit de supposer que les modes de f_1 sont en réalité très proches l'un de l'autre (ou même, à la limite, qu'il n'y a qu'un seul mode). Dans cette situation, l'algorithme RWM classique est très efficace et le fait de ne proposer, pour la première composante, que des grands pas dont l'amplitude ne dépend pas de d ne fait que ralentir le processus limite continu engendré par chacune des composantes (pour ce résultat, voir [2], théorème 3.2.1). Dans ce cas précis, il est alors évident que la vitesse du processus

limite est maximisée lorsque p approche 0. De même, lorsque la distance entre les modes croît vers ∞, la probabilité qu'un grand pas proposé soit dans le mode visé, puis qu'il soit accepté, décroît vers 0, donc la valeur optimale de p devrait approcher 1 afin d'obtenir suffisamment de changements de mode. On en déduit que la meilleure valeur de p ne devrait pas être systématiquement 1, mais devrait plutôt dépendre de la forme de la densité f_1. Plus précisément, elle ne devrait dépendre environ que des quantités $\lambda_{1,2}$ et $\lambda_{2,1}$ qui ont été définies précédemment et qui, elles, ne dépendent que de la distance entre les deux modes et de la largeur de chaque mode.

Nous présentons donc une règle empirique afin de choisir numériquement les valeurs de c et de p. Supposons que l'on souhaite implanter l'algorithme en présence d'une distribution cible de dimension fixée, disons d_0, et que l'on veuille obtenir un échantillon de taille n. L'espérance du nombre de changements de mode qu'on observera dans l'échantillon est alors npq, où q représente la probabilité, sachant que l'on utilise la densité instrumentale g_1, de proposer et d'accepter un changement de mode. Notons qu'ici, la densité g_1 est une uniforme avec paramètre d'échelle c. Pour une valeur de c donnée, cette probabilité peut être estimée facilement en utilisant une approche similaire à celle utilisée pour ajuster l : il suffit de simuler un algorithme RWM sur $\mathbb{R}$ pour un petit nombre d'itérations avec distribution cible f_1, distribution instrumentale g_1 et probabilité d'acceptation (2.4.2) puis de calculer le taux de changements de mode. Bien sûr, puisque nous n'avons pas défini rigoureusement ce qu'est un changement de mode, nous nous contentons pour le moment de mentionner qu'il suffit de calculer la proportion des incréments dont l'amplitude est supérieure à une certaine valeur, ou encore de définir les modes comme deux sous-ensembles qui partitionnent le domaine de f_1. La valeur de c qu'il faut choisir est celle qui maximise cette probabilité. Par la suite, si on note $\hat{q}$ l'estimation de q, on a que le nombre espéré de changements de mode dans notre échantillon est d'environ $np\hat{q}$. Il suffit donc de choisir p afin que cette espérance soit suffisamment grande. Par exemple, on verra que, pour les exemples numériques qu'on considèrera, observer 1 000 changements de mode semble amplement suffisant pour bien estimer la taille de chaque mode. Ainsi, dans cette situation, on posera $p = \frac{1\,000}{n\hat{q}}$. Or, remarquons que le nombre 1 000 dépend de la précision souhaitée, et aussi fort probablement de la pondération relative de chacun des deux modes. Il serait intéressant, pour de futures recherches, d'étudier (théoriquement ou numériquement) la relation entre la pondération relative des modes et le nombre de changements de mode requis afin d'obtenir un certain niveau de précision.

Notons que l'estimation $\hat{q}$ qu'on obtient devrait dépendre des rapports $\lambda_{1,2}$ et $\lambda_{2,1}$. En utilisant cette approche, notre choix pour la valeur de p sera donc également dépendant de ces deux rapports, tel que nous l'avons conjecturé.

2.4.2. Un critère plus objectif

Dans cette sous-section, nous présentons une idée, ou plutôt une avenue qui pourrait être explorée afin de dériver une stratégie d'optimisation un peu plus objective pour le nouvel algorithme. Cette stratégie est basée sur le critère de déplacement quadratique moyen déjà vu au début de ce chapitre. On cherche à maximiser le déplacement quadratique moyen du processus limite engendré par chaque composante sur un intervalle de temps de longueur, disons, h. Rappelons que le processus limite engendré par la première composante est donné par Z_{ML} tel que défini dans le théorème 2.3.1, et que le processus limite engendré par chacune des autres composantes est égal en loi à Z_L tel que défini dans le théorème 2.3.2. Le critère d'efficacité devra donc être une combinaison convexe du déplacement quadratique moyen de Z_{ML} et du déplacement quadratique moyen de Z_L. Or, quel poids devrait-on donner à chacune de ces deux quantités ?

En fait, comme on le sait, la première composante à elle seule peut faire échouer complètement l'exploration du domaine. En effet, si la première composante ne se déplace pas assez pour qu'on observe des changements de mode, l'algorithme n'explore que la moitié du domaine de la distribution cible. De même, on peut considérer que les composantes 2 à d forment un seul groupe, puisque la bonne exploration de toutes ces composantes dépend du même paramètre : l. Si l'une de ces $d-1$ composantes explore bien son support, on devine que c'est parce que l a été bien choisi, ce qui implique que les $d-2$ autres composantes explorent également bien leur support respectif. Il y a donc deux situations à éviter absolument : celle où la première composante ne se déplace pas assez pour changer de mode, et celle où l'exploration locale des $d-1$ autres composantes est trop lente. Puisque ces situations ne sont pas plus souhaitables l'une que l'autre, nous décidons d'accorder le même poids à l'efficacité de chacun des processus limites Z_{ML} et Z_L. La quantité à maximiser est donc

$$Q(l, \beta, g_1) = \mathbb{E}\left[(Z_{ML}(t+h) - Z_{ML}(t))^2\right] + \mathbb{E}\left[(Z_L(t+h) - Z_L(t))^2\right]. \tag{2.4.9}$$

Il s'avère que lorsque d croît vers l'infini, toutes les mesures d'efficacité d'un processus de diffusion de Langevin deviennent équivalentes et correspondent à la mesure de vitesse du processus limite. La deuxième espérance de (2.4.9) correspond donc à $v(l)$ tel que décrit en (2.3.2) et ne dépend que de l. Quant à la première espérance de (2.4.9), elle se décompose en deux parties : la mesure de vitesse décrite en (2.3.2) ainsi que le déplacement quadratique moyen de la portion discrète du processus asymptotique. Rappelons que cette portion discrète consiste à proposer des sauts selon la densité $g_1(\cdot \mid X_1)$, à une fréquence qui correspond à

un processus de Poisson de taux β, et qui sont acceptés avec probabilité $\alpha(l, X_1, \cdot)$, où X_1 désigne l'état actuel du processus.

La mesure de vitesse décrite en (2.3.2) est bien sûr maximisée si on choisit $l = \hat{l}$. De plus, en posant que g_1 est la densité d'une loi uniforme pour les raisons mentionnées précédemment, il ne reste qu'à choisir la valeur de p. Il suffit alors, pour une dimension d fixée, de maximiser numériquement la quantité $Q(l, \beta, g_1)$ en fonction de β (et donc de p). On peut par la suite analyser les résultats numériques obtenus pour différentes dimensions de la densité cible et essayer de dégager la tendance. Ceci constituerait évidemment la mesure d'efficacité globale la plus objective possible pour le contexte étudié. Cette avenue n'a pas pu être explorée durant la recherche qui a menée à ce mémoire. Toutefois, ce point sera abordé dans [4].

Chapitre 3

DÉMONSTRATION DES THÉORÈMES 2.3.1 ET 2.3.2

Ce chapitre, comme son titre l'indique, est entièrement dédié à la démonstration des théorèmes 2.3.1 et 2.3.2, qui établissent le comportement asymptotique de la chaîne de Markov engendrée par le nouvel algorithme lorsque la dimension du support de la distribution cible croît vers ∞. Bien que ce chapitre se veuille le plus autonome possible, la démonstration, fortement dépendante de quelques résultats fondamentaux sur la convergence des processus markoviens qui peuvent être trouvés dans [10], est en grande partie basée sur ce qui a été fait dans [3].

3.1. Générateurs

Pour démontrer la convergence faible (dans la topologie de Skorokhod) d'une suite de processus stochastiques univariés (rappelons qu'on considère chaque composante de l'algorithme de façon indépendante) vers un certain processus limite, on va représenter chaque processus par son générateur. Le générateur d'un processus markovien X est un opérateur qui, lorsqu'appliqué à une fonction test h, renvoie l'accroissement infinitésimal moyen de $h(X(t))$ par rapport à t.

Définition 3.1.1. *Soit X un processus markovien univarié à temps continu sur S et $h : S \mapsto \mathbb{R}$ une fonction test. On définit le générateur de X comme l'opérateur G défini par*

$$Gh(x) = \frac{d}{dt}\mathbb{E}\left[h\left(X(t)\right)\middle| X(s) = x\right]\Big|_{t=s}.$$

Rappelons qu'au chapitre 2, on a défini le processus accéléré $\mathbf{Z}^{(d)}$ par

$$\mathbf{Z}^{(d)}(t) = \mathbf{X}^{(d)}\left(\lfloor d \cdot t \rfloor\right)$$

et sa j-ème composante par $Z_j^{(d)}$. Cette version accélérée du processus ne préserve cependant pas la propriété markovienne. Dans ce qui suit, il sera pratique de travailler avec une version accélérée du processus qui est à la fois continue et markovienne. Dans cette optique, nous définissons la version continue du processus initial $\mathbf{X}^{(d)}$ comme étant un processus où les candidats sont proposés selon un processus de Poisson de taux 1, plutôt qu'à chaque pas de temps discret comme c'était le cas précédemment. Cela implique que dans le processus accéléré $\mathbf{Z}^{(d)}$, les candidats sont proposés selon un processus de Poisson de taux d. Notons que cette modification permet de préserver la propriété markovienne de la version accélérée du processus, et ne cause aucun souci en terme de convergence faible dans la topologie de Skorokhod, puisque celle-ci admet de petites distortions du temps et de l'espace.

Quelques résultats de convergence tirés de [**10**] nous garantissent une condition suffisante pour la convergence faible de la suite $\{Z_j^{(d)} : d \in \mathbb{N}\}$ des j-ème composantes des processus $\mathbf{Z}^{(d)}$, qui représente une suite de processus stochastiques univariés stationnaires à temps continu. Nous référons le lecteur au théorème 8.2 du chapitre 4 de cet ouvrage, ainsi qu'au corollaire 8.6 du même chapitre. Moyennant certaines conditions techniques sur le processus stochastique limite, ces résultats combinés établissent que si on a la convergence dans L^1 des générateurs de $\left\{Z_j^{(d)} : d \in \mathbb{N}\right\}$ vers celui d'un processus Z_j (représentée par la condition (8.11) du chapitre 4 de [**10**]) et certaines conditions de rigidité sur les processus $Z_j^{(d)}$ (représentées par les conditions (8.9) et (8.34) du chapitre 4 de [**10**]), alors on a la convergence faible de la suite $\left\{Z_j^{(d)} : d \in \mathbb{N}\right\}$ vers Z_j. Ces conditions de rigidité signifient essentiellement que l'amplitude du déplacement instantané du processus $Z_j^{(d)}$, ainsi que l'amplitude de son déplacement moyen sur un intervalle de temps fini, seront finies en espérance. Nous ne vérifierons ici que la convergence des générateurs. Pour une vérification rigoureuse de toutes les autres conditions, nous référons le lecteur à [**2**].

Plus précisément, si $G_j^{(d)}$ dénote le générateur de $Z_j^{(d)}$ et G_j le générateur de Z_j, alors il suffit de s'assurer que si $h \in \mathcal{C}_c^\infty$, l'espace des fonctions infiniment différentiables à support compact, et si $R^{(d)}$ (respectivement R) est une variable aléatoire distribuée selon la distribution stationnaire de $Z_j^{(d)}$ (respectivement Z_j), alors $G_j^{(d)}\left(R^{(d)}\right) \xrightarrow{L^1} G_j(R)$. Nous présentons maintenant un résultat introduisant le générateur du processus accéléré unidimensionnel $Z_j^{(d)}$.

Proposition 3.1.1. *Soit $\mathbf{Z}^{(d)}$ le processus accéléré où les candidats sont proposés selon un processus de Poisson de taux d. Le générateur de $Z_j^{(d)}$, G_j, est donné par*

$$G_j h(d,x) = d\mathbb{E}\left[(h(Y_j^{(d)}) - h(x))\alpha\left(\mathbf{X}^{(d)}(n), \mathbf{Y}^{(d)}\right)\middle| X_j^{(d)}(n) = x\right], \tag{3.1.1}$$

où $\mathbf{X}^{(d)}(n)$ désigne l'état actuel de la chaîne et $\mathbf{Y}^{(d)}$ désigne la proposition.

Démonstration. Par la définition de générateur, on a

$$\begin{aligned} G_j h(d,x) &= \frac{d}{dt}\mathbb{E}\left[h\left(Z_j^{(d)}(t)\right)\middle| Z_j^{(d)}(s) = x\right]\Big|_{t=s} \\ &= \frac{d}{dt}\mathbb{E}\left[h\left(X_j^{(d)}(d\cdot t)\right)\middle| X_j^{(d)}(d\cdot s) = x\right]\Big|_{t=s} \\ &= \lim_{t\to 0^+} \frac{\mathbb{E}\left[h\left(X_j^{(d)}(d(s+t))\right)\middle| X_j^{(d)}(d\cdot s) = x\right] - h(x)}{t}. \end{aligned} \tag{3.1.2}$$

L'espérance au numérateur de (3.1.2) dépend de N_t, le nombre de fois qu'un processus de Poisson de taux 1 sautera dans un intervalle de temps de longueur $d\cdot t$. Cette quantité est en fait une variable aléatoire de loi de Poisson de paramètre $d\cdot t$. Si on pose $T_1, ..., T_{N_t}$ les temps après lesquels surviendront chacun de ces sauts, où $0 = T_0 < T_1 < ... < T_{N_t} < d\cdot t$, on a

$$h\left(X_j^{(d)}(d(s+t))\right) - h(x) = \sum_{i=1}^{N_t} S_i, \tag{3.1.3}$$

où $S_i = h\left(X_j^{(d)}(d\cdot s + T_i)\right) - h\left(X_j^{(d)}(d\cdot s + T_{i-1})\right)$. De plus, en conditionnant sur le fait que $X_j^{(d)}(d\cdot s) = x$, nous avons

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{N_t} S_i\right] &= \mathbb{E}\left[S_1\right]\mathbb{P}\left(N_t = 1\right) + \mathbb{E}\left[\sum_{i=1}^{N_t} S_i\middle| N_t \geq 2\right]\mathbb{P}\left(N_t \geq 2\right) \\ &= d\cdot t e^{-d\cdot t}\mathbb{E}\left[S_1\right] + \sum_{n=2}^{\infty}\mathbb{E}\left[\sum_{i=1}^{n} S_i\right]\mathbb{P}\left(N_t = n\right) \\ &= d\cdot t e^{-d\cdot t}\mathbb{E}\left[S_1\right] + e^{-d\cdot t}\sum_{n=2}^{\infty}\frac{(d\cdot t)^n}{n!}\mathbb{E}\left[\sum_{i=1}^{n} S_i\right]. \end{aligned} \tag{3.1.4}$$

Or, puisque $h \in \mathcal{C}_c^\infty$, on peut borner cette fonction en valeur absolue par une constante $K < \infty$, et donc $|S_i|$ par $2K$. On peut donc borner le deuxième terme de (3.1.4) comme

suit :

$$\begin{aligned}\left|e^{-d\cdot t}\sum_{n=2}^{\infty}\frac{(d\cdot t)^n}{n!}\mathbb{E}\left[\sum_{i=1}^{n}S_i\right]\right| &\leq 2Ke^{-d\cdot t}\sum_{n=2}^{\infty}\frac{(d\cdot t)^n}{n!}n\\ &= 2Ke^{-d\cdot t}\left(\sum_{n=1}^{\infty}\frac{(d\cdot t)^n}{n!}n - d\cdot t\right)\\ &= 2Ke^{-d\cdot t}\left(d\cdot te^{d\cdot t} - d\cdot t\right)\\ &= 2Kd\cdot t\left(1-e^{-d\cdot t}\right)\\ &= o(t).\end{aligned}$$

En insérant ce résultat ainsi que (3.1.3) et (3.1.4) dans (3.1.2), on trouve finalement

$$G_jh(d,x) = \lim_{t\to 0^+}\frac{d\cdot te^{-d\cdot t}\mathbb{E}\left[S_1|\, X_j^{(d)}(d\cdot s)=x\right]+o(t)}{t} = d\mathbb{E}\left[S_1|\, X_j^{(d)}(d\cdot s)=x\right].$$

Conditionnellement à l'événement $X_j^{(d)}(d\cdot s)=x$, la loi de S_1 n'implique qu'un seul pas de la chaîne (effectué par la j-ème composante). Nous pouvons donc réexprimer le générateur en fonction de la chaîne initiale et obtenir

$$\begin{aligned}G_jh(d,x) &= d\mathbb{E}\left[h\left(X_j^{(d)}(n+1)\right)-h(x)\middle|\, X_j^{(d)}(n)=x\right]\\ &= d\mathbb{E}\left[\mathbb{E}\left[h\left(X_j^{(d)}(n+1)\right)-h(x)\middle|\, X_j^{(d)}(n)=x,\mathbf{Y}^{(d)}\right]\middle|\, X_j^{(d)}(n)=x\right]\\ &= d\mathbb{E}\left[(h(Y_j^{(d)})-h(x))\alpha\left(\mathbf{X}^{(d)}(n),\mathbf{Y}^{(d)}\right)\right.\\ &\quad\left.+(h(x)-h(x))\left(1-\alpha\left(\mathbf{X}^{(d)}(n),\mathbf{Y}^{(d)}\right)\right)\middle|\, X_j^{(d)}(n)=x\right]\\ &= d\mathbb{E}\left[(h(Y_j^{(d)})-h(x))\alpha\left(\mathbf{X}^{(d)}(n),\mathbf{Y}^{(d)}\right)\middle|\, X_j^{(d)}(n)=x\right].\end{aligned}$$

□

À partir de maintenant, par souci d'allégement des démonstrations, quelques changements seront apportés à la notation. Premièrement, nous allons renommer l'état actuel de la chaîne, $\mathbf{X}^{(d)}(n)$, par $\mathbf{X}^{(d)}$ tout simplement. En effet, il est inutile de conserver la notion du temps puisque la chaîne est homogène et stationnaire. De façon encore plus simple, X_j dénotera la j-ème composante de $\mathbf{X}^{(d)}$. De façon analogue, nous allons désigner par $\mathbf{Y}^{(d)}$ le vecteur aléatoire qui représente la proposition générée à partir de $\mathbf{X}^{(d)}$ et Y_j sa j-ème composante. De plus, nous introduisons une nouvelle notation pour les espérances conditionnelles. Il y aura, dans les prochaines sections, beaucoup de conditionnement sur des événements ainsi que sur des variables aléatoires et nous souhaitons souligner la distinction entre ces deux notions. Lorsqu'une espérance sera conditionnelle à un événement et à une variable aléatoire, on

notera l'événement en indice et le conditionnement par rapport à la variable aléatoire de la façon habituelle. Ainsi, l'espérance d'une variable aléatoire Y conditionnellement à une autre variable aléatoire X et à la réalisation d'un certain événement A sera dénotée $\mathbb{E}_A[Y|X]$.

Notons finalement qu'on va utiliser l'expression suivante pour G_j, qui sera utilisée ultérieurement lors de la décomposition de la démonstration :

$$\begin{aligned} G_j h(d,x) &= d\mathbb{E}_{X_j=x}\left[(h(Y_j)-h(x))\left(1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right] \\ &= d\mathbb{E}_{X_j=x}\left[\mathbb{E}_{X_j=x}\left[(h(Y_j)-h(x))\left(1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\middle|\,\delta\right]\right]. \end{aligned}$$

En se rappelant que $\delta \sim \text{Bernoulli}(p(d))$, on trouve

$$\begin{aligned} G_j h(d,x) &= (1-p(d))\cdot d\mathbb{E}_{\delta=0,X_j=x}\left[(h(Y_j)-h(x))\left(1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right] \\ &\quad + p(d)\cdot d\mathbb{E}_{\delta=1,X_j=x}\left[(h(Y_j)-h(x))\left(1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right] \\ &= (1-p(d))G_j h(d,0,x) + p(d)G_j h(d,1,x), \end{aligned}$$

où

$$G_j h(d,k,x) \equiv d\mathbb{E}_{\delta=k,X_j=x}\left[(h(Y_j)-h(x))\left(1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right]. \tag{3.1.5}$$

Le terme $G_j h(d,k,\cdot)$ décrit en quelque sorte le comportement de la j-ème composante de l'algorithme lorsque $\delta = k$, où δ représente le type de pas proposé pour la première composante : $\delta = 0$ représente un petit pas alors que $\delta = 1$ représente un grand pas.

3.2. Résultats préliminaires

Nous présentons ici quelques lemmes sur le comportement asymptotique de $\mathbf{X}^{(d)}$ et $\mathbf{Y}^{(d)}$, où $\mathbf{X}^{(d)}$ est distribué selon la densité cible π_d définie par (1.4.1) et $\mathbf{Y}^{(d)}$ est la proposition telle que définie dans la section 2.3. Rappelons qu'on peut supposer que $\mathbf{X}^{(d)}$ est distribué selon la distribution cible, puisque cette distribution est stationnaire pour la chaîne de Markov et puisqu'on a supposé que le point de départ de la chaîne, $\mathbf{X}^{(d)}(0)$, était distribué selon π_d.

D'abord, définissons les quantités

$$\mathbf{X}^- = \mathbf{X}^{(d)} \backslash \{X_j\} \text{ et } \pi_d^-(\mathbf{X}^-) = \prod_{i \neq j} f_i(X_i),$$

et définissons $\mathbf{Y}^-$ et $\pi_d^-(\mathbf{Y}^-)$ de façon analogue, où j dénote la composante d'intérêt (qui sera 1 lorsqu'on s'intéresse à la première composante de l'algorithme et 2 lorsqu'on s'intéresse à la seconde composante de l'algorithme).

Lemme 3.2.1. *Soient $k \geq 2$ un entier et $j \in \{1, 2\}$. Alors on a*

$$d\mathbb{E}_{\delta=0}\left[|h(Y_j) - h(X_j)||Y_j - X_j|^k \middle| X_j\right] \longrightarrow 0$$

lorsque $d \to \infty$.

Démonstration. Par un développement de Taylor d'ordre 1,

$$|h(Y_j) - h(X_j)| = |h'(V)||Y_j - X_j|,$$

avec V entre X_j et Y_j. On a donc

$$\begin{aligned} d\mathbb{E}_{\delta=0}\left[|h(Y_j) - h(X_j)||Y_j - X_j|^k \middle| X_j\right] &= d\mathbb{E}_{\delta=0}\left[|h'(V)||Y_j - X_j|^{k+1} \middle| X_j\right] \\ &\leq dK\mathbb{E}_{\delta=0}\left[|Y_j - X_j|^{k+1} \middle| X_j\right], \end{aligned}$$

où K est la constante qui majore h et h'. Par le lemme A.0.2, on trouve

$$\begin{aligned} d\mathbb{E}_{\delta=0}\left[|h(Y_j) - h(X_j)||Y_j - X_j|^k \middle| X_j\right] &\leq Kd\frac{2^{\frac{k+1}{2}}}{\sqrt{\pi}}\Gamma\left(\frac{k}{2}+1\right)\frac{l^{k+1}}{d^{\frac{k+1}{2}}} \\ &= K\frac{2^{\frac{k+1}{2}}}{\sqrt{\pi}}\Gamma\left(\frac{k}{2}+1\right)\frac{l^{k+1}}{d^{\frac{k-1}{2}}}, \end{aligned}$$

qui converge vers 0 lorsque $d \to \infty$.

□

Lemme 3.2.2. *Soit $j \in \{1, 2\}$. On a*

$$\begin{aligned} \mathbb{E}_{\delta=0}\Bigg[\Bigg|&\sum_{i \neq j}(\log f_i(Y_i) - \log f_i(X_i)) \\ &- \left(\sum_{i \neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i \neq j}((\log f_i)'(X_i))^2\right)\Bigg|\Bigg] \longrightarrow 0 \end{aligned}$$

lorsque $d \to \infty$.

Démonstration. D'abord, par un développement de Taylor d'ordre 2 des fonctions $\log f_i$ par rapport à Y_i et autour de X_i, on a

$$\sum_{i\neq j}(\log f_i(Y_i) - \log f_i(X_i)) = \sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) + \frac{1}{2}\sum_{i\neq j}(\log f_i)''(U_i)(Y_i - X_i)^2,$$

pour U_i entre X_i et Y_i. On peut donc réécrire l'espérance qui nous intéresse comme

$$\begin{aligned}
&\mathbb{E}_{\delta=0}\left[\left|\sum_{i\neq j}(\log f_i(Y_i) - \log f_i(X_i))\right.\right.\\
&\quad\left.\left. - \left(\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right)\right|\right]\\
&= \mathbb{E}_{\delta=0}\left[\left|\left(\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) + \frac{1}{2}\sum_{i\neq j}(\log f_i)''(U_i)(Y_i - X_i)^2\right)\right.\right.\\
&\quad\left.\left. - \left(\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right)\right|\right]\\
&= \mathbb{E}_{\delta=0}\left[\left|\frac{1}{2}\sum_{i\neq j}(\log f_i)''(U_i)(Y_i - X_i)^2 + \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right|\right].
\end{aligned}$$

En ajoutant, puis soustrayant, le terme $\frac{1}{2}\sum_{i\neq j}(\log f_i)''(X_i)(Y_i - X_i)^2$ dans cette espérance, celle-ci peut être bornée, en utilisant l'inégalité du triangle, par

$$\begin{aligned}
&\mathbb{E}_{\delta=0}\left[\left|\frac{1}{2}\sum_{i\neq j}(\log f_i)''(X_i)(Y_i - X_i)^2 + \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right|\right]\\
&\quad + \mathbb{E}_{\delta=0}\left[\left|\frac{1}{2}\sum_{i\neq j}(\log f_i)''(U_i)(Y_i - X_i)^2 - \frac{1}{2}\sum_{i\neq j}(\log f_i)''(X_i)(Y_i - X_i)^2\right|\right]. \qquad (3.2.1)
\end{aligned}$$

Nous allons montrer que les deux termes de (3.2.1) sont asymptotiquements nuls. Premièrement, par l'inégalité de Jensen, on obtient

$$\begin{aligned}
&\mathbb{E}_{\delta=0}\left[\left|\frac{1}{2}\sum_{i\neq j}(\log f_i)''(X_i)(Y_i - X_i)^2 + \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right|\middle|\mathbf{X}^-\right]\\
&\leq \left(\mathbb{E}_{\delta=0}\left[\left(\frac{1}{2}\sum_{i\neq j}(\log f_i)''(X_i)(Y_i - X_i)^2 + \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right)^2\middle|\mathbf{X}^-\right]\right)^{\frac{1}{2}}
\end{aligned}$$

$$= \frac{1}{2}\left(\mathbb{V}\mathrm{ar}_{\delta=0}\left(\sum_{i\neq j}(\log f_i)''(X_i)(Y_i-X_i)^2 + \frac{l^2}{d}\sum_{i\neq j}((\log f_i)'(X_i))^2\middle|\mathbf{X}^-\right)\right.$$
$$\left. + \mathbb{E}_{\delta=0}\left[\sum_{i\neq j}(\log f_i)''(X_i)(Y_i-X_i)^2 + \frac{l^2}{d}\sum_{i\neq j}((\log f_i)'(X_i))^2\middle|\mathbf{X}^-\right]^2\right)^{\frac{1}{2}}.$$

En utilisant le fait que, conditionnellement aux X_i, les Y_i sont mutuellement indépendants, on trouve

$$\frac{1}{2}\left(\sum_{i\neq j}((\log f_i)''(X_i))^2\mathbb{V}\mathrm{ar}_{\delta=0}\left((Y_i-X_i)^2\middle|X_i\right)\right.$$
$$\left. + \left(\frac{l^2}{d}\left(\sum_{i\neq j}(\log f_i)''(X_i) + \sum_{i\neq j}((\log f_i)'(X_i))^2\right)\right)^2\right)^{\frac{1}{2}}$$
$$\leq \frac{1}{2}\left(\sum_{i\neq j}((\log f_i)''(X_i))^2\mathbb{E}_{\delta=0}\left[(Y_i-X_i)^4\middle|X_i\right] + \frac{l^4}{d^2}\left(\sum_{i\neq j}\frac{f_i''(X_i)}{f_i(X_i)}\right)^2\right)^{\frac{1}{2}}$$
$$= \frac{1}{2}\sqrt{3\frac{l^4}{d^2}\sum_{i\neq j}((\log f_i)''(X_i))^2 + \frac{l^4}{d^2}\left(\sum_{i\neq j}\frac{f_i''(X_i)}{f_i(X_i)}\right)^2}$$
$$\leq \frac{l^2}{2d}\sqrt{3K^2(d-1) + \left(\sum_{i\neq j}\frac{f_i''(X_i)}{f_i(X_i)}\right)^2},$$

où $K>0$ borne les fonctions $(\log f_1)''$ et $(\log f)''$. Par concavité de la fonction racine carrée, ce terme est borné par

$$\frac{\sqrt{3}Kl^2}{2}\frac{\sqrt{d-1}}{d} + \frac{l^2}{2d}\left|\sum_{i\neq j}\frac{f_i''(X_i)}{f_i(X_i)}\right|.$$

On borne donc le premier terme de (3.2.1) de la façon suivante :

$$\mathbb{E}_{\delta=0}\left[\left|\frac{1}{2}\sum_{i\neq j}(\log f_i)''(X_i)(Y_i-X_i)^2 + \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right|\right]$$
$$= \mathbb{E}_{\delta=0}\left[\mathbb{E}_{\delta=0}\left[\left|\frac{1}{2}\sum_{i\neq j}(\log f_i)''(X_i)(Y_i-X_i)^2 + \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right|\middle|\mathbf{X}^-\right]\right]$$
$$\leq \frac{\sqrt{3}Kl^2}{2}\frac{\sqrt{d-1}}{d} + \frac{l^2}{2d}\mathbb{E}\left[\left|\sum_{i\neq j}\frac{f_i''(X_i)}{f_i(X_i)}\right|\right]$$
$$= \frac{\sqrt{3}Kl^2}{2}\frac{\sqrt{d-1}}{d} + \frac{l^2}{2}\mathbb{E}\left[|S_d|\right], \tag{3.2.2}$$

en posant $S_d = \frac{1}{d}\sum_{i\neq j}\frac{f_i''(X_i)}{f_i(X_i)}$. Le premier terme de (3.2.2), bien sûr, n'est pas aléatoire et il converge vers 0 lorsque $d \to \infty$. Il faut donc montrer que le second terme converge également vers 0, ce qui est équivalent à $S_d \xrightarrow{L_1} 0$. Or, par le lemme A.0.1, on a $\forall i$

$$\mathbb{E}\left[\frac{f_i''(X_i)}{f_i(X_i)}\right] = 0.$$

De plus, on a supposé que

$$\mathbb{E}\left[\left(\frac{f_i''(X_i)}{f_i(X_i)}\right)^2\right] < \infty.$$

La loi faible des grands nombres nous assure donc que S_d, qui est une moyenne, converge en probabilité vers 0. Pour établir la convergence dans L^1, il suffit que la suite $\{S_d : d \in \mathbb{N}\}$ soit uniformément intégrable. Or, pour $a > 0$,

$$\begin{aligned}
\mathbb{E}\left[|S_d| \cdot \mathbb{I}\left\{|S_d| > a\right\}\right] &\leq \mathbb{E}\left[\frac{|S_d|}{a}|S_d| \cdot \mathbb{I}\left\{|S_d| > a\right\}\right] \\
&= \frac{1}{a}\mathbb{E}\left[S_d^2 \cdot \mathbb{I}\left\{|S_d| > a\right\}\right] \\
&\leq \frac{1}{a}\mathbb{E}\left[S_d^2\right] \\
&= \frac{1}{a}\frac{1}{d^2}\left(\mathbb{E}\left[\sum_{i\neq j}\left(\frac{f_i''(X_i)}{f_i(X_i)}\right)^2\right] + \mathbb{E}\left[\sum_{i\neq j}\sum_{k\neq j, k\neq i}\frac{f_i''(X_i)}{f_i(X_i)}\frac{f_k''(X_k)}{f_k(X_k)}\right]\right) \\
&= \frac{1}{a}\frac{1}{d^2}\mathbb{E}\left[\sum_{i\neq j}\left(\frac{f_i''(X_i)}{f_i(X_i)}\right)^2\right],
\end{aligned}$$

puisque les X_i sont mutuellement indépendants. Puisque les termes de cette somme sont positifs, on a

$$\begin{aligned}
\mathbb{E}\left[|S_d| \cdot \mathbb{I}\left\{|S_d| > a\right\}\right] &\leq \frac{1}{a}\frac{1}{d^2}\mathbb{E}\left[\sum_{i=1}^{d}\left(\frac{f_i''(X_i)}{f_i(X_i)}\right)^2\right] \\
&\leq \frac{1}{a}\left(\frac{1}{d^2}\mathbb{E}\left[\left(\frac{f_1''(X_1)}{f_1(X_1)}\right)^2\right] + \frac{d-1}{d^2}\mathbb{E}\left[\left(\frac{f_2''(X_2)}{f_2(X_2)}\right)^2\right]\right) \\
&\leq \frac{1}{a}\left(\mathbb{E}\left[\left(\frac{f_1''(X_1)}{f_1(X_1)}\right)^2\right] + \mathbb{E}\left[\left(\frac{f_2''(X_2)}{f_2(X_2)}\right)^2\right]\right),
\end{aligned}$$

qui est indépendant de la dimension d. On a donc finalement

$$\lim_{a\to\infty}\sup_{d\geq 1}\mathbb{E}\left[|S_d| \cdot \mathbb{I}\left\{|S_d| > a\right\}\right] \leq \lim_{a\to\infty}\frac{1}{a}\left(\mathbb{E}\left[\frac{f_1''(X_1)}{f_1(X_1)}\right] + \mathbb{E}\left[\frac{f_2''(X_2)}{f_2(X_2)}\right]\right) = 0.$$

On en conclut que $S_d \xrightarrow{L_1} 0$ et donc que (3.2.2) converge vers 0 lorsque $d \to \infty$.

Pour ce qui est du second terme de (3.2.1), il est inférieur ou égal, toujours par l'inégalité du triangle, à

$$\frac{1}{2}\mathbb{E}_{\delta=0}\left[\sum_{i\neq j}\left|(\log f_i)''(U_i)-(\log f_i)''(X_i)\right|(Y_i-X_i)^2\right].$$

Lorsqu'on suppose que $\delta=0$, on sait que $Y_i=X_i+\frac{l}{\sqrt{d}}Z_i$, où $Z_i\sim N(0;1)$ et où les X_i et les Z_i sont indépendants de d. On peut donc réécrire la borne pour le second terme de (3.2.1) comme

$$\begin{aligned}&\frac{1}{2}\mathbb{E}_{\delta=0}\left[\sum_{i\neq j}\left|(\log f_i)''(U_i)-(\log f_i)''(X_i)\right|\frac{l^2}{d}Z_i^2\right]\\&\quad=\frac{l^2}{2d}\sum_{i\neq j}\mathbb{E}_{\delta=0}\left[\left|(\log f_i)''(U_i)-(\log f_i)''(X_i)\right|Z_i^2\right],\end{aligned}\tag{3.2.3}$$

où $U_i\in\left[X_i\pm\frac{l}{\sqrt{d}}Z_i\right]$. Il suffit donc de montrer que (3.2.3) converge vers 0 lorsque $d\to\infty$ pour que le second terme de (3.2.1) converge lui aussi vers 0. Pour ce faire, remarquons que les termes 2 à d de cette somme sont égaux puisque la distribution de $\left|(\log f_i)''(U_i)-(\log f_i)''(X_i)\right|Z_i^2$ est la même pour tout $i\geq 2$. Ainsi, si $j=1$, (3.2.3) est égal à

$$\frac{l^2(d-1)}{2d}\mathbb{E}\left[\left|(\log f)''(U_2)-(\log f)''(X_2)\right|Z_2^2\right],$$

alors que si $j=2$, (3.2.3) est égal à

$$\frac{l^2}{2d}\mathbb{E}_{\delta=0}\left[\left|(\log f_1)''(U_1)-(\log f_1)''(X_1)\right|Z_1^2\right]+\frac{l^2(d-2)}{2d}\mathbb{E}\left[\left|(\log f)''(U_2)-(\log f)''(X_2)\right|Z_2^2\right].$$

Dans ce deuxième cas, en se rappelant que la fonction $(\log f_1)''$ est bornée en valeur absolue par la constante K, on voit facilement que la première espérance est bornée par

$$\frac{Kl^2}{d}\mathbb{E}\left[Z_1^2\right]=\frac{Kl^2}{d}\longrightarrow 0$$

lorsque $d\to\infty$. Ainsi, on a que peu importe la valeur de $j\in\{1,2\}$, la limite de (3.2.3) est égale à celle de

$$\frac{l^2}{2}\mathbb{E}\left[\left|(\log f)''(U_2)-(\log f)''(X_2)\right|Z_2^2\right].\tag{3.2.4}$$

Afin de montrer que l'espérance (3.2.4) converge vers 0 lorsque $d\to\infty$, nous allons utiliser le théorème de convergence dominée. En effet, en utilisant la borne pour $(\log f)''$, il est facile de voir que le terme à l'intérieur de l'espérance est borné en valeur absolue par la variable aléatoire $2KZ_2^2\in L^1$, qui est indépendante de d. Puisque $U_2\in\left[X_2\pm\frac{l}{\sqrt{d}}Z_2\right]$ et puisque X_2 et Z_2 sont indépendants de d, on a également que $U_2\to X_2$ presque sûrement lorsque $d\to\infty$.

Puisque $f \in \mathcal{C}^2$, on a que

$$(\log f)'' = \left(\frac{f'}{f}\right)' = \frac{f''}{f} - \left(\frac{f'}{f}\right)^2$$

est nécessairement continue, puisque f, f' et f'' sont continues. Ainsi, le «Continuous Mapping Theorem» nous assure que $(\log f)''(U_2) \to (\log f)''(X_2)$ presque sûrement lorsque $d \to \infty$. Ceci établit la convergence vers 0 de (3.2.4), ce qui implique la convergence vers 0 de (3.2.3), et donc également de (3.2.1).

□

Lemme 3.2.3. *Soient $j \in \{1, 2\}$,*

$$B_d = \frac{1}{d} \sum_{i \neq j} ((\log f_i)'(X_i))^2 \tag{3.2.5}$$

et

$$B = \mathbb{E}\left[((\log f)'(X_2))^2\right].$$

On a alors $B_d \xrightarrow{D} B$.

Démonstration. Nous vérifions d'abord que B_d converge en probabilité en étudiant ses deux premiers moments :

$$\begin{aligned}
\mathbb{E}[B_d] &= \frac{1}{d} \sum_{i \neq j} \mathbb{E}\left[((\log f_i)'(X_i))^2\right] \\
&= \frac{1}{d}\mathbb{E}\left[((\log f_{3-j})'(X_{3-j}))^2\right] + \frac{d-2}{d}\mathbb{E}\left[((\log f)'(X_2))^2\right] \\
&\longrightarrow 0 + \mathbb{E}\left[((\log f)'(X_2))^2\right] = B
\end{aligned}$$

et

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(B_d) &= \frac{1}{d^2} \sum_{i \neq j} \mathbb{V}\mathrm{ar}\left(((\log f)'(X_i))^2\right) \\
&= \frac{1}{d^2}\mathbb{V}\mathrm{ar}\left(((\log f_{3-j})'(X_{3-j}))^2\right) + \frac{d-2}{d^2}\mathbb{V}\mathrm{ar}\left(((\log f)'(X_2))^2\right) \\
&\longrightarrow 0
\end{aligned}$$

lorsque $d \to \infty$, car, par hypothèse, $(\log f_1)'(X_1), (\log f)'(X_2) \in L^4$. La convergence en probabilité étant équivalente à la convergence en distribution lorsque la limite est une constante, on en déduit le résultat.

□

3.3. Le comportement asymptotique lors des petits pas

Cette section présente un lemme qui établit le comportement asymptotique d'une composante quelconque de l'algorithme lorsqu'un petit pas est proposé, c'est à dire lorsque $\delta = 0$. Ce comportement est celui d'un processus de diffusion Langevin.

Lemme 3.3.1. *Soient $\mathbf{X}^{(d)}$ distribué selon π_d et $\mathbf{Y}^{(d)}$ la proposition telle que définie dans la section 2.3. Soit*

$$G_{L,j}h(x) = \frac{v(l)}{2}\left(h''(x) + h'(x)\left(\log f_j\right)'(x)\right), \tag{3.3.1}$$

où

$$v(l) = 2l^2\Phi\left(-\frac{l\sqrt{B}}{2}\right),$$

B est défini comme précédemment et $j \in \{1, 2\}$. Notons que $G_{L,j}$ représente le générateur d'un processus de diffusion Langevin avec mesure de vitesse $v(l)$ et fonction de dérive $(\log f_j)'$. On a alors

$$G_j h(d, 0, X_j) \xrightarrow{L^1} G_{L,j}h(X_j)$$

lorsque $d \to \infty$, où $G_j h(d, 0, \cdot)$ est défini par (3.1.5).

Démonstration. Rappelons que

$$G_j h(d, 0, x) = d\mathbb{E}_{\delta=0, X_j=x}\left[(h(Y_j) - h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right]$$

et définissons

$$\begin{aligned}\tilde{G}_j h(d, 0, x) = l^2\Bigg(&\frac{h''(x)}{2}\mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right] + h'(x)\left(\log f_j\right)'(x) \\ &\cdot\mathbb{E}_{\delta=0}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1\right\}\right]\Bigg),\end{aligned} \tag{3.3.2}$$

où $\mathbb{I}\{A\}$ est l'indicatrice qui vaut 1 si l'évènemt A se réalise et 0 sinon. Par l'inégalité du triangle, on a

$$\begin{aligned}&\mathbb{E}\left[\left|G_j h(d, 0, X_j) - G_{L,j}h(X_j)\right|\right] \\ &\quad\leq \mathbb{E}\left[\left|G_j h(d, 0, X_j) - \tilde{G}_j h(d, 0, X_j)\right|\right] + \mathbb{E}\left[\left|\tilde{G}_j h(d, 0, X_j) - G_{L,j}h(X_j)\right|\right].\end{aligned} \tag{3.3.3}$$

Les lemmes 3.3.2 et 3.3.3 qui suivent affirment que ces deux espérances convergent vers 0 lorsque $d \to \infty$, ce qui établit le résultat voulu.

□

Lemme 3.3.2. *Sous les conditions du lemme 3.3.1, on a que*

$$\mathbb{E}\left[\left|G_j h(d,0,X_j) - \tilde{G}_j h(d,0,X_j)\right|\right] \longrightarrow 0$$

lorsque $d \to \infty$, où G_j et $\tilde{G}_j$ sont définis par (3.1.5) et (3.3.2), respectivement.

Démonstration. Notons que Y_j est indépendant de tous les autres Y_i ainsi que de tous les X_i, pour $i \neq j$. On a donc

$$\begin{aligned} G_j h(d,0,x) &= d\mathbb{E}_{\delta=0,X_j=x}\left[(h(Y_j)-h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right] \\ &= d\mathbb{E}_{\delta=0,X_j=x}\left[\mathbb{E}_{\delta=0,X_j=x}\left[(h(Y_j)-h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\middle| Y_j\right]\right] \\ &= d\mathbb{E}_{\delta=0,X_j=x}\left[(h(Y_j)-h(x))\,\mathbb{E}_{\delta=0,X_j=x}\left[1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle| Y_j\right]\right]. \end{aligned} \tag{3.3.4}$$

Le minimum entre deux fonctions différentiables étant différentiable partout sauf peut-être aux points où les deux fonctions sont égales, la fonction ψ définie par

$$\psi(\mathbf{x}^{(d)},\mathbf{y}^{(d)}) = 1 \wedge \frac{\pi_d(\mathbf{y}^{(d)})}{\pi_d(\mathbf{x}^{(d)})}$$

est différentiable presque partout en y_j. De plus, on a

$$\begin{aligned} \frac{\partial}{\partial y_j}\psi(\mathbf{x}^{(d)},\mathbf{y}^{(d)}) &= \frac{f_j'(y_j)}{f_j(x_j)}\frac{\pi_d^-(\mathbf{y}^-)}{\pi_d^-(\mathbf{x}^-)}\mathbb{I}\left\{\frac{f_j(y_j)}{f_j(x_j)}\frac{\pi_d^-(\mathbf{y}^-)}{\pi_d^-(\mathbf{x}^-)} < 1\right\} \\ \text{et } \frac{\partial^2}{\partial y_j^2}\psi(\mathbf{x}^{(d)},\mathbf{y}^{(d)}) &= \frac{f_j''(y_j)}{f_j(x_j)}\frac{\pi_d^-(\mathbf{y}^-)}{\pi_d^-(\mathbf{x}^-)}\mathbb{I}\left\{\frac{f_j(y_j)}{f_j(x_j)}\frac{\pi_d^-(\mathbf{y}^-)}{\pi_d^-(\mathbf{x}^-)} < 1\right\}. \end{aligned}$$

On peut donc réécrire, par un développement de Taylor d'ordre 2 de la fonction ψ par rapport à y_j et autour de x_j, l'espérance à l'intérieur de $G_j h(d,0,x)$ comme

$$\begin{aligned} &\mathbb{E}_{\delta=0,X_j=x}\left[1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle| Y_j\right] \\ &= \mathbb{E}_{\delta=0,X_j=x}\left[\psi(\mathbf{X}^{(d)},\mathbf{Y}^{(d)})\middle| Y_j\right] \\ &= \mathbb{E}_{\delta=0,X_j=x}\left[\psi(\mathbf{X}^{(d)},\mathbf{Y}^{(d)})\Big|_{Y_j=x} + \frac{\partial}{\partial Y_j}\psi(\mathbf{X}^{(d)},\mathbf{Y}^{(d)})\Big|_{Y_j=x}(Y_j-x)\right. \\ &\quad \left. + \frac{1}{2}\frac{\partial^2}{\partial Y_j^2}\psi(\mathbf{X}^{(d)},\mathbf{Y}^{(d)})\Big|_{Y_j=U}(Y_j-x)^2\middle| Y_j\right] \\ &= \mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right] + (Y_j-x)(\log f_j)'(x)\mathbb{E}_{\delta=0}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1\right\}\right] \end{aligned}$$

$$+\frac{1}{2}(Y_j-x)^2\frac{f_j''(U)}{f_j(x)}\mathbb{E}_{\delta=0,X_j=x}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{f_j(U)}{f_j(x)}\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}<1\right\}\middle|Y_j\right]$$

pour un certain U entre x et Y_j. En substituant cette expression dans (3.3.4), on obtient

$$\begin{aligned}
G_jh(d,0,x) &= d\mathbb{E}_{\delta=0,X_j=x}\left[(h(Y_j)-h(x))\,\mathbb{E}_{\delta=0,X_j=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle|Y_j\right]\right]\\
&= d\mathbb{E}_{\delta=0,X_j=x}\left[h(Y_j)-h(x)\right]\mathbb{E}_{\delta=0}\left[1\wedge\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right]\\
&\quad+d(\log f_j)'(x)\mathbb{E}_{\delta=0,X_j=x}\left[(h(Y_j)-h(x))\,(Y_j-x)\right]\\
&\quad\cdot\mathbb{E}_{\delta=0}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}<1\right\}\right]\\
&\quad+\frac{d}{2}\mathbb{E}_{\delta=0,X_j=x}\left[\frac{f_j''(U)}{f_j(x)}\,(h(Y_j)-h(x))\,(Y_j-x)^2\right.\\
&\quad\left.\cdot\mathbb{E}_{\delta=0,X_j=x}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{f_j(U)}{f_j(x)}\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}<1\right\}\middle|Y_j\right]\right].
\end{aligned}\tag{3.3.5}$$

Notons maintenant que, par un développement de Taylor d'ordre 3 de la fonction h par rapport à Y_j et autour de x, on a

$$h(Y_j)-h(x)=h'(x)(Y_j-x)+\frac{1}{2}h''(x)(Y_j-x)^2+\frac{1}{6}h'''(Z)(Y_j-x)^3,$$

pour un certain Z entre x et Y_j. Par le lemme A.0.2, on trouve alors

$$\begin{aligned}
\mathbb{E}_{\delta=0,X_j=x}\left[h(Y_j)-h(x)\right] &= h'(x)\mathbb{E}_{\delta=0,X_j=x}\left[Y_j-x\right]+\frac{1}{2}h''(x)\mathbb{E}_{\delta=0,X_j=x}\left[(Y_j-x)^2\right]\\
&\quad+\frac{1}{6}\mathbb{E}_{\delta=0,X_j=x}\left[h'''(Z)(Y_j-x)^3\right]\\
&= \frac{l^2h''(x)}{2d}+\frac{1}{6}\mathbb{E}_{\delta=0,X_j=x}\left[h'''(Z)(Y_j-x)^3\right]
\end{aligned}$$

et

$$\begin{aligned}
\mathbb{E}_{\delta=0,X_j=x}\left[(h(Y_j)-h(x))\,(Y_j-x)\right] &= h'(x)\mathbb{E}_{\delta=0,X_j=x}\left[(Y_j-x)^2\right]\\
&\quad+\frac{1}{2}h''(x)\mathbb{E}_{\delta=0,X_j=x}\left[(Y_j-x)^3\right]\\
&\quad+\frac{1}{6}\mathbb{E}_{\delta=0,X_j=x}\left[h'''(Z)(Y_j-x)^4\right]\\
&= \frac{l^2h'(x)}{d}+\frac{1}{6}\mathbb{E}_{\delta=0,X_j=x}\left[h'''(Z)(Y_j-x)^4\right].
\end{aligned}$$

En substituant ces expressions dans (3.3.5), on obtient

$$\begin{aligned}
G_j h(d,0,x) = &\left(l^2 \frac{h''(x)}{2} + \frac{d}{6} \mathbb{E}_{\delta=0, X_j = x} \left[h'''(Z)(Y_j - x)^3 \right] \right) \mathbb{E}_{\delta=0} \left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \right] \\
&+ (\log f_j)'(x) \left(l^2 h'(x) + \frac{d}{6} \mathbb{E}_{\delta=0, X_j = x} \left[h'''(Z)(Y_j - x)^4 \right] \right) \\
&\quad \cdot \mathbb{E}_{\delta=0} \left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \mathbb{I} \left\{ \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1 \right\} \right] \\
&+ \frac{d}{2} \mathbb{E}_{\delta=0, X_j = x} \left[\frac{f_j''(U)}{f_j(x)} \left(h(Y_j) - h(x) \right) (Y_j - x)^2 \right. \\
&\quad \left. \cdot \mathbb{E}_{\delta=0, X_j = x} \left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \mathbb{I} \left\{ \frac{f_j(U)}{f_j(x)} \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1 \right\} \middle| Y_j \right] \right].
\end{aligned}$$

En se rappelant (3.3.2), on voit apparaître $\tilde{G}$ dans cette expression. On obtient donc

$$\begin{aligned}
&G_j h(d,0,x) = \\
&= \tilde{G}_j h(d,0,x) + \frac{d}{6} \mathbb{E}_{\delta=0, X_j = x} \left[h'''(Z)(Y_j - x)^3 \right] \mathbb{E}_{\delta=0} \left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \right] \\
&\quad + \frac{d}{6} (\log f_j)'(x) \mathbb{E}_{\delta=0, X_j = x} \left[h'''(Z)(Y_j - x)^4 \right] \mathbb{E}_{\delta=0} \left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \mathbb{I} \left\{ \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1 \right\} \right] \\
&\quad + \frac{d}{2} \mathbb{E}_{\delta=0, X_j = x} \left[\frac{f_j''(U)}{f_j(U)} \left(h(Y_j) - h(x) \right) (Y_j - x)^2 \right. \\
&\qquad \left. \cdot \mathbb{E}_{\delta=0, X_j = x} \left[\frac{f_j(U)}{f_j(x)} \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \mathbb{I} \left\{ \frac{f_j(U)}{f_j(x)} \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1 \right\} \middle| Y_j \right] \right].
\end{aligned}$$

La différence absolue entre $G_j h(d,0,x)$ et $\tilde{G}_j h(d,0,x)$ satisfait donc

$$\begin{aligned}
&\left| G_j h(d,0,x) - \tilde{G}_j h(d,0,x) \right| \\
&= \left| \frac{d}{6} \mathbb{E}_{\delta=0, X_j = x} \left[h'''(Z)(Y_j - x)^3 \right] \mathbb{E}_{\delta=0} \left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \right] \right. \\
&\quad + \frac{d}{6} (\log f_j)'(x) \mathbb{E}_{\delta=0, X_j = x} \left[h'''(Z)(Y_j - x)^4 \right] \mathbb{E}_{\delta=0} \left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \mathbb{I} \left\{ \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1 \right\} \right] \\
&\quad + \frac{d}{2} \mathbb{E}_{\delta=0, X_j = x} \left[\frac{f_j''(U)}{f_j(U)} \left(h(Y_j) - h(x) \right) (Y_j - x)^2 \right. \\
&\qquad \left. \left. \cdot \mathbb{E}_{\delta=0, X_j = x} \left[\frac{f_j(U)}{f_j(x)} \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \mathbb{I} \left\{ \frac{f_j(U)}{f_j(x)} \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1 \right\} \middle| Y_j \right] \right] \right|.
\end{aligned}$$

Par l'inégalité du triangle, on obtient

$$\begin{aligned}
&\left|G_j h(d,0,x) - \tilde{G}_j h(d,0,x)\right| \\
&\leq \frac{d}{6}\mathbb{E}_{\delta=0,X_j=x}\left[\left|h'''(Z)(Y_j-x)^3\right|\right]\mathbb{E}_{\delta=0}\left[1\wedge\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right] \\
&\quad + \frac{d}{6}\left|(\log f_j)'(x)\right|\mathbb{E}_{\delta=0,X_j=x}\left[\left|h'''(Z)(Y_j-x)^4\right|\right]\mathbb{E}_{\delta=0}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}<1\right\}\right] \\
&\quad + \frac{d}{2}\mathbb{E}_{\delta=0,X_j=x}\left[\left|\frac{f_j''(U)}{f_j(U)}\right||h(Y_j)-h(x)|(Y_j-x)^2\right. \\
&\qquad \left.\cdot\mathbb{E}_{\delta=0,X_j=x}\left[\frac{f_j(U)}{f_j(x)}\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{f_j(U)}{f_j(x)}\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}<1\right\}\middle|Y_j\right]\right].
\end{aligned}$$

Puisque $h\in\mathcal{C}_c^\infty(\mathbb{R})$ et $(\log f_j)'$ est continue au sens de Lipschitz, on peut supposer que h, h', h'', h''' et $(\log f_j)''$ sont toutes majorées en valeur absolue par une constante $K<\infty$. Ceci nous mène à

$$\begin{aligned}
&\left|G_j h(d,0,x) - \tilde{G}_j h(d,0,x)\right| \\
&\leq \frac{K}{6}d\mathbb{E}_{\delta=0,X_j=x}\left[|Y_j-x|^3\right] + \frac{K}{6}d\left|(\log f_j)'(x)\right|\mathbb{E}_{\delta=0,X_j=x}\left[(Y_j-x)^4\right] \\
&\quad + \frac{d}{2}\mathbb{E}_{\delta=0,X_j=x}\left[\left|\frac{f_j''(U)}{f_j(U)}\right||h(Y_j)-h(x)|(Y_j-x)^2\right],
\end{aligned} \tag{3.3.6}$$

puisque les trois espérances $\mathbb{E}_{\delta=0}\left[1\wedge\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right]$, $\mathbb{E}_{\delta=0}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}<1\right\}\right]$ et $\mathbb{E}_{\delta=0,X_j=x}\left[\frac{f_j(U)}{f_j(x)}\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\mathbb{I}\left\{\frac{f_j(U)}{f_j(x)}\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}<1\right\}\middle|Y_j\right]$ sont toutes inférieures ou égales à 1. En utilisant le lemme A.0.2, on peut calculer les deux premières espérances dans (3.3.6), et obtenir

$$\begin{aligned}
\left|G_j h(d,0,x) - \tilde{G}_j h(d,0,x)\right| &\leq \frac{K}{6}\sqrt{\frac{8}{\pi}}d\frac{l^3}{d^{\frac{3}{2}}} + \frac{K}{2}d\left|(\log f_j)'(x)\right|\frac{l^4}{d^2} \\
&\quad + \frac{d}{2}\mathbb{E}_{\delta=0,X_j=x}\left[\left|\frac{f_j''(U)}{f_j(U)}\right||h(Y_j)-h(x)|(Y_j-x)^2\right] \\
&= \frac{K}{6}\sqrt{\frac{8}{\pi}}\frac{l^3}{\sqrt{d}} + \frac{K}{2}\left|(\log f_j)'(x)\right|\frac{l^4}{d} \\
&\quad + \frac{d}{2}\mathbb{E}_{\delta=0,X_j=x}\left[\left|\frac{f_j''(U)}{f_j(U)}\right||h(Y_j)-h(x)|(Y_j-x)^2\right].
\end{aligned} \tag{3.3.7}$$

Il ne reste qu'à travailler un peu le troisième terme de (3.3.7) avant de pouvoir conclure. Nous avons

$$
\begin{aligned}
&\frac{d}{2}\mathbb{E}_{\delta=0,X_j=x}\left[\left|\frac{f_j''(U)}{f_j(U)}\right||h(Y_j)-h(x)|(Y_j-x)^2\right]\\
&=\frac{d}{2}\mathbb{E}_{\delta=0,X_j=x}\left[\left|(\log f_j)''(U)+((\log f_j)'(U))^2\right||h(Y_j)-h(x)|(Y_j-x)^2\right]\\
&\leq\frac{d}{2}\mathbb{E}_{\delta=0,X_j=x}\left[\left|K+((\log f_j)'(U))^2\right||h(Y_j)-h(x)|(Y_j-x)^2\right].
\end{aligned}
$$

En effectuant une linéarisation de Taylor de la fonction $(\log f_j)'$ autour du point x, on obtient

$$
\begin{aligned}
K+((\log f_j)'(U))^2&=K+\left((\log f_j)'(x)+(\log f_j)''(W)(U-x)\right)^2\\
&\leq K+\left(|(\log f_j)'(x)|+K\,|U-x|\right)^2\\
&\leq K+\left(|(\log f_j)'(x)|+K\,|Y_j-x|\right)^2\\
&=K+\left((\log f_j)'(x)\right)^2+2K\,|(\log f_j)'(x)|\,|Y_j-x|\\
&\quad+K^2(Y_j-x)^2,
\end{aligned}
$$

pour W entre x et U. On peut donc borner le troisième terme de (3.3.7) par

$$
\begin{aligned}
&\frac{d}{2}\left(K+\left((\log f_j)'(x)\right)^2\right)\mathbb{E}_{\delta=0,X_j=x}\left[|h(Y_j)-h(x)|(Y_j-x)^2\right]\\
&\quad+dK\,|(\log f_j)'(x)|\,\mathbb{E}_{\delta=0,X_j=x}\left[|h(Y_j)-h(x)||Y_j-x|^3\right]\\
&\quad+\frac{d}{2}K^2\mathbb{E}_{\delta=0,X_j=x}\left[|h(Y_j)-h(x)|(Y_j-x)^4\right].
\end{aligned}
\tag{3.3.8}
$$

Nous sommes maintenant prêts à étudier $\left|G_jh(d,0,X_j)-\tilde{G}_jh(d,0,X_j)\right|$, la variable aléatoire d'intérêt. En utilisant (3.3.7) et (3.3.8), on a

$$
\begin{aligned}
\left|G_jh(d,0,X_j)-\tilde{G}_jh(d,0,X_j)\right|\leq{}&\frac{K}{6}\sqrt{\frac{8}{\pi}}\frac{l^3}{\sqrt{d}}+\frac{K}{2}\,|(\log f_j)'(X_j)|\,\frac{l^4}{d}\\
&+\frac{d}{2}\left(K+\left((\log f_j)'(X_j)\right)^2\right)\mathbb{E}_{\delta=0}\left[|h(Y_j)-h(X_j)|(Y_j-X_j)^2\middle|X_j\right]\\
&+dK\,|(\log f_j)'(X_j)|\,\mathbb{E}_{\delta=0}\left[|h(Y_j)-h(X_j)||Y_j-X_j|^3\middle|X_j\right]\\
&+\frac{d}{2}K^2\mathbb{E}_{\delta=0}\left[|h(Y_j)-h(X_j)|(Y_j-X_j)^4\middle|X_j\right].
\end{aligned}
\tag{3.3.9}
$$

Il est facile de voir que le premier terme de (3.3.9) est non aléatoire et converge vers 0. Le deuxième, lui, converge vers 0 sûrement et est borné par $\frac{K}{2}l^4\,|(\log f_j)'(X_j)|\in L^1$. De plus, par le lemme 3.2.1, les trois derniers termes convergent vers 0 sûrement et sont

asymptotiquement bornés par, respectivement, $\frac{1}{2}\left(K + ((\log f_j)'(X_j))^2\right)$, $K\,|(\log f_j)'(X_j)|$ et $\frac{K^2}{2}$ qui sont tous trois dans L^1. Alors, par le théorème de convergence dominée, ces cinq termes convergent vers 0 dans L^1, ainsi que leur somme. Cela signifie que

$$\mathbb{E}\left[\left|G_j h(d,0,X_j) - \tilde{G}_j h(d,0,X_j)\right|\right] \longrightarrow 0. \tag{3.3.10}$$

□

Lemme 3.3.3. *Sous les conditions du lemme 3.3.1, on a que*

$$\mathbb{E}\left[\left|\tilde{G}_j h(d,0,X_j) - G_{L,j} h(X_j)\right|\right] \longrightarrow 0$$

lorsque $d \to \infty$, où $\tilde{G}_j$ et $G_{L,j}$ sont définis par (3.3.2) et (3.3.1), respectivement.

Démonstration. De par les définitions de $\tilde{G}$ et de G_L, soit (3.3.2) et (3.3.1), nous avons

$$\begin{aligned}\left|\tilde{G}_j h(d,0,x) - G_{L,j} h(x)\right| = \Bigg|\frac{h''(x)}{2}\left(l^2 \mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right] - v(l)\right) + h'(x)(\log f_j)'(x) \\ \cdot \left(l^2 \mathbb{E}_{\delta=0}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \mathbb{I}\left\{\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1\right\}\right] - \frac{v(l)}{2}\right)\Bigg|.\end{aligned} \tag{3.3.11}$$

Or, en appliquant le lemme A.0.3 aux vecteurs aléatoires $\mathbf{X}^-$ et $\mathbf{Y}^-$, on obtient

$$\mathbb{E}_{\delta=0}\left[\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} \mathbb{I}\left\{\frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} < 1\right\}\right] = \frac{1}{2}\mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right],$$

et donc, par substitution, (3.3.11) devient

$$\begin{aligned}&\left|\tilde{G}_j h(d,0,x) - G_{L,j} h(x)\right| = \\ &\frac{1}{2}\left|h''(x) + h'(x)(\log f_j)'(x)\right| \cdot \left|l^2 \mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right] - v(l)\right|.\end{aligned} \tag{3.3.12}$$

On peut directement en déduire l'espérance de la différence absolue entre les deux variables aléatoires qui nous intéressent :

$$\begin{aligned}&\mathbb{E}\left[\left|\tilde{G}_j h(d,0,X_j) - G_{L,j} h(X_j)\right|\right] = \\ &\frac{1}{2}\mathbb{E}\left[\left|h''(X_j) + h'(X_j)(\log f_j)'(X_j)\right|\right] \cdot \left|l^2 \mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right] - v(l)\right|.\end{aligned}$$

Puisque $h \in \mathcal{C}_c^\infty$ et $(\log f_j)'(X_j) \in L^1$, la première espérance est finie. Il suffit donc de montrer que la valeur absolue converge vers 0 lorsque $d \to \infty$. D'abord, notons que

$$1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} = 1 \wedge \exp\left\{\sum_{i\neq j}(\log f_i(Y_i) - \log f_i(X_i))\right\}.$$

Or, par le lemme 3.2.2, on a

$$\mathbb{E}_{\delta=0}\left[\left|\sum_{i\neq j}(\log f_i(Y_i) - \log f_i(X_i)) - \left(\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right)\right|\right] \longrightarrow 0$$

lorsque $d \to \infty$. Par les lemmes A.0.4 et A.0.5, on en déduit donc que

$$\mathbb{E}_{\delta=0}\left[\left|1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)} - 1 \wedge \exp\left\{\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right\}\right|\right] \longrightarrow 0$$

lorsque $d \to \infty$. Ainsi, si la limite

$$\lim_{d\to\infty} \mathbb{E}_{\delta=0}\left[1 \wedge \exp\left\{\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right\}\right] \tag{3.3.13}$$

existe, alors elle sera égale à celle qui nous intéresse, c'est à dire

$$\lim_{d\to\infty} \mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right]. \tag{3.3.14}$$

Soit B_d la variable aléatoire définie comme en (3.2.5). On voit que l'intérieur de l'exponentielle dans (3.3.13) est une variable aléatoire dont la distribution conditionnelle à $\delta = 0$ et à $\mathbf{X}^-$ est normale :

$$\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2 \,\Big|\, \delta = 0, \mathbf{X}^- \sim N\left(-\frac{l^2}{2}B_d; l^2 B_d\right).$$

Par le lemme A.0.6, on a donc

$$\begin{aligned}
&\mathbb{E}_{\delta=0}\left[1 \wedge \exp\left\{\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right\}\,\middle|\,\mathbf{X}^-\right] \\
&= \Phi\left(\frac{-\frac{l^2}{2}B_d}{l\sqrt{B_d}}\right) + \exp\left\{-\frac{l^2}{2}B_d + \frac{l^2 B_d}{2}\right\}\Phi\left(-l\sqrt{B_d} - \frac{-\frac{l^2}{2}B_d}{l\sqrt{B_d}}\right) \\
&= \Phi\left(-\frac{l\sqrt{B_d}}{2}\right) + \exp\{0\}\Phi\left(-\frac{l\sqrt{B_d}}{2}\right)
\end{aligned}$$

$$= 2\Phi\left(-\frac{l\sqrt{B_d}}{2}\right)$$
$$\equiv \rho_0(B_d).$$

Clairement, ρ_0 est une fonction continue et bornée par 1 en valeur absolue, puisque $\forall x \geq 0$, $\rho_0(x)$ représente l'espérance du minimum entre 1 et une variable aléatoire positive. De plus, le lemme 3.2.3 assure que

$$B_d \xrightarrow{D} B.$$

Par le théorème porte-manteau, on en conclut que

$$\begin{aligned}
&\mathbb{E}_{\delta=0}\left[1 \wedge \exp\left\{\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right\}\right] \\
&= \mathbb{E}_{\delta=0}\left[\mathbb{E}_{\delta=0}\left[1 \wedge \exp\left\{\sum_{i\neq j}(\log f_i)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq j}((\log f_i)'(X_i))^2\right\}\middle| \mathbf{X}^-\right]\right] \\
&= \mathbb{E}\left[\rho_0(B_d)\right] \\
&\longrightarrow \mathbb{E}\left[\rho_0(B)\right] = \rho_0(B) = 2\Phi\left(-\frac{l\sqrt{B}}{2}\right)
\end{aligned}$$

lorsque $d \to \infty$. Puisqu'on a établi que cette espérance était asymptotiquement égale à (3.3.14), on a

$$\mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right] \longrightarrow 2\Phi\left(-\frac{l\sqrt{B}}{2}\right)$$

lorsque $d \to \infty$. On n'a qu'à remplacer cette expression dans (3.3.12) pour obtenir

$$\begin{aligned}
&\mathbb{E}\left[\left|\tilde{G}_j h(d,0,X_j) - G_{L,j}h(X_j)\right|\right] = \\
&\quad \frac{1}{2}\mathbb{E}\left[|h''(X_j) + h'(X_j)(\log f_j)'(X_j)|\right] \cdot \left|l^2\mathbb{E}_{\delta=0}\left[1 \wedge \frac{\pi_d^-(\mathbf{Y}^-)}{\pi_d^-(\mathbf{X}^-)}\right] - v(l)\right| \\
&\quad \longrightarrow \quad \frac{1}{2}\mathbb{E}\left[|h''(X_j) + h'(X_j)(\log f_j)'(X_j)|\right] \cdot \left|2l^2\Phi\left(-\frac{l\sqrt{B}}{2}\right) - v(l)\right| \\
&\quad = 0.
\end{aligned}$$

□

3.4. Le processus limite formé par la première composante de l'algorithme

Pour connaître entièrement le comportement asymptotique d'une certaine composante de l'algorithme, il faut également étudier ce qui se produit lorsqu'un grand pas est proposé, c'est-à-dire lorsque $\delta = 1$. Dans cette section, on établit donc le comportement asymptotique de la première composante de l'algorithme alors que, dans la prochaine section, le comportement asymptotique des autres composantes sera étudié.

Lemme 3.4.1. *Soient $\mathbf{X}^{(d)}$ distribué selon π_d et $\mathbf{Y}^{(d)}$ la proposition définie à la section 2.3. Soit*

$$G_{MH}h(x) = \mathbb{E}_{\delta=1,X_1=x}\left[(h(Y_1) - h(x))\,\alpha(l,x,Y_1)\right],$$

où

$$\alpha(l,x,y) = \Phi\left(\frac{A(x,y) - \frac{l^2}{2}B}{l\sqrt{B}}\right) + e^{A(x,y)}\Phi\left(\frac{-A(x,y) - \frac{l^2}{2}B}{l\sqrt{B}}\right),$$

$$A(x,y) = \log\left(\frac{f_1(y)}{f_1(x)}\right)$$

et posons B tel que défini dans les théorèmes 2.3.1 et 2.3.2.

Notons que G_{MH} représente le générateur d'un algorithme de Metropolis-Hastings avec probabilité d'acceptation $\alpha(l,\cdot,\cdot)$. On a alors

$$d^{-1}G_1h(d,1,X_1) \xrightarrow{L^1} G_{MH}h(X_1)$$

lorsque $d \to \infty$, où $G_jh(d,1,\cdot)$ est défini par (3.1.5).

Démonstration. Rappelons que

$$G_1h(d,1,x) = d\mathbb{E}_{\delta=1,X_1=x}\left[(h(Y_1) - h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right],$$

et donc

$$\begin{aligned}
&\left|d^{-1}G_1h(d,1,x) - G_{MH}h(x)\right| \\
&= \left|\mathbb{E}_{\delta=1,X_1=x}\left[(h(Y_1) - h(x))\left(\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right) - \alpha(l,x,Y_1)\right)\right]\right| \\
&= \left|\mathbb{E}_{\delta=1,X_1=x}\left[\mathbb{E}_{\delta=1,X_1=x}\left[(h(Y_1) - h(x))\left(\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right) - \alpha(l,x,Y_1)\right)\middle| Y_1\right]\right]\right| \\
&= \left|\mathbb{E}_{\delta=1,X_1=x}\left[(h(Y_1) - h(x))\left(\mathbb{E}_{\delta=1,X_1=x}\left[1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle| Y_1\right] - \alpha(l,x,Y_1)\right)\right]\right|.
\end{aligned}$$

En utilisant l'inégalité du triangle et le fait que h est bornée en valeur absolue par la constante K, on trouve

$$\begin{aligned}
&\left|d^{-1}G_1h(d,1,x) - G_{MH}h(x)\right| \\
&\leq \mathbb{E}_{\delta=1,X_1=x}\left[|h(Y_1)-h(x)|\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right|Y_1\right] - \alpha(l,x,Y_1)\right|\right] \\
&\leq 2K\mathbb{E}_{\delta=1,X_1=x}\left[\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right|Y_1\right] - \alpha(l,x,Y_1)\right|\right].
\end{aligned}$$

Il suffit de développer un peu le terme $\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle|Y_1\right]$ pour voir qu'il converge presque sûrement vers $\alpha(l,x,Y_1)$. Comme dans la démonstration du lemme 3.3.3, on peut écrire

$$\begin{aligned}
1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})} &= 1\wedge\frac{f_1(Y_1)}{f_1(X_1)}\prod_{i\neq 1}\frac{f(Y_i)}{f(X_i)} \\
&= 1\wedge\exp\left\{A(X_1,Y_1)+\sum_{i\neq 1}\left(\log f(Y_i)-\log f(X_i)\right)\right\}.
\end{aligned}$$

Or, en utilisant le lemme 3.2.2 avec $j=1$, on a

$$\begin{aligned}
&\mathbb{E}_{\delta=1,X_1=x}\left[\left|\left(A(x,Y_1)+\sum_{i\neq 1}\left(\log f(Y_i)-\log f(X_i)\right)\right)\right.\right. \\
&\quad\left.\left.-\left(A(x,Y_1)+\sum_{i\neq 1}(\log f)'(X_i)(Y_i-X_i)-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right)\right|\right|Y_1\Bigg] \\
&= \mathbb{E}\left[\left|\sum_{i\neq 1}\left(\log f(Y_i)-\log f(X_i)\right)\right.\right. \\
&\quad\left.\left.-\left(\sum_{i\neq 1}(\log f)'(X_i)(Y_i-X_i)-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right)\right|\right],
\end{aligned}$$

ce qui converge vers 0 lorsque $d\to\infty$. Par les lemmes A.0.4 et A.0.5, on en déduit donc que

$$\begin{aligned}
&\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle|Y_1\right] - \mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\exp\left\{A(x,Y_1)\right.\right.\right. \\
&\qquad\left.\left.\left.+\sum_{i\neq 1}(\log f)'(X_i)(Y_i-X_i)-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right\}\middle|Y_1\right]\right| \\
&\leq \mathbb{E}_{\delta=1,X_1=x}\left[\left|1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})} - 1\wedge\exp\left\{A(x,Y_1)\right.\right.\right.
\end{aligned}$$

$$+\sum_{i\neq 1}(\log f)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\Bigg\}\Bigg|\Bigg| Y_1\Bigg]$$
$$\longrightarrow 0 \tag{3.4.1}$$

lorsque $d \to \infty$. Cela signifie que

$$\begin{aligned}
&\lim_{d\to\infty}\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle| Y_1\right] - \alpha(l,x,Y_1)\right| \\
&\leq \lim_{d\to\infty}\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle| Y_1\right] - \mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\exp\left\{A(x,Y_1)\right.\right.\right. \\
&\quad \left.\left.\left.+\sum_{i\neq 1}(\log f)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right\}\middle| Y_1\right]\right| \\
&\quad + \lim_{d\to\infty}\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\exp\left\{A(x,Y_1) + \sum_{i\neq 1}(\log f)'(X_i)(Y_i - X_i)\right.\right.\right. \\
&\quad \left.\left.\left.-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right\}\middle| Y_1\right] - \alpha(l,x,Y_1)\right|.
\end{aligned}$$

En utilisant (3.4.1), nous obtenons alors

$$\begin{aligned}
&\lim_{d\to\infty}\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle| Y_1\right] - \alpha(l,x,Y_1)\right| \\
&\leq 0 + \lim_{d\to\infty}\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\exp\left\{A(x,Y_1) + \sum_{i\neq 1}(\log f)'(X_i)(Y_i - X_i)\right.\right.\right. \\
&\quad \left.\left.\left.-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right\}\middle| Y_1\right] - \alpha(l,x,Y_1)\right| \\
&= \left|\lim_{d\to\infty}\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\exp\left\{A(x,Y_1) + \sum_{i\neq 1}(\log f)'(X_i)(Y_i - X_i)\right.\right.\right. \\
&\quad \left.\left.\left.-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right\}\middle| Y_1\right] - \alpha(l,x,Y_1)\right|.
\end{aligned} \tag{3.4.2}$$

On voit que, si on considère que seulement $\mathbf{Y}^-$ est aléatoire (donc si on conditionne sur les valeurs de $\mathbf{X}^-$ et de Y_1), alors le terme à l'intérieur de l'exponentielle dans 3.4.2 est normalement distribué :

$$A(x,Y_1) + \sum_{i\neq 1}(\log f)'(X_i)(Y_i - X_i) - \frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\Big|\mathbf{X}^-, Y_1$$

$$\sim N\left(A(x,Y_1)-\frac{l^2}{2}B_d; l^2B_d\right),$$

où B_d est défini par (3.2.5). Par le lemme A.0.6, on a donc

$$\begin{aligned}
&\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\exp\left\{A(x,Y_1)+\sum_{i\neq 1}(\log f)'(X_i)(Y_i-X_i)-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right\}\middle|\,\mathbf{X}^-,Y_1\right]\\
&=\Phi\left(\frac{A(x,Y_1)-\frac{l^2}{2}B_d}{l\sqrt{B_d}}\right)+\exp\left\{A(x,Y_1)-\frac{l^2}{2}B_d+\frac{l^2}{2}B_d\right\}\\
&\quad\cdot\Phi\left(-l\sqrt{B_d}-\frac{A(x,Y_1)-\frac{l^2}{2}B_d}{l\sqrt{B_d}}\right)\\
&=\Phi\left(\frac{A(x,Y_1)-\frac{l^2}{2}B_d}{l\sqrt{B_d}}\right)+\exp\left\{A(x,Y_1)\right\}\Phi\left(\frac{-A(x,Y_1)-\frac{l^2}{2}B_d}{l\sqrt{B_d}}\right)\\
&\equiv\rho_1(B_d,x,Y_1).
\end{aligned}$$

Cela signifie que

$$\begin{aligned}
&\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\exp\left\{A(x,Y_1)+\sum_{i\neq 1}(\log f)'(X_i)(Y_i-X_i)-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right\}\middle|\,Y_1\right]\\
&=\mathbb{E}_{\delta=1,X_1=x}\left[\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\exp\left\{A(x,Y_1)+\sum_{i\neq 1}(\log f)'(X_i)(Y_i-X_i)\right.\right.\right.\\
&\qquad\left.\left.\left.-\frac{l^2}{2d}\sum_{i\neq 1}((\log f)'(X_i))^2\right\}\middle|\,\mathbf{X}^-,Y_1\right]\middle|\,Y_1\right]\\
&=\mathbb{E}_{\delta=1,X_1=x}\left[\rho_1(B_d,x,Y_1)\middle|\,Y_1\right].
\end{aligned}$$

Clairement, $\forall x,y,\ \rho_1(\cdot,x,y)$ est une fonction continue et bornée. Par le lemme 3.2.3 et par le théorème porte-manteau, on a donc

$$\mathbb{E}_{\delta=1,X_1=x}\left[\rho_1(B_d,x,Y_1)\middle|\,Y_1\right]\longrightarrow\rho_1(B,x,Y_1)=\alpha(l,x,Y_1)$$

presque sûrement lorsque $d\to\infty$. On peut en déduire que, $\forall x$, on a

$$\begin{aligned}
&\lim_{d\to\infty}\left|\mathbb{E}_{\delta=1,X_1=x}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle|\,Y_1\right]-\alpha(l,x,Y_1)\right|\\
&\leq\lim_{d\to\infty}\left|\mathbb{E}_{\delta=1,X_1=x}\left[\rho_1(B_d,x,Y_1)\middle|\,Y_1\right]-\alpha(l,x,Y_1)\right|=0
\end{aligned}$$

presque sûrement. On peut finalement conclure que

$$\mathbb{E}\left[\left|d^{-1}G_1h(d,1,X_1)-G_{MH}h(X_1)\right|\right]$$

$$
\begin{aligned}
&\leq 2K\mathbb{E}_{\delta=1}\left[\mathbb{E}_{\delta=1}\left[\left|\mathbb{E}_{\delta=1}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle| X_1,Y_1\right]-\alpha(l,X_1,Y_1)\right|\middle| X_1\right]\right]\\
&= 2K\mathbb{E}_{\delta=1}\left[\left|\mathbb{E}_{\delta=1}\left[1\wedge\frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\middle| X_1,Y_1\right]-\alpha(l,X_1,Y_1)\right|\right] \qquad (3.4.3)\\
&\longrightarrow 0
\end{aligned}
$$

lorsque $d\to\infty$ par le théorème de convergence dominée, puisque l'intérieur de l'espérance dans (3.4.3) est une variable aléatoire bornée qui converge vers 0 presque sûrement.

□

Les lemmes 3.3.1 et 3.4.1 nous permettent d'obtenir le processus limite formé par la première composante de l'algorithme. Le comportement asymptotique de cette composante est résumé dans le théorème suivant.

Théorème 3.4.1. *Soient $\mathbf{X}^{(d)}$ distribué selon π_d et $\mathbf{Y}^{(d)}$ la proposition telle que définie à la section 2.3. Soient G_1 le générateur de la chaîne accélérée formée par la première composante de l'algorithme et*

$$G_{ML}h(x)=G_{L,1}h(x)+\beta G_{MH}h(x).$$

On a alors

$$G_1h(d,X_1)\xrightarrow{L^1}G_{ML}h(X_1)$$

lorsque $d\to\infty$.

Démonstration. À la section 3.1, nous avons établi que

$$G_1h(d,x)=(1-p(d))G_1h(d,0,x)+p(d)G_1h(d,1,x).$$

Puisqu'on s'intéresse au comportement asymptotique, nous supposons, sans perte de généralité, que $d\geq\beta$. Cette hypothèse nous permet de poser $p(d)=\frac{\beta}{d}$, ce qui nous mène à

$$G_1h(d,x)=\left(1-\frac{\beta}{d}\right)G_1h(d,0,x)+\frac{\beta}{d}G_1h(d,1,x). \qquad (3.4.4)$$

En utilisant (3.4.4) et la définition de G_{ML}, on a

$$
\begin{aligned}
&|G_1h(d,x)-G_{ML}h(x)|\\
&=\left|\left(1-\frac{\beta}{d}\right)G_1h(d,0,x)-G_{L,1}h(x)+\frac{\beta}{d}G_1h(d,1,x)-\beta G_{MH}h(x)\right|
\end{aligned}
$$

$$= \left| G_1 h(d,0,x) - G_{L,1} h(x) - \frac{\beta}{d} G_1 h(d,0,x) + \frac{\beta}{d} G_1 h(d,1,x) - \beta G_{MH} h(x) \right|$$
$$\leq |G_1 h(d,0,x) - G_{L,1} h(x)| + \beta \left| d^{-1} G_1 h(d,1,x) - G_{MH} h(x) \right| + \frac{\beta}{d} |G_1 h(d,0,x)| .$$

À l'aide des lemmes 3.3.1 et 3.4.1, on obtient

$$\begin{aligned}
\lim_{d\to\infty} \mathbb{E}\left[|G_1 h(d, X_1) - G_{ML} h(X_1)| \right] &\leq \lim_{d\to\infty} \frac{\beta}{d} \mathbb{E}\left[|G_1 h(d,0,X_1)| \right] \\
&= \beta \lim_{d\to\infty} \mathbb{E}\left[\left| \mathbb{E}_{\delta=0}\left[(h(Y_1) - h(X_1)) \left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})} \right) \right| X_1 \right] \right] \\
&\leq \beta \lim_{d\to\infty} \mathbb{E}_{\delta=0}\left[|h(Y_1) - h(X_1)| \left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})} \right) \right] \\
&\leq \beta \lim_{d\to\infty} \mathbb{E}_{\delta=0}\left[|h(Y_1) - h(X_1)| \right] .
\end{aligned}$$

Un développement de Taylor d'ordre 1 par rapport à Y_1 et autour de X_1 nous donne, pour un certain U entre X_1 et Y_1,

$$\begin{aligned}
\lim_{d\to\infty} \mathbb{E}\left[|G_1 h(d, X_1) - G_{ML} h(X_1)| \right] &\leq \beta \lim_{d\to\infty} \mathbb{E}_{\delta=0}\left[|h'(U)(Y_1 - X_1)| \right] \\
&\leq \beta K \lim_{d\to\infty} \mathbb{E}_{\delta=0}\left[|Y_1 - X_1| \right] .
\end{aligned}$$

Par le lemme A.0.2, nous concluons que

$$\lim_{d\to\infty} \mathbb{E}\left[|G_1 h(d, X_1) - G_{ML} h(X_1)| \right] \leq \beta K \lim_{d\to\infty} \sqrt{\frac{\pi}{2}} \frac{l}{\sqrt{d}} = 0.$$

□

3.5. Le processus limite formé par la seconde composante de l'algorithme

Contrairement au cas de la première composante de l'algorithme, pour connaître entièrement le comportement asymptotique de la seconde composante, on n'a pas besoin d'étudier ce qui se passe lorsque $\delta = 1$. En effet, le processus limite est continu lorsque des petits pas sont proposés alors que de grands pas ne sont proposés qu'un nombre fini de fois par intervalle de temps. Il suffit donc de s'assurer que le processus reste continu à ces endroits. Le lemme suivant le vérifie.

Lemme 3.5.1. *Soient $\mathbf{X}^{(d)}$ distribué selon π_d et $\mathbf{Y}^{(d)}$ la proposition définie à la section 2.3. On a*

$$d^{-1}\left|G_2h(d,1,X_2) - G_2h(d,0,X_2)\right| \xrightarrow{L^1} 0$$

lorsque $d \to \infty$.

Démonstration. Rappelons que

$$G_2h(d,k,x) = d\mathbb{E}_{\delta=k,X_2=x}\left[(h(Y_2)-h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right].$$

On a alors

$$\begin{aligned}
&d^{-1}\left|G_2h(d,1,x) - G_2h(d,0,x)\right| \\
&= \left|\mathbb{E}_{\delta=1,X_2=x}\left[(h(Y_2)-h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right]\right. \\
&\quad \left. -\mathbb{E}_{\delta=0,X_2=x}\left[(h(Y_2)-h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right]\right| \\
&\leq \left|\mathbb{E}_{\delta=1,X_2=x}\left[(h(Y_2)-h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right]\right| \\
&\quad + \left|\mathbb{E}_{\delta=0,X_2=x}\left[(h(Y_2)-h(x))\left(1 \wedge \frac{\pi_d(\mathbf{Y}^{(d)})}{\pi_d(\mathbf{X}^{(d)})}\right)\right]\right|.
\end{aligned}$$

Puisque la probabilité d'acceptation est positive et inférieure ou égale à 1, on obtient la borne suivante :

$$\begin{aligned}
d^{-1}\left|G_2h(d,1,x) - G_2h(d,0,x)\right| &\leq \mathbb{E}_{\delta=1,X_2=x}\left[|h(Y_2)-h(x)|\right] + \mathbb{E}_{\delta=0,X_2=x}\left[|h(Y_2)-h(x)|\right] \\
&= 2\mathbb{E}_{X_2=x}\left[|h(Y_2)-h(x)|\right],
\end{aligned}$$

où l'égalité vient du fait que la variable aléatoire Y_2 est indépendante de δ (les composantes 2 à d ne font toujours que des petits pas, peu importe la valeur de δ). Par une linéarisation de Taylor de h par rapport à Y_2 et autour de X_2, ceci nous permet de conclure que

$$\mathbb{E}\left[d^{-1}\left|G_2h(d,1,X_2) - G_2h(d,0,X_2)\right|\right] \leq 2\mathbb{E}\left[|h(Y_2)-h(X_2)|\right] = 2\mathbb{E}\left[|h'(U)(Y_2-X_2)|\right],$$

pour un certain U entre X_2 et Y_2. Puisque h' est bornée en valeur absolue par la constante K, on a donc

$$\mathbb{E}\left[d^{-1}\left|G_2h(d,1,X_2) - G_2h(d,0,X_2)\right|\right] \leq 2K\mathbb{E}\left[|Y_2-X_2|\right] = 2K\sqrt{\frac{\pi}{2}}\frac{l}{\sqrt{d}}$$

par le lemme A.0.2. Le résultat est obtenu en notant que cette dernière expression converge vers 0 lorsque $d \to \infty$.

□

Nous sommes maintenant prêts à obtenir le processus limite formé par la seconde composante de l'algorithme. Il s'agit bien sûr d'un processus de diffusion continu de Langevin.

Théorème 3.5.1. *Soient* $\mathbf{X}^{(d)}$ *distribué selon* π_d *et* $\mathbf{Y}^{(d)}$ *la proposition définie précédemment. Soient* G_2 *le générateur de la chaîne accélérée formée par la seconde composante de l'algorithme et* $G_{L,2}$ *défini tel que dans le lemme 3.3.1. On a alors*

$$G_2 h(d, X_2) \xrightarrow{L^1} G_{L,2} h(X_2)$$

lorsque $d \to \infty$.

Démonstration. On rappelle qu'on a établi, à la section 3.1, que

$$G_2 h(d,x) = \left(1 - \frac{\beta}{d}\right) G_2 h(d,0,x) + \frac{\beta}{d} G_2 h(d,1,x), \tag{3.5.1}$$

en supposant bien sûr que $d \geq \beta$. En utilisant (3.5.1), on a

$$\begin{aligned}
&|G_2 h(d,x) - G_{L,2} h(x)| \\
&= \left| \left(1 - \frac{\beta}{d}\right) G_2 h(d,0,x) - G_{L,2} h(x) + \frac{\beta}{d} G_2 h(d,1,x) \right| \\
&= \left| G_2 h(d,0,x) - G_{L,2} h(x) - \frac{\beta}{d} G_2 h(d,0,x) + \frac{\beta}{d} G_2 h(d,1,x) \right| \\
&\leq |G_2 h(d,0,x) - G_{L,2} h(x)| + \frac{\beta}{d} |G_2 h(d,1,x) - G_2 h(d,0,x)| .
\end{aligned}$$

On obtient donc

$$\begin{aligned}
&\mathbb{E}\left[|G_2 h(X_2) - G_{L,2} h(X_2)|\right] \\
&\leq \mathbb{E}\left[|G_2 h(d,0,X_2) - G_{L,2} h(X_2)|\right] + \beta \mathbb{E}\left[d^{-1} |G_2 h(d,1,X_2) - G_2 h(d,0,X_2)|\right] .
\end{aligned}$$

Ces deux termes convergent vers 0 lorsque $d \to \infty$, par les lemmes 3.3.1 et 3.5.1, repectivement.

□

Chapitre 4

SIMULATIONS

Le présent chapitre est dédié à la mise en pratique du nouvel algorithme selon la stratégie décrite au chapitre 2. Nous présentons également une brève étude numérique de la robustesse de ce nouvel algorithme. Nous considérons, dans l'exemple de la section 4.2, une distribution cible qui viole les conditions imposées par les résultats du chapitre 2.

Notons que pour évaluer la performance de l'algorithme, nous nous concentrons principalement sur l'estimation de la densité cible (à l'aide de la fonction «density» en R, qui utilise la méthode des noyaux gaussiens). La bonne estimation de la densité traduit généralement un échantillon représentatif de la distribution cible. Or, dans certains cas, une simple espérance peut être estimée avec précision à l'aide de l'estimateur de Monte Carlo (1.1.1) même si l'échantillon n'est pas parfaitment représentatif. Ainsi, il pourrait également être intéressant d'étudier la performance de l'estimateur de Monte Carlo en utilisant le nouvel algorithme.

Les simulations présentées ici ne sont pas une étude complète, mais de simples exemples d'application du nouvel algorithme ainsi que de l'algorithme RWM. Bien sûr, il serait intéressant de mener une étude de simulation afin de comparer la performance de cet algorithme à celle d'autres méthodes (par exemple celles abordées dans l'introduction) en terme de précision et de temps de calcul. Il faudrait de plus considérer plusieurs formes pour la distribution cible, ainsi que plusieurs valeurs différentes pour la dimension d du problème. Ce sont des points qui seront abordés ultérieurement (notamment dans [4]).

4.1. Retour sur l'exemple normal bimodal

Considérons à nouveau l'exemple bimodal de la section 1.3.3, mais cette fois avec une dimension $d = 100$. On a donc

$$\mathbf{X} \sim \begin{cases} N_{100}(-\mu; 9I_{100}), \text{ avec probabilité } \frac{1}{2} \\ N_{100}(\mu; 9I_{100}), \text{ avec probabilité } \frac{1}{2} \end{cases}, \tag{4.1.1}$$

où $\mu = (15, 0, ..., 0)^\top \in \mathbb{R}^{100}$ et I_{100} représente la matrice identité de format 100×100. Il est évident que les composantes 2 à d de $\mathbf{X}$ sont distribuées selon une loi $N(0; 9)$, alors que la première composante de $\mathbf{X}$ est distribuée selon la densité f_1, donnée par

$$f_1(x) = \frac{1}{2}\left(\frac{1}{3}\phi\left(\frac{x+15}{3}\right) + \frac{1}{3}\phi\left(\frac{x-15}{3}\right)\right),$$

où ϕ représente la densité de la loi normale standard.

Le fait d'augmenter de telle façon la dimension de la distribution cible réduit encore plus l'efficacité de l'algorithme RWM classique, puisque l'écart-type instrumental requis pour que le taux d'acceptation soit raisonnable est petit (on se rappelle que l'écart-type instrumental optimal est proportionnel à $d^{-\frac{1}{2}}$). Les chances de changer de mode tout en ayant un taux d'acceptation raisonnable sont donc presque nulles. C'est pourquoi on utilise le nouvel algorithme afin d'obtenir notre échantillon, qu'on veut de taille $n = 1\,000\,000$.

4.1.1. Choix des paramètres

Tel que proposé dans le chapitre 2, on utilise un écart-type instrumental de $\sigma = \frac{\hat{l}}{\sqrt{d}}$ pour les petits pas, où $\hat{l}$ est la valeur optimale trouvée par [**20**]. Afin de trouver cette valeur, on applique l'algorithme RWM classique à la distribution cible donnée par (4.1.1) avec une taille échantillonnale de $n_{\text{test}} = 1\,000$, et ce, pour chaque valeur de $l \in \left\{\frac{1}{10}, \frac{2}{10}, ..., \frac{99}{10}, 10\right\}$. En représentant le taux d'acceptation obtenu à partir de chacune de ces 100 simulations (qui comportent chacune 1 000 itérations) sur la figure 4.1, on voit que, sur cet intervalle, celui-ci semble diminuer linéairement avec l'augmentation de l. Il suffit donc de choisir la valeur de l qui nous donne le taux d'acceptation le plus près de 0,234, qui est ici 7,2. On va donc utiliser l'écart-type $\sigma = \frac{\hat{l}}{\sqrt{d}} = 0{,}72$. Notons qu'en pratique, il aurait suffit d'essayer plusieurs valeurs de l l'une après l'autre jusqu'à en trouver une qui produit un taux d'acceptation suffisamment proche de 0,234, mais nous souhaitions ici visualiser la relation entre l et le taux d'acceptation.

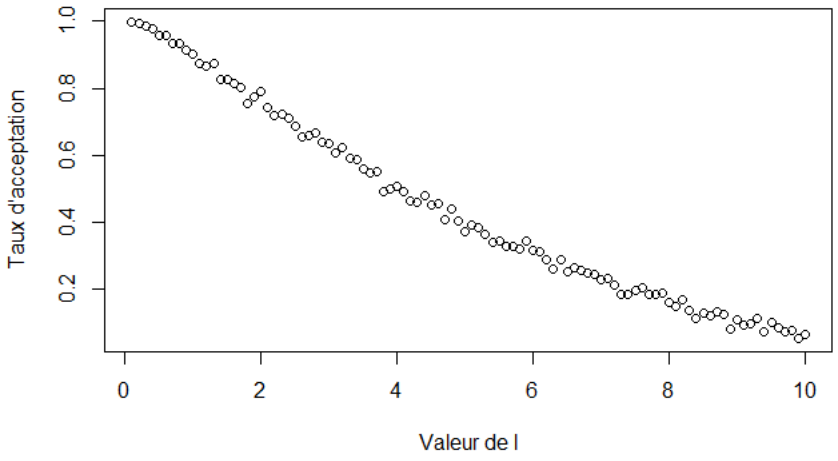

FIGURE 4.1. Taux d'acceptation des petits pas proposés en fonction du paramètre d'échelle l.

Remarquons qu'ici, puisque la loi des $d-1$ composantes iid est normale, la valeur optimale de l aurait pu être calculée théoriquement. En effet, supposons que f est la densité de la loi normale d'espérance μ et de variance τ^2, et supposons que X est une variable aléatoire distribuée selon cette loi. La valeur de B est alors τ^{-2} :

$$B = \mathbb{E}\left[((\log f)'(X))^2\right] = \mathbb{E}\left[\left(\frac{d}{dX}\left(-\frac{1}{2\tau^2}(X-\mu)^2\right)\right)^2\right] = \mathbb{E}\left[\left(\frac{X-\mu}{\tau^2}\right)^2\right] = \frac{1}{\tau^2}.$$

Ainsi, dans notre situation, puisque $\tau^2 = 9$, on trouverait $\hat{l} = \frac{2{,}38}{\sqrt{B}} = 2{,}38\tau = 7{,}14$. Or, puisque l'objectif de ce chapitre est d'étudier la performance de l'algorithme et puisqu'en général on doit se baser sur une approximation numérique pour $\hat{l}$, on va utiliser la valeur qu'on a trouvée numériquement, soit $\hat{l} = 7{,}2$.

Pour ce qui est de la densité instrumentale g_1, on retient deux choix : la loi uniforme sur un intervalle de la forme $[X_1 \pm c]$, puisqu'il a été démontré qu'il s'agit de la «meilleure», ainsi que la loi normale centrée à X_1 de variance σ_0^2 suffisamment grande, où X_1 désigne l'état actuel de la chaîne.

Afin de choisir le paramètre c (ou σ_0^2 pour la loi normale) et le paramètre p, on doit estimer la probabilité dénotée q dans la section 2.4. Celle-ci représente la probabilité de changer de mode (donc de proposer un changement de mode et de l'accepter selon la probabilité (2.4.2)), sachant que l'on propose un candidat avec la densité instrumentale g_1. Pour ce faire, on simule plusieurs algorithmes RWM sur $\mathbb{R}$ avec distribution cible f_1 et distribution instrumentale uniforme (pour différentes valeurs de c), ainsi que la probabilité d'acceptation

(2.4.2). Nous répétons ensuite le même processus, en remplaçant la loi uniforme par une loi normale (pour différentes valeurs de σ_0^2). Notons que la stratégie proposée pour les valeurs de c qui sont ainsi «testées» est de les choisir autour de la distance entre les modes (qui est ici 30), puisque c'est en moyenne la largeur des grands pas que l'on souhaite proposer. Avec une taille échantillonnale de $n_{\text{test}} = 100\,000$, cette fois, on obtient les probabilités de changement de mode représentées par les figures 4.2 et 4.3. On observe que la distribution

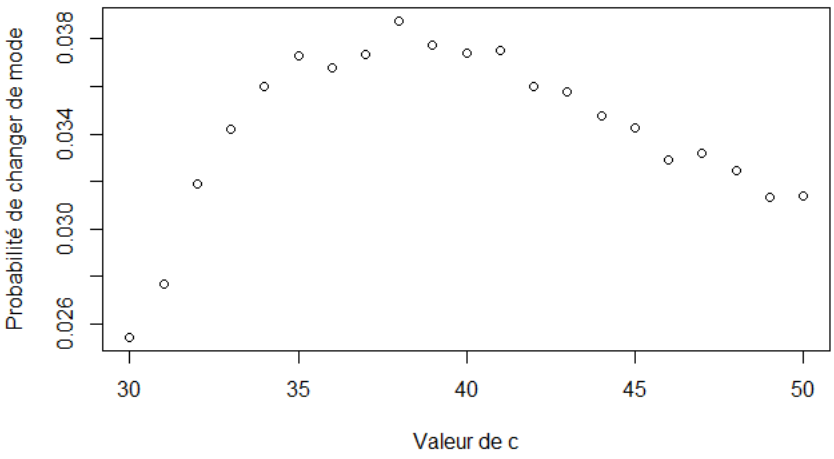

FIGURE 4.2. Probabilité estimée de changement de mode lorsqu'un grand pas est proposé selon une loi uniforme en fonction du paramètre d'échelle c.

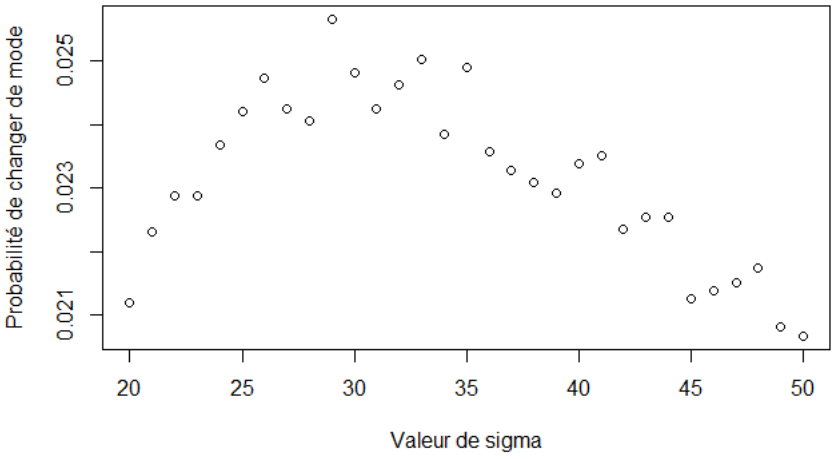

FIGURE 4.3. Probabilité estimée de changement de mode lorsqu'un grand pas est proposé selon une loi normale en fonction du paramètre d'échelle σ.

instrumentale qui maximise la valeur de $\hat{q}$ semble être la loi uniforme avec une valeur de c préférablement dans l'intervalle $[35; 41]$, pour $\hat{q} \approx 0{,}037$. Par conséquent, en utilisant une densité g_1 uniforme sur $[X_1 \pm 38]$, où X_1 désigne l'état actuel de la première composante de la chaîne, et $p = \frac{1\,000}{n\hat{q}} \approx 0{,}027$, notre algorithme est entièrement spécifié. Nous sommes donc prêts à lancer la vraie simulation.

4.1.2. Résultats

En appliquant l'algorithme avec les paramètres mentionnés et avec $n = 1\,000\,000$ itérations, on obtient la chaîne dont la trace est représentée à la figure 4.4. On remarque qu'il y a énormément de déplacements sur le domaine. Cela se traduit par une bonne estimation de la densité cible f_1 (la densité de la première composante seulement) comme on peut le constater à la figure 4.5. Notons que le nombre d'observations qui ont dû être simulées pour le choix des paramètres est d'environ $5\,000\,000$, ce qui est supérieure à la taille de l'échantillon désiré. Or, ce nombre n'est pas fonction de la dimension d du problème, puisque les simulations effectuées pour choisir le paramètre c optimal sont faites uniquement sur la première composante de la distribution cible. Ainsi, pour un problème de très grande dimension ($d >> 100$), où le nombre d'itérations requis sera possiblement bien plus grand que $1\,000\,000$, le temps de calcul requis pour le choix des paramètres deviendra négligeable.

4.2. Une distribution cible avec structure de dépendance

Bien que les résultats aient été démontrés pour une forme bien précise de la densité cible π_d, la méthode présentée dans ce mémoire se veut applicable à des cas plus généraux. C'est pourquoi nous allons considérer une légère modification de la distribution cible (4.1.1) qui consiste à y ajouter une structure simple de corrélation : supposons que, pour $i \neq j$, la corrélation entre les i-ème et j-ème termes est de $\frac{1}{2}$. Cette structure peut être représentée par la matrice de corrélation sphérique

$$\Sigma_{100} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & \cdots & \frac{1}{2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{100\times 100}.$$

De plus, le poids accordé à chaque mode, qui a toujours été $\frac{1}{2}$ jusqu'ici, sera maintenant de $\frac{1}{3}$ pour le mode négatif et de $\frac{2}{3}$ pour le mode positif.

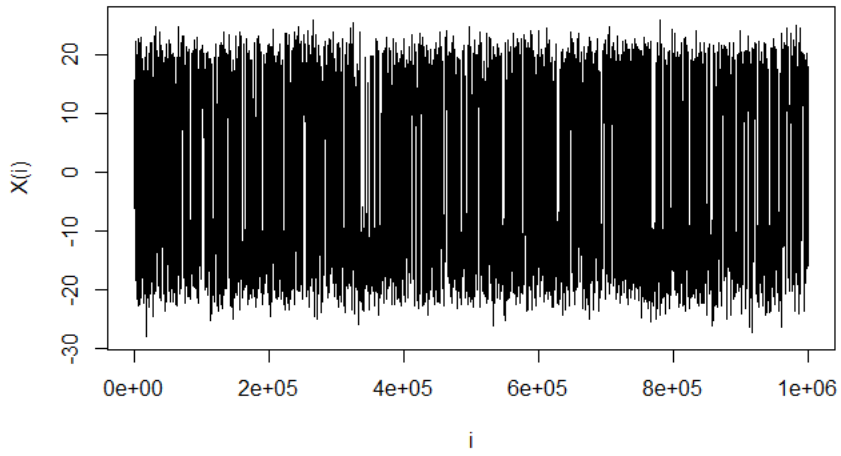

FIGURE 4.4. Trace de la chaîne engendrée par la première composante de l'algorithme.

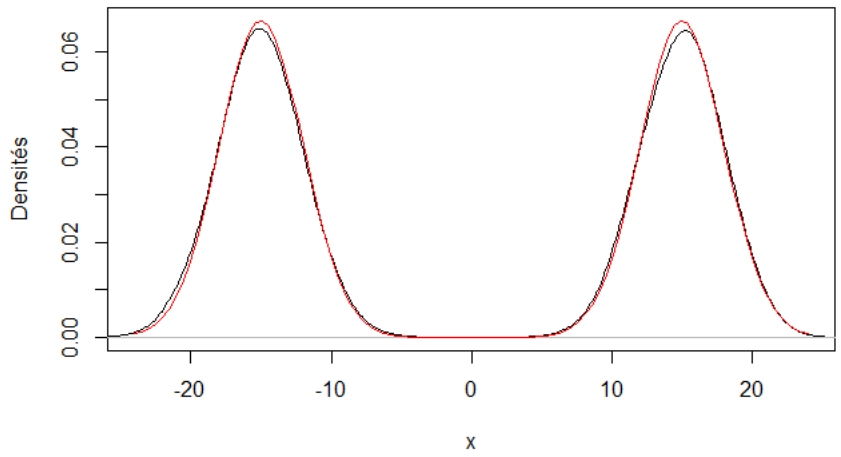

FIGURE 4.5. Estimation de la densité cible f_1.

La distribution à estimer est donc celle de la variable aléatoire

$$\mathbf{X} \sim \begin{cases} N_{100}(-\mu; 9\Sigma_{100}), \text{ avec probabilité } \frac{1}{3} \\ N_{100}(\mu; 9\Sigma_{100}), \text{ avec probabilité } \frac{2}{3} \end{cases},$$

où $\mu = (15, 0, ..., 0)^\top \in \mathbb{R}^{100}$. Remarquons que pour l'implantation de l'algorithme, on a besoin de la forme de la densité f_1 de la première composante de $\mathbf{X}$. Or, on peut facilement voir que la loi de cette composante est toujours un mélange de deux lois normales de variance

9 et d'espérance ± 15. Sa densité est donc donnée par

$$f_1(x) = \frac{1}{3}\left(\frac{1}{3}\phi\left(\frac{x+15}{3}\right)\right) + \frac{2}{3}\left(\frac{1}{3}\phi\left(\frac{x-15}{3}\right)\right), \tag{4.2.1}$$

où ϕ représente toujours la densité de la loi normale standard.

4.2.1. Choix des paramètres

Comme précédemment, on souhaite trouver une valeur de l qui nous donne un taux d'acceptation d'environ 0,234. On utilise donc la même stratégie que pour l'exemple précédent ; pour chaque valeur de $l \in \left\{\frac{1}{10}, \frac{2}{10}, ..., \frac{99}{10}, 10\right\}$, on estime le taux d'acceptation en simulant un algorithme RWM classique avec une taille échantillonnale de $n_{\text{test}} = 1\,000$ et on choisit celui qui est le plus près de 0,234. Cette fois, la décroissance du taux d'acceptation par rapport à l ne semble pas linéaire, comme l'indique la figure 4.6. Avec nos données, on obtient $\hat{l} = 5{,}2$.

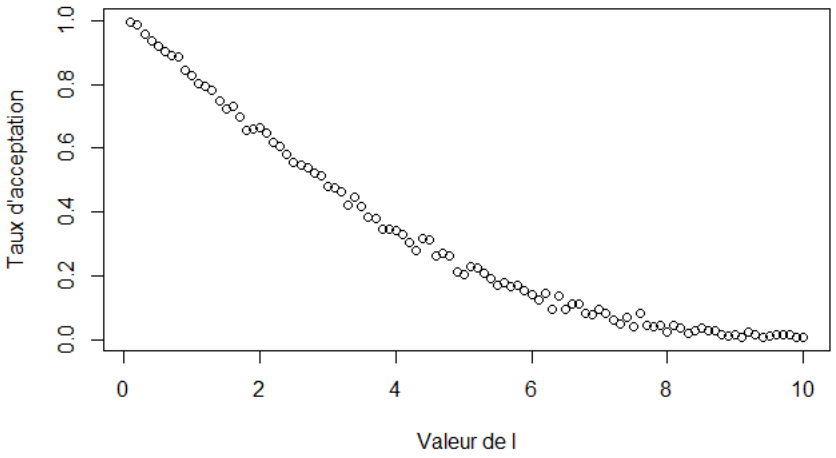

FIGURE 4.6. Taux d'acceptation des petits pas proposés en fonction du paramètre d'échelle l.

On va donc utiliser l'écart-type instrumental $\sigma = \frac{\hat{l}}{\sqrt{d}} = 0{,}52$.

Pour ce qui est de la densité instrumentale g_1, sans surprise, le meilleur choix (entre uniforme et normale) semble toujours être la densité uniforme avec une valeur de c autour de 37. La probabilité de changement de mode estimée avec le choix optimal de g_1, $\hat{q}$, est d'environ 0,036, ce qui est très semblable à précédemment. En posant g_1 la densité de la loi uniforme sur $[X_1 \pm 37]$, où X_1 désigne l'état actuel de la première composante de la chaîne,

et $p = \frac{1\,000}{n\hat{q}} \approx 0{,}028$, notre algorithme est entièrement spécifié. Nous sommes donc prêts à lancer la vraie simulation.

4.2.2. Résultats

Avec $n = 1\,000\,000$ itérations, la première composante de la chaîne obtenue est représentée à la figure 4.7. Il y a évidemment un aussi grand nombre de changements de mode

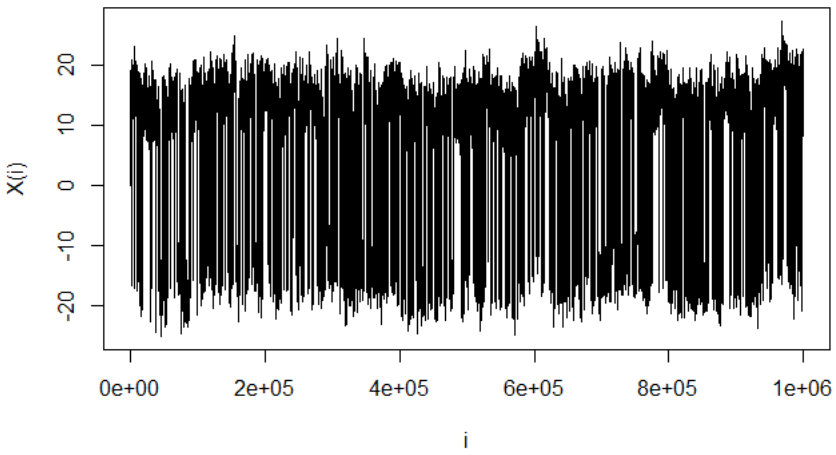

FIGURE 4.7. Trace de la chaîne engendrée par la première composante de l'algorithme.

qu'auparavant, puisque notre méthode fixe l'espérance de ce nombre. Or, la chaîne passe plus de temps dans le mode positif que dans le mode négatif. L'estimation de la densité f_1 semble légèrement moins précise que précédemment, comme on le voit à la figure 4.8.

4.2.3. Comparaison avec les résultats obtenus sur le modèle de base

Afin de déterminer l'effet sur l'estimation de certains paramètres des modèles considérés dans les deux exemples précédents, nous comparons la simulation de la section 4.2 avec celle effectuée sur le modèle sans corrélation (section 4.1). Notons d'abord que, dans le premier exemple où les d composantes du modèle étaient indépendantes, la pondération théorique du mode positif, c'est-à-dire la probabilité $\mathbb{P}(X_1 > 0)$, était de $\frac{1}{2}$. En estimant cette pondération à l'aide de la proportion des valeurs de X_1 positives dans notre échantillon, on obtenait environ 0,505.

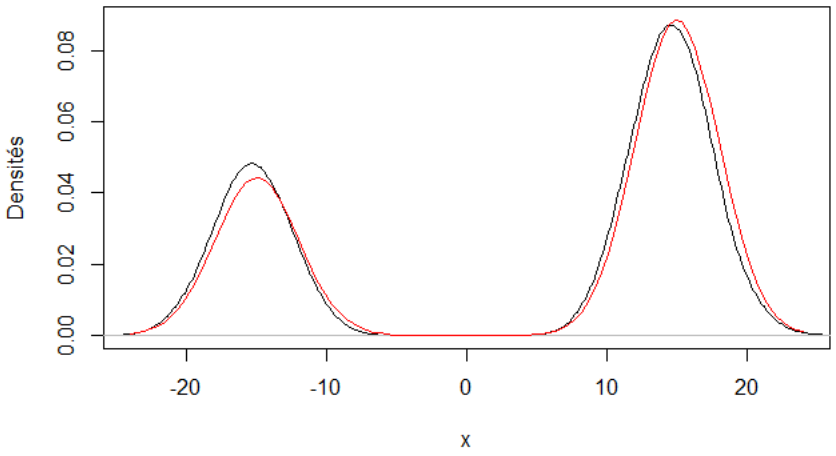

FIGURE 4.8. Estimation de la densité cible f_1.

Pour le second exemple, cette probabilité théorique n'est plus la même. À l'aide d'un calcul direct, on trouve qu'elle est égale à $\frac{1}{3}(1 + \Phi(5)) \approx \frac{2}{3}$, où Φ représente toujours la fonction de répartition de la loi normale standard. En estimant cette probabilité à l'aide de la proportion des valeurs de X_1 positives dans notre échantillon, on obtient cette fois environ 0,652. L'erreur absolue pour l'estimation du poids des modes a donc légèrement augmenté.

On pourrait également s'intéresser, tout simplement, à l'estimation de $\mathbb{E}[\mathbf{X}]$. Il est facile de voir que cette espérance est 0 dans le premier exemple et $(5, 0, ..., 0)^\top$ dans le second. En calculant l'erreur quadratique dans chacun des deux cas, c'est-à-dire

$$\left\| \bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}] \right\|_2^2, \tag{4.2.2}$$

où $\bar{\mathbf{X}} \in \mathbb{R}^{100}$ représente le vecteur des moyennes des 100 composantes calculées sur l'échantillon généré par l'algorithme, on obtient environ 0,32 dans le premier exemple et 16,33 dans le second. Cette augmentation par un facteur d'environ 50 est assez drastique et s'explique vraisemblablement par l'ajout de la structure de corrélation. En effet, en calculant l'erreur quadratique due à chacune des d composantes, on peut voir que l'erreur due à la première composante n'a pas changé, alors que l'erreur due aux $d - 1$ composantes iid est principalement responsable de l'augmentation de l'erreur totale (4.2.2). On ne peut donc pas attribuer cette augmentation à la pondération de chaque mode, qui a également changé entre les sections 4.1 et 4.2. Notons finalement que dans le second exemple, chacune des 100 composantes de $\bar{\mathbf{X}}$ est inférieure à la moyenne théorique. Cela s'explique par la corrélation positive entre chacune des composantes de la distribution cible, ce qui entraîne une corrélation positive entre chacune des composantes de $\bar{\mathbf{X}}$.

En somme, la perte de précision dans l'estimation est attribuable à l'ajout de la structure de corrélation dans la distribution cible, alors que l'algorithme a été développé pour une distribution cible dont les composantes sont indépendantes. Une telle perte aurait également été observée avec l'algorithme RWM.

4.3. Une distribution cible presque unimodale

Lorsqu'on est en présence d'une fonction de densité cible dont l'expression est compliquée, on peut avoir du mal à estimer la distance entre les modes. Il peut arriver, toutefois, qu'on ait de bonnes raisons de croire que cette distance est suffisamment petite pour que le «trou» entre les modes n'en soit pas un, c'est-à-dire qu'il n'y ait pas de région de très faible densité entre les deux modes. Dans cette zone grise où la densité cible n'est pas unimodale mais n'est pas non plus composée de deux modes complètement séparés, il est raisonnable de croire que l'algorithme RWM classique sera efficace. Bien sûr, le nouvel algorithme, avec un choix raisonnable des paramètres, le sera également. De plus, l'utilisation du nouvel algorithme serait un choix plus prudent, considérant que si on a légèrement sous-estimé la profondeur du trou entre les deux modes, il devrait tout de même être efficace. Une question s'impose. Perd-on beaucoup, en terme de précision de l'estimation, en utilisant le nouvel algorithme sur une distribution cible qui peut être estimée à l'aide de l'algorithme RWM classique ?

Supposons que la distribution cible est celle de la section précédente, sauf que la pondération de chacun des deux modes est à nouveau $\frac{1}{2}$ et supposons qu'on a rapproché les modes. On veut donc estimer la distribution de la variable aléatoire

$$\mathbf{X} \sim \begin{cases} N_{100}(-\mu; 9\Sigma_{100}), \text{ avec probabilité } \frac{1}{2} \\ N_{100}(\mu; 9\Sigma_{100}), \text{ avec probabilité } \frac{1}{2} \end{cases},$$

où la matrice Σ_{100} est définie comme auparavant et où $\mu = (4, 0, ..., 0)^\top$. On voit facilement que la densité de la première composante est désormais donnée par

$$f_1(x) = \frac{1}{2}\left(\frac{1}{3}\phi\left(\frac{x+4}{3}\right) + \frac{1}{3}\phi\left(\frac{x-4}{3}\right)\right),$$

où ϕ représente la densité de la loi normale standard. Elle est représentée à la figure 4.9. On va utiliser l'algorithme RWM ainsi que le nouvel algorithme, toujours avec une taille échantillonnale de $n = 1\,000\,000$, afin d'estimer cette distribution.

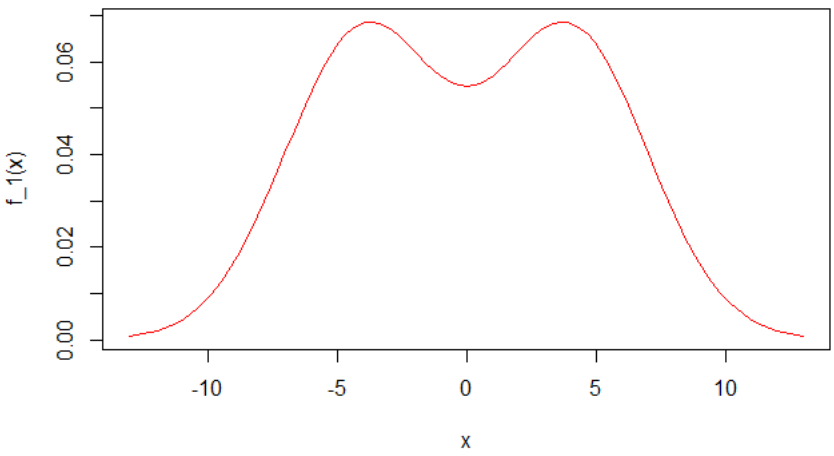

FIGURE 4.9. Densité de la première composante de $\mathbf{X}$.

4.3.1. Choix des paramètres

Comme toujours, nous commençons par trouver $\hat{l}$ en simulant l'algorithme RWM avec plusieurs valeurs différentes de l et une taille échantillonnale de $n_{\text{test}} = 1\,000$, puis en calculant le taux d'acceptation de chaque simulation. Cette fois, la valeur de l qui produit le taux d'acceptation le plus près de 0,234 est $\hat{l} = 4{,}78$. L'écart-type instrumental sera donc de $\sigma = 0{,}478$.

On utilise la même stratégie qu'auparavant pour choisir la densité cible g_1, c'est-à-dire en estimant, avec $n_{\text{test}} = 100\,000$, la probabilité de changer de mode lorsqu'un grand pas est proposé selon une loi uniforme ou normale avec plusieurs valeurs différentes pour le paramètre d'échelle. Par contre, puisqu'ils ne sont plus aussi distincts qu'auparavant, il importe de préciser ici que les modes sont définis comme étant les deux régions qui sont séparées par l'hyperplan $\left\{x = (x_1, ..., x_d)^\top \in \mathbb{R}^d : x_1 = 0\right\}$. C'est la même définition qu'auparavant. Ainsi, un changement de mode est défini par un changement de signe de la première composante de l'algorithme. Le meilleur choix semble toujours être la loi uniforme, mais bien sûr, la valeur optimale du paramètre c est inférieure à celle qu'on utilisait précédemment, puisque les modes ont été rapprochés. Cette fois, la valeur optimale semble être $c = 13$, pour une probabilité estimée de changement de mode de $\hat{q} = 0{,}07$. Notons que l'augmentation dans la valeur de $\hat{q}$ est également due au rapprochement entre les modes. En posant g_1 la densité de la loi uniforme sur $[X_1 \pm 13]$, où X_1 désigne l'état actuel de la première composante de

la chaîne, et $p = \frac{1\,000}{n\hat{q}} \approx 0{,}014$, notre algorithme est entièrement spécifié. Nous sommes donc prêts à lancer la vraie simulation.

4.3.2. Résulats obtenus avec l'algorithme classique

L'algorithme RWM classique affiche un bon déplacement sur tout le domaine (la chaîne ne semble pas rester «prisonnière» des modes), comme on le voit à la figure 4.10. Cependant, cet algorithme n'estime pas à la perfection la densité (voir la figure 4.11). L'erreur quadratique commise lors de l'estimation de $\mathbb{E}[\mathbf{X}]$, qui est ici $\mathbf{0}$, est d'environ 50,87. L'estimation de la probabilité associée au mode positif, $\mathbb{P}(X_1 > 0)$, qu'on sait égale à $\frac{1}{2}$, est d'environ 0,464.

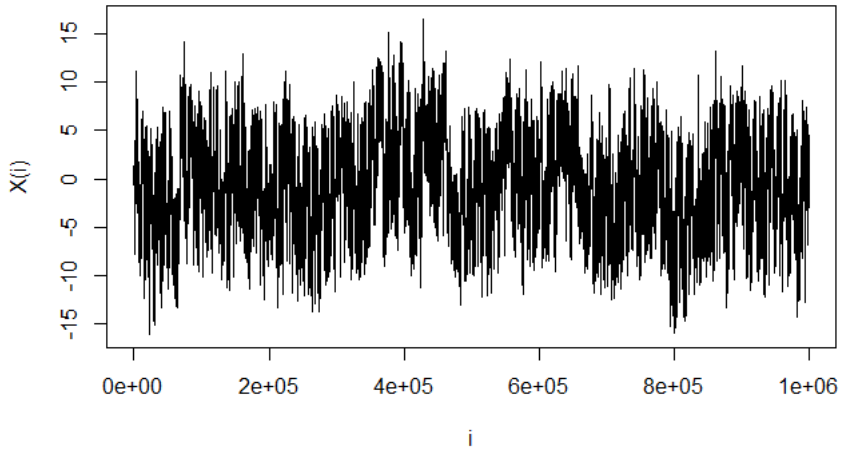

FIGURE 4.10. Trace de la chaîne engendrée par la première composante (algorithme RWM).

4.3.3. Résultats obtenus avec le nouvel algorithme

Avec le nouvel algorithme, on voit à la figure 4.12 qu'il semble y avoir encore plus de mouvements qu'avec l'algorithme RWM. En ce qui concerne la qualité de l'estimation de la densité f_1, celle-ci semble visuellement comparable lorsqu'on se réfère aux figures 4.11 et 4.13. Par contre, l'erreur quadratique commise lors de l'estimation de $\mathbb{E}[\mathbf{X}]$ est d'environ 12,87, soit environ quatre fois moindre qu'avec l'algorithme classique. De plus, l'estimation de la probabilité associée au mode positif est d'environ 0,494, ce qui constitue une erreur absolue environ six fois moindre qu'avec l'algorithme classique. Ainsi, même pour cette densité cible

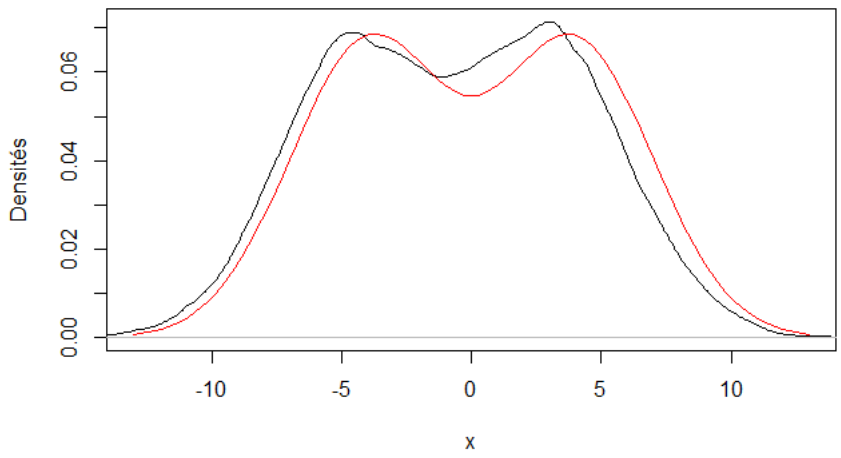

FIGURE 4.11. Estimation de la densité cible f_1 (algorithme RWM).

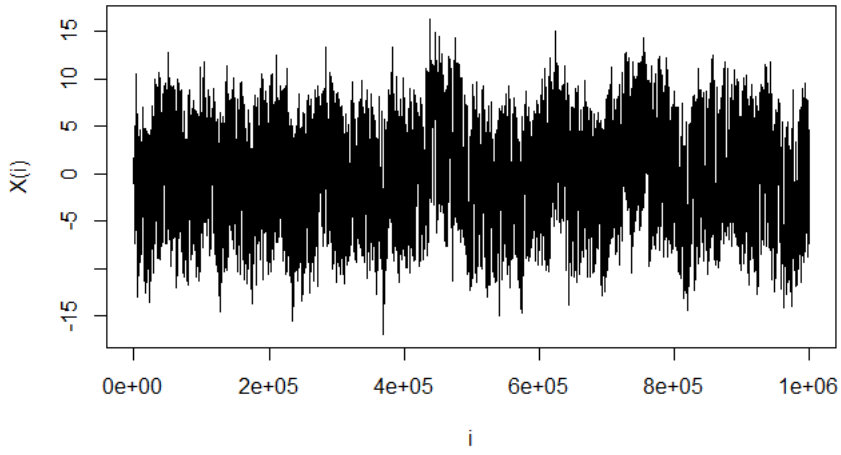

FIGURE 4.12. Trace de la chaîne engendrée par la première composante (nouvel algorithme).

presque unimodale, où l'algorithme RWM est à première vue efficace, le nouvel algorithme estime mieux la distribution.

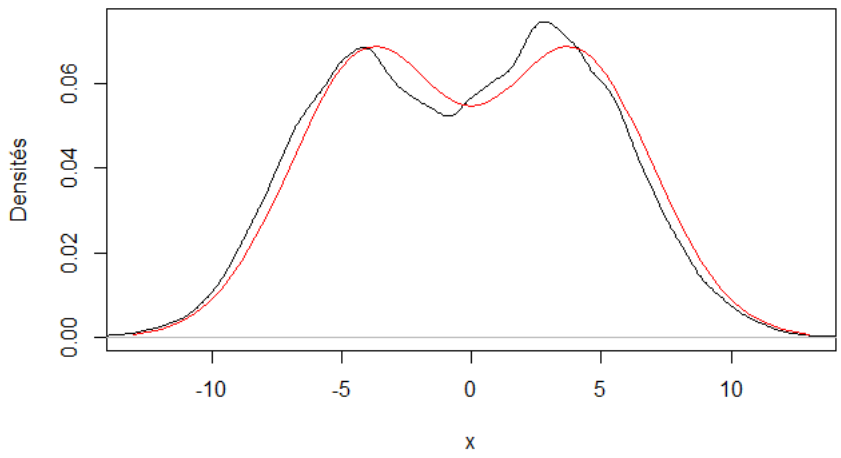

FIGURE 4.13. Estimation de la densité cible f_1 (nouvel algorithme).

4.4. EFFET DE LA DISTANCE ENTRE LES MODES

L'objet de cette section est une courte étude de l'influence de la distance entre les deux modes de la distribution cible sur la valeur optimale du paramètre d'échelle de la distribution instrumentale, c, ainsi que sur la probabilité de proposer un grand pas, p. Au chapitre 2, nous avons établi qu'à partir d'un point x appartenant à l'un des modes de la première composante de la densité cible, f_1, la valeur de c qui maximise la probabilité de proposer un changement de mode est de la forme $|b_2 - x|$, où b_2 représente la borne du mode visé la plus éloignée de x. Ceci peut être vu comme une fonction affine (dont la pente est 1) de la distance entre x et le mode visé. On aimerait savoir si la valeur de c qui maximise les changements de mode est également une fonction affine de la distance entre les deux modes.

Afin de tester ces hypothèses, considérons le modèle de la section 4.1, mais laissons varier la distance entre les deux modes. Le modèle est donc donné par

$$\mathbf{X} \sim \begin{cases} N_{100}(-\mu; 9I_{100}), \text{ avec probabilité } \frac{1}{2} \\ N_{100}(\mu; 9I_{100}), \text{ avec probabilité } \frac{1}{2} \end{cases}, \tag{4.4.1}$$

où $\mu = (\mu_1, 0, ..., 0)^\top \in \mathbb{R}^{100}$, $\mu_1 > 0$, et I_{100} représente la matrice identité de format 100×100. On souhaite utiliser le nouvel algorithme afin d'estimer la distribution de $\mathbf{X}$, à l'aide de la distribution instrumentale g_1 qui correspond à la loi uniforme.

Pour chaque valeur de $\mu_1 \in \{1, ..., 100\}$, on utilise la procédure décrite précédemment pour estimer la meilleure valeur de c, en terme de la probabilité de changer de mode, qu'on appelle c_{opt}. On vérifie ensuite si le coefficient de corrélation linéaire entre μ_1 et c_{opt} est grand puis, s'il l'est, on effectue une régression linéaire simple entre ces deux variables. Si notre hypothèse est vérifiée, le coefficient de variation devrait être près de 2, puisque la distance entre les deux modes est de l'ordre de $2\mu_1$.

On obtient la figure 4.14. Le coefficient de corrélation de Pearson entre μ_1 et c_{opt} est

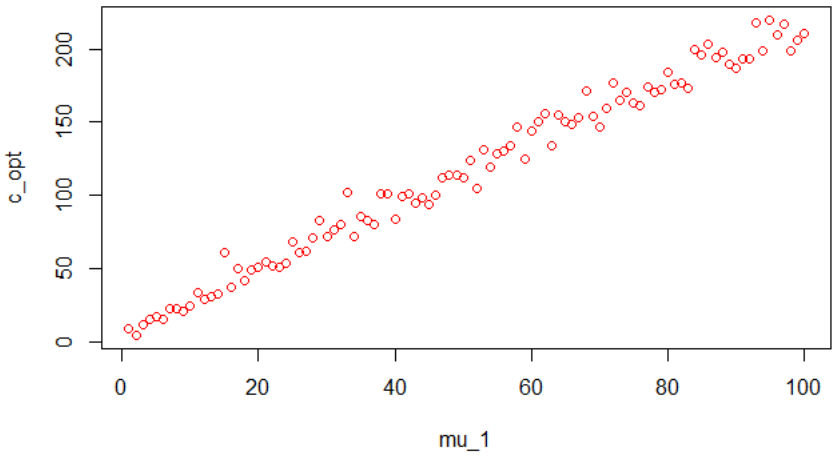

FIGURE 4.14. Valeur de c_{opt} en fonction de μ_1.

$r \approx 0{,}94$. De plus, la pente estimée par la régression linéaire est d'environ 2,3. La petite différence entre ce nombre et la valeur espérée (qui est 2) peut être expliquée de la façon suivante. Le résultat théorique qui justifie la valeur 2 est basé sur la probabilité de proposer un changement de mode à partir d'un certain point fixé. Or, supposons sans perte de généralité que l'on se trouve dans le mode négatif (c'est-à-dire que la première composante de notre algorithme se trouve autour de la valeur $-\mu_1$). Si on se trouve dans la partie la plus éloignée de l'autre mode, c'est-à-dire $X_1 < -\mu_1$, proposer un pas selon la loi uniforme avec $c = 2\mu_1$ n'est pas du tout optimal, puisque cette distribution instrumentale ne couvre alors qu'une petite partie du mode visé (on ne peut pas proposer un pas plus loin que $\mu_1 + (X_1 + \mu_1) < \mu_1$). Il est de loin préférable de choisir une valeur de c plus grande que $2\mu_1$. À l'inverse, il se peut que l'on se trouve dans la partie la plus rapprochée du mode visé ($X_1 > -\mu_1$) et que X_1 soit suffisamment grand pour que l'intervalle $[X_1 \pm 2\mu_1]$ couvre complètement le mode visé. Dans cette situation, une augmentation de la valeur de c résulte clairement en une diminution de la probabilité de proposer un changement de mode, mais cette diminution sera négligeable

par rapport à l'augmentation de la probabilité lorsqu'on se trouve plus loin du mode visé. C'est pourquoi, en général, il est préférable d'utiliser une valeur de c supérieure à $2\mu_1$.

De même, pour chaque valeur de μ_1, on estime la probabilité de changer de mode (conditionnellement au fait qu'on utilise la densité g_1 pour générer le candidat) obtenue avec la valeur c_{opt}, qu'on appelle $\hat{q}$. Rappelons que, dans les exemples précédents, on utilisait $p = \frac{1\,000}{n\hat{q}} = \frac{0{,}001}{\hat{q}}$. À l'aide de cette règle, on calcule p pour chaque valeur de μ_1. On obtient la figure 4.15. On remarque que l'augmentation semble presque linéaire : graphiquement,

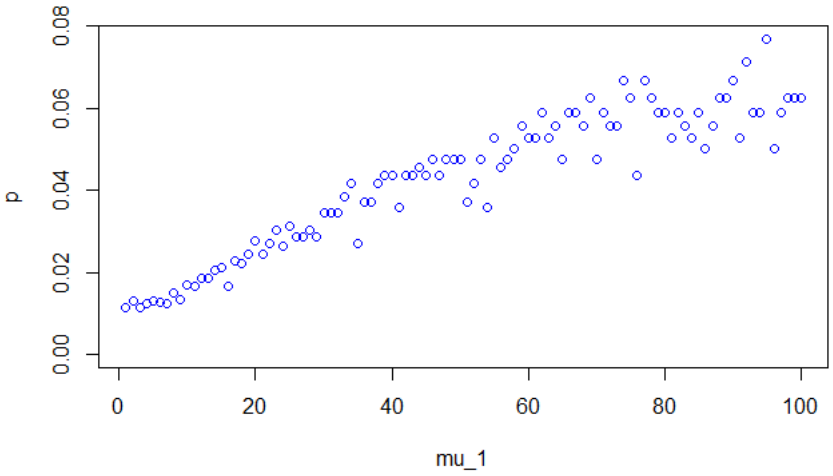

FIGURE 4.15. Valeur de p en fonction de μ_1.

lorsque $\mu_1 \to 0$, p semble tendre vers 0,01. Cela corrobore les résultats théoriques du chapitre 2 ; nous avions alors établi que la valeur optimale de q était proportionnelle à l'inverse de la distance entre les modes, ce qui signifie que p, qui est inversement proportionnel à q, devrait donc être en relation linéaire avec la distance entre les modes. De plus, cela suggère que lorsque $\mu_1 \to \infty$, la valeur choisie de p tend vers 1, ce qui confirme l'hypothèse affirmée au chapitre 2. Il devient de plus en plus difficile de passer d'un mode à l'autre lorsque $\mu_1 \to \infty$ et donc, pour contrer ceci, la valeur optimale de p devrait approcher 1 afin d'obtenir suffisamment de changements de mode.

CONCLUSION

Dans ce mémoire, nous avons non seulement présenté un nouvel algorithme, mais également développé une méthode pratique pour implanter cet algorithme afin de simuler d'une distribution multidimensionnelle dont l'une des composantes est bimodale. Ce nouvel algorithme est une combinaison de deux algorithmes de type RWM différents ; bien que l'idée soit basée sur des algorithmes existants, la paramétrisation que nous lui avons donnée dans ce mémoire (en laissant la pondération p donnée à l'un des algorithmes RWM dépendre de la dimension d, par exemple) et les résultats de convergence faible ainsi obtenus sont inédits.

L'objectif visé lorsqu'on étudie la convergence d'une suite d'algorithmes MCMC est bien sûr d'obtenir un processus limite qui est facile à optimiser. Comme nous avons pu le constater, il y a plusieurs paramètres qui entrent en jeu dans le contexte considéré. Dans cette optique, nous espérons poursuivre nos travaux de recherche dans cette voie et utiliser les résultats de convergence faible obtenus afin d'automatiser l'implémentation de cet algorithme. Bien que l'algorithme de départ requiert de l'information au sujet de la distance entre les modes et de leur dispersion afin de bien performer, il n'est pas impossible que l'on puisse utiliser les résultats asymptotiques associés afin de proposer une approche adaptative qui permettrait de se défaire des contraintes associées à l'implémentation de la méthode actuelle.

Néanmoins, la stratégie empirique que nous avons développée tente de rendre l'implantation le plus simple possible. Les ressources requises sont peu contraignantes ; il suffit de pouvoir générer des variables aléatoires de loi normale et de loi uniforme pour mettre en place l'algorithme, ainsi que de pouvoir générer un algorithme RWM classique pour choisir numériquement la valeur des paramètres à utiliser. Le choix du paramètre l se fait à l'aide de la règle habituelle du 0,234, et le choix des paramètres c et p à l'aide d'une simple maximisation basée sur des simulations.

La qualité des résultats qu'elle offre, bien que difficile à quantifier puisqu'il n'y a pas d'unique mesure d'efficacité, est visiblement supérieure à celle des résultats offerts par l'algorithme RWM pour les distributions cibles considérées. Ce dernier, en effet, produit une chaîne qui est tout à fait incapable d'explorer complètement le support d'une distribution bimodale dont les modes sont suffisamment éloignés et étroits, à moins de choisir un paramètre d'échelle énorme, ce qui a un effet désastreux sur le taux d'acceptation. Avec le nouvel algorithme, la même contrainte est rencontrée : si les modes sont très éloignés et étroits, il faut utiliser une grande valeur pour le paramètre d'échelle c, mais puisqu'une proportion $1-p$ des pas proposés sont échelonnés de façon optimale (au sens de la mesure de vitesse du processus limite continu), le taux d'acceptation moyen n'est jamais inférieur à $0{,}234 \cdot (1-p)$. Bien sûr, il faut noter que le paramètre p augmente lorsque les modes s'éloignent l'un de l'autre, ce qui a pour effet de diminuer cette borne inférieure pour le taux d'acceptation. Or, cette faible diminution n'a rien à voir avec la diminution du taux d'acceptation de l'algorithme RWM si on l'échelonnait de façon à obtenir des changements de mode. Nous n'avons donc pas à faire de compromis entre une bonne exploration locale des modes de la distribution cible et une bonne exploration globale, qui correspond à bien estimer le poids de chacun des modes (ce qui nous donne une vision d'ensemble sur la distribution).

Il serait intéressant, dans la recherche future, de comparer la performance de cet algorithme à celle d'autres méthodes spécifiquement conçues pour échantillonner de densités bimodales, en tenant non seulement compte de différentes mesures d'efficacité, mais également de l'effort computationnel requis. Il pourrait également être intéressant de choisir un critère «défendable», comme par exemple le déplacement quadratique moyen des processus limites, et de trouver une stratégie plus objective d'optimisation du nouvel algorithme. Par contre, il est vraisemblable que, si une telle stratégie puisse être trouvée, elle dépende fortement de la forme de la distribution cible (en particulier de sa composante bimodale), ce qui n'est heureusement pas le cas avec la méthode proposée dans ce mémoire. Il serait également pertinent de trouver de nouvelles façons de choisir la distribution instrumentale associée aux grands pas ; la loi uniforme est le meilleur choix lorsqu'on impose une certaine condition de monotonicité, mais dans certains contextes où la distance entre les modes est connue avec certitude, cette condition devient désuète. Finalement, comme il a été mentionné, il serait important d'étudier plus rigoureusement l'effet de la pondération relative des modes sur l'efficacité de l'estimation de cette pondération à l'aide de notre nouvel algorithme. Ces points seront considérés dans un article qui est actuellement en préparation ([4]).

Bibliographie

[1] Y. Bai, R.V. Craiu et F. Di Narzo : Divide and conquer : a mixture-based approach to regional adaptation for MCMC. *Journal of Computational and Graphical Statistics*, 20:63–79, 2011.

[2] M. Bédard : *On the Robustness of Optimal Scaling for Random Walk Metropolis Algorithms*. Thèse de doctorat, University of Toronto, 2006.

[3] M. Bédard : Weak convergence of Metropolis algorithms for non-iid target distributions. *The Annals of Applied Probability*, 17:1222–1244, 2007.

[4] M. Bédard et M. Lalancette : On a Metropolis algorithm involving local and global strategies for sampling from bimodal densities. *En préparation*, 2017.

[5] P. Billingsley : *Convergence of Probability Measures*. Wiley, 1968.

[6] M.K. Cowles et B.P. Carlin : Markov chain Monte Carlo convergence diagnostics : A comparative review. *Journal of the American Statistical Association*, 91:883–904, 1996.

[7] R.V. Craiu, J. Rosenthal et C. Yang : Learn from thy neighbor : Parallel-chain and regional adaptative MCMC. *Journal of the American Statistical Association*, 104:1454–1466, 2009.

[8] P. Diaconis et D. Freedman : On Markov chains with continuous state space (no. 501). Rapport technique, University of California, Berkeley, 1997.

[9] D. Elal-Olivero, H.W. Gómez et F.A. Quintana : Bayesian modeling using a class of bimodal skew-elliptical distributions. *Journal of Statistical Planning and Inference*, 139:1484–1492, 2009.

[10] S.N. Ethier et T.G. Kurtz : *Markov Processes : Characterization and Convergence*. Wiley, 1986.

[11] W.K. Hastings : Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[12] S.C. Kou, Q. Zhou et W.H. Wong : Equi-energy sampler with applications in statistical inference and statistical mechanics. *The Annals of Statistics*, 34:1581–1619, 2006.

[13] E. Marinari et G. Parisi : Simulated tempering : A new Monte Carlo scheme. *Europhysics Letters*, 19:451–462, 1992.

[14] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller et E. Teller : Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953.

[15] S.P. Meyn et R.L. Tweedie : *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[16] E. Nummelin : *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, 1984.

[17] C. Pasarica et A. Gelman : Adaptively scaling the Metropolis algorithm using squared jumped distance. *Statistica Sinica*, 20:343–364, 2010.

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

[19] S. Resnick : *A Probability Path*. Birkhäuser, 1998.

[20] G.O. Roberts, A. Gelman et W.R. Gilks : Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

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

[22] G.O. Roberts et J.S. Rosenthal : General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

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

Annexe A

LEMMES UTILISÉS DANS LE CHAPITRE 3

Nous présentons dans cette annexe six lemmes qui sont utilisés dans les démonstrations du chapitre 3. Ils apparaissent ici dans le même ordre que dans ce chapitre. Les preuves de ces lemmes sont complètes, sauf celle du lemme A.0.1 qui repose sur un résultat tiré de [3].

Lemme A.0.1. *Soit $f \in \mathcal{C}^2$ une densité, et soit $(\log f)'$ continue au sens de Lipschitz. Alors, si X est une variable aléatoire de densité f,*

$$\mathbb{E}\left[\frac{f''(X)}{f(X)}\right] = 0.$$

Démonstration. On a

$$\mathbb{E}\left[\frac{f''(X)}{f(X)}\right] = \int_{-\infty}^{\infty} f''(x)\, dx = \lim_{x\to\infty} f'(x) - \lim_{x\to-\infty} f'(x).$$

Selon le lemme 12 de [3], puisque $f \in \mathcal{C}^2$ et puisque $(\log f)'$ est continue au sens de Lipschitz, on a $\lim_{x\to\infty} f'(x) = \lim_{x\to-\infty} f'(x) = 0$.

□

Lemme A.0.2. *Soit $Z \sim N(0;\sigma^2)$ et $k \in \mathbb{N}$. On a*

$$\mathbb{E}\left[|Z|^k\right] = \frac{2^{\frac{k}{2}}}{\sqrt{\pi}}\Gamma\left(\frac{k+1}{2}\right)\sigma^k.$$

Démonstration.

$$\mathbb{E}\left[|Z|^k\right] = \int_{\mathbb{R}} |z|^k \phi(z)\, dz$$

$$= 2\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_0^\infty z^k \exp\left\{-\frac{z^2}{2\sigma^2}\right\}\, dz.$$

À l'aide du changement de variable $u = z^2$, on obtient

$$\begin{aligned}
\mathbb{E}\left[|Z|^k\right] &= 2\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_0^\infty u^{\frac{k}{2}} \exp\left\{-\frac{u}{2\sigma^2}\right\} \frac{1}{2} u^{-\frac{1}{2}}\, du \\
&= \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_0^\infty u^{\frac{k-1}{2}} \exp\left\{-\frac{u}{2\sigma^2}\right\}\, du \\
&= \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{(2\sigma^2)^{-\frac{k+1}{2}}} \\
&= \frac{2^{\frac{k}{2}}}{\sqrt{\pi}} \Gamma\left(\frac{k+1}{2}\right) \sigma^k.
\end{aligned}$$

□

Lemme A.0.3. *Soient $\mathbf{U}, \mathbf{V} : \Omega \mapsto \mathbb{R}^n$ deux vecteurs aléatoires absolument continus dont les fonctions de densité marginales et conjointe sont f_U, f_V et $f_{U,V}$, respectivement. Supposons également que la fonction de densité marginale $f_{V|U}$ est symétrique en ses arguments, c'est-à-dire que $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^n$,*

$$f_{V|U}(\mathbf{v} \mid \mathbf{u}) = f_{V|U}(\mathbf{u} \mid \mathbf{v}).$$

Alors on a

$$\mathbb{E}\left[\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} \mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} < 1\right\}\right] = \frac{1}{2}\mathbb{E}\left[1 \wedge \frac{f_U(\mathbf{V})}{f_U(\mathbf{U})}\right].$$

Démonstration.

$$\begin{aligned}
\mathbb{E}\left[\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} \mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} < 1\right\}\right] &= \iint_{\mathbb{R}^{2n}} \mathbb{I}\left\{\frac{f_U(\mathbf{v})}{f_U(\mathbf{u})} < 1\right\} \frac{f_U(\mathbf{v})}{f_U(\mathbf{u})} f_{U,V}(\mathbf{u}, \mathbf{v})\, \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\
&= \iint_{\mathbb{R}^{2n}} \mathbb{I}\left\{\frac{f_U(\mathbf{v})}{f_U(\mathbf{u})} < 1\right\} \frac{f_U(\mathbf{v})}{f_U(\mathbf{u})} f_U(\mathbf{u}) f_{V|U}(\mathbf{v} \mid \mathbf{u})\, \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\
&= \iint_{\mathbb{R}^{2n}} \mathbb{I}\left\{\frac{f_U(\mathbf{v})}{f_U(\mathbf{u})} < 1\right\} f_U(\mathbf{v}) f_{V|U}(\mathbf{v} \mid \mathbf{u})\, \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\
&= \iint_{\mathbb{R}^{2n}} \mathbb{I}\left\{\frac{f_U(\mathbf{v})}{f_U(\mathbf{u})} < 1\right\} f_U(\mathbf{v}) f_{V|U}(\mathbf{u} \mid \mathbf{v})\, \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v},
\end{aligned}$$

par symétrie de $f_{V|U}$. En échangeant les intégrandes $\mathbf{u}$ et $\mathbf{v}$, on trouve

$$\mathbb{E}\left[\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} \mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} < 1\right\}\right] = \iint_{\mathbb{R}^{2n}} \mathbb{I}\left\{\frac{f_U(\mathbf{u})}{f_U(\mathbf{v})} < 1\right\} f_U(\mathbf{u}) f_{V|U}(\mathbf{v} \mid \mathbf{u})\, \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u},$$

$$= \iint_{\mathbb{R}^{2n}} \mathbb{I}\left\{\frac{f_U(\mathbf{v})}{f_U(\mathbf{u})} > 1\right\} f_{U,V}(\mathbf{u},\mathbf{v})\ \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}$$
$$= \mathbb{E}\left[\mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} > 1\right\}\right]$$

et donc

$$\begin{aligned}
\mathbb{E}\left[1 \wedge \frac{f_U(\mathbf{V})}{f_U(\mathbf{U})}\right] &= \mathbb{E}\left[\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})}\mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} < 1\right\}\right] + \mathbb{E}\left[\mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} \geq 1\right\}\right] \\
&= \mathbb{E}\left[\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})}\mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} < 1\right\}\right] + \mathbb{E}\left[\mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} > 1\right\}\right] \\
&= 2\mathbb{E}\left[\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})}\mathbb{I}\left\{\frac{f_U(\mathbf{V})}{f_U(\mathbf{U})} < 1\right\}\right].
\end{aligned}$$

□

Lemme A.0.4. *Soit $f(x) = 1 \wedge e^x$. La fonction $f : \mathbb{R} \mapsto \mathbb{R}$ ainsi définie est continue au sens de Lipschitz d'ordre 1.*

Démonstration. Sans perdre de généralité, on va supposer $x \leq y$. Il y a alors trois cas possibles.

Premier cas : x,$y \leq 0$

$$\begin{aligned}
|f(x) - f(y)| &= |e^x - e^y| \\
&\leq |x - y| \sup_{x \leq t \leq y} \left|\frac{d}{dt}e^t\right| \\
&\leq |x - y| \sup_{x \leq t \leq y} \left|e^t\right| \\
&\leq |x - y|
\end{aligned}$$

par le théorème des valeurs moyennes, puisque la fonction exponentielle est partout différentiable.

Second cas : x,$y \geq 0$

$$|f(x) - f(y)| = |1 - 1| = 0.$$

Troisième cas : $x \leq 0$, $y \geq 0$

$$\begin{aligned}
|f(x) - f(y)| &= |e^x - 1| \\
&= \left|e^x - e^0\right| \\
&\leq |x| \sup_{x \leq t \leq 0} \left|\frac{d}{dt}e^t\right|,
\end{aligned}$$

par le théorème des valeurs moyennes. En dérivant la fonction exponentielle, on trouve

$$\begin{aligned} |f(x) - f(y)| &\leq |x| \sup_{x \leq t \leq 0} \left| e^t \right| \\ &\leq |x| \\ &\leq |x - y|. \end{aligned}$$

Dans tous les cas, on a donc $|f(x) - f(y)| \leq |x - y|$.

□

Lemme A.0.5. *Soient $\{X_n\}_{n\geq 1}$ et $\{Y_n\}_{n\geq 1}$ deux suites de variables aléatoires telles que $\mathbb{E}[|X_n - Y_n|] \longrightarrow 0$ lorsque $n \to \infty$ et soit g une fonction continue au sens de Lipschitz d'ordre K. Alors*

$$\mathbb{E}[|g(X_n) - g(Y_n)|] \longrightarrow 0$$

lorsque $n \to \infty$.

Démonstration.

$$\mathbb{E}[|g(X_n) - g(Y_n)|] \leq K\mathbb{E}[|X_n - Y_n|] \longrightarrow 0.$$

□

Lemme A.0.6. *Soit $Z \sim N(\mu; \sigma^2)$. On a*

$$\mathbb{E}\left[1 \wedge e^Z\right] = \Phi\left(\frac{\mu}{\sigma}\right) + \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \Phi\left(-\sigma - \frac{\mu}{\sigma}\right),$$

où Φ représente la fonction de répartition de la loi normale standard.

Démonstration. D'abord,

$$\mathbb{E}\left[1 \wedge e^Z\right] = \mathbb{P}(Z > 0) + \mathbb{E}\left[e^Z \mathbb{I}\{Z \leq 0\}\right].$$

Il suffit d'évaluer ces deux termes. Premièrement,

$$\mathbb{P}(Z > 0) = \int_0^\infty \frac{1}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right)\, dz = \int_{-\frac{\mu}{\sigma}}^\infty \phi(z)\, dz = 1 - \Phi\left(-\frac{\mu}{\sigma}\right) = \Phi\left(\frac{\mu}{\sigma}\right).$$

De plus,

$$\mathbb{E}\left[e^Z \mathbb{I}\{Z \leq 0\}\right] = \int_{-\infty}^0 e^z \frac{1}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right)\, dz$$

$$= \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left\{z - \frac{1}{2\sigma^2}(z-\mu)^2\right\}\, dz$$
$$= \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left\{-\frac{1}{2\sigma^2}\left(z^2 - 2\left(\mu+\sigma^2\right)z + \mu^2\right)\right\}\, dz.$$

En complétant le carré à l'intérieur de l'exponentielle, cette quantité devient

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left\{-\frac{1}{2\sigma^2}\left(z^2 - 2\left(\mu+\sigma^2\right)z + \left(\mu+\sigma^2\right)^2\right) + \mu + \frac{\sigma^2}{2}\right\}\, dz$$
$$= \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \int_{-\infty}^{0} \exp\left\{-\frac{1}{2\sigma^2}\left(z - \left(\mu+\sigma^2\right)\right)^2\right\}\, dz$$
$$= \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \int_{-\infty}^{-\frac{\mu+\sigma^2}{\sigma}} \sigma \exp\left\{-\frac{z^2}{2}\right\}\, dz$$
$$= \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \int_{-\infty}^{-\sigma-\frac{\mu}{\sigma}} \phi(z)\, dz$$
$$= \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \Phi\left(-\sigma - \frac{\mu}{\sigma}\right),$$

ce qui complète la démonstration.

□

Annexe B

CODES R

B.1. Code pour implémenter le nouvel algorithme

Dans cette section, nous présentons le code R utilisé dans l'exemple de la section 4.1. Notons que le code utilisé pour les sections 4.2 et 4.3 est très similaire.

```
################## NOUVEL ALGORITHME: EXEMPLE MULTINORMAL ##############



library(mvtnorm)

d = 100
mu = 15
sigma2 = 9

pi1 <- function(x) 0.5 * (dnorm(x, mean = -mu, sd = sqrt(sigma2)) +
 dnorm(x, mean = mu, sd = sqrt(sigma2)))
pi <- function(x) pi1(x[1]) * dmvnorm(x[2:d], mean = rep(0, d-1),
 sigma = sigma2*diag(d-1))
alpha <- function(x, y) pi(y)/pi(x)
```

```
############## Choix de l ############



n_test = 1000
J = 100
taux_accept = rep(0, J)

for (j in 1:J) {

l = j/10
X = matrix(rep(0, d*(n_test + 1)), nrow = d)
nbaccept = 0

for (i in 1:n_test) {

U = runif(1)

Y = rnorm(d, mean = X[,i], sd = l/sqrt(d))

if (U <= alpha(X[,i], Y)) {
X[,i+1] = Y
nbaccept = nbaccept + 1
} else {
X[,i+1] = X[,i]
}

}

taux_accept[j] = nbaccept/n_test
}

plot((1:J)/10, taux_accept, xlab = "Valeur de l",
 ylab = "Taux d'acceptation")

b = lsfit((1:J)/10, taux_accept, intercept = TRUE)$coef
```

```
l = (0.234 - b[1])/b[2]



############## Choix de p ##############



n_test = 100000

A <- function(x, y) log(pi1(y)/pi1(x))

probaccept <- function(x, y) pnorm(A(x, y)/2.38 - 1.19) +
 exp(A(x, y)) * pnorm(-A(x, y)/2.38 - 1.19)

cmin = 30
cmax = 50
sigmin = 20
sigmax = 50
prob_changement_unif = rep(0, cmax - cmin + 1)
prob_changement_norm = rep(0, cmax - cmin + 1)



for(c in cmin:cmax) {

X = rep(0, n_test + 1)
nbchangements = 0

for (i in 1:n_test) {

U = runif(1)

Y = runif(1, X[i] - c, X[i] + c)

if (U <= probaccept(X[i], Y)) {
X[i+1] = Y
} else {
```

```
X[i+1] = X[i]
}

nbchangements = nbchangements + max(0, -sign(X[i]*X[i+1]))
}

prob_changement_unif[c + 1 - cmin] = nbchangements/n_test
}

plot(cmin:cmax, prob_changement_unif, xlab = "Valeur de c",
 ylab = "Probabilité de changer de mode")



for(sig in sigmin:sigmax) {

X = rep(0, n_test + 1)
nbchangements = 0

for (i in 1:n_test) {

U = runif(1)

Y = rnorm(1, mean = X[i], sd = sig)

if (U <= probaccept(X[i], Y)) {
X[i+1] = Y
} else {
X[i+1] = X[i]
}

nbchangements = nbchangements + max(0, -sign(X[i]*X[i+1]))
}

prob_changement_norm[sig + 1 - sigmin] = nbchangements/n_test
}

plot(sigmin:sigmax, prob_changement_norm, xlab = "Valeur de sigma",
```

```
 ylab = "Probabilité de changer de mode")

c = 38
q = 0.037



############## Vraie simulation #############



n = 1000000
p = 1000/(n*q)



densite <- function(X) {

plot(density(X[1, 2:(n+1)]), xlim = c(-1, 1)*(mu + 3*sqrt(sigma2)),
ylim = c(0, max(pi1(mu), max(density(X[1, 2:(n+1)])$y))), xlab = "x",
ylab = "Densités", main = "")

plot(pi1, col = 'red', xlim = c(-1, 1)*(mu + 3*sqrt(sigma2)), add = TRUE)
}

trace <- function(X) {

plot(1:(n+1), X[1,], xlab = "i", ylab = "X(i)", type = "l")
}



temps = proc.time()

X = matrix(rep(0, d*(n+1)), nrow = d)

for (i in 1:n) {
```

```
U = runif(2)

if( U[1] <= p) {
Y1 = runif(1, X[1,i] - c, X[1,i] + c)
} else {
Y1 = rnorm(1, mean = X[1,i], sd = l/sqrt(d))
}

Y = c(Y1, rnorm(d-1, mean = X[2:d,i], sd = l/sqrt(d)))

if (U[2] <= alpha(X[,i], Y)) {
X[,i+1] = Y
} else {
X[,i+1] = X[,i]
}

print(i)
}

proc.time() - temps
```

B.2. Code pour l'étude de la distance entre les modes

Dans cette section, nous présentons le code R utilisé dans la section 4.4.

```
######### ESTIMATION DE LA RELATION ENTRE mu ET LE beta OPTIMAL ######


library(mvtnorm)
```

```
d = 100
sigma2 = 9

n_test = 1000
mumin = 1
mumax = 100

copt = rep(0, mumax - mumin + 1)
qhat = copt



temps = proc.time()

for (mu in mumin:mumax) {

pi1 <- function(x) 0.5 * (dnorm(x, mean = -mu, sd = sqrt(sigma2))
+ dnorm(x, mean = mu, sd = sqrt(sigma2)))

A <- function(x, y) log(pi1(y)/pi1(x))

probaccept <- function(x, y) pnorm(A(x, y)/2.38 - 1.19)
+ exp(A(x, y)) * pnorm(-A(x, y)/2.38 - 1.19)

cmin = max(1, 2*mu - 12*sqrt(sigma2))
cmax = 2*mu + 100*sqrt(sigma2)
prob_changement = rep(0, cmax - cmin + 1)



for(c in cmin:cmax) {

X = rep(0, n_test + 1)
nbchangements = 0

for (i in 1:n_test) {

U = runif(1)
```

```
Y = runif(1, X[i] - c, X[i] + c)

if (U <= probaccept(X[i], Y)) {
X[i+1] = Y
} else {
X[i+1] = X[i]
}

nbchangements = nbchangements + max(0, -sign(X[i]*X[i+1]))
}

prob_changement[c + 1 - cmin] = nbchangements/n_test
}

copt[mu - mumin + 1] = which.max(prob_changement) + cmin - 1
qhat[mu - mumin + 1] = prob_changement[which.max(prob_changement)]

print(mu)
}

proc.time() - temps



plot(mumin:mumax, copt, col = "red", xlab = "mu_1", ylab = "c_opt")
cor(mumin:mumax, copt)
lsfit(mumin:mumax, copt, intercept = TRUE)$coef

plot(mumin:mumax, qhat, col = "blue", xlab = "mu_1", ylab = "q")
plot(mumin:mumax, 0.001*qhat^(-1), ylim = c(0, max(0.001*qhat^(-1))),
 col = "blue", xlab = "mu_1", ylab = "p")
```