$N°$ d'ordre :

République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et la Recherche Scientifique

Université Djillali Liabès de Sidi Bel Abbès

Faculté des Sciences Exactes

Département de Probabilités Statistique

Mémoire Master

présenté par

BELAROUCI IMED

Domaine : MATHÉMATIQUES - INFORMATIQUE

Filière : MATHÉMATIQUE.

Parcours : Probabilité et applications.

Intitulée

MCMC(Monte Carlo par chaines de Markov) :

Application en Machine learning

Soutenue le 4 Juin 2022 devant le jury composé de :

Pr. GHERIBALLAH ABDELKADER	Université Djillali Liabès	Président
Dr. HAMMAD MALIKA	Université Djillali Liabès	Examinateur
Dr. RIGHI ALI	Université Djillali Liabès	Encadreur

Résumé

Dans ce mémoire, nous étudions les méthodes MCMC et leur application en Apprentissage automatique . Plus précisement, nous nous intéressons à la théorie des chaines de Markov pour la justfication théorique des algorithmes MCMC et nous utiliserons ces algorithmes dans certains modéles d'apprentissage automatique .
Mots- clés : MCMC ; Chaines de Markov ;Monte Carlo ;Apprentissage automatique.

ملخص

في هذه الرسالة ، ندرس طرق بثبث وتطبيقاتها في التعلم الآلي. وبشكل أكثر تحديدًا ، نحن مهتمون بنظرية سلسلة ماركوف للتبرير النظري لخوارزميات بثبث وسنستخدم هذه الخوارزميات في بعض نماذج التعلم الآلي. تشتبف الكلمات الرئيسية : بثبث سلاسل ماركوف مونتي كارلو التعلم الآلي.

Remerciement

Je tiens à remercier Monsieur RIGHI Ali pour son encadrement , ainsi que la qualité de son enseignement durant le cursus qui est exemplaire , et qui permet un transfert de connaissance optimale et résulte en un processus d'appentissage trés agréable .

Je tiens aussi à remercier le département de Probabilité et Statistique de tout simplement existé et pouvoir offrir des formations en mathématiques appliquées qui sont existentielles pour le développement de l'économie et surtout de l'industrie ,
car aussi majestueux que ça puisse être une bouteille de Klein à 4 dimensions sans intérieur et sans extérieur ne peut combattre la soif dans le sens propre propre et figuré

TABLE DES MATIÈRES

TABLE DES FIGURES

INTRODUCTION

- Cet exposé concerne les méthodes monte carlo par chaines de markov (mcmc) et leur application en apprentissage automatique .
- La version récente du modéle d'intélligence artificielle de l'entreprise californienne Open-AI , GPT-4 contient selon certaines sources plus de 100 mille-milliards de paramétres , et dans le futur il est probable que les modéles dépassent largement ce nombres en termes de paramétres .
- Un autre domaine qui contient des nombres faramineux de paramétres est la physique statitstique , et pour essayer de faire des calculs , des optimisations ou des simulations dans ce genre de modéles , les méthodes mcmc se sont révélés un choix assez pertinent .
- En effet ces méthodes ont été dévelopés à la base par les physiciens qui trvailler sur la bombe à hydrogéne , qui n'est ni plus ni moins qu'un probléme de physique statistique .
- Donc l'utilisation des méthodes mcmc à l'avenir en apprentissage automatique est quasi-certaine .

- On ne peut pas parler de méthodes mcmc sans parler de statistique bayésienne , car c'est grace aux méthodes mcmc que le bayésisanisme à pu passer d'une philosophie à une méthode pratiques .
- Sans les méthodes mcmc et le développement des ordinateurs , le calcul des lois a posteriori au centre du paradigme bayésien étaient trés diffcile .
- On retrouve aussi le paradigme bayésien en apprentissage automatique , et dans ce cas l'usage aux méthodes mcmc est doublement justifiable .

- Toutefois dans cet exposé il ne sera pas question de GPT-4 ou de modéles d'apprentissage automatique sophistiqué , mais de quelques exemples simples , dans lesquels l'utilisation des méthodes mcmc n'est pas vraiment indispensable mais , le fait est que à un niveau plus élevé , s'en passer n'est pas vraiment une option .
- Nous illustrerons les méthodes mcmc dans des modéles utilisés dans l'apprentissage automatique à savoir : la régression polynomiale , la régression logistique , les modéles de markov cachés et les modéles à variables latentes .
- pour les exemples accés pratiques , ils ne seront pas réalisé pour l'apprentissage automatique , enfin pas directement , mais le principe reste le méme , en effet la régression peut étre utilisé en économétrie , en biologie et en apprentissage automatique mais la méthode est la méme seul le sens des données différent .
- Les exemples d'apprentissage présentés requirent la simulation de lois continues , ce qui fait appel aux méthodes mcmc aux cas général c'est à dire avec des chaines de Markov à

espace d'état continu .

- Le premier chapitre concerne justement la théorie des chaines de Markov pour le cas général .

- La théorie des chaines de Markov sera présentés d'une façon superficielle c'est à dire que la totalité des résultats seront donnnées sans preuves , et beaucoup de notions introduites n'oront pas d'impact dans l'application des méthodes mcmc mais sont utiles dans les constructions théoriques méme si nous n'allons pas l'illustrer , tout simplement parceque ce n'est pas l'objet de cet exposé et que vu la difficulté du sujet , les traités à leur juste valeur nésseciterais un exposé tout entier , mais malheuresement nous ne pouvions pas non plus ne pas parler d'un minimum de justification théorique des méthodes mcmc , surtout qu'il s'agit d'un exposé de probabilité .

- Concernant les modéles d'apprentissage automatique , nous rentrerons pas non plus dans les détails par souci d'espace , par exemple les modéles de Markov cachés seront présentés en effleurons à peine le sujet au vu de leur théorie colossale et de leur porté pratique quasi-infini .

1 CHAINES DE MARKOV

1.1 Chaînes de Markov et apprentissage automatique :

- Les chaines de Markov ont une place importante en apprentissage automatique , on les retrouve entre autre en :
modéle de Markov caché , processus de décision Markovien et l'apprentissage par renforcement .
- Nous allons considérer l'application de chaine de Markov à espace d'état discret pour l'apprentissage automatique .

1.1.1 Application : Modélisation du langage :

- Une application importante des chaines de Markov est de créer des modèles de langage statistique, qui sont des distributions de probabilité sur des séquences de mots.
- Nous définissons l'espace d'état comme étant tous les mots en anglais (ou dans une autre langue).
- Les probabilités marginales $p(X_t = k)$ sont appelées unigramme statistiques.
- Si on utilise une chaine de Markov du premier ordre, alors $p(X_t = k|X_{t-1} = j)$ est appelé modèle bigramme .
- Si on utilise une chaine de Markov du second ordre, alors $p(X_t = k|X_{t-1} = j, X_{t-2} = i)$ est appelé un modèle de trigramme. Et ainsi de suite. En général, ceux-ci sont appelés modèles n-grammes.
- Par exemple, La figure ci-dessous montre des comptages à 1 gramme et à 2 grammes pour les lettres $\{a, ..., z, -\}$ (où - représente espace) estimée à partir de l'origine des espèces de Darwin , L'image 2D sur la droite est un diagramme de Hinton de la distribution conjointe. La taille des carrés blancs est proportionnelle à la valeur de l'entrée dans le vecteur/matrice correspondant. Basé sur (MacKay 2003, p22).
- Les modèles de langage peuvent être utilisés pour plusieurs choses, telles que les suivantes :
1) Achèvement de la phrase : Un modèle de langage peut prédire le mot suivant compte tenu du mot précédent dans une phrase.
Cela peut être utilisé pour réduire la quantité de frappe requise, ce qui est particulièrement important pour les utilisateurs handicapés , ou les utilisations d'appareils mobiles.
2) Compression des données : Tout modèle de densité peut être utilisé pour définir un schéma d'encodage, en attribuons des mots de code courts à des chaînes plus probables. Plus le modèle prédictif est précis, moins il faut de bits pour stocker les données.

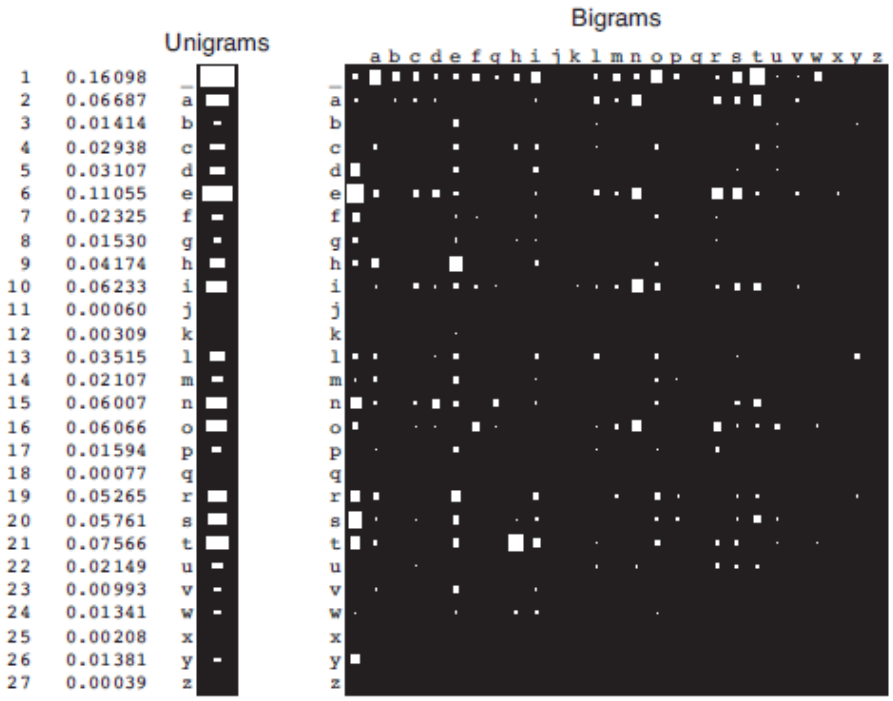

FIGURE 1.1 – résultat d'unigrammes et de bigrammes de Darwin sur l'origine des espèces

3) Classification textuelle : Tout modèle de densité peut être utilisé comme densité conditionnelle de classe et donc transformé en un classificateur (génératif).
Notez qu'en utilisant une densité conditionnelle de classe de 0 gramme (c'est-à-dire, seules statistiques unigrammes) équivaudrait à un classificateur bayésien naïf .
4) Rédaction automatique d'un essai : On peut échantillonner à partir de $p(x_{1:t})$ pour générer un texte artificiel.
C'est une façon d'évaluer la qualité du modèle.
- voici un exemple de texte généré à partir d'un modèle de 4 grammes, entraîné sur un corpus de 400 millions de mots , entrainé à l'aide du lissage d'interruption sur la diffusion sur le Corpus d'actualités , Les 4 premiers mots sont spécifiés à la main, le modèle génère le 5ème mot, puis les résultats sont réinjectés dans le modèle :

```
    SAYS IT'S NOT IN THE CARDS LEGENDARY RECONNAISSANCE BY ROLLIE
DEMOCRACIES UNSUSTAINABLE COULD STRIKE REDLINING VISITS TO PROFIT
BOOKING WAIT HERE AT MADISON SQUARE GARDEN COUNTY COURTHOUSE WHERE HE
HAD BEEN DONE IN THREE ALREADY IN ANY WAY IN WHICH A TEACHER
```

- ((Tomas et al. 2011) décrit un bien meilleur modèle de langage, basé sur des réseaux de

neurones récurrents, qui génère un texte beaucoup plus plausible sémantiquement) .

1.2 Chaînes de Markov

-la nature même de la simulation nous conduit à ne considérer que les processus stochastiques à temps discret, $(X_n)_{n\epsilon N}$.
-En effet, Hastings (1970) note que l'utilisation de générateurs pseudo-aléatoires et la représentation de nombres dans un ordinateur impliquent que les chaînes de Markov liées aux méthodes de Monte Carlo par chaînes de Markov sont, en fait, des chaînes de Markov à espace d'état fini.
-Cependant, nous considérons également des chaînes de Markov d'espace d'état arbitraires pour permettre des distributions de support continues et pour éviter de traiter le problème d'approximation de ces distributions avec des distributions de support discrèts, puisque une telle approximation dépend à la fois des spécificités matérielles et algorithmiques d'une technique donnée .

1.3 Essentiels pour MCMC

-Dans le cadre des algorithmes MCMC, les chaînes de Markov sont construites à partir d'un noyau de transition K , une densité de probabilité conditionnelle telle que

$$X_{n+1} \sim K(X_n, X_{n+1})$$

1.3.1 Marche aléatoire

-Une suite de variables aléatoires (X_n) est une marche aléatoire si elle satisfait :

$$X_{n+1} = X_n + \epsilon_n$$

où ϵ_n est généré indépendamment de X_n, X_{n-1}, ...
-Si la distribution de ϵ_n est symétrique en zéro, la séquence est appelée marche aléatoire symétrique .
-les marches aléatoires jouent un rôle clé dans de nombreux algorithmes MCMC, en particulier ceux basés sur l'algorithme de Metropolis-Hastings .
-Les chaines rencontrées en MCMC jouissent d'une très forte propriété de stabilité, à savoir une distribution de probabilité stationnaire existe par construction ; soit une distribution π telle que :

$$\text{si } X_n \sim \pi, \text{ alors } X_{n+1} \sim \pi$$

, si le noyau K permet des déplacements libres dans tout l'espace d'état.
-Cette liberté est appelée irréductibilité dans la théorie des chaînes de Markov et est formalisée en l'existence de $n \in N$ tel que :

$$P(X_n \in A|X_0) > 0 \text{ pour tout A tel que } \pi(A) > 0$$

-Cette propriété garantit également que la plupart des chaînes impliquées dans les algorithmes MCMC sont récurrentes (c'est-à-dire que le nombre moyen de visites à un ensemble arbitraire A est infini , ou même Harris récurrente (c'est-à-dire tel que la probabilité d'un nombre infini de retour à A est 1) .
-La récurrence de Harris garantit que la chaîne a le même comportement limite pour chaque valeur de départ au lieu de presque chaque valeur de départ. ((Par conséquent, c'est l'équivalent en chaîne de Markov de la notion de continuité pour les fonctions.)
-Puisque la plupart des algorithmes sont lancés à partir d'un point arbitraire x_0 , nous sommes en effet entrain de démarrer l'algorithme à partir d'un ensemble de mesure zéro (sous une mesure dominante continue).
Ainsi,s'assurer que la chaîne converge pour presque chaque point de départ ne suffit pas, et nous avons besoin de la récurrence de Harris pour garantir la convergence à partir de chaque point de départ.
-La distribution stationnaire est aussi une distribution limite dans le sens où la distribution limite de X_{n+1} est π sous la norme de variation totale ,pour toute valeur initiale de X_0.
-Dans une configuration de simulation, la conséquence la plus intéressante de cette propriété de convergence est que la moyenne $\frac{1}{N}\sum_{i=1}^{N} h(X_n)$ converge presque surement vers l'espérence $E_\pi(h(X))$.
-Lorsque la chaîne est réversible (c'est-à-dire lorsque le noyau de transition est symétrique), Le théorème central limite est également valable pour cette moyenne.

1.4 Notions de base

1.4.1 Noyau de transition

- Un noyau de transition est une fonction K définie sur $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ telle que :
i) $\forall x \in \mathcal{X}$, $K(x, .)$ est une mesure de probabilité .
ii) $\forall A \in \mathcal{B}(\mathcal{X})$, $K(., A)$ est une application mesurable .
-Lorsque $\mathcal{X}$ est discret, le noyau de transition est simplement une matrice (de transition) K avec des éléments :

$$P_{xy} = P(X_n = y|X_{n-1} = x), x, y \in \mathcal{X}$$

-Dans le cas continu, le noyau désigne aussi la densité conditionnelle $K(x, x^{'})$ de la transition $K(x, .)$; c'est-à-dire

$$P(X \in A|x) = \int_A K(x, x^{'})dx^{'}$$

-La chaîne (X_n) est généralement définie pour $n \in N$ plutôt que pour $n \in Z$. Par conséquent, la distribution de X_0 , l'état initial de la chaîne, joue un rôle important.
-Dans le cas discret, où le noyau K est une matrice de transition, étant donné une distribution initiale $\mu = (w_1, w_2, ...)$, la distribution de probabilité marginale de X_1 est obtenu à partir de la multiplication matricielle

$$\mu_1 = \mu K$$

et par multiplication consécutive

$$X_n \sim \mu_n = \mu K^n$$

-De même, dans le cas continu , si μ désigne la distribution initiale de la chaîne, à savoir si $X_0 \sim \mu$, alors nous notons P_μ la loi de probabilité de (X_n) sous la condtion $X_0 \sim \mu$.
-Lorsque X_0 est fixé, en particulier pour μ égal à la masse de Dirac δ_{x_0} , on utilise la notation alternative P_{x_0} .

1.4.2 Chaine de Markov

-Étant donné un noyau de transition K, une suite $X_0, X_1, ..., X_n$, ... de variables aléatoires est une chaîne de Markov, notée (X_n), si, pour tout t, la distribution conditionnelle de X_t sachant $x_{t-1}, x_{t-2}, ..., x_0$ est la même que la distribution de X_t étant donné x_{t-1} ; c'est à dire :

$$P(X_{k+1} \in A|x_0, x_1, ..., x_k) = P(X_{k+1} \in A|x_k) = \int_A K(x_k, dx)$$

-La chaîne est homogène dans le temps, ou simplement homogène, si la distribution de $(X_{t_1}, ..., X_{t_k})$ étant donné x_{t_0} est la même que la distribution de $(X_{t_1-t_0}, X_{t_2-t_0}, ..., X_{t_k-t_0})$ étant donné x_0 pour tout k et tout (k + 1)-uplet
$t_0 \leq t_1 \leq ... \leq t_k$.
-Ainsi, dans le cas d'une chaîne de Markov, si la distribution initiale ou l'état initiale est connu, la construction de la chaîne de Markov (X_n) est entièrement déterminée par sa transition, à savoir par la distribution de X_n conditionnellement à X_{n-1}.
-L'étude des chaînes de Markov est presque toujours restreinte au cas d'homogèneité dans le temps et nous omettons cette désignation dans ce qui suit.

-Cependant, il est important de noter qu'une implémentation incorrecte de la chaîne de Markov dans Les algorithmes de Monte Carlo peut facilement produire des chaînes de Markov non homogènes pour lesquels les propriétés de convergence standard ne s'appliquent pas.

Exemple : Models AR(1)

-Les modèles AR(1) illustrent des chaînes de Markov sur un espace d'état continu.
-Si

$$X_n = \theta X_{n-1} + \epsilon_n, \theta \in R$$

avec $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$, et si les ϵ_n sont indépendants , X_n est en effet indépendent de $X_{n-2}, X_{n-3}, ...$ conditionnellement à X_{n-1} .
-Les propriétés Markoviennes d'un processus AR(q) peuvent être dérivé en considérant le vecteur $(X_n, ..., X_{n-q+1})$·
-D'autre part, les modèles ARMA(p, q) ne rentrent pas dans le cadre Markovien.

-Dans le cas général, le fait que le noyau K détermine les propriétés de la chaîne (X_n) peut être déduit des relations :

$$P_x(X_1 \in A_1) = K(x, A_1)$$

$$P_x((X_1, X_2) \in A_1 \times A_2) = \int_{A_1} K(y, A_2) K(x, dy_1)$$

...

$$P_x((X_1, ..., X_n) \in A_1 \times ... \times ... \times A_n) = \int_{A_1} ... \int_{A_{n-1}} K(y_{n-1}, A_n) K(x, dy_1) ... K(y_{n-2}, dy_{n-1})$$

-si on note $K^1(x, A) = K(x, A)$, le noyau pour n transitions est donné par :

$$K^n(x, A) = \int_{\mathcal{X}} K^{n-1}(y, A) K(x, dy), (n > 1)$$

1.4.3 Équations de Chapman-Kolmogorov

-Pour tout $(m, n) \in N^2$, $x \in \mathcal{X}$, $A \in \mathcal{B}(\mathcal{X})$:

$$K^{m+n}(x, A) = \int_{\mathcal{X}} K^n(y, A) K^m(x, dy)$$

-Dans un sens très informel, les équations de Chapman-Kolmogorov indiquent que pour passer de x à A en $m + n$ pas, il faut passer par un y le nième pas.
-Dans le cas discret,l'équation est simplement interprété comme un produit matriciel.

-Dans le cas général, il faut considérer K comme opérateur sur l'espace des fonctions intégrables ; c'est-à-dire que nous définissons

$$Kh(x) = \int h(y)K(x, dy), h \in \mathcal{L}_1(\lambda)$$

λ étant la mesure dominante du modèle.
-K^n est alors la nième composition de P, à savoir $K^n = KoK^{n-1}$.

1.4.4 Résolvante

-Une résolvante associé au noyau P est un noyau de la forme :

$$K_\epsilon = (1-\epsilon)\sum_{i=0}^{\infty} \epsilon^i K^i(x, A), 0 < \epsilon < 1$$

et la chaine avec le noyau K_ϵ est une chaine K_ϵ.

-Étant donné une distribution initiale μ, on peut associer au noyau K_ϵ une chaîne (X_n^ϵ) qui correspond formellement à une sous-chaîne de la chaîne d'origine (X_n), où les indices de la sous-chaîne sont générés à partir d'une distribution géométrique avec le paramètre $1-\epsilon$.

1.4.5 Propriété faible de Markov

-Si $E_\mu(.)$ désigne l'espérance associée à la distribution P_μ, La propriété de Markov {faible} peut être écrite comme le résultat suivant, qui reformule les propriétés de mémoire limitée d'une chaîne de Markov :

Pour toute distribution initiale μ et pour tout (n+1) échantillon $(X_0, ..., X_n)$,

$$E_\mu(h(X_{n+1}, X_{n+2}, ...)|x_0, ..., x_n) = E_{x_n}(h(X_1, X_2, ...))$$

à condition que les espérences existent .

1.4.6 Temps d'arrêt

-Soit $A \in \mathcal{B}(\mathcal{X})$, le premier n pour lequel la chaine rentre dans l'ensemble A est notée par :

$$\tau_A = inf\{n \geq 1; X_n \in A\}$$

et est appelé temps d'arrêt en A avec, par convention, $\tau_A = +\infty$, si $X_n \notin A \ \forall n$.

-Plus généralement, une fonction $\zeta(x_1, x_2, ...)$ est appelée règle d'arrêt si l'ensemble $\{\zeta = n\}$ est mesurable pour la tribu induite par $(X_0, ..., X_n)$· Associé à l'ensemble A,
On définit également

$$\eta_A = \sum_{n=1}^{\infty} \mathbb{1}_A(X_n)$$

le nombre de passage de (X_n) sur A .

-Les quantités $E_x(\eta_A)$ et $P_x(\tau_A < \infty)$, qui sont le nombre moyen de passages sur A et la probabilité de retour à A en un nombre fini de pas sont particuliérement importantes .
-Nous nous intéresserons surtout aux règles d'arrêt de la forme $\tau_A = inf\{n \geq 1; X_n \in A\}$ qui expriment le fait que τ_A prend la valeur n quand aucune des valeurs de $X_0, X_1, ..., X_{n-1}$ sont dans l'état (ou ensemble) donné A, mais la nième valeur l'est.

1.4.7 Propriété forte de Markov

-La propriété de Markov forte correspond au résultat suivant, dont la preuve découle de la propriété faible de Markov et du conditionnement sur $\{\zeta = n\}$:

Pour toute distribution initiale μ et pour tout temps d'arret ζ qui est presque surement fini ,

$$E_\mu(h(X_{\zeta+1}, X_{\zeta+2}, ...)|x_1, ..., x_\zeta) = E_{x_\zeta}(h(X_1, X_2, ...))$$

à condition que les espérences existent .

- On peut ainsi conditionner sur un nombre aléatoire d'instants tout en gardant les propriétés fondamentales d'une chaîne de Markov.

1.5 Irréductibilité, atomes et petits ensembles

1.5.1 Irréductibilité

-La propriété d'irréductibilité est une première mesure de la sensibilité de la chaîne de Markov aux conditions initiales, x_0 ou μ

Elle est crucial dans la configuration des Algorithmes MCMC, car elle conduit à une garantie de convergence,
Donc évite une étude détaillée de l'opérateur de transition, sans laquelle il serait nécessaire de spécifier des conditions initiales "acceptables".

-Dans la cas discret , la chaine est irréductible si tous les états communiquent , c'est à dire :

$$P_x(\tau_y < \infty) > 0, \forall x, y \in \mathcal{X}$$

- Dans de nombreux cas, $P_x(\tau_y < \infty)$ est uniformément égal à zéro, et il est nécessaire d'introduire une mesure auxiliaire φ sur $\mathcal{B}(\mathcal{X})$ pour bien définir la notion d'irréductibilité.

1.5.2 φ Irréductibilité

Etant donnée une mesure φ ,
la chaine de Markov (X_n) avec un noyau de transition $K(x, y)$ est φ irréductible si

$$\forall A \in \mathcal{B}(\mathcal{X}), \text{ avec } \varphi(A) > 0 : \exists n \text{ tel que } K^n(x, A) > 0, \forall x \in \mathcal{X} \ (P_x(\tau_A < \infty) > 0)$$

-La chaine est fortement φ irréductible si $n = 1$ pour tout ensemble mesurable A .

1.5.3 Théoréme

-La chain (X_n) est φ irréductible si et seulement si pour tout $x \in \mathcal{X}$ et pour tout $A \in \mathcal{B}(\mathcal{X})$ tel que $\varphi(A) > 0$, l'une des propriétés suivantes est valable :

$$1) \ \exists n \in N^* \text{ tel que } K^n(x, A) > 0$$

$$2) \ E_x(\eta_A) > 0$$

$$3) \ K_\epsilon(x, A) > 0 \text{ pour un } 0 < \epsilon < 1$$

-L'introduction de la chaîne K_ϵ permet alors de créer un noyau strictement positif dans le cas d'une chaîne φ irréductible .

-La mesure φ ne joue aucun role crucial dans le sens ou l'irréductibilité est une propriété intrinsec de (X_n) et ne dépend pas de φ .

1.5.4 Théoréme

-Si (X_n) est φ irréductible , alors il existe une mesure de probabilité ψ telle que :

1) la chaine (X_n) est ψ irréductible

2) si $\exists \xi$ mesure telle que (X_n) est ξ irréductible alors ξ est dominé par ψ

3) si $\psi(A) > 0$ alors $\psi(\{y; P_y(\tau_A < \infty) > 0\}) = 0$

4) la mesure ψ est équivalente à : $\psi_0(A) = \int_{\mathcal{X}} K_{1/2}\varphi(dx), \forall A \in \mathcal{B}(\mathcal{X})$

-Ce résultat fournit une méthode constructive pour déterminer la mesure maximale d'irréductibilité ψ à travers une mesure candidate φ, qui doit encore être défini.

1.5.5 Exemple

-Quand $X_{n+1} = \theta X_n + \epsilon_{n+1}$ et les ϵ_n sont des variables normales indépendantes , la chaine est irréductible , la mesure de référence étant la mesure de Lebesgue λ
(en fait , $K(x, A) > 0$, $\forall x \in R$, $\forall A$ tel que $\lambda(A) > 0$)

-Par contre si ϵ_n est uniforme sur [-1,1] et $|\theta| > 1$, la chaine n'est plus irréductible ,
En praticulier , si $\theta > 1$ alors :

$$X_{n+1} - X_n \geq (\theta - 1)X_n - 1 \geq 0$$

pour $X_n \geq 1/(\theta - 1)$. La chaîne est donc monotone croissante et évidemment il est impossible de visiter les valeurs précédentes .

1.5.6 Atomes et petits ensembles

-Dans le cas discret, le noyau de transition est nécessairement atomique dans le sens usuel ;
c'est-à-dire qu'il existe des points dans l'espace d'état avec une masse positive.
-L'extension de cette notion au cas général par Nummelin (1978) est suffisament puissante pour permettre un contrôle de la chaîne presque aussi "précis" que dans le cas discret.

1.5.7 Atomes

-La chaine de Markov (X_n) a un atome $\alpha \in \mathcal{B}(\mathcal{X})$, s'il existe une mesure associée non nulle ν telle que

$$K(x, A) = \nu(A), \forall x \in \alpha, \forall A \in \mathcal{B}(\mathcal{X})$$

-Si (X_n) est ψ irréductible , l'atome est accessible si $\psi(\alpha) > 0$.
- Bien qu'elle s'applique trivialement à toutes les valeurs possibles de X_n dans le cas discret,
cette notion est souvent trop forte pour être utile dans le cas continu puisqu'elle implique que le noyau de transition est constant sur un ensemble de mesure positive .
-Une généralisation plus puissante est la condition dite de minorisation, à savoir qu'il existe un ensemble $C \in \mathcal{B}(\mathcal{X})$, une constante $\epsilon > 0$, et une mesure de probabilité ν telle que :

$$K(x, A) \geq \epsilon\nu(A), \forall x \in C, \forall A \in \mathcal{B}(\mathcal{X})$$

-La mesure de probabilité ν apparaît donc comme une composante constante du noyau de transition sur C.

1.5.8 Petits ensembles

-Un ensemble C est petit si il existe $m \in N^*$ et une mesure non nulle ν_m telle que :

$$K^m(x, A) \geq \nu_m(A), \forall x \in C, \forall A \in \mathcal{B}(\mathcal{X})$$

1.5.9 Théoréme

-Soit (X_n) une chaîne ψ irréductible.
Pour tout ensemble $A \in \mathcal{B}(\mathcal{X})$ tel que $\psi(A) > 0$, il existe $m \in N^*$ et un petit ensemble $C \subset A$ tel que la mesure minorisante associée vérifie $\nu_m(C) > 0$.
De plus, $\mathcal{X}$ peut être décomposé en une partition dénombrable de petits ensembles.

-La décomposition de $\mathcal{X}$ en union dénombrable de petits ensembles est basée sur un petit ensemble arbitraire C et la suite :

$$C_{nm} = \{y; K^n(y, C) > 1/m\}$$

-Les petits ensembles sont évidemment plus faciles à exposer que les atomes, étant donné la liberté autorisé par la condition de minorization $K^m(x, A) \geq \nu_m(A)$.

-Pour des chaînes de Markov suffisamment régulières (au sens topologique), chaque ensemble compact est petit.

-Les atomes, bien qu'il s'agisse d'un cas particulier de petits ensembles, ont des propriétés de stabilité plus fortes puisque la probabilité de transition est invariante sur α.

- Cependant, les méthodes de fractionnement (voir ci-dessous) offrent la possibilité d'étendre la plupart de ces propriétés au cas général .

-Si la condition de minorisation est vraie pour (X_n), il y a deux façons de dériver une chaîne de Markov compagnon (X_n) qui partage de nombreuses propriétés avec (X_n) et posséde un atome α.

-La première méthode est appelée fractionnement de Nummelin et construit une chaîne composée de deux copies de (X_n) .

-Une deuxième méthode utilise un temps d'arrêt pour créer un atome.
Nous préférons privilégier cette dernière méthode car elle est liée à des notions de temps de renouvellement, qui sont également utiles dans le contrôle des algorithmes MCMC .

1.5.10 Temps de renouvellement

-Un temps de renouvellement (ou temps de régénération) est une règle d'arrêt τ avec la propriété que :

$$(X_\tau, X_{\tau+1}, ...) \text{ est indépendant de } (X_{\tau-1}, X_{\tau-2}, ...)$$

-Les visites d'atomes sont des temps de renouvellement, dont les caractéristiques sont assez séduisantes en contrôle de convergence pour les Algorithmes MCMC .

-Si $K(x, A) \geq \epsilon\nu(A) \in C$, $\forall A \in \mathcal{B}(\mathcal{X})$, est valable avec la probabilité $P_x(\tau_C < \infty)$ de retour à C en un temps fini est identiquement égal à 1 sur $\mathcal{X}$,
Alors on peut modifier le noyau de transition quand $X_n \in C$, en simulant X_{n+1} par :

$$X_{n+1} \sim \begin{cases} \nu & \text{avec probabilité } \epsilon \\ \frac{K(X_n,.)-\epsilon\nu(.)}{1-\epsilon} & \text{avec probabilité } 1-\epsilon \end{cases}$$

-C'est-à-dire en simulant X_{n+1} à partir de ν avec probabilité ϵ chaque fois que X_n est dans

C.

-Cette modification ne change pas la distribution marginale de X_{n+1} conditionnellement sur X_n, puisque :

$$\epsilon\nu(A) + (1-\epsilon)\frac{K(x_n, A) - \epsilon\nu(A)}{1-\epsilon} = K(x_n, A), \forall A \in \mathcal{B}(\mathcal{X})$$

mais elle produit des temps de renouvellement pour chaque temps j tels que $X_j \in C$ et $X_{j+1} \sim \nu$.

-Maintenant, nous voyons clairement comment les temps de renouvellement se traduisent pour des chaînes indépendantes.

-Lorsque $X_{j+1} \sim \nu$, cet événement est totalement indépendant de tout passé , comme l'état actuel de la chaîne n'a aucun effet sur la mesure ν .

-La condition de minorization nous permet de créer la chaîne fractionée avec la même distribution marginale que la chaîne d'origine .

-On note la séquence des temps de renouvellement par :

$$\tau_j = inf\{n \geq \tau_{j-1}; X_n \in C \text{ et } X_{n+1} \sim \nu\}, (j > 0)$$

avec $\tau_0 = 0$.

-On peut introduire la chaîne augmentée, également appelée chaîne fractionnée $X_n^\vee = (X_n, \omega_,^\vee)$,
avec $\omega_n^\vee = 1$ quand $X_n \in C$ et X_{n+1} est générée depuis ν .

-On peut montrer que l'ensemble $\alpha^\vee = C \times \{1\}$ est un atome de la chaine $X_n^\vee$,
La sous-chaine résultante (X_n) étant encore une chaine de Markov avec noyau de transition $K(x_n, .)$.

-La notion de petit ensemble n'est utile que dans des contextes finis et discrets lorsque les probabilités individuelles des états sont trop faibles pour permettre un rythme de renouvellement.
-Dans ces cas, de petits ensembles sont constitués de collections d'états avec ν défini au minimum.

-Sinon, de petits ensembles réduits à une seule valeur sont également des atomes.

1.6 Cycles et apériodicité

-Le comportement de (X_n) peut parfois être limité par des contraintes déterministes sur les déplacements de X_n vers X_{n+1}.
-Nous formalisons ici ces contraintes et nous allons montrer que les chaînes produites par Les algorithmes MCMC n'affichent pas ce comportement et, par conséquent, ne souffrent pas des inconvénients associés.

1.6.1 Période

-Dans le cas discret, la période d'un état $\omega \in \mathcal{X}$ est définie comme :

$$d(\omega) = p.g.c.d\{m \geq 1; K^m(\omega, \omega) > 0\}$$

-La valeur de la période est constante sur tous les états qui communiquent avec ω.

-Dans le cas d'une chaîne irréductible sur un espace fini $\mathcal{X}$, la matrice de transition peut s'écrire (avec une éventuelle réorganisation des états) sous forme de bloc de matrices :

$$P = \begin{pmatrix} 0 & D_1 & 0 & ... & 0 \\ 0 & 0 & D_2 & ... & 0 \\ ... & ... & ... & ... & ... \\ D_d & 0 & ... & ... & 0 \end{pmatrix}$$

ou les blocks D_i sont des matrices stochastiques .

-Cette représentation illustre bien le passage forcé d'un groupe d'états à un autre, avec un retour au groupe initial se produisant à chaque dième étape.

-Si la chaîne est irréductible (donc tout les états communiquent), il n'y a qu'une seule valeur pour la période.

-Une chaîne irréductible est apériodique si elle est de période 1.

-L'extension au cas général nécessite l'existence d'un petit ensemble.

1.6.2 Cycle

-Une chaîne ψ-irréductible (X_n) a un cycle de longueur d s'il existe un petit ensemble C , un entier associé M , et une distribution de probabilité ν_M tel que d est le p.g.c.d. de

$$\{m \geq 1; \exists \delta_M > 0 \text{ tel que } C \text{ est un petit ensemble pour } \nu_m \geq \delta_M \nu_M\}$$

-Une décomposition comme $P = \begin{pmatrix} 0 & D_1 & 0 & ... & 0 \\ 0 & 0 & D_2 & ... & 0 \\ ... & ... & ... & ... & ... \\ D_d & 0 & ... & ... & 0 \end{pmatrix}$ peut être établie en général.

-On peut montrer que le nombre d est indépendant du petit ensemble C et que ce nombre caractérise intrinsèquement la chaîne (X_n)·

-La période de (X_n) est alors définie comme le plus grand entier d satisfaisant la définition précédente
$d = max_i\ d_i$ tel que :
$d_i = p.g.c.d\{m \geq 1; \exists \delta_M > 0$ tel que C est un petit ensemble pour $\nu_m \geq \delta_M \nu_M\}$

-(X_n) est apériodique si $d = 1$.

-S'il existe un petit ensemble A et une mesure minorizante ν_1 tels que $\nu_1(A) > 0$ (il est donc possible de passer de A à A en une seule étape), la chaîne est dite fortement apériodique .

- la chaîne K_ϵ peut être utilisée pour transformer une chaîne apériodique en une chaîne fortement apériodique.

-Dans les configurations discrètes, si un état $x \in \mathcal{X}$ satisfait $P_{xx} > 0$, la chaîne (X_n) est apériodique, bien que ce ne soit pas une condition nécessaire .

-Lorsque la chaîne est continue et que le noyau de transition a une composante qui est absolument continue par rapport à la mesure de Lebesgue, avec densité $f(.|x_n)$,
Une condition suffisante pour l'apériodicité est que $f(.|x_n)$ soit positif au voisinage de x_n·

-La chaîne peut alors rester dans ce voisinage pendant un nombre arbitraire d'instants

avant de visiter un ensemble A.

-Les algorithmes MCMC conduisent à des chaînes apériodiques, éventuellement via l'introduction d'étapes supplémentaires.

1.7 Récurrence et transience

1.7.1 Classification des chaînes irréductibles

-D'un point de vue algorithmique, une chaîne de Markov doit jouir d'une bonne stabilité pour garantir une approximation acceptable du modèle simulé.

-En effet, l'irréductibilité assure que tout ensemble A sera visité par la chaine de Markov mais cette propriété est trop faible pour assurer que la trajectoire de (X_n) entrera A assez souvent.

-Formaliser cette stabilité des chaînes Markov conduit à différentes notions de récurrence.

-Dans une configuration discrète, la récurrence d'un état équivaut à une garantie de retour sûr.

-Cette notion est donc nécessairement satisfaite pour les chaînes irréductibles sur un espace fini .

1.7.2 Etats transients et récurrents

-Dans un espace d'état fini $\mathcal{X}$, un état $\omega \in \mathcal{X}$ est transient si le nombre moyen de visites de ω , $E_\omega(\eta_\omega)$ est fini ,
et récurrent si $E_\omega(\eta_\omega) = \infty$.

-Pour les chaînes irréductibles, les propriétés de récurrence et de transience sont des propriétés de la chaîne, pas d'un état particulier.

-Par conséquent, si η_A désigne le nombre de visites définies ,
pour tout $(x, y) \in \mathcal{X}^2$ soit $E_x(\eta_y) < \infty$ dans cas le transitoire
ou $E_x(\eta_y) = \infty$ dans le cas récurrent.

-On dit alors que la chaîne est transitoire ou récurrente, l'une des deux propriétés étant nécessairement satisfaite dans le cas irréductible.

-Le traitement du cas général (c'est-à-dire non discret) repose sur des chaînes avec des atomes,
l'extension aux chaînes générales (avec de petits ensembles) se base sur le fractionement .

-Commençons par étendre les notions de récurrence et de transience .

1.7.3 Ensembles transients et récurrents

-Un ensemble A est dit récurrent si $E_x(\eta_A) = +\infty$ pour tout $x \in A$.

-L'ensemble A est uniformément transitoire s'il existe une constante M telle que $E_x(\eta_A) < M$ pour tout $x \in A$.

-Il est transitoire s'il existe un recouvrement de $\mathcal{X}$ par ensembles uniformément transitoires ;
c'est-à-dire une collection dénombrable d'esembles uniforméments transitoires B_i tel que

$$A = \bigcup_i B_i$$

1.7.4 Théoréme

-Soit (X_n) une chaine de Markov ψ-irréductible avec un atome accessible α alors :

1) si α est récurrent alors tout ensemble $A \in \mathcal{B}(\mathcal{X})$ tel que $\psi(A) > 0$ est récurrent

2) si α est transient alors $\mathcal{X}$ est transient

-La propriété 1) est la plus pertinente dans la configuration MCMC et peut être dérivé des équations de Chapman-Kolmogorov.

1.7.5 Chaine récurrente et transiente :

-Une chaine de Markov (X_n) est récurrente si :
i) il existe une mesure ψ telle que (X_n) est ψ – irréductible et

ii) $\forall A \in \mathcal{B}(\mathcal{X})$ tel que $\psi(A) > 0, E_x(\eta_A) = \infty, \forall x \in A$

-La chaine (X_n) est transiente si (X_n) est ψ-irréductible et si $\mathcal{X}$ est transient .

-Le résultat de classification du Théorème précédent peut être facilement étendu aux chaînes fortement apériodiques puisqu'elles satisfont une condition minorisante ,et peuvent donc être fractionées , tandis que la chaîne (X_n) et sa version fractionée $(X_n^{\vee})$ sont toutes les deux récurrentes ou toutes les deux transitoires.

-La généralisation à une chaîne irréductible arbitraire découle des propriétés de la K_ϵ chaîne qui est fortement apériodique, par la relation

$$\sum_{n=0}^{\infty} K_\epsilon^n = \frac{1-\epsilon}{\epsilon} \sum_{n=0}^{\infty} K^n$$

$$\text{vu que } E_x(\eta_A) = \sum_{n=0}^{\infty} K^n(x, A) = \frac{\epsilon}{1-\epsilon} \sum_{n=0}^{\infty} K_\epsilon^n(x, A)$$

1.7.6 Théoréme

-une chaine ψ-irréductible est soit récurrente soit transiente .

1.8 Critères de récurrence

-Une chaine (X_n) ψ-irréductible est récurrente si :
$\exists C$ un petit ensemble avec $\psi(C) > 0$ tel que $P_x(\tau_c < \infty) = 1, \forall x \in C$.

1.9 Harris récurrence

-Il est en effet possible de renforcer les propriétés de stabilité d'une chaîne (X_n) en exigeant non seulement un nombre moyen infini de visites à chaque petit ensemble, mais aussi un nombre infini de visites pour chaque chemin de la chaîne de Markov .

1.9.1 Ensemble harris récurrent

-Un ensemble A est harris récurrent si :

$$P_x(\eta_A = \infty) = 1, \forall x \in A$$

1.9.2 Chaine harris récurrente

-Une chaine (X_n) est harris récurrente si :

il existe une mesure ψ telle que (X_n) est ψ – irréductible , et

pour tout ensemble A tel que $\psi(A) > 0, A$ est harris récurrent

-Pour rappel la récurrence correspond à $E_x(\eta_\alpha) = \infty$, une condition plus faible que $P_x(\eta_A = \infty) = 1$.

-Si pour tout $A \in \mathcal{B}(\mathcal{X})$, $P_x(\tau_A < \infty) = 1$, $\forall x \in A$ alors :
$P_x(\eta_A = \infty) = 1, \forall x \in \mathcal{X}$ et
(X_n) est harris récurrente .

-La propriété de récurrence de Harris n'est nécessaire que lorsque $\mathcal{X}$ est non dénombrable.

-Si $\mathcal{X}$ est fini ou dénombrable, on peut en effet montrer que $E_x(\eta_x) = \infty$ si et seulement si $P_x(\tau_x < \infty) = 1, \forall x \in \mathcal{X}$.

-Dans le cas général, c'est possible de prouver que si (X_n) est harris récurrente , alors :
$P_x(\eta_B = \infty) = 1$ pour tout $x \in \mathcal{X}$ et $B \in \mathcal{B}(\mathcal{X})$ tel que $\psi(B) > 0$.

1.9.3 Théoréme

-si (X_n) est une chaine de Markov ψ-irréductible avec un petit ensemble C tel que $P_x(\tau_C < \infty) = 1$, $\forall x \in C$ alors :
(X_n) est harris récurrente .

-la récurrence d'Harris est valable pour la plupart des algorithmes MCMC .

1.10 Mesures invariantes

1.10.1 Chaînes stationnaires

-Un niveau accru de stabilité de la chaîne (X_n) est atteint si la distribution marginale de X_n est indépendante de n .
-Plus formellement, il s'agit d'une exigence pour l'existence d'une loi de probabilité π

telle que :

$$X_{n+1} \sim \pi \text{ si } X_n \sim \pi$$

et les méthodes MCMC sont basées sur le fait que cette exigence, qui définit un type particulier de récurrence appelé récurrence positive, peut être rencontré.

1.10.2 Mesure invariante

-Une mesure σ finie π est invariante pour le noyau de transition $K(.,.)$ si :

$$\pi(B) = \int_{\mathcal{X}} K(x, B)\pi(dx), \ \forall B \in \mathcal{B}(\mathcal{X})$$

1.10.3 Chaine positive

-Lorsqu'il existe une mesure de probabilité invariante pour une chaine ψ-irréductible (donc récurrente), la chaîne est positive.

1.10.4 Chaine récurrente nulle

-les Chaînes récurrentes qui ne permettent pas une mesure invariante finie sont appelées récurrentes nulles.

-La distribution invariante est également dite stationnaire si π est une mesure de probabilité ,
puisque $X_0 \sim \pi$ implique que $X_n \sim \pi$ pour tout n ;
Ainsi, la chaîne est stationnaire en distribution. (Notez que le cas alternatif où π n'est pas fini est plus difficile à interpréter en termes de comportement de la chaîne.)
-Il est facile de montrer que si la chaîne est irréductible et admet une mesure invariante σ-finie, cette mesure est unique, à un facteur multiplicatif près .

-si la chaine (X_n) est positive alors elle est récurrente .

1.10.5 Théorème de Kac

-Un résultat classique sur les chaînes de Markov irréductibles à espace d'état discret est que la distribution stationnaire, lorsqu'elle existe, est donnée par :

$$\pi_x = (E_x(\tau_x))^{-1} \ , \forall x \in \mathcal{X}$$

-on peut interpréter $E_x(\tau_x)$ comme le nombre moyen d'excursions entre deux passages en x. (parfois on parle de théorème de Kac) .
-Il s'ensuit également que $(E_x(\tau_x))^{-1}$ est le vecteur propre associé à la valeur propre 1 pour la matrice de transition P .

-Nous établissons maintenant ce résultat dans le cas plus général où (X_n) possède un atome α .

1.10.6 Théoréme

-Soit (X_n) une chaine de Markov ψ-irréductible avec un atome α alors :

la chaine est positive si et seulement si $E_\alpha(\tau_\alpha) < \infty$

et dans ce cas , la distribution invariante π de (X_n) satisfait : $\pi(\alpha) = (E_\alpha(\tau_\alpha))^{-1}$

-La notation $E_\alpha(.)$ est légitime dans ce cas puisque le noyau de transition est le même pour tout $x \in \alpha$.
-Ce théoréme indique que la positivité est une propriété de stabilité plus forte que la récurrence .
-En fait cette derniére correspond à $P_\alpha(\tau_\alpha = \infty) = 0$ qui est une condition nécessaire pour $E_\alpha(\tau_\alpha) < \infty$.

-le cas générale peut etre traité en fractionant (X_n) en $(X_n^\vee)$ qui a un atome , et la mesure invariante de $(X_n^\vee)$ induit une mesure invariante pour (X_n) par marginalisation .

1.10.7 Théoréme

- Si (X_n) est une chaîne récurrente, alors :
il existe une mesure invariante $\sigma - fini$ qui est unique à un facteur multiplicatif près.

1.10.8 Exemple : marche aléatoire sur R

-on considére la marche aléatoire su R :
$X_{n+1} = X_n + W_n$ ou W_n a pour densité Γ .
-Puisque $K(x, .)$ est la distribution avec comme densité $\Gamma(y - x)$,
la distribution de X_{n+1} est invariante par translation, et ceci implique que la mesure de Lebesgue est une mesure invariante :

On a $\lambda(A) = \lambda(A + x)$ et $K(x, A + x) = \int_A \Gamma(dy)$ d'ou :

$$\int K(x, A + x)\lambda(dx) = \int \int_A \Gamma(dy)\lambda(dx) = \int \Gamma(dy) \int_A \lambda(dx) = \lambda(A) = \lambda(A + x)$$

-De plus, l'invariance de λ et l'unicité de la mesure invariante impliquent que la chaîne (X_n) ne peut pas être récurrente positive (en fait, elle est récurrente nulle).

1.10.9 Exemple : continuation (exemple AR(1))

-Étant donné que le noyau de transition correspond à une dsitribution $\mathcal{N}(\theta x_{n-1}, \sigma^2)$,
une distribution normale $\mathcal{N}(\mu, \tau^2)$ est stationnaire pour la chaine AR(1) seulement si :

$$\mu = \theta\mu \text{ et } \tau^2 = \tau^2\theta^2 + \sigma^2$$

-Ces conditions impliquent que : $\mu = 0$ et $\tau^2 = \sigma^2/(1 - \theta^2)$,
qui peut etre possible que si $|\theta| < 1$.
-Dans ce cas $\mathcal{N}(0, \sigma^2/(1 - \theta^2))$ est en effet l'unique distribution stationnaire de la chaine AR(1) .
-Donc si $|\theta| < 1$, la distribution marginale de la chaîne est une densité indépendante de n , et la chaîne est positive (donc récurrente) .
-Nous allons illustrer cet exemple en montrons les trajectoires bidimensionnelles d'une chaîne AR(1), où chaque coordonnée est une chaîne AR(1) univariée. (Nous utilisons deux dimensions pour mieux illustrer graphiquement le comportement de la chaîne.)
-Voici le code R (pour $\sigma = 1$ et $\theta = 0.4$) :

```
sigma=1;theta=1.001;N=10^3;X=NULL;epsX=rnorm(N,0,sigma)
X[1]=epsX[1]
for(n in 1:(N-1)){
X[n+1]=theta*X[n]+epsX[n+1]};Y=NULL;epsY=rnorm(N,0,sigma)
Y[1]=epsY[1]
for(n in 1:(N-1)){
Y[n+1]=theta*Y[n]+epsY[n+1]};plot(X,Y)
```

-Nous avons le résultat de l'histogramme ci-dessous ;
ainsi que le résultat (pour $\sigma = 1$ et $\theta = 0.99$); et de meme le résultat (pour $\sigma = 1$ et $\theta = 1.001$) .
-pour les 2 premiéres chaines on voit θ augmenter, mais les 2 sont positives récurrentes .

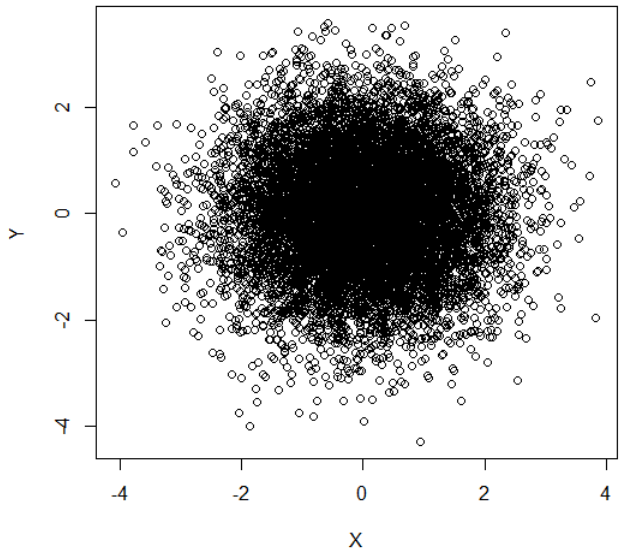

FIGURE 1.2 – trajectoire d'une chaîne AR(1) pour $\theta = 0.4$ et $\sigma = 1$

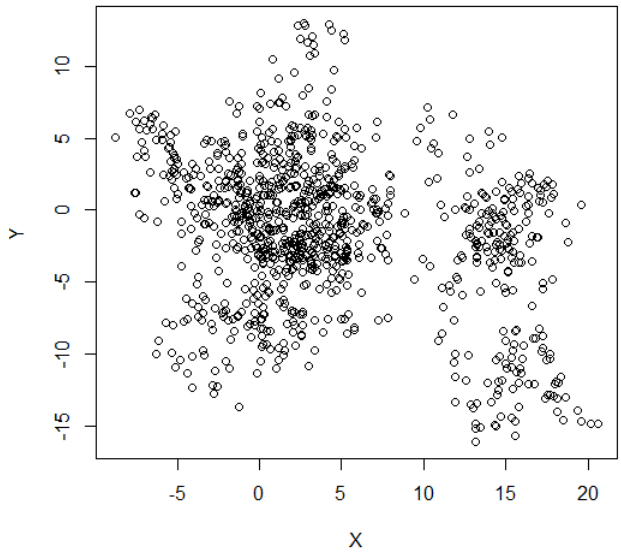

FIGURE 1.3 – trajectoire d'une chaîne AR(1) pour $\theta = 0.99$ et $\sigma = 1$

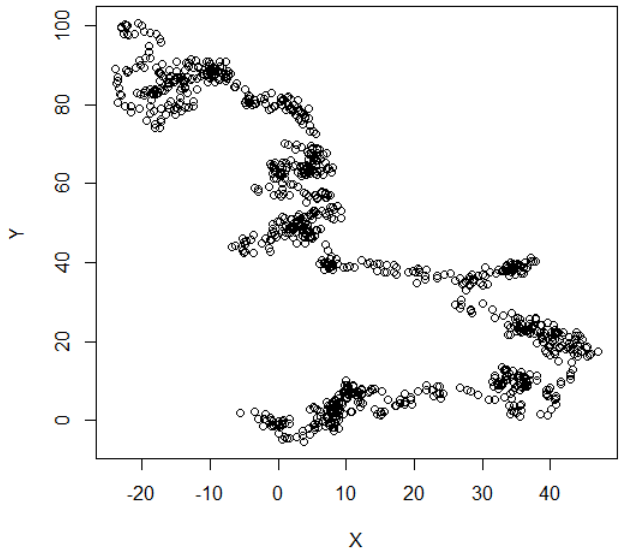

FIGURE 1.4 – trajectoire d'une chaîne AR(1) pour θ = 1.001 et σ = 1

Il en résulte que la chaîne "remplit" l'espace ; et nous pouvons voir que quand θ augmente, il y a moins de remplissage dense.
-Enfin, la troisième chaîne, avec θ = 1.001, est transitoire, et non seulement elle ne remplit pas l'espace, mais elle s'échappe et ne revient jamais.

-Lorsque nous utilisons une chaîne de Markov pour explorer un espace, nous voulons qu'elle remplisse l'espace.
Ainsi, nous voulons que nos chaînes MCMC soient récurrentes positives.
-il existe des Chaînes de Markov transitoires à mesures stationnaires.
Par exemple, les marches aléatoires dans R^3 et Z^3 , sont toutes deux transitoire et ont la mesure de Lebesgue et la mesure de comptage comme mesures stationnaires .

1.10.10 Réversibilité et condition d'équilibre ponctuel

-La propriété de stabilité inhérente aux chaînes stationnaires peut être liée à une autre propriété de stabilité appelée réversibilité, qui stipule que la direction du temps n'a pas d'importance dans la dynamique de la chaîne .

1.10.11 Chaine réversible

-Une chaîne de Markov stationnaire (X_n) est réversible si la distribution de X_{n+1} conditionnellement à $X_{n+2} = x$ est la même que la distribution de X_{n+1} conditionnellement à

$X_n = x$.

-En fait, la réversibilité peut être liée à l'existence d'une mesure stationnaire π avec condition .

1.10.12 Condition d'équilibre ponctuel

-Une chaine de Markov avec noyau de transition K satisfait la condition d'équilibre ponctuel si :
Il existe une fonction f satisfaisant :

$$K(y,x)f(y) = K(x,y)f(x) \ , \forall (x,y)$$

-Bien que cette condition ne soit pas nécessaire pour que f soit une mesure stationnaire associé au noyau de transition K, elle fournit une condition suffisante qui est souvent facile à vérifier et peut être utilisé pour la plupart des algorithmes MCMC.
- La condition d'équilibre ponctuel exprime un équilibre dans le flux de la chaîne de Markov , à savoir que la probabilité d'être en x et de se déplacer vers y est la même que la probabilité d'être en y et de revenir en x.
-Lorsque f est une densité, cela implique également que la chaîne est réversible.

1.10.13 Théoréme

-On suppose que la chaine de Markov avec noyau de transition K satisfait la condition d'équilibre ponctuel avec une densité de probabilité π alors :
1) la densité π est une densité invariante pour la chaine .
2) la chaine est réversible .

-Si $f(x,y)$ est une densité jointe , alors on peut écrire :

$$f(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$$

Et donc la condition d'équilibre ponctuel require :
$f_X = f_Y$ et $f_{X|Y} = f_{Y|X}$,
Donc il y a symétrie dans les conditionnels et les marginaux sont les mêmes .

1.11 Ergodicité et convergence

1.11.1 Ergodicité

-Considérant la chaîne de Markov (X_n) d'un point de vue temporel, il est naturel (et important) d'établir le comportement limite de X_n ; C'est à dire vers quoi la chaîne converge ?
-L'existence (et l'unicité) d'une distribution invariante π fait de cette distribution un candidat naturel pour la distribution limite , et nous nous tournons maintenant vers la recherche de conditions suffisantes sur (X_n) pour que X_n soit distribué asymptotiquement selon π .

1.11.2 Atome ergodique

-soit (X_n) une chaine Harris récurrente positive , avec π distribution invariante ,
un atome α est ergodique si :

$$\lim_{n\to\infty} |K^n(\alpha,\alpha) - \pi(\alpha)| = 0$$

-Dans le cas dénombrable, l'existence d'un atome ergodique est, en fait, suffisante pour établir la convergence selon la norme de variation totale,

$$||\mu_1 - \mu_2||_{VT} = \sup_A |\mu_1(A) - \mu_2(A)|$$

-Si (X_n) est Harris positive sur $\mathcal{X}$ dénombrable et s'il existe un atome ergodique $\alpha \subset \mathcal{X}$ alors :

$$\lim_{n\to\infty} ||K^n(x,.) - \pi||_{VT} = 0 \ , \forall x \in \mathcal{X}$$

-Nous développons une extension ce qui nous permet de traiter le cas général en utilisant des techniques de couplage.
(Ces techniques sont également utiles dans l'évaluation de la convergence pour les algorithmes MCMC).
-Le principe de couplage utilise deux chaînes (X_n) et (X'_n) associées aux même noyau, l'événement de "couplage" ayant lieu lorsqu'ils se rencontrent en α ;
c'est à dire , au premier instant n_0 tel que $X_{n_0} \in \alpha$ et $X'_{n_0} \in \alpha$.
-Après cet instant, les propriétés probabilistes de (X_n) et (X'_n) sont identiques et si l'une des deux chaînes est stationnaire, il n'y a plus de dépendance aux conditions initiales pour l'autre chaîne.
-Par conséquent, si nous pouvons montrer que le temps de couplage (c'est-à-dire le temps qu'il faut pour que les deux chaînes se rencontrent), est fini pour presque chaque point de

départ , l'ergodicité de la chaîne suit.

-Pour un atome récurrent α dans un espace dénombrable $\mathcal{X}$,
On note $\tau_\alpha(k)$ la k'iéme visit de α (k=1,2,...) , et on note $p = (p(1), p(2), ...)$ la distribution du temps d'excursion : $S_k = \tau_\alpha(k+1) - \tau_\alpha(k)$, entre 2 visites de α .
-Si $q = (q(0), q(1), ...)$ représente la distribution de $\tau_\alpha(1)$ qui dépend de la condition initiale x_0 ou μ , alors :
La distribution de $\tau_\alpha(n+1)$ est donnée par le produit de convolution $q \star p^{n^\star}$,
C'est à dire la distribution de la some de n variables aléatoires i.i.d distribué selon p et une variable aléatoire distribuée selon q ,
Vu que :

$$\tau_\alpha(n+1) = S_n + ... + S_1 + \tau_\alpha(1)$$

-On considére 2 suites (S_i) et (S_i') telles que $S_1, S_2, ...$ et $S_1', S_2', ...$ sont i.i.d suivant p avec $S_0 \sim q$ et $S_0' \sim r$.
-On introduit les fonctions indicatrices :

$$Z_q(n) = \sum_{j=0}^{n} \mathbb{1}_{S_1+...+S_j=n} \text{ et } Z_r(n) = \sum_{j=0}^{n} \mathbb{1}_{S_1'+...+S_j'=n}$$

qui correspondent aux évenements que les chaines (X_n) et (X_n') visitent α au temps n .
-Le temps de couplage est donné par :

$$T_{qr} = min\{j\,;\, Z_q(j) = Z_r(j) = 1\}$$

-Si le temps d'excursion satisfait : $m_p = \sum_{n=0}^{\infty} np(n) < \infty$,
Et si p est aperiodique (le p.g.c.d du support de p est 1) , alors :
Le temps de couplage T_{pq} est presque surement fini , c'est à dire :
$P(T_{pq} < \infty) = 1$, $\forall q$

-Si p est aperiodique avec une moyenne finie m_p , alors : Z_q satisfait :

$$\lim_{n\to\infty} |P(Z_q(n) = 1) - m_q^{-1}| = 0$$

-La probabilité de visiter α au temps n est donc asymptotiquement indépendante de la distribution initiale .

1.11.3 Théoréme

-Pour une chaîne de Markov apériodique récurrente positive sur un espace dénombrable ,

pour tout état initial x :

$$\lim_{n\to\infty} ||K^n(x,.) - \pi||_{VT} = 0$$

-Pour un espace d'états générale $\mathcal{X}$, la récurrence de Harris est néanmoins nécessaire dans la dérivation de la convergence de K^n vers π.
(Notez qu'une autre caractérisation de récurrence de Harris est la convergence de $||K_x - \pi||_{VT}$ vers 0 pour toute valeur x, au lieu de presque toutes les valeurs.)

1.11.4 Théoréme

-Si (X_n) est Harris positive et aperiodique alors :

$$\lim_{n\to\infty} || \int K^n(x,.)\mu(dx) - \pi||_{VT} = 0$$

pour toute distribution initiale μ .

-Ce résultat découle d'une extension du cas dénombrable aux chaînes Harris positive fortement apériodiques par fractionement, puisque ces chaînes autorisent toujours les petits ensembles ,
Il est alors possible de passer à des chaînes arbitraires par le résultat suivant.

-Si π est une distribution invariante de P , alors :

$$|| \int K^n(x,.)\mu(dx) - \pi||_{VT} \text{ est décroissante en } n$$

-L'équivalence $\mu_{VT} = \frac{1}{2}\sum_{|h|\leq 1} |\int h(x)\mu(dx)|$ implique la convergence :
$\lim_{n\to\infty} |E_\mu(h(X_n)) - E^\pi(h(X))| = 0$, pour toute fonction bornée h .
-Cette équivalence est, en effet, souvent prise comme la condition de définition pour la convergence des distributions .
-On peut cependant conclure $\lim_{n\to\infty} |E_\mu(h(X_n)) - E^\pi(h(X))| = 0$ à partir d'un ensemble d'hypothèses plus faible, où nous n'avons pas besoin de toute la force de la récurrence de Harris .
-L'extension de $\lim_{n\to\infty} |E_\mu(h(X_n)) - E^\pi(h(X))| = 0$ à des fonctions plus générales h est appelée h-ergodicité .

1.11.5 Théoréme

-Soit (X_n) un chaine positive , récurrente et aperiodique alors :
a) Si $E^\pi(|h(X)|) = \infty$ alors : $E_x(|h(X)|) \to \infty$, $\forall x$.

b) Si $\int |h(x)|\pi(dx) < \infty$ alors :
$\lim_{n\to\infty} \sup_{|m(x)|\leq|h(x)|} |E_y(m(X_n)) - E^\pi(m(X))| = 0$, pour tout petit ensemble C tel que $\sup_{y\in C} E_y(\sum_{t=0}^{\tau_C-1} h(X_t)) < \infty$.

-Des conditions similaires apparaissent comme des conditions nécessaires pour le Théorème centrale limite .
-La condition $\sup_{y\in C} E_y(\sum_{t=0}^{\tau_C-1} h(X_t)) < \infty$ concerne un argument de couplage , dans le sens que l'influence de la condition initiale s'annule " assez rapidement ,"

1.11.6 Convergence géométrique

-La convergence $\lim_{n\to\infty} \sup_{|m(x)|\leq|h(x)|} |E_y(m(X_n)) - E^\pi(m(X))| = 0$ de l'espérance de $h(x)$ au temps n vers l'espérance de $h(x)$ sous la distribution stationnaire π assure en quelque sorte le bon comportement de la chaîne (X_n) quelle que soit la valeur initiale X_0 (ou sa distribution).
-Une description plus précise des propriétés de convergence passe par l'étude des vitesse de convergence de K^n vers π.
-Une évaluation de cette vitesse est importante pour les Algorithmes MCMC dans le sens où il se rapporte aux règles d'arrêts pour ces algorithmes ;
Une vitesse de convergence minimale est également une exigence pour l'application du théorème central limite.
-Pour étudier de plus près la vitesse de convergence, nous introduisons d'abord une extension de la norme de variation totale, notée $||.||_h$, qui permet une borne supérieure autre que 1 sur les fonctions.
-La généralisation est définie par :

$$||\mu||_h = \sup_{|g|\leq h} |\int g(x)\mu(dx)|$$

1.11.7 Chaine géometriquement h-ergodique

-Une chaine (X_n) est géometriquement h-ergodique avec $h \geq 1$ sur $\mathcal{X}$ si :
(X_n) est Harris positive avec une distribution stationaire π ,
et $E^\pi(h) < \infty$,
et il existe $r_h > 1$ tel que : $\sum_{n=1}^{\infty} r_h^n ||K^n(x,.) - \pi||_h < \infty$, $\forall x \in \mathcal{X}$.

-Le cas h = 1 correspond à l'ergodicité géométrique de (X_n) .
-La h-ergodicité géométrique signifie que $||K^n(x,.) - \pi||_h$ décroît au moins à une vitesse géométrique, puisque :

$\sum_{n=1}^{\infty} r_h^n ||K^n(x,.) - \pi||_h < \infty$ implique : $||K^n(x,.) - \pi||_h \leq M r_h^{-n}$ avec
$M = \sum_{n=1}^{\infty} r_h^n ||K^n(x,.) - \pi||_h$.
-Si (X_n) a un atome α , $\sum_{n=1}^{\infty} r_h^n ||K^n(x,.) - \pi||_h < \infty$ implique que pour un réel $r > 1$,
$E_x(\sum_{n=1}^{\tau_\alpha} h(X_n) r^n) < \infty$ et $\sum_{n=1}^{\infty} |P_\alpha(X_n \in \alpha) - \pi(\alpha)| r^n < \infty$.
-La série associée à $|P_\alpha(X_n \in \alpha) - \pi(\alpha)| r^n$ converge en dehors du cercle unitaire si la série entière associée à $P_\alpha(\tau_\alpha = n)$ converge pour les valeurs de $|r|$ strictement supérieur à 1.

1.11.8 Atome géometriquement ergodique

-Un atome accessible α est géometriquement ergodique si :
Il existe $r > 1$ tel que : $\sum_{n=1}^{\infty} |K^n(\alpha,\alpha) - \pi(\alpha)| r^n < \infty$.

1.11.9 Atome de Kendall

-α est un atome de Kendall si :
Il existe $\kappa > 1$ tel que : $E_\alpha(\kappa^{\tau_\alpha}) < \infty$.

-Si α est un atome de Kendall alors :
α est géometriquement ergodique et assure l'ergodicité géometrique de (X_n) .

1.11.10 Théoreme

-Si (X_n) est ψ-irréductible avec distribution invariante π , et s'il existe un atome ergodique α alors :
Il existe $r > 1, \kappa > 1, R < \infty$ tels que :
Pour presque tout $x \in \mathcal{X}$: $\sum_{n=1}^{\infty} r^n ||K^n(x,.) - \pi||_{VT} < R E_x(\kappa^{\tau_\alpha}) < \infty$.

1.11.11 Ergodicité uniforme

-La propriété d'ergodicité uniforme est plus forte que l'ergodicité géométrique dans le sens que le taux de convergence géométrique doit être uniforme sur tout l'espace .
-Elle est utilisé dans le théorème central limite .

1.11.12 Chaine uniformément ergodique

-La chaine de Markov (X_n) est uniformément ergodique si :

$$\lim_{n \to \infty} \sup_{x \in \mathcal{X}} ||K^n(x,.) - \pi||_{VT} = 0$$

1.11.13 Théoréme

-Les conditions suivantes sont équivalentes :

1) (X_n) est uniformément ergodique

2) Il existe $R < \infty, r > 1$ tels que $||K^n(x,.) - \pi||_{VT} < Rr^{-n}, \forall x \in \mathcal{X}$

3) (X_n) est apériodique et $\mathcal{X}$ est un petit ensemble

4) (X_n) est apériodique et il existe un petit ensemble C et un réel $\kappa > 1$ tels que $\sup_{x\epsilon\mathcal{X}} E_x(\kappa^{\tau_C}) < \infty$

-Si tout l'espace $\mathcal{X}$ est petit , alors il existe une distribution de probabilité ϕ sur $\mathcal{X}$ et des constantes $\epsilon < 1, \delta > 0$ et n tels que :
Si $\phi(A) > \epsilon$ alors $\inf_{x\epsilon\mathcal{X}} K^n(x,A) > \delta$.
- Cette propriété est parfois appelée condition de Doeblin.
-Cette exigence montre la force de l'ergodicité uniforme et suggère des difficultés sur la vérification.
-Dans le cas fini, l'ergodicité uniforme peut être dérivée du fait que $\mathcal{X}$ est petit puisque la condition :
$P(X_{n+1} = y|X_n = x) \geq \inf_z P_{zy} = \rho_y, , y \in \mathcal{X}$, conduit au choix de la mesure minorizante ν comme :
$\nu(y) = \frac{\rho_y}{\sum_{z\epsilon\mathcal{X}} \rho_z}$; tant que $\rho_y > 0$ pour un certain $y \in \mathcal{X}$. (si (X_n) est récurrente et apériodique , Cette condition de positivité peut etre obtenue par une sous chaine $(Y_m) = (X_{nd})$ pour d suffisament grand .

1.12 Théorèmes limites

-Les différents résultats de convergence (ergodicité) obtenus ne traitent que de la mesure de probabilité P_x^n (par normes différentes), qui est en quelque sorte un "instantané" de la chaîne (X_n) au temps n.
-Ainsi, ils déterminent les propriétés probabilistes du comportement moyen de la chaîne à un instant fixe.
-De telles propriétés, même si elles justifient les méthodes de simulation, sont de moindre importance pour le contrôle de convergence d'une simulation donnée, où les propriétés de la réalisation (x_n) de la chaîne sont les seules caractéristiques qui comptent vraiment.
-Nous devons considérer la différence entre l'analyse probabiliste, qui décrit le comportement moyen d'échantillons, et l'inférence statistique, qui doit raisonner par induction à

partir de l'échantillon observé.
-Alors que les propriétés probabilistes peuvent justifier ou réfuter certaines approches statistiques , cela ne contredit pas le fait que l'analyse statistique doit être faite en fonction de l'échantillon observé.
-Une telle considération peut conduire à l'approche bayésienne dans un cadre statistique .
-Dans la configuration des chaînes de Markov, une analyse conditionnelle peut tirer parti des propriétés de convergence de P_x^n vers π uniquement pour vérifier la convergence, à une quantité d'intérêt, de fonctions du chemin observé de la chaîne.
-En effet, le fait que $||P_x^n - \pi||$ est proche de 0, voir converge géométriquement et rapidement vers 0 avec vitesse ρ^n $(0 < \rho < 1)$, n'apporte pas d'information directe sur l'unique observation disponible de P_x^n , à savoir X_n.
-Les problèmes d'application directe des théorèmes de convergence classiques (Loi des grands nombres, loi du logarithme itéré, théorème central limite, etc.) à l'échantillon $(X_1, ..., X_n)$ sont dus à la fois à la structure de dépendance markovienne entre les observations X_i et à la non-stationnarité de la séquence.
(Ce n'est que si $X_0 \sim \pi$, la distribution stationnaire de la chaîne, que la chaîne soit stationnaire. Puisque cela équivaut à intégrer sur les conditions initiales, ce qui élimine le besoin d'une analyse conditionnelle. Un tel événement, en particulier en MCMC , est quelque peu rare.)
-On suppose donc que la chaîne commence à partir d'un point X_0 dont la distribution n'est pas la distribution stationnaire de la chaîne, et donc nous traitons directement avec des chaînes non stationnaires.

1.12.1 Théorèmes ergodiques

-Étant donné les observations $X_1, ..., X_n$ d'une chaîne de Markov, nous examinons maintenant le comportement limite des sommes partielles :

$$S_n(h) = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

quand n tend vers l'infini ,
en revenant au cas iid par le renouvellement lorsque (X_n) a un atome.

1.12.2 Fonction harmonique

-Une fonction mesurable h est harmonique pour la chaine (X_n) si :

$$E(h(X_{n+1})|x_n) = h(x_n)$$

-Ces fonctions sont invariantes pour le noyau de transition (dans le sens fonctionnelle) et elles caractérisent la récurence d'Harris .

-Pour une chaine de Markov positive , si les seules fonctions harmoniques bornées sont les fonctions constantes alors : la chaine est Harris récurente .

-La fonction $Q(x, A) = P_x(\eta_A = \infty)$ qui est la probabilité d'un nombre infini de retour , décrit un évenement de queue , un événement qui ne dépend pas de $X_1, ..., X_m$ pour un m finie .
- De tels événements obéissent en général à une loi 0-1 , c'est à dire leur probabilitées d'occurence sont soit 0 soit 1 .
-Si π est une mesure invariante et $\pi(A) > 0$, le cas $Q(x, A) = 0$ est impossible .
-Pour voir cela , on suppose que $Q(x, A) = 0$, il s'ensuit que la chaine visite presque surement A seulement un nombre fini de fois et la moyenne : $\frac{1}{N} \sum_{i=1}^{n} \mathbb{1}_A(X_i)$ ne va pas converger vers $\pi(A)$, en contradiction avec la loi des grands nombres .
-Donc , $\forall x \; Q(x, A) = 1$, ce qui montre que la chaine est une chaine de Harris .

-La proposition (pour une chaine de Markov positive , si les seules fonctions harmoniques bornées sont les fonctions constantes alors : la chaine est Harris récurente) peut etre interprété comme une propriété de continuité pour la transition fonctionnelle $Kh(x) = E_x(h(X_1))$ dans le sens suivant .
-Par induction une fonction harmonique h satisfait $h(x) = E_x(h(X_n))$, et par virtue d'un théoréme précedent , $h(x)$ est presque surement égal à $E^\pi(h(X))$; donc elle est constante partout .
-Pour les chaine Harris récurrente , la proposition (pour une chaine de Markov positive , si les seules fonctions harmoniques bornées sont les fonctions constantes alors : la chaine est Harris récurente) montre que , ce si implique que $h(x)$ est constante partout .

-Pour les chaînes de Markov Harris récurrentes , les constantes sont les seules fonctions harmoniques bornées .
-Une conséquence de ça est que si (X_n) est Harris positive avec une distribution stationnaire π et si $S_n(h)$ converge μ_0- presque surement vers $\int_{\mathcal{X}} h(x)\pi(dx)$, pour une distribution initale μ_0 , cette convergence se produit pour toute distribution initiale .
-En effet , la probabilité de convergence $P_x(S_n(h) \to E^\pi(h))$ est alors harmonique .

-Encore une fois, cela montre que la récurrence d'Harris est un type de stabilité supérieure dans le sens où la convergence presque sûre est remplacée par convergence en tout point.

-On sait que si des fonctions autres que les fonctions bornées sont harmoniques, la chaîne n'est pas Harris récurrente .

1.12.3 Théoréme : théoréme ergodique

-Si (X_n) a une mesure invariante σ-finie π alors les 2 propositions suivantes sont équivalentes :

$$1)\ \text{Si} f, g \in L^1(\pi) \text{ avec } \int g(x)d\pi(x) \neq 0 \text{ alors : } \lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f(x)d\pi(x)}{\int g(x)d\pi(x)}$$

$$2)\ \text{La chaine de Markov } (X_n) \text{ est Harris récurrente}$$

-Un aspect important de ce théorème est que π n'a pas besoin d'être une mesure de probabilité et, par conséquent, qu'il peut y avoir un certain type de forte stabilité même si la chaîne est récurrente nulle.
-Dans la configuration des Algorithmes MCMC , ce résultat est parfois invoqué pour justifier l'utilisation des mesures a posteriori impropres, même si nous ne voyons pas la pertinence de ce genre d'argumentation .

1.12.4 Théorémes central limite

1.12.5 le cas discret

1.12.6 Théoréme

-Si (X_n) est Harris positive avec un atome α tel que :

$$E_\alpha(\tau_\alpha^2) < \infty \text{ et } E_\alpha((\sum_{n=1}^{\tau_\alpha} |h(X_n)|)^2) < \infty$$

$$\text{et } \gamma_h^2 = \pi(\alpha)E_\alpha((\sum_{n=1}^{\tau_\alpha} (h(X_n) - E^\pi(h)))^2) > 0$$

Alors le théoréme central limite s'applique , c'est à dire :

$$\frac{1}{\sqrt{N}}(\sum_{n=1}^{N} (h(X_n) - E^\pi(h))) \rightsquigarrow^{\mathcal{L}} \mathcal{N}(0, \gamma_h^2)$$

-Ce résultat indique qu'une extension du théorème central limite au cas non atomique sera plus délicat que pour le théorème ergodique ;
Les Conditions $E_\alpha(\tau_\alpha^2) < \infty$ et $E_\alpha((\sum_{n=1}^{\tau_\alpha} |h(X_n)|)^2) < \infty$ sont en effet exprimés en termes de chaîne fractionnée $(X_n^\vee)$.

1.12.7 Réversibilité

1.12.8 Théoréme

-Si (X_n) est aperiodique , irreductible et réversible avec distribution invariante π , le théoréme central limit s'applique quand :

$$0 < \gamma_g^2 = E_\pi(\overline{g}^2(X_0)) + 2\sum_{k=1}^{\infty} E_\pi(\overline{g}(X_0)\overline{g}(X_k)) < \infty$$

ou $\overline{g} = g - E^\pi(g)$.

-Le point principal ici est que même si la réversibilité est une hypothése très restrictif en général, il est souvent facile de l'imposer dans les Algorithmes MCMC en introduisant des étapes de simulation supplémentaires .

1.12.9 Exemple : continuation de l'exemple AR(1)

-Pour la chaine AR(1) le noyau de transition correspond à la distribution : $\mathcal{N}(\theta x_{n-1}, \sigma^2)$, et la distribution stationnaire est $\mathcal{N}(0, \sigma^2/(1-\theta^2))$.

-On peut vérifier que la chaine est reversible .
-Pour montrer que la chaine est réversible , il suffit de montrer qu'elle vérifie la condition d'équilibre ponctuel donc que :

$$K(y,x)\pi(y) = K(x,y)\pi(x), \forall (x,y)$$

-Or

$$K(y,x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\theta y)^2}{2\sigma^2}\right) \text{ et } \pi(y) = \frac{\sqrt{1-\theta^2}}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}(1-\theta^2)\right)$$

-Donc

$$\log\left(\frac{K(y,x)}{K(x,y)}\right) = -\frac{1}{2\sigma^2}((x-\theta y)^2 - ((y-\theta x)^2))$$

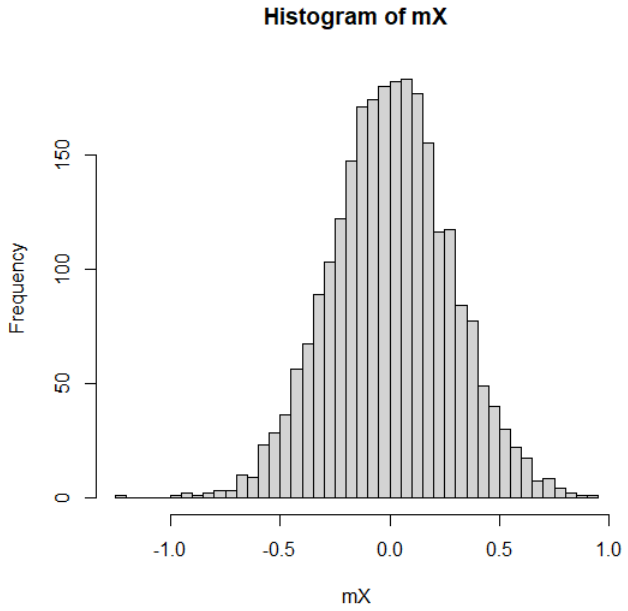

FIGURE 1.5 – histogramme des moyennes pour $\theta = 0.5$

Et

$$\log\left(\frac{\pi(x)}{\pi(y)}\right) = -\frac{1}{2\sigma^2}(x^2(1-\theta^2) - y^2(1-\theta^2))$$

-Or

$$(x-\theta y)^2 - ((y-\theta x)^2) = x^2(1-\theta^2) - y^2(1-\theta^2)$$

. -D'ou la chaine AR(1) satisfait la condition d'équilibre ponctuel donc elle est réversible

-On souhaite simuler la densité des moyennes de 2 chaine AR(1) avec $\theta = 0.5$ et $\theta = 2$.
-Voici le code R pour un histogramme de 2500 moyennes chacun basé sur 50 observations :

```
sigma=1;theta=0.5;n=50;N=2500;eps=NULL;
X=NULL;mX=NULL;
for(j in 1:N){eps=rnorm(n,0,sigma);X[1]=eps[1];
for(i in 1:(n-1)){X[i+1]=theta*X[i]+eps[i+1]};mX[j]=mean(X)}
hist(mX,nclass=50)
```

-On a ci-dessus et ci-dessous les résultats pour $\theta = 0.5$, et pour le résultat pour $\theta = 2$.
- Dans le cas $\theta = 0.5$ on a une chaine chaine positive récurrente qui satisfait les conditions du théoréme central limite .

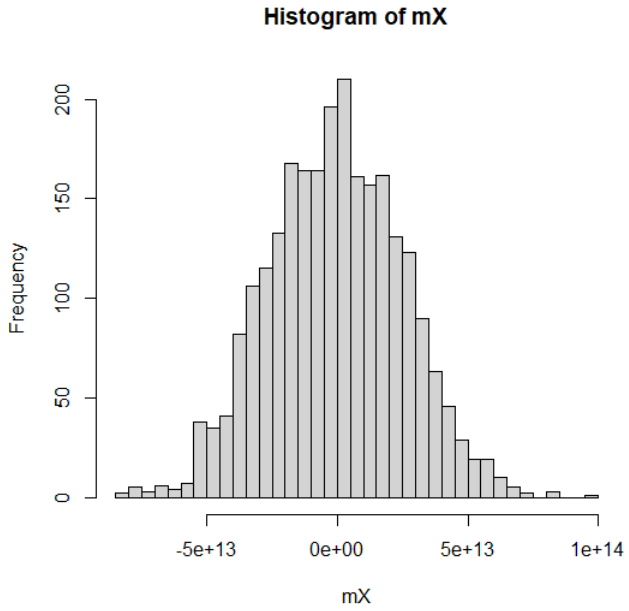

FIGURE 1.6 – histogramme des moyennes pour $\theta = 2$

-Dans le second cas $\theta = 2$, la chaine est transiente cependant, l'histogramme des moyennes "à l'aire" d'assez bien se comportés, ne donnant aucun signe que la chaîne ne converge pas.
-Il peut arriver que des chaînes nulles récurrentes et transitoires aient souvent un bon comportement en apparence lorsqu'elles sont examiné graphiquement à travers une sortie.
-Par contre en regardant les trajectoires des moyennes cumulatives et des écarts types , pour une chaine AR(1) de taille 1000 , on constate clairement la divergence pour le cas à peine transient en prenant $\theta = 1.01$, et la convergence pour le cas $\theta = 0.5$.
-voici le code R :

```
sigma=1;theta=0.5;n=1000;eps=NULL;X=NULL;
eps=rnorm(n,0,1);X[1]=eps[1];
for(i in 1:(n-1)){X[i+1]=theta*X[i]+eps[i+1]};
scX=cumsum(X);mcX=NULL;
for(i in 1:n){mcX[i]=(1/i)*scX[i]};plot(mcX,type="b");
m=mean(X);Y=(X-m)^2;scY=cumsum(Y);mcY=NULL;
for(i in 1:n){mcY[i]=(1/i)*scY[i]}
windows();mcY=sqrt(mcY);plot(mcY,type="b")
```

-Nous avons les résultats ci-dessous pour $\theta = 0.5$ et $\theta = 1.01$

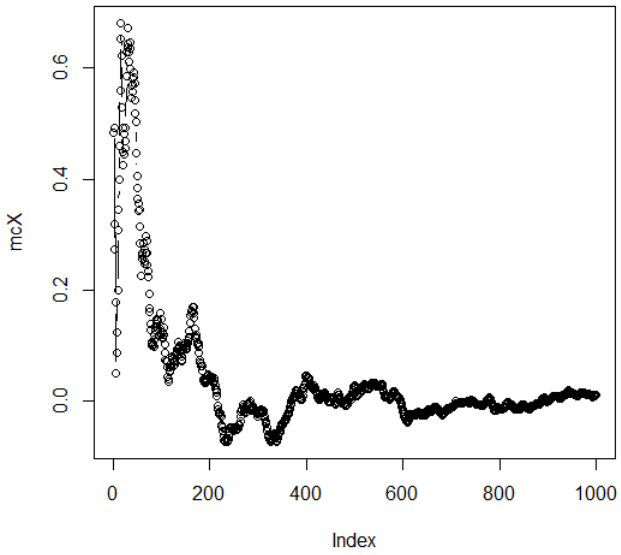

FIGURE 1.7 – trajectoire des moyennes cumulées pour $\theta = 0.5$

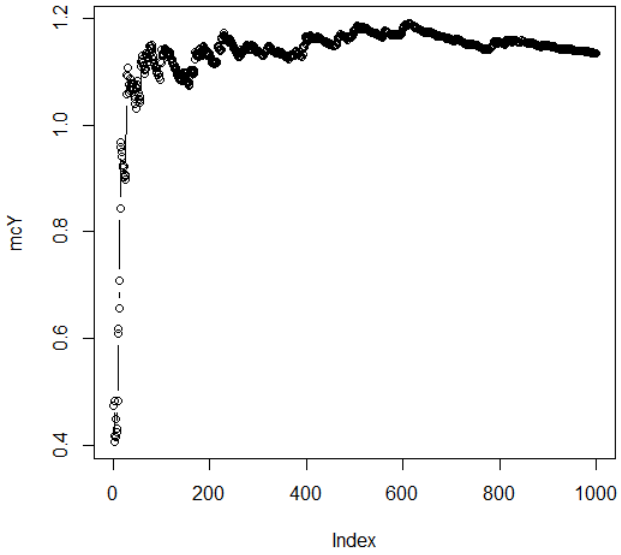

FIGURE 1.8 – trajectoire des écart-types cumulées pour $\theta = 0.5$

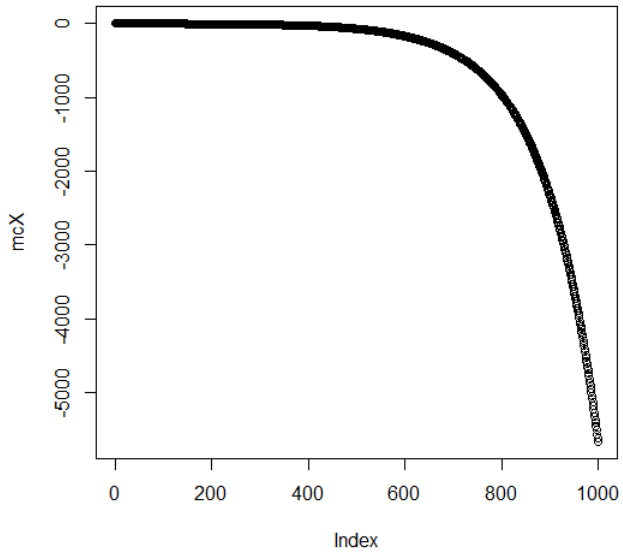

FIGURE 1.9 – trajectoire des moyennes cumulées pour $\theta = 1.01$

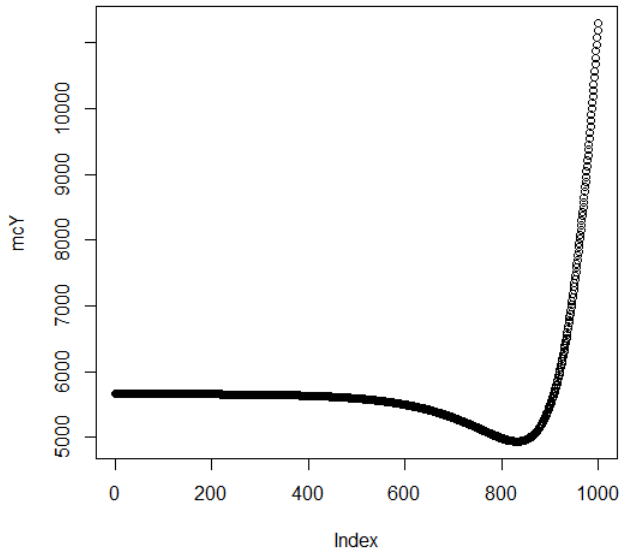

FIGURE 1.10 – trajectoire des écart-types cumulées pour $\theta = 1.01$

1.12.10 Ergodicité géométrique et régénération

-Il existe encore une autre approche du théorème central limite pour les Chaînes de Markov .
-Elle repose sur l'ergodicité géométrique, une condition de moment de type Liapounov sur la fonction h, et un argument de régénération.

1.12.11 Théoréme

-Si (X_n) est apériodique, irréductible, Harris récurrente positif , de loi invariante π et géométriquement ergodique, et si, de plus,

$$E^{\pi}(|h(X)|^{2+\epsilon}) < \infty \text{ pour un } \epsilon > 0$$

Alors :

$$\sqrt{n}(\frac{S_n(h)}{n} - E^{\pi}(h(X))) \rightsquigarrow^{\mathcal{L}} \mathcal{N}(0, \gamma_h^2)$$

-Nous avons besoin de conditions spécifiques pour que ce théoréme s'applique , à savoir une condition de Liapounov et une estimation consistante de γ_h^2 .
-Trouver de telles estimations est difficile, car les approximations moyennes du lot fixe ne sont pas cohérentes lorsque la taille du lot est fixe.
-Nous pouvons, cependant, utilisez la régénération lorsqu'elle est disponible ; c'est quand une condition de minorisation est vraie :
il existe une fonction $0 < s(x) < 1$ et une mesure de probabilité Q telle que,
pour tout $x \in \mathcal{X}$, pour tout ensemble mesurable A ,

$$P(x, A) \geq s(x)Q(A)$$

-nous pouvons alors construire des erreurs asymptotiques légitimes en contournant l'estimation de γ_g^2 .
-La démarche consiste à introduire les temps de régénération $0 = \tau_0 < \tau_1 < \tau_2 < \ldots$ associée à la chaîne de Markov (X_t) et d'écrire $S_n(h)$ en fonction du temps de régénération,
à savoir, si la chaîne est démarrée en tant que $X_0 \sim Q$ et arrêtée après la T-ème régénération,

$$S_{\tau_T}(h) = \sum_{t=1}^{T} \sum_{j=\tau_{t-1}}^{\tau_t - 1} h(X_j) = \sum_{t=1}^{T} \tilde{S}_t$$

ou les $\tilde{S}_t$ sont des sommes partielles qui sont i.i.d .

-Si l'on définit les longueurs inter-régénération $N_t = \tau_t - \tau_{t-1}$ alors :

$$\overline{h}_{\tau_t} = \frac{\sum_{t=1}^{T} \tilde{S}_t}{\sum_{t=1}^{T} N_t} = \frac{\overline{S}_t}{\overline{N}_t} = \frac{1}{\tau_T} \sum_{j=0}^{\tau_T - 1} g(X_j)$$

converge presque surement vers $E^{\pi}(h(X))$ quand T tend vers l'infini , par virtue du théoréme érgodique , vu que τ_T converge presque surement vers l'infini .

-Par un théoréme précedent $E^{Q}(N_1) = \frac{1}{E^{\pi}(s(X))}$ qui est supposée fini .
-Il s'ensuit par la loi forte des grands nombres que $\overline{N}$ converge presque surement vers $E^{Q}(N_1)$, et vu que $\overline{h} = \tau_t \frac{\overline{S}_t}{\overline{N}_t}$,on a $\overline{S}_t$ qui converge presque surement vers $E^{Q}(N_1)E^{\pi}(h(X))$.
-Ceci implique en particulier que $E^{Q}(|\tilde{S}_1|) < \infty$ et $E^{Q}(\tilde{S}_1) = E^{Q}(N_1)E^{\pi}(h(X))$.
-Donc les variables aléatoires $\tilde{S}_t - N_t E^{\pi}(h(X))$ sont i.i.d et centrées .

1.12.12 Théoréme

-Si $E^{Q}(\tilde{S}_1^2)$ et $E^{Q}(N_1^2)$ sont tous les 2 finis , le théoréme central limite s'applique :

$$\sqrt{R}(\overline{h}_{\tau_R} - E^{\pi}(h(X))) \rightsquigarrow^{\mathcal{L}} \mathcal{N}(0, \sigma_h^2)$$

$$\text{ou } \sigma_h^2 = \frac{E^{Q}((\tilde{S}_1 - N_1 E^{\pi}(h(X)))^2)}{(E^{Q}(N_1))^2}$$

-L'avantage en utilisant cette approche est que σ_h^2 peut être estimé beaucoup plus facilement en raison de la structure indépendante sous-jacente.
-Par exemple,

$$\hat{\sigma}_h^2 = \frac{\sum_{t=1}^{R} (\tilde{S}_t - \overline{h}_{\tau_R} N_t)}{R\overline{N}^2}$$

est une estimation consistante de σ_h^2 .
-De plus, les conditions $E^{Q}(\tilde{S}_1^2)$ et $E^{Q}(N_1^2)$ apparaissant dans ce théorème sont minimes en le sens qu'elles sont vérifiés lorsque les conditions du théorème précédent sont vérifiées .

2.1 Le principe MCMC

-Il n'est pas nécessaire de simuler directement un échantillon de la distribution f pour approximer l'intégrale :

$$\mathcal{I} = \int h(x)f(x)dx$$

puisque d'autres approches comme l'échantillonnage préférentielle peuvent être utilisées.
-Il est possible d'obtenir un échantillon $X_1, ..., X_n$ distribués approximativement suivant f sans simuler directement à partir de f en utilisant une chaîne de Markov ergodique avec comme distribution stationnaire f .
-Le principe des algorithmes MCMC est donc le suivant :
Pour une valeur de démarrage arbitraire $x^{(0)}$, une chaîne $(X^{(t)})$ est générée à l'aide d'un noyau de transition avec pour distribution stationnaire f , qui assure la convergence en distribution de $(X^{(t)})$ vers une variable aléatoire de f .
(Étant donné que la chaîne est ergodique, la valeur de départ $x^{(0)}$ est, en principe, sans importance.)

2.2 Monte Carlo par Chaines de Markov

-Une méthode de Monte Carlo par chaîne de Markov (MCMC) pour la simulation d'une distribution f est toute méthode produisant une chaîne de Markov ergodique $(X^{(t)})$ dont la distribution stationnaire est f.
-Ici ont reposent sur des propriétés de convergence asymptotique plus complexes qu'une simple loi de grands nombres, car nous générons des dépendances au sein de l'échantillon qui ralentissent la convergence de l'approximation de $\mathcal{I}$.
-Ainsi, le nombre d'itérations nécessaire pour obtenir une bonne approximation semble a priori beaucoup plus important qu'avec une méthode de Monte Carlo standard.
-Il y a de nombreux cas où un algorithme MCMC spécifique domine, en termes de variance, la proposition de Monte Carlo correspondante.
-Par exemple, alors que l'échantillonnage préférentielle est une méthode quasi universelle, son efficacité repose sur des choix adéquats de la fonction d'importance et ce choix devient de plus en plus difficile à mesure que la dimension augmente, il s'agit une réalisation pratique de la malédiction de la dimensionnalité.
- A premier niveau, certains algorithmes génériques, comme les algorithmes de Metropolis-Hastings, utilise également des simulations à partir de presque n'importe quelle densité

arbitraire g pour générer réellement à partir d'une densité donnée également arbitraire f .
-Mais à un deuxième niveau, puisque ces algorithmes tiennent compte de la dépendance de g à la simulation précédente, le choix de g ne demande pas a priori une construction particulièrement élaborée mais peut tirer parti des caractéristiques locales de la distribution stationnaire.
-De plus, même lorsqu'un algorithme d'acceptation-Reject est disponible, il est parfois plus efficace d'utiliser la paire (f, g) à travers une chaîne de Markov .
-Il faut souligner que la (re)découverte des méthodes de Monte Carlo par chaînes de Markov par les statisticiens dans les années 1990 a produit des progrès considérables dans l'inférence basée sur la simulation et, en particulier, dans l'inférence bayésienne, puisqu'elle a permis l'analyse d'une multitude de modèles trop complexes pour être traités de manière satisfaisante par schémas précédents.

2.3 Méthodes de Monte Carlo par Chaine de Markov

-L'utilisation d'une chaîne $(X^{(t)})$ produite par un algorithme de Monte Carlo par chaîne de Markov avec distribution stationnaire f est fondamentalement identique à l'utilisation d'un échantillon iid au sens où le théorème ergodique garantit la convergence (presque sûre) de la moyenne empirique :

$$\mathcal{I}_T = \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$$

vers la quantitée $E_f(h(X))$.
-Une suite $(X^{(t)})$ produite par algorithme de Monte Carlo par chaîne de Markov peut donc être utilisé comme un échantillon iid.
- Une seule réalisation (ou chemin) d'une chaîne de Markov est suffisante pour assurer une bonne approximation de $\mathcal{I}$ par estimations comme $\mathcal{I}_T = \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$ pour les fonctions d'intérêt h.
-Évidemment, la gestion de cette séquence est un peu plus ardu que dans le cas iid en raison de la structure de dépendance .

-On peut proposer une infinité d'implémentations pratiques comme celles, par exemple, utilisées dans la physique.
Les algorithmes de Metropolis-Hastings ont l'avantage d'imposer des exigences minimales sur la densité cible f et permettent un large choix d'implémentations possibles.
-En revanche, l'échantillonneur de Gibbs est plus restrictif, en ce sens qu'il nécessite une certaine connaissance de la densité cible pour dériver certaines densités conditionnelles,

mais il peut aussi être plus efficace qu'un algorithme Metropolis-Hastings générique .

2.4 L'algorithme de Metropolis-Hastings

-L'algorithme de Metropolis-Hastings commence par la densité objective (cible) f ,
Une densité conditionnelle $q(y|x)$, définie par rapport à la mesure dominante pour le modèle, est alors choisi .
-L'algorithme de Metropolis-Hastings peut être implémenté en pratique lorsque $q(.|x)$ est facile à simuler et est soit explicitement disponible (jusqu'à une constante multiplicative indépendante de x) ou symétrique ; c'est-à-dire tel que $q(x|y) = q(y|x)$.
-La densité cible f doit être disponible dans une certaine mesure : une exigence générale est que le ratio

$$f(y)/q(y|x)$$

est connue à une constante près indépendante de x.
-L'algorithme Metropolis-Hastings associé à la densité objectif (cible) f et la densité conditionnelle q produit une chaîne de Markov $(X^{(t)})$ par la transition suivante.

2.4.1 Algorithme : Metropolis-Hastings

Etant donné $x^{(t)}$
1. Générer $Y_t \sim q(y|x^{(t)})$
2. Prendre

$$X^{(t+1)} = \begin{cases} Y_t \text{ avec probabilité } \rho(x^{(t)}, Y_t) \\ x^{(t)} \text{ avec probabilité } 1 - \rho(x^{(t)}, Y_t) \end{cases}$$

$$\text{avec } \rho(x, y) = min\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\}$$

-La distribution q est appelée distribution instrumentale (ou de proposition) et la probabilité $\rho(x, y)$ est la probabilité d'acceptation de Metropolis-Hastings .
-Cet algorithme accepte toujours des valeurs y_t telles que le rapport $f(y_t)/q(y_t|x^{(t)})$ est augmentée, par rapport à la valeur précédente $f(x^{(t)})/(q(x^{(t)}|y_t)$.
-C'est seulement dans le cas symétrique où l'acceptation est déterminée par le rapport objectif $f(y_t)/f(x^{(t)})$.
-Une caractéristique importante de l'algorithme de Metropolis-Hastings est qu'il peut accepter des valeurs y_t telles que le rapport soit diminué, similaire aux méthodes d'optimisation stochastique .

-Comme la méthode d'Acceptation-Rejet, L'algorithme de Metropolis-Hastings ne dépend que des rapports

$$f(y_t)/f(x^{(t)}) \text{ et } q(x^{(t)}|y_t)/q(y_t|x^{(t)})$$

et est donc indépendant des constantes de normalisation, en supposant, encore une fois, que $q(.|x)$ est connu à une constante près indépendante de x .

-Évidemment, la probabilité $\rho(x^{(t)}, y_t)$ n'est définie que lorsque $f(x^{(t)}) > 0$.
-Cependant, si la chaîne commence par une valeur $x^{(0)}$ telle que $f(x^{(0)}) > 0$, il s'ensuit que $f(x^{(t)}) > 0$ pour tout $t \in N$ puisque les valeurs de y_t telles que $f(y_t) = 0$ conduisent à $\rho(x^{(t)}, y_t) = 0$ et sont donc rejetées par l'algorithme.
-Nous ferons la convention selon laquelle le rapport $\rho(x, y)$ est égal à 0 lorsque $f(x)$ et $f(y)$ sont nulles, afin d'éviter des difficultés théoriques.

-Il existe des similitudes entre l'algorithme de Metropolis-Hastings et les méthodes d'Acceptation-Rejet , et il est possible d'utiliser l'algorithme de Metropolis-Hastings comme alternative à un algorithme d'Acceptation-Reject pour une paire donnée (f, g).
-Cependant, un échantillon produit par Metropolis-Hastings diffère d'un échantillon iid pour une raison
, un tel échantillon peut impliquer des occurrences répétées de même valeur, puisque le rejet de y_t entraîne la répétition de $X^{(t)}$ à l'instant $t + 1$ (une occurrence impossible dans des paramètres iid absolument continus).
-Alors que l'Algorithme de Metropolis-Hastings est un algorithme générique, défini pour tous les f et q , il est néanmoins nécessaire d'imposer des conditions minimales de régularité à la fois sur f et sur la distribution conditionnelle q pour que f soit la distribution limite de la chaîne $(X^{(t)})$ produite par l'Algorithme de Metropolis-Hastings.
-Par exemple, il est plus facile si ξ, le support de f, est connexe ;
un support ξ qui n'est pas connexe peut invalider l'algorithme de Metropolis-Hastings .
-Pour de tels supports, il est nécessaire de procéder sur une composante connexe à la fois et montrer que les différentes composantes connexes de ξ sont liés par le noyau de l'Algorithme de Metropolis-Hastings.
-Si le support ξ est tronqué par q , c'est à dire , s'il existe $A \subset \xi$ tel que :

$$\int_A f(x)dx > 0 \text{ et } \int_A q(y|x)dy = 0 \ \forall x \in \xi$$

l'Algorithme de Metropolis-Hastings n'a pas f comme distribution limite ,
vu que pour $x^{(0)} \notin A$, la chaine $X^{(t)}$ ne visitera jamais A .

-Ainsi, une condition nécessaire minimale est que :

$$\bigcup_{x \in supp\ f} supp\ q(.|x) \supset supp\ f$$

-Pour voir que f est la distribution stationnaire de la chaîne Metropolis-Hastings , nous examinons d'abord de plus près le noyau de Metropolis-Hastings pour constater qu'il satisfait les propriété d'équilibre ponctuel .

2.4.2 Théoréme

-Soit $X^{(t)}$ une chaine produite par l'Algorithme de Metropolis-Hastings ;
Pour toute distribution q dont le support inclut ξ , on a :
i) Le noyau de la chaine satisfait la condition d'équilibre ponctuel avec f .
ii) f est la distribution stationnaire de la chaine .

-La stationnarité de f est donc établie pour presque toute distribution conditionnelle q, un fait qui indique l'universalité de l'algorithme de Metropolis-Hastings .

2.5 Propriétés de convergence

-Pour montrer que la chaîne de Markov de l'algorithme de Metropolis-Hastings converge bien vers la distribution stationnaire et que $\mathcal{I}_T = \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$ est une approximation convergente de $\mathcal{I}$, nous devons appliquer davantage la théorie des chaines de Markov .
-Comme la chaîne de Markov de l'algorithme de Metropolis-Hastings a, par construction, une distribution de probabilité invariante f,
Si c'est aussi une chaîne de Harris apériodique , alors le théorème ergodique s'applique pour établir un résultat comme la convergence de $\mathcal{I}_T = \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$ vers $\mathcal{I}$.

-Une condition suffisante pour que la chaîne de Markov de l'algorithme de Metropolis-Hastings soit apériodique est que l'algorithme autorise des événements tels que $\{X^{(t+1)} = X^{(t)}\}$;
C'est-à-dire que la probabilité de tels événements n'est pas nulle, et donc :

$$P(\ f(X^{(t)})q(Y_t|X^{(t)}) \leq f(Y_t)q(X^{(t)}|Y_t)\ ) < 1$$

-Le fait que l'algorithme de Metropolis-Hastings ne fonctionne que lorsque
$P(\ f(X^{(t)})q(Y_t|X^{(t)}) \leq f(Y_t)q(X^{(t)}|Y_t)\ ) < 1$ est satisfait n'est pas trop gênant, puisqu'il indique simplement qu'il est inutile de perturber davantage une chaîne de Markov

de noyau de transition q si cette derniere converge déjà vers la distribution f.
-Il suffit alors d'étudier directement la chaîne associée à q .
-La propriété d'irréductibilité de la chaîne de Metropolis-Hastings $(X^{(t)})$ découle de conditions suffisantes telles que la positivité de la densité conditionnelle q ; c'est à dire :

$$q(y|x) > 0 \ , \forall (x, y) \in \xi \times \xi$$

Puisqu'il s'ensuit alors que tout ensemble de ξ de mesure de Lebesgue positive peut être atteint en une seule étape.
-Comme la densité f est la mesure invariante de la chaîne, la chaîne est positive et cela implique que la chaîne est récurrente.

-On peut aussi établir le résultat suivant plus fort pour la chaîne de Metropolis-Hastings.

-si la chaine de Metropolis-Hastings $(X^{(t)})$ est f-irréductible alors elle est Harris récurrente .

2.5.1 Théoréme

-Supposons que la chaine de Metropolis-Hastings $(X^{(t)})$ est f-irréductible alors :
i) Si $h \in L^1(f)$ alors :

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h(X^{(t)}) = \int h(x) f(x) dx \ f\text{-presque surement}$$

ii) Si en plus $(X^{(t)})$ est apériodique alors :

$$\lim_{n\to\infty} || \int K^n(x, .)\mu(dx) - f||_{VT} = 0$$

pour toute distribution initiale μ .

-les conclusions du théoréme précedent restent valables si la chaine de Markov de Metropolis-Hastings $(X^{(t)})$ a une densité conditionnelle $q(x|y)$ satisfaisant :

$$P(\ f(X^{(t)})q(Y_t|X^{(t)}) \leq f(Y_t)q(X^{(t)}|Y_t)\ ) < 1 \text{ et } q(y|x) > 0, \forall (x, y) \in \xi \times \xi$$

-Bien que la condition $q(y|x) > 0, \forall (x, y) \in \xi \times \xi$ puisse sembler restrictive, elle est souvent satisfaite en pratique.

- Supposons que f est bornée et positive sur tout ensemble compact de son support ξ ;

S'il existe des nombres positifs ϵ et δ tels que :

$$q(y|x) > \epsilon \text{ si } |x - y| < \delta$$

Alors :
La chaîne de Markov de Metropolis-Hastings $(X^{(t)})$ est f-irréductible et apériodique. De plus, tout ensemble compact non vide est un petit ensemble.
-La justification de ce résultat est la suivante :
Si la distribution conditionnelle $q(y|x)$ permet des déplacements dans un voisinage de $x^{(t)}$ de diamètre borné par le bas et si f est tel que $\rho(x^{(t)}, y)$ est positif dans ce voisinage, alors tout sous-ensemble de ξ peut être visité en k étapes pour k assez grand.
(Cette propriété repose évidemment sur l'hypothèse que ξ est connexe.)

-Les conclusions du théoréme précedent restent valables si la chaine de Markov de Metropolis-Hastings $(X^{(t)})$ a une densité de probabilité invariante f et une densité conditionnelle $q(y|x)$ satisfaisant : $\exists \epsilon, \delta$ tels que $q(y|x) > \epsilon$ si $|x - y| < \delta$.
-L'un des aspects les plus fascinants de l'algorithme de Metropolis-Hastings est son universalité ;
C'est-à-dire le fait qu'une distribution conditionnelle arbitraire q avec support ξ peut conduire à la simulation d'une distribution arbitraire f sur ξ.
-D'autre part, cette universalité ne peut tenir formellement que si la distribution instrumentale q simule rarement des points dans la partie principale de ξ ; c'est-à-dire dans la région où se trouve la majeure partie de la masse de la densité f.

2.6 L'algorithme de Metropolis-Hastings indépendant :

2.6.1 Propositions fixes :

-Cette méthode apparaît comme une simple généralisation de la méthode d"acceptation-Reject en ce sens que la distribution instrumentale q est indépendante de $X^{(t)}$ et est noté g par analogie.

2.6.2 Algorithme : Metropolis-Hastings indépendent :

- Etant donné $x^{(t)}$:
1. Génerer $Y_t \sim g(y)$.
2. Prendre $X^{(t+1)} = \begin{cases} Y_t \text{ avec probabilité } min\{\frac{f(Y_t)}{f(x^{(t)})}\frac{g(x^{(t)})}{g(Y_t)}, 1\} \\ x^{(t)} \text{ autrement} \end{cases}$

- Bien que les y_t soient générés indépendamment, l'échantillon résultant n'est pas iid ; par exemple, la probabilité d'acceptation de Y_t dépend de $X^{(t)}$ (sauf dans le cas trivial où $f = g$) .
-Les propriétés de convergence de la chaîne $(X^{(t)})$ découlent des propriétés de la densité g au sens où $(X^{(t)})$ est irréductible et apériodique (donc ergodique) si et seulement si g est presque partout positif sur le support de f .

2.7 Théoréme

- L'algorithme de Metropolis-Hastings indépendent produit une chaine uniformément ergodique si :
Il existe une constante M tel que :

$$f(x) \leq Mg(x) \ , \forall x \in supp\ f$$

Dans ce cas :

$$\|K^n(x,.) - f\|_{TV} \leq 2(1 - \frac{1}{M})^n$$

-D'autre part, si pour tout M, il existe un ensemble de mesure positive où
'$f(x) \leq Mg(x) \ , \forall x \in supp\ f$' ne tient pas, $(X^{(t)})$ n'est même pas géométriquement ergodique.

-Cette classe particulière d'algorithmes de Metropolis-Hastings suggère naturellement une comparaison avec les méthodes d'Acceptation-Reject puisque chaque paire (f, g) qui satisfait $f(x) \leq Mg(x) \ , \forall x \in supp\ f$ peut aussi induire un algorithme d'Acceptation-Reject.

-si $f(x) \leq Mg(x) \ , \forall x \in supp\ f$:
la probabilité d'acceptation espéré avec l'algorithme de Metropolis-Hastings indépendent est au moins $\frac{1}{M}$ quand la chaine est stationnaire .

-Ainsi, l'algorithme de Metropolis-Hastings indépendant est plus efficace que l'algorithme d'Acceptation-Reject dans son traitement de l'échantillon produit par g, puisque, en moyenne, il accepte plus de valeurs proposées.

2.7.1 Exemple : génération de variables gamma

- on souhaite comparer l'algorithme de Metropolis-Hastings et d'acceptation-rejet pour génerer une loi cible $\mathcal{G}a(\alpha,\beta)$ à partir d'une loi candidate $\mathcal{G}a([\alpha],b)$. , avec $[\alpha]$: partie entiére de α .
- on a $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}\exp(-\beta x)x^{\alpha-1}$, pour $x \geq 0$.
- de meme $q(y|x) = g(y) = \frac{b^{[\alpha]}}{\Gamma([\alpha)]}\exp(-by)y^{[\alpha]-1}$, pour $y \geq 0$.
- on a : $\rho(x,y) = min\{\exp(-(\beta-b)(y-x))(\frac{y}{x})^{\alpha-[\alpha]},1\}$.
- pour simplifier on prendra $\beta = 1$ et $b = [\alpha]/\alpha$.
- on a $\frac{f(x)}{g(x)} = \frac{1}{\Gamma(\alpha)}\frac{\Gamma([\alpha])}{b^{[\alpha]}}\exp((b-1)x)x^{\alpha-[\alpha]}$;
donc : $\frac{f(x)}{g(x)} = \frac{\Gamma([\alpha])}{\Gamma(\alpha)}(\frac{\alpha}{[\alpha]})^{[\alpha]}e^{(\frac{[\alpha]}{\alpha}-1)x}x^{\alpha-[\alpha]}$;
ce qui donne : $\frac{f(x)}{g(x)} = \frac{\Gamma([\alpha])}{\Gamma(\alpha)}(\frac{\alpha}{[\alpha]})^{[\alpha]}e^{-\frac{1}{\alpha}(\alpha-[\alpha])x}x^{\alpha-[\alpha]}$.
- soit $h(x) = e^{-\frac{1}{\alpha}(\alpha-[\alpha])x}x^{\alpha-[\alpha]}$.
-on a $h'(x) = (\alpha-[\alpha])e^{-\frac{1}{\alpha}(\alpha-[\alpha])}x^{\alpha-[\alpha]-1}(-\frac{x}{\alpha}+1)$.
-donc clairement $h(x) \leq h(\alpha) = e^{-(\alpha-[\alpha])}\alpha^{\alpha-[\alpha]}$.
-d'ou : $\frac{f(x)}{g(x)} \leq \frac{\Gamma([\alpha])}{\Gamma(\alpha)}(\frac{\alpha}{[\alpha]})^{[\alpha]}e^{-(\alpha-[\alpha])}\alpha^{\alpha-[\alpha]}$.
-donc on prend $M = \frac{\Gamma([\alpha])}{\Gamma(\alpha)}(\frac{\alpha}{[\alpha]})^{[\alpha]}e^{-(\alpha-[\alpha])}\alpha^{\alpha-[\alpha]}$.
-en remplacant $\beta = 1$ et $b = [\alpha]/\alpha$,
on obtient : $\rho(x,y) = min\{\exp(-(1-[\alpha]/\alpha)(y-x)(\frac{y}{x})^{\alpha-[\alpha]}),1\}$.
-voici le code R pour la l'algorithme acceptation-rejet :

```
alpha=2.43;alphae=floor(alpha);
M=gamma(alphae)/gamma(alpha)*(alpha/alphae)^alphae*
exp(-(alpha-alphae))*alpha^(alpha-alphae)
f=function(x){return(1/gamma(alpha)*exp(-x)*x^(alpha-1))}
g=function(x){return(
1/gamma(alphae)*(alphae/alpha)^alphae*exp(-alphae*x/alpha)*
x^(alphae-1))};
n=10^4;X=NULL;Y=NULL;
for(i in 1:n){Y[i]=rgamma(1,alphae,alphae/alpha)
U=runif(1)
while(U>f(Y[i])/(M*g(Y[i]))){Y[i]=rgamma(1,alphae,alphae/alpha)};
X[i]=Y[i]};hist(X)
```

-Voici le code R pour l'algorithme de Metropolis-Hastings :

```
alpha=2.43;alphae=floor(alpha);
M=gamma(alphae)/gamma(alpha)*(alpha/alphae)^alphae*
exp(-(alpha-alphae))*alpha^(alpha-alphae)
```

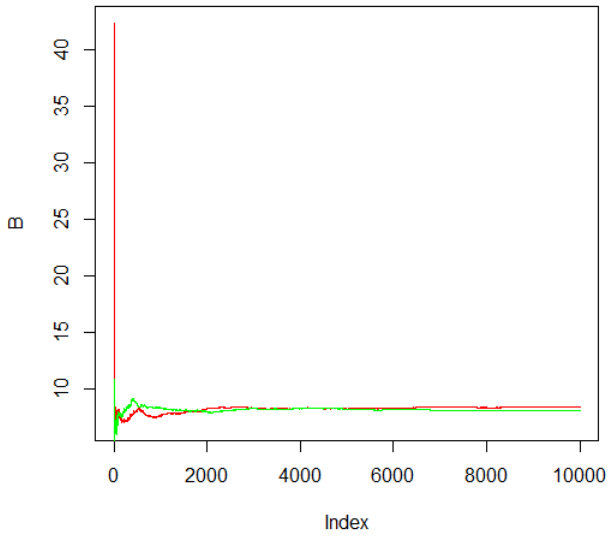

FIGURE 2.1 – convergence de l'estimateur de $E[X^2]$

```
f=function(x){return(1/gamma(alpha)*exp(-x)*x^(alpha-1))}
g=function(x){
return(1/gamma(alphae)*(alphae/alpha)^alphae*exp(-alphae*x/alpha)*
x^(alphae-1))};n=10^4;rho=function(x,y){return(
min(1,exp(-(1-alphae/alpha)*(y-x))*(y/x)^(alpha-alphae)))}
Ym=NULL;Xm=NULL;Xm[1]=rgamma(1,alphae,alphae/alpha)
for(i in 1:(n-1)){U=runif(1);
Ym[i]=rgamma(1,alphae,alphae/alpha)
Xm[i+1]=(Ym[i]-Xm[i])*as.numeric(U<=rho(Xm[i],Ym[i]))+Xm[i]};
hist(Xm)
```

-Nous avons ci-dessu le résulatat de la convergence de l'estimateur de $E[X^2]$ pour la méthode acceptation-rejet en vert et la méthode mcmc en rouge .
- le nombre t de valeur acceptées par la méthode d'acceptation-rejet est aléatoire pour un échantillon de proposition de taille fixe n $(y_1, ..., y_n)$, D'autre part, le nombre t de valeur acceptées par la méthode de Metropolis-Hastings peut etre fixé .

-Notez cependant que les deux comparaisons sont biaisées.
Dans le premier cas, l'échantillon de $X^{(i)}$ produit par la méthode acceptation-rejet n'a pas la distribution f et,
Dans le second cas, l'échantillon des Y_i dans la méthode de Metropolis-Hastings n'est pas iid.

Dans les deux cas, cela est dû à l'utilisation d'une règle d'arrêt qui modifie la répartition des échantillons.

- Notez que la classification habituelle de "Hastings" pour l'algorithme de Metropolis-Hastings Indépendent est quelque peu inapproprié, puisque Hastings (1970) considère l'algorithme de Metropolis-Hastings en général, en utilisant des marches aléatoires plutôt que des distributions indépendantes dans ses exemples.
-Il est également intéressant de rappeler que Hastings (1970) propose une justification théorique de ces méthodes pour des chaînes de Markov d'espace d'états fini basées sur la représentation finie des nombres réels dans un ordinateur.
-Cependant, une justification complète de cette discrétisation physique doit tenir compte de l'effet de l'approximation dans l'ensemble de l'analyse.
-En particulier, il faut vérifier que le choix informatique de l'approximation discrète de la distribution continue n'a aucun effet sur la distribution stationnaire résultante ou l'irréductibilité de la chaîne.
-Vu que Hastings (1970) n'entre pas dans ces détails, mais s'en tient au niveau de la simulation, ont préfèrent étudier les propriétés théoriques de ces algorithmes en contournant la représentation finie des nombres dans un ordinateur et en supposant sans défaut des générateurs pseudo-aléatoires, à savoir des algorithmes produisant des variables qui sont uniformément distribué sur [0, 1].

-Une dernière remarque sur les algorithmes de Metropolis-Hastings indépendants est qu'ils ne peuvent pas être omniscients ;
Il existe des contextes où une proposition indépendante ne fonctionne pas bien en raison de la complexité de la distribution cible.
-Vu que le but principal des algorithmes MCMC est de fournir une technique de simulation grossière mais facile , il est difficile d'imaginer passer beaucoup de temps sur la conception de la distribution de la proposition.
-Ceci est particulièrement pertinent dans les modèles de grandes dimensions où la capture des principales caractéristiques de la distribution cible est impossible le plus souvent.
-Il y a donc une limitation de la proposition indépendante, qui peut être perçu comme une proposition globale, et il y a une nécessité d'utiliser d'avantage de propositions locales moins sensibles à la distribution cible .
-Une autre possibilité est de valider des algorithmes adaptatifs qui apprennent des performances en cours des propositions actuels pour affiner leur construction. Mais cette solution est délicate, tant d'un point de vue théorique (« L'ergodicité s'applique-t-elle ? ») et algorithmique (« Comment accorde-t-on l'adaptation ? »).

2.8 Marches aléatoire

-Une approche naturelle pour la construction pratique d'un algorithme de Metropolis-Hastings est de prendre en compte la valeur précédemment simulée pour générer la valeur suivante ; c'est-à-dire d'envisager une exploration locale du voisinage de la valeur courante de la chaîne de Markov.
-Puisque le candidat g dans l'algorithme de Metropolis-Hastings est autorisé à dépendre de l'état courant $X^{(t)}$, un premier choix à considérer est de simuler Y_t selon :

$$Y_t = X^{(t)} + \epsilon_t$$

où ϵ_t est une perturbation aléatoire de distribution g, indépendante de $X^{(t)}$.
-en termes de l'algorithme de Metropolis-Hastings, $q(y|x)$ est maintenant de la forme $g(y-x)$.
-La chaîne de Markov associée à q est une marche aléatoire sur ξ .

-Les résultats de convergence précedents s'appliquent naturellement dans ce cas particulier .
-Si g est positif au voisinage de 0, la chaîne $(X^{(t)})$ est f-irréductible et apériodique, donc ergodique.
- Les distributions g Les plus communes dans cette configuration sont les distributions uniformes sur les sphères centrées à l'origine ou des distributions standard comme la distribution normale et celle de Student .
-Toutes ces distributions doivent généralement être mises à l'échelle .
-Notons que le choix d'une fonction symétrique g (c'est-à-dire telle que $g(-t) = g(t)$), conduit à l'expression original suivante de l'algorithme de Metropolis-Hastings , comme proposé par Metropolis et al (1953).

2.8.1 Algorithme : Metropolis-Hastings par marche aléatoire .

-Etant donné $x^{(t)}$,
1. Générer $Y_t \sim g(|y - x^{(t)}|)$.
2. Prendre $X^{(t+1)} = \begin{cases} Y_t \text{ avec probabilité } min\{1, \frac{f(Y_t)}{f(x^{(t)})}\} \\ x^{(t)} \text{ autrement} \end{cases}$

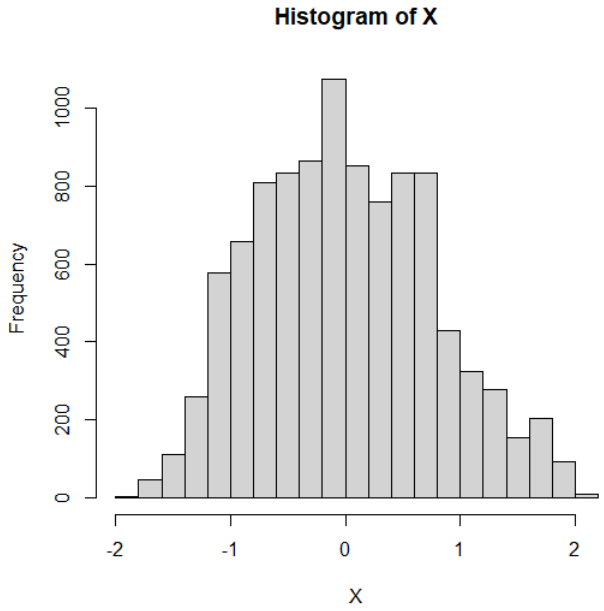

FIGURE 2.2 – densité $\mathcal{N}(0,1)$ à partir de $\mathcal{U}_{[-\delta,\delta]}$ pour $\delta = 0.1$

2.8.2 Exemple : Lois normales à partir de lois uniformes :

-L'exemple historique de Hastings (1970) considère le problème formel de génération d'une loi normale $\mathcal{N}(0,1)$ fondée sur une marche aléatoire comme proposition, égale à la loi uniforme sur $[-\delta,\delta]$.
-on a $X \sim \mathcal{N}(0,1)$ donc $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}$, pour $x \in R$.
-et $q(y|x) = g(y-x) = \frac{1}{2\delta}$, pour $y - x \in [-\delta,\delta]$.
-La probabilité d'acceptation est alors $\rho(x^{(t)}, y_t) = min\{\exp -\frac{1}{2}(y^2 - x^2), 1\}$.
-Voici le code R correspondant :

```
rho=function(x,y){return(min(1,exp(-0.5*(y^2-x^2))))};
delta=1;X=NULL;Y=NULL ;X[1]=runif(1,-delta,delta);n=10000 ;
for(t in 1:(n-1)){Y[t]=runif(1,-delta,delta)+X[t];
X[t+1]=sample(c(Y[t],X[t]),size=1,prob=c(rho(X[t],Y[t]),
1-rho(X[t],Y[t])))} ;hist(X);acf(X);A=cumsum(X)/1:n;
plot(A,type="s")
```

-Nous obtenons différents résultats d'histogrammes , de fonctions d'autocovariances et de convergence de la moyenne arithmétique pour $\delta = 0.1, 1, 10$
-Une proposition trop serrée ou trop large (c'est-à-dire une valeur trop petite ou trop grande de δ) a pour résultat une autocovariance plus grande et une convergence plus lente.
-Pour $\delta = 0.1$ la chaîne de Markov se déplace à chaque itération, mais très lentement, tandis que pour $\delta = 10$, elle reste constante sur de longs intervalles de temps.

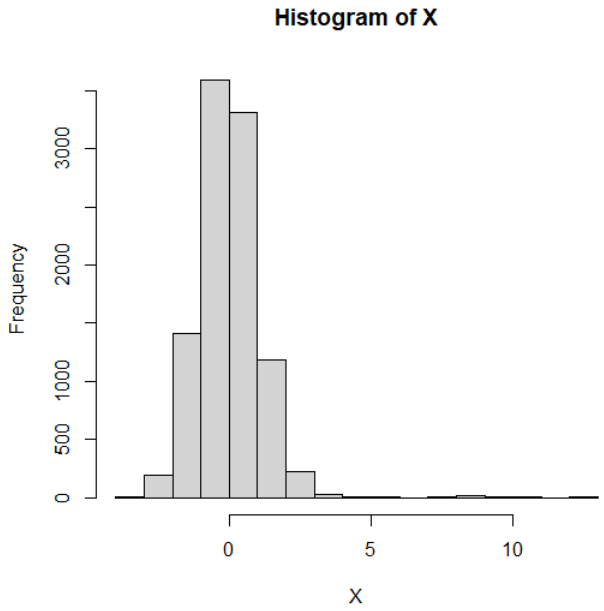

FIGURE 2.3 – densité $\mathcal{N}(0,1)$ à partir de $\mathcal{U}_{[-\delta,\delta]}$ pour $\delta = 1$

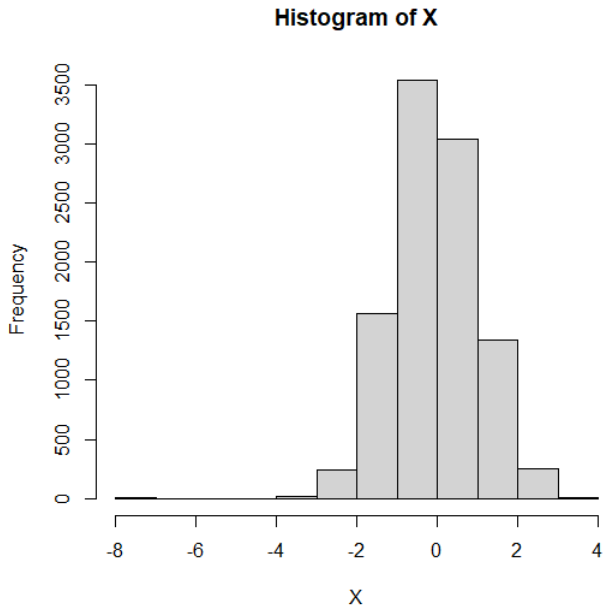

FIGURE 2.4 – densité $\mathcal{N}(0,1)$ à partir de $\mathcal{U}_{[-\delta,\delta]}$ pour $\delta = 10$

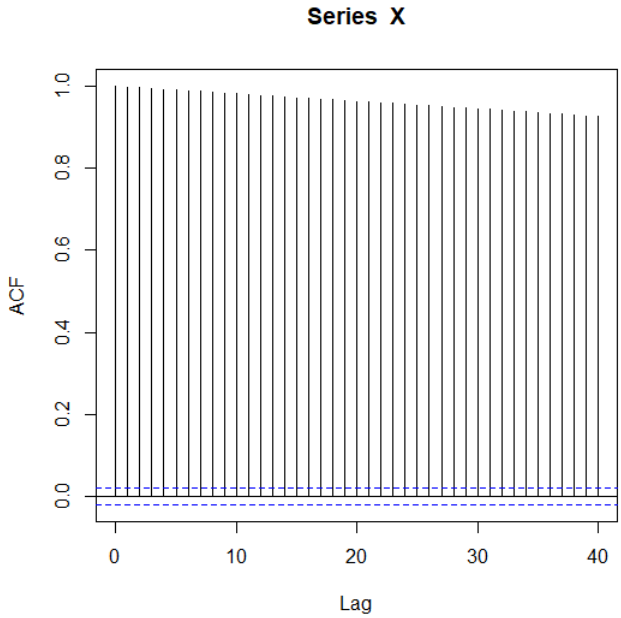

FIGURE 2.5 – autocovariance pour $\delta = 0.1$

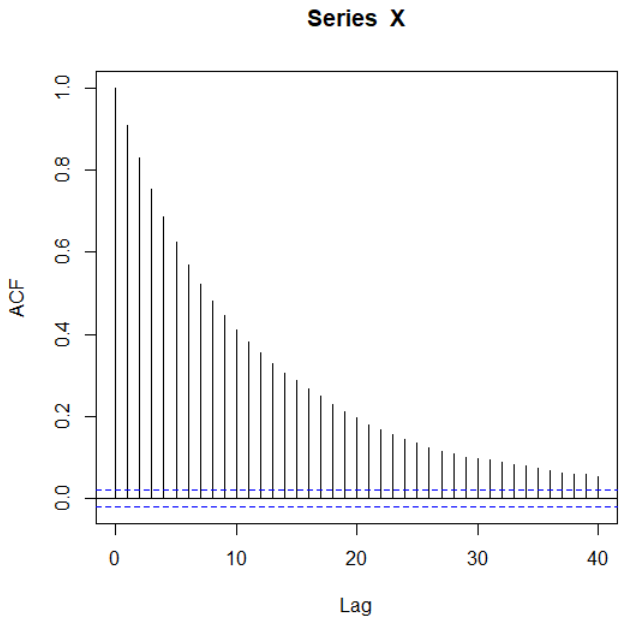

FIGURE 2.6 – autocovariance pour $\delta = 1$

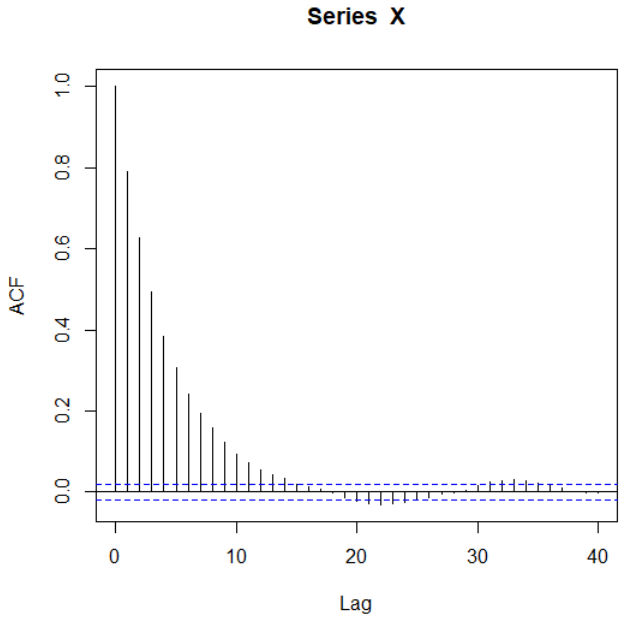

FIGURE 2.7 – autocovariance pour $\delta = 10$

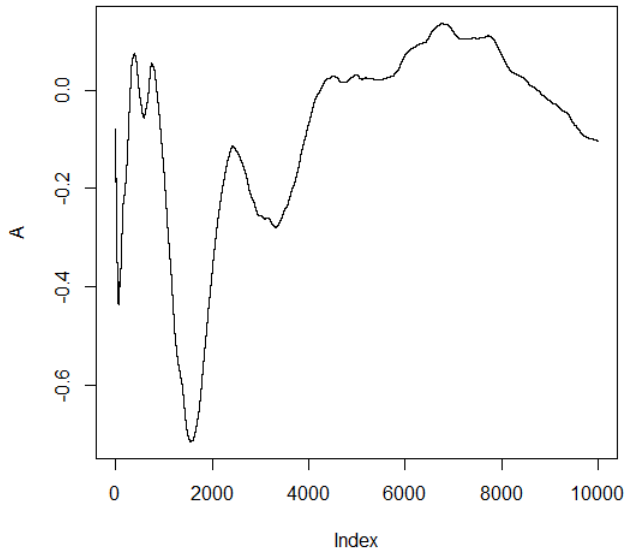

FIGURE 2.8 – convergence de la moyenne arithmétique pour $\delta = 0.1$

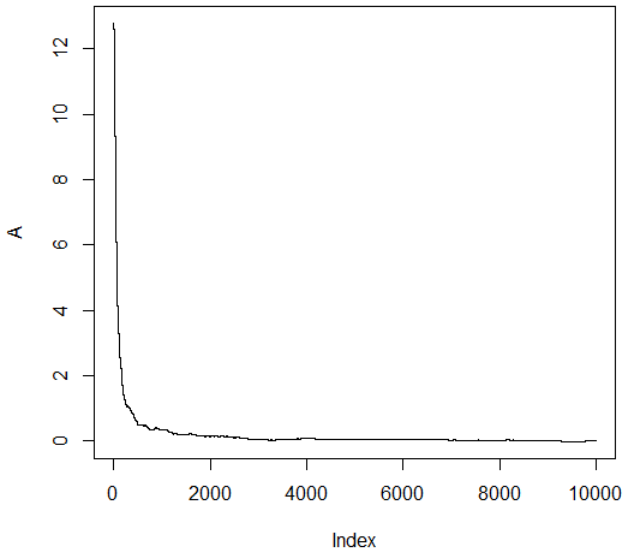

FIGURE 2.9 – convergence de la moyenne arithmétique pour $\delta = 1$

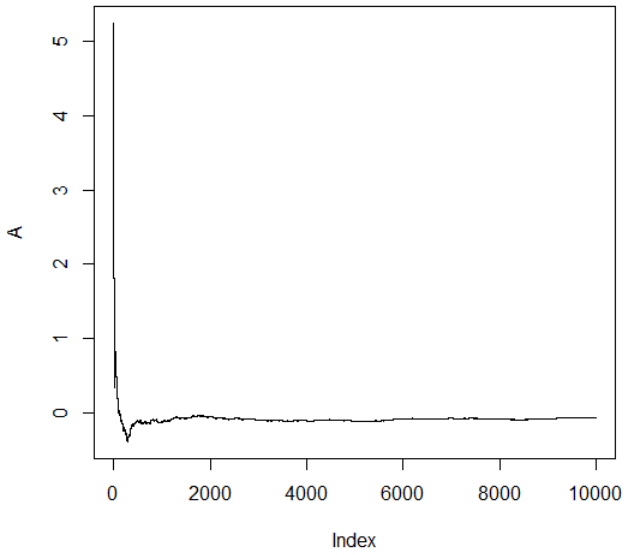

FIGURE 2.10 – convergence de la moyenne arithmétique pour $\delta = 10$

-Calibrer l'échelle δ de la marche aléatoire est crucial pour obtenir une bonne approximation de la loi cible en un nombre d'itérations raisonnable .

2.8.3 Ergodicité uniforme pour L'algorithme de Metropolis- Hastings par marche aléatoire :

-Malgré sa simplicité et ses caractéristiques naturelles, L'algorithme de Metropolis- Hastings par marche aléatoire ne bénéficie pas des propriétés d'ergodicité uniformes.
-Mengersen et Tweedie (1996) ont montré que dans le cas où $supp\ f = R$, cet algorithme ne peut pas produire une chaîne de Markov uniformément ergodique sur R.
-Il s'agit d'une caractéristique plutôt peu surprenante si l'on considère le caractère local de la proposition de marche aléatoire, centrée sur la valeur actuelle de la chaîne de Markov.

2.8.4 Ergodicité géometrique pour L'algorithme de Metropolis- Hastings par marche aléatoire :

-Bien qu'une ergodicité uniforme ne puisse pas être obtenue avec un Algorithmes Metropolis-Hastings par marche aléatoire, , il est possible de dériver des conditions nécessaires et suffisantes d'ergodicité géométrique.
-Mengersen et Tweedie (1996) ont proposé une condition basée sur la concavité logarithmique dans les queues ;
C'est-à-dire s'il existent $\alpha > 0$ et x_1 tels que :

$$\log(f(x)) - \log(f(y)) \geq \alpha|y - x| \text{ pour } y < x < -x_1 \text{ ou } x_1 < x < y$$

2.8.5 Théoréme

-Considérons une densité symétrique f qui est log-concave avec α une constante associée dans la concavité logarithmique pour $|x|$ assez grand.
Si la densité g est positive et symétrique,
La chaîne $(X^{(t)})$ de L'algorithme de Metropolis- Hastings par marche aléatoire est géométriquement ergodique.
-Si f n'est pas symétrique, une condition suffisante pour l'ergodicité géométrique est que $g(t)$ soit bornée par $b\exp(-\alpha|t|)$ pour une constante b suffisamment grande.

-Tierney (1994) a proposé une modification de l'algorithme précédent avec une densité

de proposition de la forme $g(y - a - b(x - a))$; c'est à dire :

$$y_t = a + b(x^{(t)} - a) + z_t \ , z_t \sim g$$

-Cette représentation autorégressive peut être vue comme un intermédiaire entre la version indépendante $(b = 0)$ et la version marche aléatoire $(b = 1)$ de l'Algorithme de Metropolis-Hastings.
-De plus, lorsque $b < 0$, $X(t)$ et $X^{(t+1)}$ sont corrélés négativement, ce qui peut permettre des excursions plus rapides sur la surface de f si le point de symétrie a est bien choisi .
-Hastings (1970) considère également une alternative à la distribution uniforme sur $[x^{(t)} - \delta, x^{(t)} + \delta]$ avec la distribution uniforme sur $[-x^{(t)} - \delta, -x^{(t)} + \delta]$;
La convergence de la moyenne empirique à 0 est alors plus rapide dans ce cas, mais le choix de 0 comme le centre de symétrie est évidemment crucial et nécessite quelques informations a priori sur la distribution f.
-Dans un cadre général, a et b peuvent être calibrés pendant les premières itérations .

2.9 Optimisation et contrôle :

-On a établi la validité théorique des algorithmes de Metropolis-Hastings en montrant que sous des conditions convenables (et peu restrictives) sur le noyau de transition, la chaîne produite par l'algorithme est ergodique et, par conséquent, que la moyenne $\mathcal{I}_T = \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$ converge vers l'espérance $E_f(h(X))$.
-Cependant, nous avons montré que les algorithmes les plus commun ne jouissent que rarement de fortes propriétés d'ergodicité (ergodicité géométriques ou uniformes) ;
En particulier, il existe des exemples simples qui montrent à quel point la convergence peut être lente.
-Ici on aborde le problème du choix du noyau de transition $q(y|x)$.

2.9.1 Optimiser le taux d'acceptation :

-En ne considérant que les classes d'algorithmes décrites précédemment , les alternatives les plus courantes consistent à utiliser :
i) une densité instrumentale g qui est une approximation de f, telle que f/g soit borné pour que l'ergodicité uniforme s'applique à l'algorithme de Metropolis-Hastings indépendent
ii) une marche aléatoire comme dans L'algorithme de Metropolis- Hastings par marche aléatoire .

-Dans les deux cas, le choix de g est beaucoup plus critique, car il détermine les performances de l'algorithme de Metropolis-Hastings résultant .
-Comme nous le verrons plus loin, les quelques conseils disponibles sur les choix de g sont, en fait, contraires ! ;
Selon le type de l'algorithme de Metropolis-Hastings sélectionné, on voudrait des taux d'acceptation élevés dans le cas i) et de faibles taux d'acceptation dans le cas ii) .

-Considérons d'abord l'algorithme de Metropolis-Hastings indépendant ;
Sa similarité avec l'algorithme Acceptation-Reject suggère un choix de g qui maximise le taux d'acceptation moyen :

$$\rho = E(min\{\frac{f(Y)}{f(X)}\frac{g(X)}{g(Y)}, 1\}) = 2P(\frac{f(Y)}{g(Y)} \geq \frac{f(X)}{g(X)}) \ , X \sim f, Y \sim g$$

-En effet, l'optimisation associée au choix de g est lié à la vitesse de convergence de $\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$ vers $E_f(h(X))$
Et, par conséquent, à la capacité de l'algorithme de Metropolis-Hastings indépendant à explorer rapidement toute complexité de f .

-Si cette optimisation doit être générique (c'est-à-dire indépendante de h), g doit reproduire le plus fidèlement possible la densité f, ce qui implique la maximisation de ρ.
-Par exemple, une densité g qui est soit beaucoup plus petite soit beaucoup plus concentré, par rapport à f, produit un rapport

$$\frac{f(y)}{f(x)}\frac{g(x)}{g(y)} \wedge 1$$

ayant d'énormes variations et, par conséquent, conduit à un faible taux d'acceptation .

-Le taux d'acceptation ρ est généralement impossible à calculer, et une solution est d'utiliser le résultat de minorisation $\rho \geq 1/M$ pour minimiser M comme dans le cas de l'algorithme d'Acceptation-Reject.
-Alternativement, on peut envisager une approche plus empirique qui consiste à choisir une distribution instrumentale paramétrée $g(.|\theta)$ et d'ajuster les paramètres correspondants θ basés sur le taux d'acceptation évalué, actuel $\hat{\rho}(\theta)$;
C'est-à-dire, de choisir d'abord une valeur initiale pour les paramètres, θ_0, et d'estimer le taux d'acceptation correspondant, $\hat{\rho}(\theta_0)$, basé sur m itérations de l'algorithme de Metropolis-Hastings indépendant ,pour alors modifier θ_0 pour obtenir une augmentation de $\hat{\rho}$.

-Dans les cas les plus simples, θ_0 se réduira à un paramètre d'échelle qui est augmenté ou diminué selon le comportement de $\hat{\rho}(\theta)$.
-Dans des contextes multidimensionnels, θ_0 peut également comporter un paramètre de position ou une matrice jouant le rôle de paramètre d'échelle, ce qui rend l'optimisation de $\rho(\theta)$ plus complexe.

-Notez que peut $\hat{\rho}(\theta)$ être obtenu par simple comptage des acceptations ou par :

$$\frac{2}{m}\sum_{i=1}^{m} 1_{\{f(y_i)g(x_i|\theta)>f(x_i)g(y_i|\theta)\}}$$

où $x_1, ..., x_m$ est un échantillon de f, obtenu par exemple à partir d'un premier algorithme MCMC, et $y_1, ..., y_m$ est un échantillon iid de $g(.|\theta)$.
-Donc, si θ est composé de paramètres de localisation et d'échelle, un échantillon $((x_1, y_1), ..., (x_m, y_m))$ correspondant à une valeur θ_0 peut être utilisé à plusieurs reprises pour évaluer différentes valeurs de θ par une modification déterministe des y_i, ce qui facilite la maximisation de $\rho(\theta)$.

-La version de marche aléatoire de l'algorithme de Metropolis-Hastings, nécessite une approche différente des taux d'acceptation, compte tenu de la dépendance de la distribution instrumentale à l'état actuel de la chaîne.
-En fait, un taux d'acceptation élevé n'indique pas nécessairement que l'algorithme se déplace correctement car cela peut indiquer que la marche aléatoire se déplace également lentement à la surface de f.
-Si $x^{(t)}$ et y_t sont proches, au sens où $f(x^{(t)})$ et $f(y_t)$ sont approximativement égaux, l'algorithme de Metropolis-Hastings par marche aléatoire conduit à l'acceptation de y avec probabilité :

$$min\{\frac{f(y_t)}{f(x^{(t)})}, 1\} \approx 1$$

-Un taux d'acceptation plus élevé peut donc correspondre à une convergence plus lente car les déplacements sur le support de f sont plus limités.
-Dans le cas particulier de densités multimodales dont les modes sont séparés par des zones de faible probabilité, l'effet négatif des mouvements limités sur la surface de f se voit clairement.
-Alors que le taux d'acceptation est assez élevé pour une distribution g avec une faible variance, la probabilité de sauter d'un mode à l'autre peut être arbitrairement petite.
-Ce phénomène se produit, par exemple, dans le cas de mélanges de distributions et dans les modèles surparamétrés .

-En revanche, si le taux d'acceptation moyen est faible, les valeurs successives de $f(y_t)$ tendent à être petites par rapport à $f(x^{(t)})$, ce qui signifie que la marche aléatoire se déplace rapidement sur la surface car elle atteint souvent les "bords" du support de f .
(ou, du moins, que la marche aléatoire explore des régions avec une faible probabilité sous f) .

-L'analyse ci-dessus semble nécessiter une connaissance avancée de la densité d'intérêt, car une distribution instrumentale g avec une plage trop étroite ralentira la vitesse de convergence de l'algorithme.
-D'autre part, une distribution g avec une large plage entraîne un gaspillage de simulations de points en dehors de la plage de f sans améliorer la probabilité de visiter tous les modes de f.
-Il est regrettable qu'un paramétrage automatisé de g ne puisse garantir des performances uniformément optimales pour l'algorithme de Metropolis-Hastings par marche aléatoire, et que les règles de choix du taux ne sont qu'heuristiques .

2.10 Schémas adaptatifs :

-Compte tenu de l'éventail des situations où les méthodes MCMC s'applique, il est irréaliste d'espérer un échantillonneur MCMC générique qui fonctionnerait dans tous les paramètres possibles.
-Les propositions les plus génériques comme les algorithmes Metropolis-Hastings à marche aléatoire sont connus pour échouer dans les supports de grande dimension et déconnectés, car ils prennent trop de temps pour explorer l'espace d'intérêt (Neal 2003).
-La raison pour ce théorème d'impossibilité est que, dans des problèmes réalistes, la complexité de la distribution à la simulation est la raison même pour laquelle la méthode MCMC est utilisé !
Alors il est est difficile de demander un avis préalable sur cette distribution, son support ou les paramètres de la distribution de proposition utilisés dans l'algorithme MCMC ; l'intuition est proche du vide dans la plupart de ces problèmes.

-Cependant, les performances des algorithmes comme la marche aléatoire de Metropolis-Hastings apporte des informations sur la distribution d'intérêt et, par conséquent, devraient être incorporés dans la conception de meilleurs et plus puissants algorithmes .
-Le problème est que nous manquons généralement de temps pour entraîner l'algorithme sur ces performances précédentes et ont recherchent le Saint Graal des procédures MCMC

automatisées !
S'il est naturel de penser que les informations apportées par les premières étapes d'un algorithme MCMC doivent être utilisées dans les étapes ultérieures, il y a un sérieux hic : Utiliser tout le passé de la "chaîne" implique que ce n'est plus une chaîne de Markov.
Par conséquent, les théorèmes de convergence habituelle ne s'appliquent pas et la validité des algorithmes correspondants est questionnable .
-De plus, il se peut que, dans la pratique, de tels algorithmes dégénèrent aux masses ponctuelles à cause d'une décroissance trop rapide de la variation de leur proposition .

-Même si la chaîne de Markov converge en distribution vers la distribution cible (lors de l'utilisation d'un schéma de mise à jour approprié, c'est-à-dire homogène dans le temps), utiliser des simulations passées pour créer une approximation non paramétrique de la distribution cible ne fonctionne pas non plus .
-Le message général est donc qu'il ne faut pas constamment adapter la distribution de proposition sur les performances passées de la chaîne simulée.
- Soit l'adaptation doit cesser après une période de rodage (à ne pas prendre en compte pour les calculs d'espérances et de quantités liées à la distribution cible), ou le schéma adaptatif doit être théoriquement évalué dans son propre droit.
-Cette dernière voie n'est pas facile et seuls quelques exemples peuvent être trouvés .

2.11 L'échantillonneur par tranches :

-Alors que de nombreux algorithmes MCMC présentés précédemment sont à la fois génériques et universels, il existe une classe spéciale d'algorithmes MCMC qui dépendent davantage du modèle dans le sens qu'ils exploitent les caractéristiques conditionnelles locales des distributions à simuler .
- Nous avons développé des techniques de simulation que l'on pourrait appeler "génériques", car elles ne nécessitent qu'une quantité limitée d'informations sur la distribution à simuler .
-Cependant, Les algorithmes de Metropolis-Hastings peuvent atteindre des niveaux d'efficacité plus élevés lorsque ils prennent en compte les spécificités de la distribution cible f, notamment par le calibrage du taux d'acceptation .

2.12 Un autre regard sur le théorème fondamental :

-La génération à partir d'une distribution de densité $f(x)$ équivaut à une génération uniforme sur le sous-graphe de f :

$$S(f) = \{(x,u); 0 \leq u \leq f(x)\}$$

quelle que soit la dimension de x, et f n'a besoin d'être connu qu'à une constante de normalisation près .
-On peut considérer la possibilité d'utiliser une chaîne de Markov de distribution stationnaire égale à cette distribution uniforme sur $S(f)$ comme moyen approximatif de simuler à partir de f.
-Une solution naturelle est d'utiliser une marche aléatoire sur $S(f)$, puisqu'une marche aléatoire sur un ensemble $\mathcal{A}$ aboutit généralement à une distribution stationnaire qui est la distribution uniforme sur $\mathcal{A}$.
-Il existe de nombreuses façons d'implémenter une marche aléatoire sur cet ensemble, mais une solution naturelle est d'aller dans une direction à la fois,
C'est-à-dire de se déplacer de manière itérative le long de l'axe u puis le long de l'axe x.
-De plus, nous pouvons utiliser des déplacement uniformes dans les deux sens, puisque, comme formellement montré ci-dessous, La chaîne de Markov sur $S(f)$ ne nécessite pas de correction de Metropolis-Hastings pour avoir la distribution uniforme sur $S(f)$ comme distribution stationnaire.
-En partant d'un point (x,u) dans $\{(x,u) : 0 < u < f(x)\}$, le déplacement le long de l'axe u correspondent à la distribution conditionnelle :

$$U|X = x \sim \mathcal{U}(\{u; u \leq f(x)\})$$

entraînant un passage du point (x,u) au point (x,u'), toujours dans $S(f)$, puis le déplacement le long de l'axe des x correspondent à la distribution conditionnelle :

$$X|U = u' \sim \mathcal{U}(\{x; u' \leq f(x)\})$$

résultant en un changement du point (x,u') vers le point (x',u') .

-Cet ensemble de propositions est la base choisie pour L'échantillonneur par tranches original de Neal (1997) (publié sous le nom de Neal 2003) et Damien et. Al. (1999), qui ainsi utilise une marche aléatoire uniforme en 2 étapes sur le sous-graphe.
-Nous l'appelons à tort l'échantillonneur par tranches 2D pour le distinguer de l'échan-

tillonneur par tranches général , même si la dimension de x est arbitraire .

2.12.1 Algorithme : échantillonneur par tranches 2D :

A l'itératon t , simuler :
1. $u^{(t+1)} \sim \mathcal{U}_{[0,f(x^{(t)})]}$;
2. $x^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$ avec $A^{(t+1)} = \{x; f(x) \geq u^{(t+1)}\}$.

-De $U|X = x \sim \mathcal{U}(\{u; u \leq f(x)\})$ il ressort également que $x^{(t)}$ fait toujours partie de l'ensemble $A^{(t+1)}$, qui est donc non vide.
-De plus, l'algorithme reste valide si $f(x) = Cf_1(x)$, et nous utilisons f_1 à la place de de f.
-Ceci est assez avantageux dans les cas où f est une densité non normalisée comme une densité a posteriori .

-La validité de l'échantillonneur par tranches 2D comme algorithme MCMC associé à f_1 découle du fait que les deux les étapes 1. et 2. de l'algorithme préservent successivement la distribution uniforme sur le sous-graphe de f :
premièrement, si $x^{(t)} \sim f(x)$ et $u^{(t+1)} \sim \mathcal{U}_{[0,f_1(x^{(t)})]}$ alors :

$$(x^{(t)}, u^{(t+1)}) \sim f(x)\frac{1_{[0,f_1(x)]}}{f_1(x)} \propto 1_{0\leq u\leq f_1(x)}$$

deuxiément si $x^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$ alors :

$$(x^{(t)}, u^{(t+1)}, x^{(t+1)}) \sim f(x^{(t)})\frac{1_{[0,f_1(x^{(t)})]}(u^{(t+1)})}{f_1(x^{(t)})}\frac{1_{A^{(t+1)}}(x^{(t+1)})}{mes(A^{(t+1)})}$$

ou $mes(A^{(t+1)})$ est la mesure de l'ensemble $A^{(t+1)}$ (généralement la mesure de Lebesgue) .
-Donc :

$$(x^{(t+1)}, u^{(t+1)}) \sim C\int 1_{0\leq u\leq f_1(x)}\frac{1_{f_1(x^{(t+1)})\geq u}}{mes(A^{(t+1)})}dx$$

$$= C1_{0\leq u\leq f_1(x^{(t+1)})}\int \frac{1_{u\leq f(x)}}{mes(A^{(t+1)}}dx$$

$$\propto 1_{0\leq u\leq f_1(x^{(t+1)})}$$

et la distribution uniforme sur $S(f)$ est bien stationnaire pour les deux étapes .

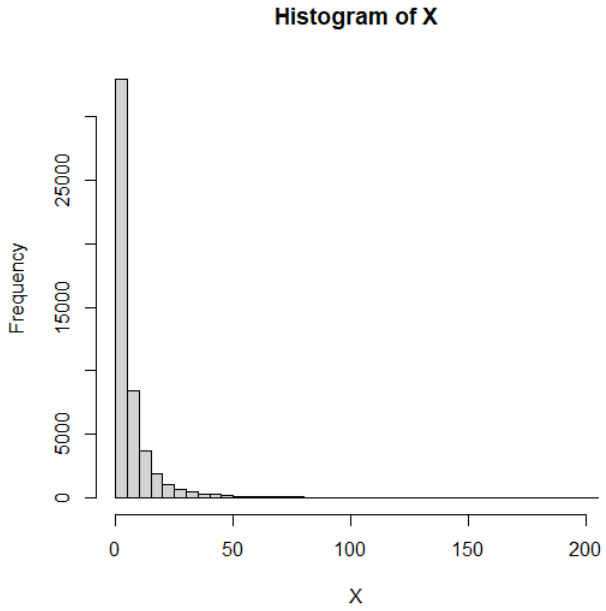

FIGURE 2.11 – densité de l'échantillonneur par tranches

2.12.2 Exemple : Échantillonneur par tranche simple :

-Soit la densité $f(x) = \frac{1}{2}\exp(-\sqrt{x})$; pour $x > 0$.
-Pour simuler f avec l'échantillonneur par tranches , il suffit d'expliciter l'ensemble A .
-Or $f(x) \geq u$ ssi $\frac{1}{2}\exp(-\sqrt{x}) \geq u$
ssi $-\sqrt{x} \geq \log(2u)$ ssi $\sqrt{x} \leq -\log(2u)$ ssi $x \leq (\log(2u)^2$.
-Donc $U|x \sim \mathcal{U}[0, \frac{1}{2}\exp(-\sqrt{x})]$ et $X|u \sim \mathcal{U}[0, (\log(2u))^2]$.
-Voici le code R correspondant :

```
    n=50000;X=NULL;U=NULL;U[1]=runif(1,0,0.5);
for(t in 1:(n-1)){X[t]=runif(1,0,(log(2*U[t]))^2);
U[t+1]=runif(1,0,0.5*exp(-sqrt(X[t])))};hist(X,nclass=50)
windows();acf(X)
```

-Nous avons ci-dessus le résultat de l'histogramme et ci-dessous le résulat de l'estimation de l'autocorrélation .

2.12.3 Exemple : Distribution normale tronquée :

-on souhaite simuler une distribution normale tronquée $\mathcal{N}(3,1)$ restreinte sur l'intervalle [0,1] ,

$$f(x) \propto f_1(x) = \exp(-(x-3)^2/2)1_{[0,1]}(x)$$

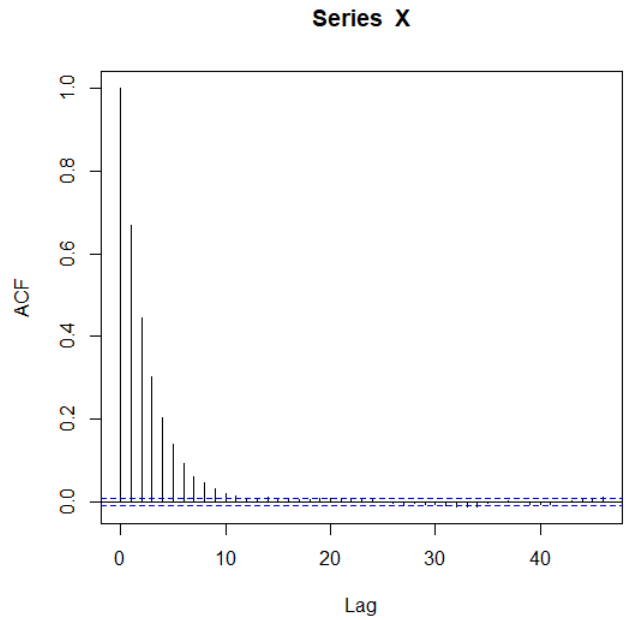

FIGURE 2.12 – autocorrélation de l'échantillonneur par tranches

-La simulation naïve d'une variable aléatoire normal $\mathcal{N}(3,1)$ jusqu'à ce que le résultat soit dans [0, 1] n'est pas optimal car il n'y a que une probabilité de 2% que cela se produise .
-Cependant, concevoir et optimiser l'algorithme d'acceptation-rejet peut être coûteux si l'algorithme doit être utilisé que quelques fois .

-L'échantillonneur par tranches appliqué à ce problème est alors associé à la tranche horizontale :

$$A = \{x\,;\, f_1(x) \geq u\}$$

$$= \{x \in [0,1]\,; \exp\left(-(x-3)^2/2\right) \geq u\}$$

$$= \{x \in [0,1]\,;\, (x-3)^2 \leq -2\log(u)\}$$

-Voici le code R :

```
f1=function(x){return(exp(-1/2*(x-3)^2)*as.numeric(x>=0 & x<=1))};
X=NULL;U=NULL;n=10^3;X[1]=0.25;for(t in 1:(n-1)){
U[t+1]=runif(1,0,f1(X[t]));x=runif(1)
while( (x-3)^2>-2*log(U[t+1]) ){x=runif(1)};X[t+1]=x};
hist(X)
```

-Nous avons ci-dessous le résultat de l'histogramme .

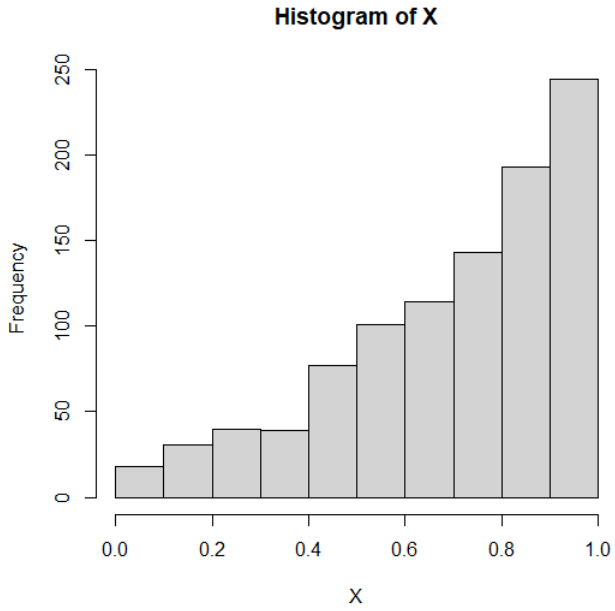

FIGURE 2.13 – densité normale tronquée par l'échantillonneur par tranches

-La principale difficulté (pratique) avec l'échantillonneur par tranches réside dans la simulation de la distribution uniforme $\mathcal{U}_{A^{(t+1)}}$, puisque la détermination de l'ensemble des y est tel que $f_1(y) \geq \omega$ peut être insoluble si f_1 est suffisamment complexe .

2.13 L'échantillonneur par tranches général :

-Comme indiqué ci-dessus, échantillonner uniformément à partir de la tranche $A^{(t)} = \{x; f_1(x) \geq \omega^{(t)}\}$ peut être complètement insoluble .
-Cette difficulté persiste à mesure que la dimension de x devient plus grande.
-Cependant, il existe une généralisation de l'échantillon par tranche 2D qui atténue partiellement cette difficulté en introduisant plusieurs tranches .

-Cet échantillonneur par tranches général peut être retracé jusqu'à l'algorithme de variable auxiliaire d'Edwards et Sokal (1988) appliqué au mode d'Ising .
- Elle repose sur la décomposition de la densité $f(x)$ comme :

$$f(x) \propto \prod_{i=1}^{k} f_i(x)$$

où les f_i sont des fonctions positives, mais pas nécessairement des densités.
-Par exemple , dans un cadre bayésien avec a priori plat, les $f_i(x)$ peuvent être choisi comme les vraisemblances individuelles .

-Cette décomposition peut alors être associée à k variables auxiliaires ω_i, plutôt qu'une comme dans le théorème fondamental, en ce sens que chaque $f_i(x)$ peut s'écrire comme une intégrale :

$$f_i(x) = \int 1_{0\leq\omega_i\leq f_i(x)} d\omega_i$$

et que f est la distribution marginale de la distribution conjointe :

$$(x, \omega_1, ..., \omega_k) \sim p(x, \omega_1, ..., \omega_k) \propto \prod_{i=1}^{k} 1_{0\leq\omega_i\leq f_i(x)}$$

-Cette démarginalisation particulière introduit une plus grande dimensionnalité au problème et induit une généralisation de la marche aléatoire qui est d'avoir des propositions uniformes une direction à la fois.
- La généralisation correspondante de l'algortihme de l'échantillonneur par tranches 2D est donc la suivante .

2.13.1 Algortihme : échantillonneur par tranches :

A l'itéation $t + 1$, simuler :
1. $\omega_1^{(t+1)} \sim \mathcal{U}_{[0,f_1(x^{(t)})]}$
...
...
k. $\omega_k^{(t+1)} \sim \mathcal{U}_{[0,f_k(x^{(t)})]}$
k+1. $x^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$ avec $A^{(t+1)} = \{x; f_i(x) \geq \omega_i^{(t+1)}$, $i = 1, ..., k\}$

2.13.2 Exemple : Un échantillonneur de tranches 3D :

-On considére la densité proportionnelle à :

$$(1 + \sin(3x)^2)(1 + \cos(5x)^4) \exp(-x^2/2)$$

-Les fonctions correspondantes sont :
$f_1(x) = (1 + \sin(3x)^2)$; $f_2(x) = (1 + \cos(5x)^4)$ et $f_3(x) = \exp(-x^2/2)$.
-Et donc dans l'itération de l'échantillonneur de tranches 3D on a :
$A = \{x\,;\, f_1(x) \geq u_1\} \cap \{x\,;\, f_2(x) \geq u_2\} \cap \{x\,;\, f_3(x) \geq u_3\}$
$= \{x\,;\, \sin(3x)^2 \geq u_1 - 1\} \cap \{x\,;\, \cos(5x)^4 \geq u_2 - 1\} \cap \{x\,;\, |x| \leq \sqrt{-2\log(u_3)}\}$
- Voici le code R :

```
f1=function(x){return(1+sin(3*x)^2)};
```

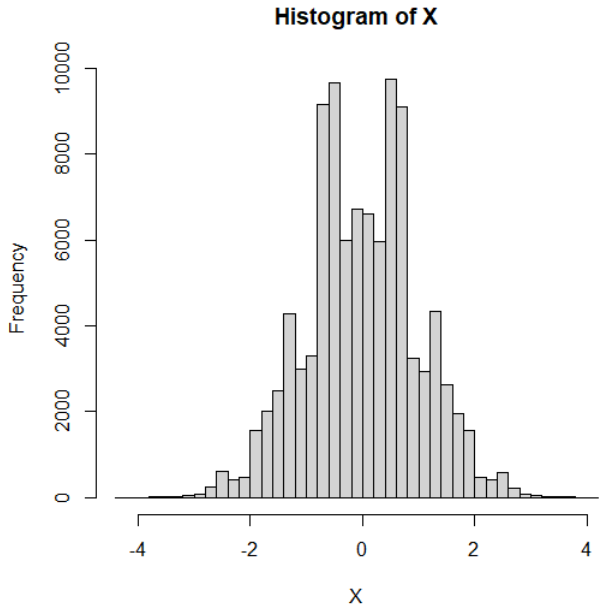

FIGURE 2.14 – densité par un échantillonneur de tranches 3D

```
f2=function(x){return(1+cos(5*x)^4)};
f3=function(x){return(exp(-x^2/2))};n=10^5
U1=NULL;U2=NULL;U3=NULL;X=NULL;X[1]=0.75;
for(t in 1:(n-1)){U1[t]=runif(1,0,f1(X[t]));
U2[t]=runif(1,0,f2(X[t]));U3[t]=runif(1,0,f3(X[t]));x=runif(1,-5,5);
while( sin(3*x)^2<U1[t]-1 | cos(5*x)^4<U2[t]-1
|abs(x)>sqrt(-2*log(U3[t])) )
{x=runif(1,-5,5)};X[t+1]=x};
hist(X,nclass=50)
```

-Nous avons ci-dessus le résultat de l'histogramme .
-Bien que la représentation $(x, \omega_1, ..., \omega_k) \sim p(x, \omega_1, ..., \omega_k) \propto \prod_{i=1}^{k} 1_{0 \leq \omega_i \leq f_i(x)}$ de f nous éloigne du théorème fondamental de la simulation, la validation de cet algorithme est la même comme dans le précédent :
chacune des $k + 1$ étapes de l'algortihme préserve la distribution $p(x, \omega_1, ..., \omega_k)$.
-Ce sera aussi la base de l'échantillonneur de Gibbs .

-Alors que l'attrait fondamental de l'utilisation de cette généralisation est que l'ensemble $A^{(t+1)}$ peut être plus facile à calculer comme étant l'intersection des tranches :

$$A_i^{(t+1)} = \{y\,;\, f_i(y) \geq \omega_i^{(t+1)}\}$$

il peut encore y avoir des problèmes de mise en oeuvre.
-Lorsque k augmente, la détermination de l'ensemble $A^{(t+1)}$ devient généralement de plus en plus complexe .

-Notez également que, dans les modèles à variables manquantes, le nombre de variables auxiliaires (c'est-à-dire de tranches) augmente avec le nombre d'observations et peut créer des blocages pour les grandes tailles de données .

2.14 Propriétés de convergence de l'échantillonneur par tranches :

-Sous des conditions , Tierney et Mira (1999) ont établi le résultat suivant d'ergodicité uniforme .

-Si f_1 est borné et $supp\ f_1$ est borné, l'échantillonneur par tranches 2D est uniformément ergodique .

2.15 L'échantillonneur de Gibbs à deux étapes :

-Nous avons présenté l'échantillonneur par tranches , un cas particulier d'algorithme de chaîne de Markov qui n'avait pas besoin d'une étape d'Acceptation-Rejet pour être valide, apparemment en raison de l'uniformité de la distribution cible.
-La raison pour laquelle l'échantillonneur par tranches marche est cependant sans rapport avec cette uniformité et nous verrons une famille beaucoup plus générale d'algorithmes qui fonctionnent sur le même principe.
-Ce principe est celui d'utiliser les vraies distributions conditionnelles associé à la distribution cible pour générer à partir de cette distribution .

-L'échantillonneur de Gibbs à deux étapes peut être dérivé comme une généralisation de l'échantillonneur par tranche.
-Les deux algorithmes partagent ainsi des propriétés de convergence supérieures qui ne s'appliquent pas à l'échantillonneur de Gibbs général à plusieurs étapes .
- L'échantillonneur de Gibbs à deux étapes s'applique naturellement dans une large gamme de modèles statistiques qui ne font pas appel à la généralité de l'échantillonneur de Gibbs général à plusieurs étapes .

2.16 Une classe générale d'algorithmes en deux étapes :

2.17 De l'échantillonnage par tranche à l'échantillonnage de Gibbs :

-Au lieu d'une densité $f_X(x)$, considérons maintenant une densité conjointe $f(x,y)$ défini sur un espace produit arbitraire, $\mathcal{X} \times \mathcal{Y}$.
-Si nous utilisons le théorème fondamental de simulation dans cette configuration, nous simulons une distribution uniforme sur l'ensemble :

$$S(f) = \{(x,y,u)\,;\, 0 \leq u \leq f(x,y)\}$$

-Étant donné que nous sommes maintenant confrontés à un cadre à trois composantes, une mise en oeuvre naturelle du principe de la marche aléatoire consiste à se déplacer uniformément dans une composante à la fois.
-Ce qui signifie que, partant d'un point (x,y,u) dans $S(f)$, on génère :
i) X le long de l'axe des x à partir de la distribution uniforme sur $\{x\,;\, u \leq f(x,y)\}$
ii) Y le long de l'axe des y à partir de la distribution uniforme sur $\{y\,;\, u \leq f(x',y)\}$
iii) U le long de l'axe des u à partir de la distribution uniforme sur $[0, f(x',y')]$.
-Notez que ceci est différent de l'échantillonneur de tranches 3D .

-Il y a deux choses importantes à noter :
(1) Générer à partir de la distribution uniforme sur $\{x\,;\, u \leq f(x,y)\}$ est équivalent à générer à partir de la distribution uniforme sur $\{x\,;\, f_{X|Y}(x|y) \geq u f_Y(y)\}$
où $f_{X|Y}$ et f_Y désignent les distributions conditionnelle et marginale de X étant donné Y et de Y, respectivement, c'est-à-dire :

$$f_Y(y) = \int f(x,y)dx\,;\, f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

(2) La séquence de générations uniformes le long des trois axes n'a pas besoin d'etre faite dans le même ordre $x - y - u$ tout le temps pour que la chaîne de Markov restent stationnaires avec comme distribution stationnaire l'uniforme sur $S(f)$.

-Ainsi, par exemple, les simulations selon les axes x et u peuvent être répétées plusieurs fois avant de passer à la simulation le long de l'axe y.
-Si nous mettons ces deux remarques ensemble et considérons le cas limite où les simulations de X et U sont répétées un nombre infini de fois avant de passer à la simulation de

Y , on aboutit à une simulation (en X) de :

$$X \sim f_{X|Y}(x|y)$$

grâce à l'échantillonneur de tranches 2D.

-Considérons maintenant la même répétition des simulations de Y et U avec X fixé à sa dernière valeur :
dans le cas limite, cela produit une simulation (en Y) de :

$$Y \sim f_{Y|X}(y|x)$$

-En supposant que ces deux distributions conditionnelles peuvent être simulées, nous pouvons implémentez donc le cas limite de l'échantillonneur par tranches et toujours maintenir la stationnarité de la distribution uniforme.
-De plus, la simulation de U devient en quelque sorte superflu puisque nous nous intéressons vraiment qu'à la génération de $f(x, y)$, plutôt que de l'uniforme sur $S(f)$.

-Bien que ce ne soit pas ainsi qu'il a été dérivé à l'origine, cette introduction souligne le lien fort entre l'échantillonneur par tranches et l'échantillonneur de Gibbs à deux étapes , parfois appelé L'augmentation des données (Tanner et Wong 1987).
-Elle souligne également comment l'échantillonneur de Gibbs à deux étapes tire davantage parti de la connaissances sur la distribution f, par rapport à l'échantillonneur par tranches qui utilise uniquement les valeurs numériques de $f(x)$.
-Cela se traduit par le fait que chaque étape de l'échantillonneur de Gibbs à deux étapes équivaut à une infinité d'étapes d'un échantillonneur par tranches.
(Notez que cela ne signifie pas que l'échantillonneur de Gibbs à deux étapes fait toujours mieux que n'importe quel échantillonneur par tranches.)

-L'implémentation algorithmique de l'échantillonneur de Gibbs à deux étapes est donc direct.
-Si les variables aléatoires X et Y ont une densité conjointe $f(x, y)$, l'échantillonneur de Gibbs à deux étapes génère une chaîne de Markov (X_t, Y_t) selon les étapes suivantes :
Prendre X_0 = x_0
Pour t = $1, 2, ...$ générer :
1. $Y_t \sim f_{Y|X}(.|x_{t-1})$
2. $X_t \sim f_{X|Y}(.|y_t)$

- Notons ici que non seulement la suite (X_t, Y_t) est une chaîne de Markov, mais aussi que chaque sous-suite (X_t) et (Y_t) est une chaîne de Markov.
-Par exemple, la chaîne (X_t) a une densité de transition :

$$K(x, x^*) = \int f_{Y|X}(y|x) f_{X|Y}(x^*|y) dy$$

qui ne dépend en effet du passé que par la dernière valeur de (X_t)·
-De plus, il est également facile de montrer que f_X est la distribution stationnaire associée à cette (sous)chaîne, puisque :

$$f_X(x') = \int f_{X|Y}(x'|y) f_Y(y) dy$$

$$= \int f_{X|Y}(x'|y) \int f_{Y|X}(y|x) f_X(x) dx dy$$

$$= \int (\int f_{X|Y}(x'|y) f_{Y|X}(y|x) dy) f_X(x) dx$$

$$= \int K(x, x') f_X(x) dx$$

2.17.1 Exemple : Gibbs pour une loi normale bivariée) :

-On peut appliquer l'échantillonneur de Gibbs pour une loi normal bivariée $(X, Y) \sim \mathcal{N}(0, \Gamma)$, avec $\Gamma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.
-$f_{X,Y}(x, y) = \frac{1}{2\pi} \frac{1}{\sqrt{det\Gamma}} \exp(-\frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} \Gamma^{-1} \begin{pmatrix} x \\ y \end{pmatrix})$.; or $\Gamma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$.
-Donc $f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy))$.
- Comme $f_{X|Y=y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ et $Y \sim \mathcal{N}(0, 1)$, on trouve
$f_{X|Y=y}(x|y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}(x - \rho y)^2)$.
-Donc $X|Y = y \sim \mathcal{N}(\rho y, 1 - \rho^2)$ et $Y|X = x \sim \mathcal{N}(\rho x, 1 - \rho^2)$.
-Voici le code R :

```
r=0.25;n=10^4;X=NULL;Y=NULL;X[1]=rnorm(1,0,1-r^2);
for(t in 1:(n-1)){Y[t]=rnorm(1,r*X[t],1-r^2);
X[t+1]=rnorm(1,r*Y[t],1-r^2)};hist(X);windows();hist(Y);windows()
```

-Cette motivation de l'échantillonneur de Gibbs à deux étapes a commencé avec une distribution conjointe $f(x, y)$;
Cependant, ce que nous avons vu pour l'échantillonneur par tranches était exactement le contraire.

-Là, nous sommes partis d'une densité marginale $f_X(x)$ et avons construit (ou complété) une densité conjointe pour aider à la simulation où la deuxième variable Y (c'est-à-dire U dans l'échantillonnage par tranches) est une variable auxiliaire qui n'est pas directement pertinente du point de vue statistique.
-Il existe de nombreux cas où une complétion naturel de $f_X(x)$ en $f(x,y)$ existe.
-L'un de ces cas est le domaine des modèles de données manquants :

$$f(x|\theta) = \int_{\mathcal{Z}} g(x,z|\theta)$$

comme, par exemple, les mélanges de distributions .

2.18 Retour à l'échantillonneur par tranches :

-Tout comme l'échantillonneur de Gibbs à deux étapes peut être considéré comme un cas limite de l'échantillonneur par tranches dans un problème à trois coordonnées, l'échantillonneur par tranches peut être interprété comme cas particulier d'échantillonneur de Gibbs à deux étapes lorsque la distribution conjointe est la distribution uniforme sur le sous-graphe $S(f)$.
-De ce point de vue, l'échantillonneur par tranches commence par $f_X(x)$ et crée une densité commune $f(x,u) = 1_{(0<u<f_X(x))}$.
-Les densitées conditionnelles associées sont :

$$f_{X|U}(x|u) = \frac{1_{(0<u<f_X(x))}}{\int 1_{(0<u<f_X(x))dx}} \text{ et } f_{U|X}(u|x) = \frac{1_{(0<u<f_X(x))}}{\int 1_{(0<u<f_X(x))du}}$$

qui sont exactement ceux utilisés dans l'échantillonneur par tranches.
-Par conséquent, la suite X est aussi une chaîne de Markov avec noyau de transition :

$$K(x,x') = \int f_{X|U}(x'|u) f_{U|X}(u|x) du$$

et avec densité stationnaire $f_X(x)$.
-Ce que l'échantillonneur par tranches nous dit, c'est que nous pouvons induire un échantillonneur de Gibbs pour toute distribution marginale $f_X(x)$ en créant une distribution conjointe qui est, formellement, arbitraire.
-A partir de $f_X(x)$, on peut prendre n'importe quelle densité conditionnelle $g(y|x)$ et créez un échantillonneur Gibbs avec :

$$f_{X|Y}(x|y) = \frac{g(y|x) f_X(x)}{\int g(y|x) f_X(x) dx} \text{ et } f_{Y|X}(y|x) = \frac{g(y|x) f_X(x)}{\int g(y|x) f_X(x) dy}$$

2.19 Le théorème de Hammersley-Clifford :

-Une des caractéristiques les plus surprenantes de l'échantillonneur de Gibbs est que les distributions conditionnelles contiennent suffisamment d'informations pour produire un échantillon à partir de la distribution conjointe.
(C'est le cas pour l'échantillonneur de Gibbs à deux étapes et à plusieurs étapes)
-Par comparaison aux problèmes de maximisation, cette approche revient à maximiser une fonction objectif successivement dans toutes les directions d'une base donnée.
-Il est bien connu que cette méthode d'optimisation ne conduit pas nécessairement au maximum global, mais peut aboutir à un point de selle.
-Il est donc quelque peu remarquable que les distributions conditionnelles complètes résument parfaitement la densité conjointe, bien que l'ensemble des distributions marginales ne parvient manifestement pas à le faire.

-Le résultat suivant montre alors que la densité conjointe peut être directement et constructivement dérivée des densités conditionnelles .

2.19.1 Théoréme :

-La distribution conjointe associée aux densités conditionnelles $f_{Y|X}(y|x)$ et $f_{X|Y}(x|y)$ a la densité conjointe :

$$f(x,y) = \frac{f_{Y|X}(y|x)}{\int f_{Y|X}(y|x)/f_{X|Y}(x|y)dy}$$

-Cette dérivation de $f(x,y)$ nécessite évidemment l'existence et le calcul de l'intégrale $\int f_{Y|X}(y|x)/f_{X|Y}(x|y)dy$.
-Cependant, ce résultat démontre clairement le caractère fondamental caractéristique que les deux densités conditionnels sont suffisamment informatifs pour retrouver la densité conjointe.
-Notez également que ce théorème fait l'hypothèse implicite que la densité conjointe $f(x,y)$ existe.

2.20 Propriétés fondamentales :

-Une caractéristique particulièrement intéressante de l'échantillonneur de Gibbs à deux étages est que cet algorithme se prête à une étude composante par composante, car les séquences associées $(X^{(t)})$ et $(Y^{(t)})$ sont des chaînes de Markov.
-Cette décomposition en deux chaînes de Markov nous permet d'évaluer plus en profondeur les propriétés de l'Échantillonneur de Gibbs à 2 étapes .

2.21 Structures probabilistes :

-Une condition suffisante pour l'irréductibilité de la chaîne de Markov de Gibbs est la condition suivante, introduite par Besag (1974) .
Nous allons l'énoncez en toute généralité puisqu'elle s'applique également au cas général de l'échantillonneur de Gibbs .

2.21.1 Positivité :

-soit $(Y_1, ..., Y_p) \sim g(y_1, ..., y_p)$ ou $g^{(i)}$ est la distribution marginale de Y_i ;
si $g^{(i)}(y_i) > 0$ pour tout $i = 1, ..., p$ implique que $g(y_1, ..., y_p) > 0$, alors :
g satisfait la condition de positivité .

-Ainsi, le support de g est le produit cartésien des supports des $g^{(i)}$.
-De plus, il s'ensuit que les distributions conditionnelles ne réduiront pas la plage des valeurs possibles de Y_i par rapport à g.
-Dans ce cas, des sous-ensembles de Borel arbitraires du support peuvent être joints en une seule itération de l'échantilonneur de Gibbs à 2 étapes .

2.21.2 Irréductibilité forte :

-Chacune des séquences $(X^{(t)})$ et $(Y^{(t)})$ produites par l'échantillonneur de Gibbs est une Chaîne de Markov avec distributions stationnaires correspondantes :

$$f_X(x) = \int f(x, y)dy \text{ et } f_Y(y) = \int f(x, y)dx$$

-Si la contrainte de positivité sur f est vérifiée, alors les deux chaînes sont fortement irréductibles .

-Si l'interet est portée seulement sur la chaîne $(X^{(t)})$ et si la condition $f_{X|Y}(x|y) > 0$ est vraie pour tout couple (X', Y), l'irréductibilité est satisfaite.
-la chaîne "dual" $(Y^{(t)})$ peut être utilisé pour établir certaines propriétés probabilistes de $(X^{(t)})$.

2.21.3 Théoréme :

-Sous la condition de positivité, si le noyau de transition :

$$K((x, y), (x', y')) = f_{X|Y}(x'|y) f_{Y|X}(y'|x')$$

est absolument continue par rapport à la mesure dominante ,
la chaîne $(X^{(t)}, Y^{(t)})$ est Harris récurrente et ergodique et a pour distribution stationnaire f .

-Ce théorème implique la convergence pour la plupart des échantillonneurs de Gibbs à deux étapes puisque le noyau $K(x, y)$ sera absolument continu dans la plupart des configurations, et il n'y aura pas de points masse à craindre .
-Les illustrations typiques de l'échantillonneur de Gibbs à deux étapes sont dans les modèles à variable manquante , où une chaîne est généralement sur un espace d'état fini .

3 PARADIGME BAYÉSIEN

3.1 Classification :

- La classification fait partie des méthodes d'apprentissage supervisé qui est un type d'apprentissage machine .
-Ici le but est d'apprendre une cartographie à partir des entrées x aux sorties y, où $y \in \{1, ..., C\}$, C étant le nombre de classes.
-Si $C = 2$, il s'agit de classification binaire (auquel cas on suppose souvent $y \in \{0, 1\}$.
-Si $C > 2$, il s'agit de classification multiclasse .
-Si les étiquettes de classe ne sont pas mutuellement exclusives (par exemple, quelqu'un peut être classé comme grand et fort), il s'agit de classification multi-étiquettes, mais il est préférable de la considérer comme une prédiction à plusieurs étiquettes de classes binaires liées (un modèle dit à sorties multiples).
-Lorsque nous utilisons le terme "classification", nous entendrons une classification multiclasse avec une seule sortie .

-Une façon de formaliser le problème est une approximation de fonction.
-Nous supposons $y = f(x)$ pour une fonction inconnue f , et le but de l'apprentissage est d'estimer la fonction f étant donné un ensemble d'apprentissage étiqueté , puis de faire des prédictions en utilisant $\hat{y} = \hat{f}(x)$.
- Notre objectif principal est de faire des prédictions sur de nouvelles entrées, c'est-à-dire celles que nous avons pas vu auparavant (c'est ce qu'on appelle la généralisation), puisque prédire la réponse sur l'ensemble d'apprentissage est facile (nous pouvons simplement rechercher la réponse) .

3.1.1 Le besoin de prédictions probabilistes :

-Nous noterons la distribution de probabilité sur les étiquettes possibles, étant donné le vecteur d'entrée x et l'ensemble d'apprentissage $\mathcal{D}$ par $p(y|x, \mathcal{D})$.
-En général, cela représente un vecteur de longueur C ;
(S'il n'y a que deux classes, il suffit de retourner le nombre unique $p(y = 1|x, \mathcal{D})$, puisque $p(y = 1|x, \mathcal{D}) + p(y = 0|x, \mathcal{D}) = 1$.)

-Étant donné une sortie probabiliste, nous pouvons toujours calculer notre « meilleure

estimation » quant à la « véritable étiquette » en utilisant :

$$\hat{y} = \hat{f}(x) = argmax_{c=1}^{C} p(y = c|x, \mathcal{D})$$

-Cela correspond à l'étiquette de classe la plus probable, et s'appelle le mode de la distribution $p(y|x, \mathcal{D})$; il est également connu sous le nom d'estimation MAP (MAP signifie maximum a posteriori).
- Utiliser l'étiquette la plus probable a un sens intuitif.

-Considérons maintenant un cas , où $p(\hat{y}|x, \mathcal{D})$ est loin de 1.0 .
-Dans un tel cas où nous ne sommes pas très sûrs de notre réponse , il serait donc préférable de dire "je ne sais pas" à la place de retourner une réponse à laquelle nous n'avons pas vraiment confiance .
-Ceci est particulièrement important dans les domaines comme la médecine et la finance où nous pouvons avoir une aversion pour le risque .

3.1.2 Exemple : diagnostic médical :

-Comme exemple d'utilisation de cette règle, considérons le problème de diagnostic médical suivant.
- Supposons que vous êtes une femme dans la quarantaine et que vous décidez de passer un test médical pour le cancer du sein appelée mammographie.
-Si le test est positif, quelle est la probabilité que vous ayez un cancer ?
ça dépend évidemment de la fiabilité du test.
-Supposons qu'on vous dise que le test a une sensibilité de 80 %, ce qui signifie que si vous avez un cancer, le test sera positif avec une probabilité de 0.8.
-En d'autre termes : $p(x = 1|y = 1) = 0.8$.
où $x = 1$ est l'événement où la mammographie est positive, et $y = 1$ est l'événement où vous avez le cancer du sein .
-Beaucoup de gens concluent qu'ils sont donc 80% susceptibles d'avoir un cancer. Mais c'est faux !
Il ignore la probabilité a priori d'avoir un cancer du sein, heureusement assez faible : $p(y = 1) = 0.004$.
-Ignorer cet a priori s'appelle l'erreur du taux de base.
-Nous devons également tenir compte du fait que le test peut être un faux positif ou une fausse alerte.

- Malheureusement, ces faux positifs sont très probable (avec la technologie de dépistage actuelle) : $p(x = 1|y = 0) = 0.1$
-En combinant ces trois termes à l'aide de la règle de Bayes, nous pouvons calculer la bonne réponse comme suit :

$$p(y = 1|x = 1) = \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)}$$

$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

où $p(y = 0) = 1 - p(y = 1) = 0.996$
- En d'autres termes, si votre test est positif, vous n'avez qu'environ 3 % de (mal) chances d'avoir un cancer du sein .

3.1.3 Exemple : Classificateurs génératifs :

- Nous pouvons généraliser l'exemple du diagonostic médicale pour classer les vecteurs de caractéristiques x de type arbitraire comme suit :

$$p(y = c|x, \theta) = \frac{p(y = c|\theta)p(x|y = c, \theta)}{\sum_{c'} p(y = c'|\theta)p(x|y = c', \theta)}$$

-C'est ce qu'on appelle un classificateur génératif, car il spécifie comment générer les données à l'aide de la densité conditionnelle de class $p(x|y = c)$ et la classe a priori $p(y = c)$.

- Une approche alternative consiste à ajuster directement la classe a posteriori, $p(y = c|x)$; c'est ce qu'on appelle un classificateur discriminant .

3.2 Modèles génératifs pour données discrètes :

- Nous avons discuté de la façon de classer un vecteur de caractéristiques x en appliquant la règle de Bayes à un classificateur génératif de la forme :

$$p(y = c|x, \theta) \propto p(x|y = c, \theta)p(y = c|\theta)$$

-La clé de l'utilisation de tels modèles est de spécifier une forme appropriée pour la densité conditionnelle de classe $p(x|y = c, \theta)$, qui définit le type de données que nous nous

attendons à voir dans chaque classe.
- Ici nous nous concentrons sur le cas où les données observées sont des symboles discrets.
-Nous discutons également de la manière de déduire les paramètres inconnus θ de tels modèles .

3.3 Apprentissage du concept bayésien :

- Considérez comment un enfant apprend à comprendre le sens d'un mot, tel que "chien".
-Probablement les parents de l'enfant signalent des exemples positifs de ce concept, en disant des choses telles que "regarde le chien mignon !", ou "attention au chien", etc.
-Cependant, il est très peu probable qu'ils fournissent des informations négatives ; par exemples, en disant "regarde ce non-chien".
-Certes, des exemples négatifs peuvent être obtenus lors d'un processus d'apprentissage actif - l'enfant dit "regarde le chien" et le parent dit "c'est un chat, chéri, pas un chien "- mais la recherche psychologique a montré que les gens peuvent apprendre des concepts à partir de seuls exemples positifs .

- Nous pouvons penser que l'apprentissage du sens d'un mot équivaut à l'apprentissage d'un concept, qui à son tour équivaut à une classification binaire.
- Pour voir cela, définissez $f(x) = 1$ si x est un exemple de concept C, et $f(x) = 0$ sinon.
- Ensuite, le but est d'apprendre la fonction indicatrice f, qui définit juste quels éléments sont dans l'ensemble C .
- En tenant compte de l'incertitude sur la définition de f, ou de manière équivalente les éléments de C , nous pouvons émuler la théorie des ensembles flous, mais en utilisant le calcul de probabilité.
- Notez que les techniques de classification binaire standard nécessitent des exemples négatifs.
- En revanche, nous allons concevoir un moyen d'apprendre uniquement à partir d'exemples positifs .

- À des fins pédagogiques, nous allons considérer un exemple très simple d'apprentissage de concept appelé le jeu des nombres, basé sur une partie de la thèse de doctorat de Josh Tenenbaum (Tenenbaum 1999).
- Le jeu procède comme suit ;
Je choisis un concept arithmétique simple C , tel que "nombre premier" ou "un nombre entre 1 et 10" ;

Je vous donne ensuite une série d'exemples positifs choisis au hasard $\mathcal{D} = \{x_1, ..., x_N\}$ tiré de C, et vous demande si un nouveau cas de test $\tilde{x}$ appartient à C , c'est-à-dire que je vous demande de classer $\tilde{x}$.

-Supposons, pour simplifier, que tous les nombres soient des nombres entiers compris entre 1 et 100.
-Supposons maintenant que je vous dise "16" est un exemple positif du concept.
-Selon vous, quels autres chiffres sont positifs ? 17 ? 6 ? 32 ? 99 ? C'est difficile à dire avec un seul exemple, donc vos prédictions seront assez vagues.
- Vraisemblablement, les nombres qui sont similaires dans un certain sens à 16 sont plus probables.
-Mais semblable en quoi ? 17 est similaire, car il est "proche", 6 est similaire car il a un chiffre en commun, 32 est similaire car il est également pair et une puissance de 2, mais 99 ne semble pas similaire.
- Ainsi certains chiffres sont plus probables que d'autres.
- Nous pouvons représenter cela comme une distribution de probabilité, $p(\tilde{x}|\mathcal{D})$, qui est la probabilité que $\tilde{x} \in C$ étant donné les données $\mathcal{D}$ pour tout $\tilde{x} \in \{1, ..., 100\}$;
il s'agit de la distribution prédictive a posteriori .

- Supposons maintenant que je vous dise que 8, 2 et 64 sont aussi des exemples positifs.
- Maintenant, vous pouvez deviner que le concept caché est "puissance de deux".
- Ceci est un exemple d'induction.
-Compte tenu de cette hypothèse, la distribution prédictive est assez spécifique et met la majeure partie de sa masse sur des puissances de 2 .

- Si à la place je vous dis que les données sont $\mathcal{D} = \{16, 23, 19, 20\}$, vous obtiendrez un différent type de gradient de généralisation .

- Nous avons ci-dessous le résultat de la distribution prédictive empirique de l'expérience sur 8 humains , les Deux premieres rangées : après avoir vu $\mathcal{D} = \{16\}$ et $\mathcal{D} = \{60\}$. Cela illustre une similarité diffuse. la Troisième rang : après avoir vu $\mathcal{D} = \{16, 8, 2, 64\}$. Cela illustre un comportement de type règle (puissances de 2). la Rangée du bas : après avoir vu $\mathcal{D} = \{16, 23, 19, 20\}$. Cela illustre une similarité focalisée (nombres proches de 20).
- Comment expliquer ce comportement et l'imiter dans une machine ?
- L'approche classique de l'induction consiste à supposer que nous avons un espace d'hypothèses de concepts, $\mathcal{H}$, tel que :
les nombres impairs, les nombres pairs, tous les nombres compris entre 1 et 100, les

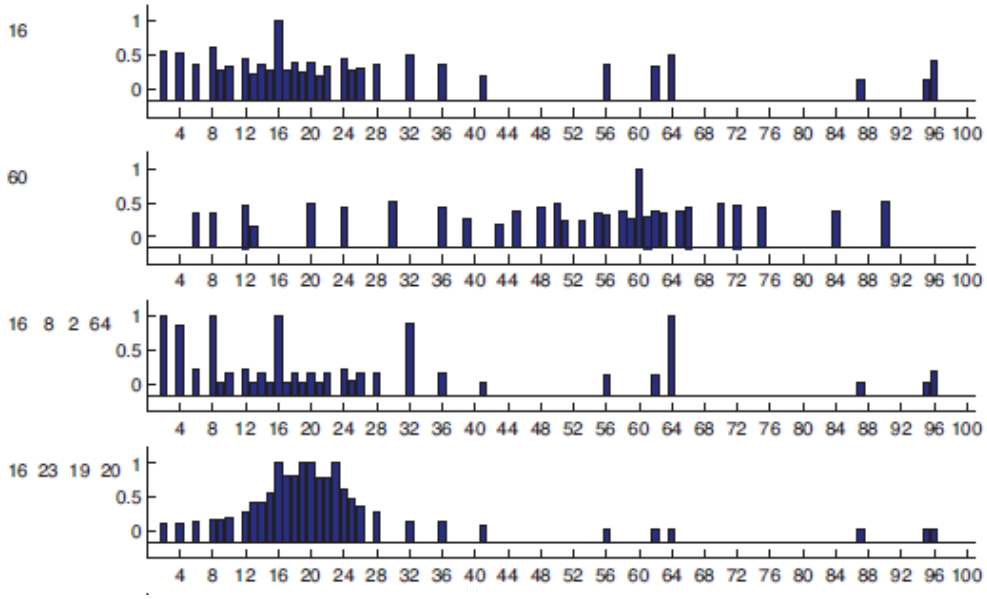

FIGURE 3.1 – distribution prédictive empirique en moyenne sur 8 humains dans le jeu des nombres.

puissances de deux, tous les nombres se terminant par j (pour $0 \leq j \leq 9$), etc.
-Le sous-ensemble de $\mathcal{H}$ qui est cohérent avec les données $\mathcal{D}$ est appelé l'espace des versions.
- Au fur et à mesure que nous voyons plus d'exemples, l'espace de version se rétrécit et nous devenons de plus en plus certains du concept .

- Cependant, l'espace de version ne résume pas toute l'histoire .
- Après avoir vu $\mathcal{D} = \{16\}$, il y a plusieurs règles cohérentes ;
comment les combine-t-on pour prédire si $\tilde{x} \in C$?
Aussi, après avoir vu $\mathcal{D} = \{16, 8, 2, 64\}$, pourquoi avez-vous choisi la règle "puissances de deux" et non, disons, "tous les nombres pairs", ou « puissances de deux sauf 32 », les deux étant également compatibles avec la preuve ?
- Nous allons donner maintenant une explication bayésienne à cela .

3.4 Vraisemblance :

- Nous devons expliquer pourquoi nous avons choisi h_{deux} = "puissances de deux" et non, disons, h_{pair} = "nombres pairs" après avoir vu $\mathcal{D} = \{16, 8, 2, 64\}$, étant donné que les deux hypothèses sont cohérentes avec les preuves.
- L'intuition clé est que nous voulons éviter les coïncidences suspectes.
- Si le vrai concept était des nombres pairs, comment se fait-il que nous n'ayons vu que

des nombres qui se trouvaient être des puissances de deux ?

- Pour formaliser cela, supposons que les exemples sont échantillonnés uniformément au hasard parmi les prolongement d'un concept .
(L'extension d'un concept n'est que l'ensemble des nombres qui appartiennent à lui, par exemple, l'extension de h_{pair} est $\{2, 4, 6, ..., 98, 100\}$; l'extension des "numéros se terminant dans par 9" est $\{9, 19, ..., 99\}$.)
- Tenenbaum appelle cela l'hypothèse d'échantillonnage forte.
- Compte tenu de cette hypothèse, la probabilité d'échantillonner indépendamment N éléments (avec remise) à partir de h est donné par :

$$p(\mathcal{D}|h) = (\frac{1}{taille(h)})^N = (\frac{1}{|h|})^N$$

- Cette équation cruciale incarne ce que Tenenbaum appelle le principe de taille, ce qui signifie que le modèle privilégie l'hypothèse la plus simple (la plus petite) compatible avec les données.
- C'est plus communément connu sous le nom de rasoir d'Occam .

- Pour voir comment cela fonctionne, laissez $\mathcal{D} = \{16\}$;
Alors $p(\mathcal{D}|h_{deux}) = 1/6$, puisqu'il n'y a que 6 puissances de deux inférieur à 100, mais $p(\mathcal{D}|h_{pair}) = 1/50$, puisqu'il y a 50 nombres pairs .
- Alors la probabilité que $h = h_{deux}$ est plus élevée que $h = h_{pair}$.
- Après 4 exemples, la vraisemblance de h_{deux} est $(1/6)^4 = 7.7 \times 10^4$, alors que la probabilité de h_{pair} est $(1/50)^4 = 1.6 \times 10^7$.
-C'est un rapport de vraisemblance de presque 5000 pour 1 en faveur de h_{deux} .
- Cela quantifie notre intuition antérieure selon laquelle $\mathcal{D} = \{16, 8, 2, 64\}$ serait une coïncidence très suspecte si elle était générée par h_{pair} .

3.5 a priori

- Supposons que $\mathcal{D} = \{16, 8, 2, 64\}$.
- Compte tenu de ces données, le concept h' = "puissances de deux sauf 32" est plus probable que h = "puissances de deux", puisque h' n'a pas besoin d'expliquer la coïncidence que 32 manque dans la série d'exemples.
- Cependant, l'hypothèse h' = "puissances de deux sauf 32" semble "conceptuellement non naturelle".

- Nous pouvons capturer une telle intution en attribuant une faible probabilité a priori à des concepts non naturels.
- Bien sûr, votre a priori pourrait être différent du mien.
- Cet aspect subjectif du raisonnement bayésien est une source de nombreuses controverses, puisqu'elle signifie, par exemple, qu'un enfant et un professeur de mathématiques arriveront à des réponses différentes.
- En fait, ils ont vraisemblablement non seulement des antécédents différents, mais aussi différents espaces d'hypothèses.
- Cependant, nous pouvons affiner cela en définissant que l'espace d'hypothèse de l'enfant et le professeur de mathématiques soient les mêmes, puis en fixant le poids antérieur de l'enfant à zéro sur certains concepts "avancés".
- Il n'y a donc pas de distinction nette entre l'antérieur et l'espace des hypothèses .

- Bien que la subjectivité de l'apriori soit controversée, elle est en fait très utile.
- Si on vous dit que les nombres proviennent d'une règle arithmétique, puis étant donné 1200, 1500, 900 et 1400, vous pouvez pensez que 400 est probable mais 1183 est peu probable.
- Mais si on vous dit que les chiffres sont des exemples de taux de cholestérol sain, vous penseriez probablement que 400 est peu probable et 1183 est probable.
- Ainsi on peut voir que l'a priori est le mécanisme par lequel les connaissances de base peuvent être amenées à influer sur un problème.
- Sans cela, un apprentissage rapide (c'est-à-dire à partir de petits échantillons) est impossible .

- Alors, quel a priori devrions-nous utiliser ?
- A titre d'illustration, utilisons un a priori simple qui attribue une probabilité uniforme à 30 concepts arithmétiques simples, tels que "nombres pairs", "nombres impairs". , " nombres premiers ", " nombres se terminant par 9 " , etc.
- Pour rendre les choses plus intéressantes, nous rendrons les concepts pairs et impairs plus probables a priori.
- Nous incluons également deux concepts "non naturels", à savoir " puissances de 2, plus 37 " et " puissances de 2, sauf 32 " , mais en leur donnant un poids a priori faible .

3.6 a posteriori

- L'apostériori est simplement la vraisemblance multipliée par l'apriori , normalisé.

- Dans ce contexte nous avons :

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h'} p(\mathcal{D}|h')p(h')} = \frac{p(h)\mathbb{1}(\mathcal{D} \in h)/|h|^N}{\sum_{h' \in \mathcal{H}} p(h')\mathbb{1}(\mathcal{D} \in h')/|h'|^N}$$

où $\mathbb{1}(\mathcal{D} \in h)$ vaut 1 ssi (si et seulement si) toutes les données sont dans le prolongement de l'hypothèse h .

- Nous voyons que l'a posteriori est une combinaison de l'a priori et de la vraisemblance.
- Dans le cas de la plupart des concepts, l'a priori est uniforme, donc la postériori est proportionnelle à la vraisemblance.
- Cependant, les concepts " non naturels " de " puissances de 2, plus 37 " et " puissances de 2, sauf 32 " ont un faible soutien postérieur, malgré leur forte vraisemblance, en raison du faible a priori.
- Inversement, le concept de nombres impairs a un faible soutien postérieure , malgré un a priori élevé, en raison de la faible vraisemblance .

- Nous avons ci-dessous les résultats de l'aposteriori de quelques exemples :
- En général, lorsque nous avons suffisamment de données, l'a posteriori $p(h|\mathcal{D})$ culmine sur un seul concept, à savoir l'estimation MAP, c'est-à-dire :

$$p(h|\mathcal{D}) \to \delta_{\hat{h}^{MAP}}(h)$$

ou $\hat{h}^{MAP} = argmax_h p(h|\mathcal{D})$ est le mode a posteriori et ou δ est la mesure de dirac .

- Notez que l'estimation MAP peut être écrite comme :

$$\hat{h}^{MAP} = argmax_h p(\mathcal{D}|h)p(h) = argmax(\log(p(\mathcal{D}|h)) + \log(p(h)))$$

Puisque le terme de vraisemblance dépend exponentiellement de N, et que l'a priori reste constant, à mesure que nous obtenons de plus en plus de données, l'estimation MAP converge vers l'estimation du maximum de vraisemblance ou MLE :

$$\hat{h}^{MLE} = argmax_h p(\mathcal{D}|h) = argmax \log(p(\mathcal{D}|h))$$

- En d'autres termes, si nous avons suffisamment de données, nous voyons que les données submerge l'apriori.
- Dans ce cas, l'estimation MAP converge vers le MLE .

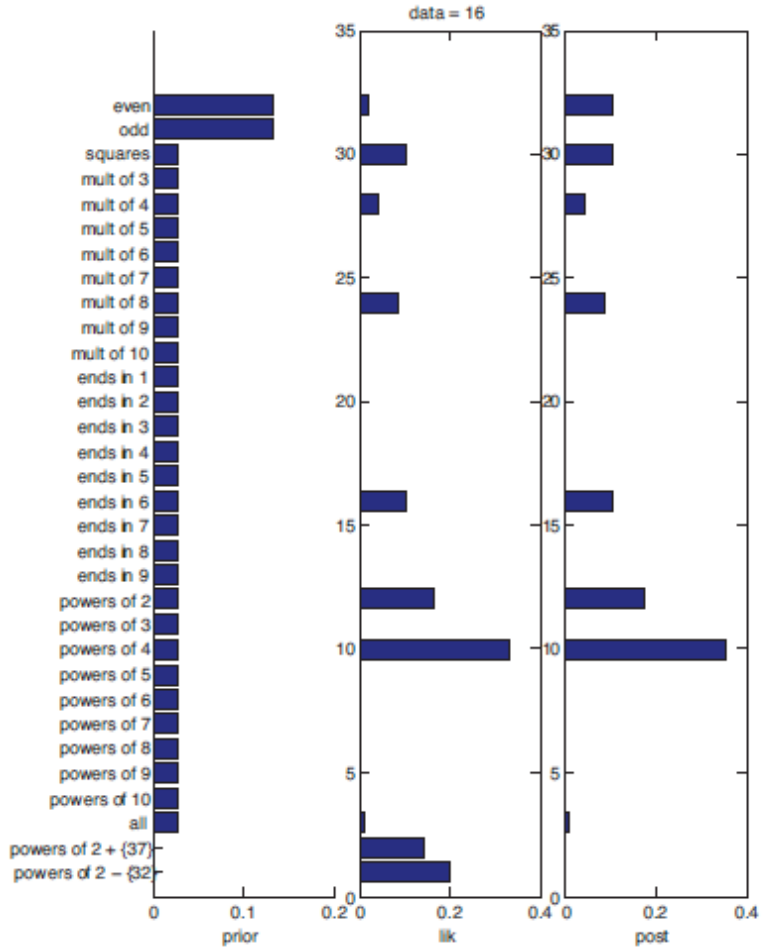

FIGURE 3.2 – A priori, vraisemblance et a posteriori pour $\mathcal{D} = \{16\}$

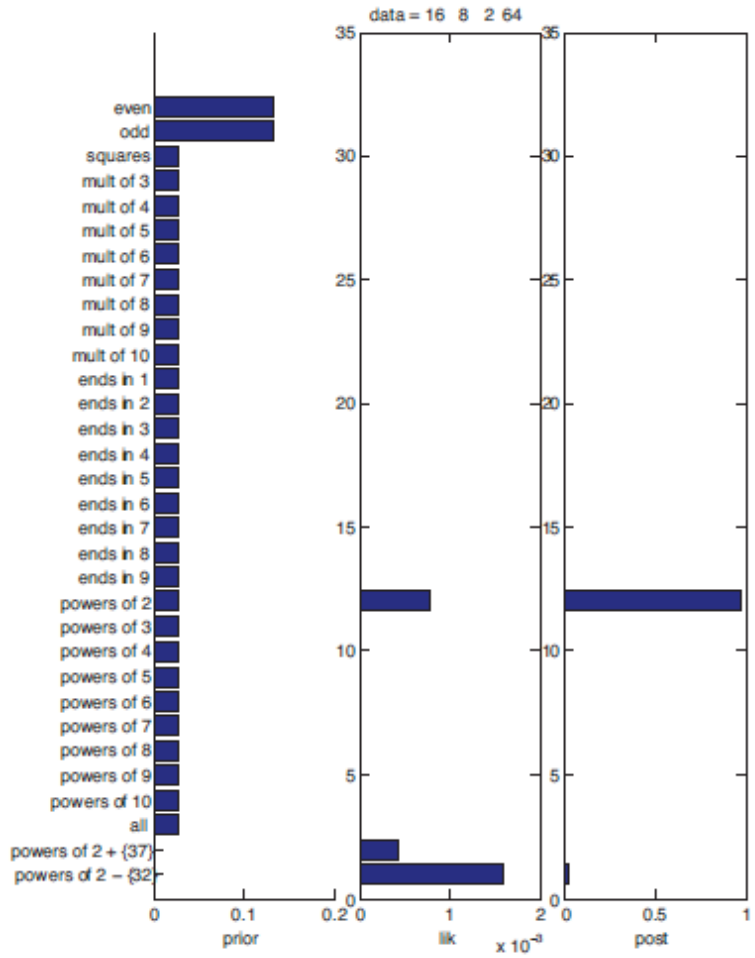

FIGURE 3.3 – A priori, vraisemblance et a posteriori pour $\mathcal{D} = \{16, 8, 2, 24\}$

- Si la vraie hypothèse est dans l'espace des hypothèses, alors l'estimation MAP/ML convergera sur cette hypothèse.
- Ainsi, nous disons que l'inférence bayésienne (et l'estimation ML) est une estimation cohérente .
- On dit aussi que l'espace des hypothèses est identifiable dans la limite, ce qui signifie que nous pouvons retrouver la vérité dans la limite des données infinies.
- Si notre classe d'hypothèses n'est pas assez riche pour représenter la "vérité" (ce qui sera généralement le cas), nous convergerons sur l'hypothèse la plus proche possible de la vérité .

3.7 Distribution prédictive postérieure :

- L'a postériori est notre état de croyance interne sur le monde .
- La façon de tester si nos croyances sont justifié est de les utiliser pour prédire des quantités objectivement observables (c'est la base de la méthode scientifique).
- Plus précisément, la distribution prédictive a posteriori dans ce contexte est donnée par :

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(y = 1|\tilde{x}, h)p(h|\mathcal{D})$$

Ceci est juste une moyenne pondérée des prédictions de chaque hypothèse individuelle et s'appelle Moyenne du modèle de Bayes .

- Lorsque nous avons un jeu de données petit et/ou ambigu, l'a posteriori $p(h|\mathcal{D})$ est vague, ce qui induit une large distribution prédictive .
- Cependant, une fois que nous avons "compris les choses", l'a postériori devient une fonction delta centrée sur l'estimation MAP.
- Dans ce cas, la distribution prédictive devient :

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(\tilde{x}|h)\delta_{\hat{h}}(h) = p(\tilde{x}|\hat{h})$$

- C'est ce qu'on appelle une approximation de branchement de la densité prédictive et elle est très largement utilisée, en raison de sa simplicité.

- Bien que l'apprentissage MAP soit simple, il ne peut pas expliquer le passage progressif d'une approche basée sur le raisonnement de similarité (avec des a postérioris incertains) au raisonnement fondé sur des règles (avec des a postérioris certains).
- Par exemple, supposons que nous observons $\mathcal{D} = \{16\}$. Si nous utilisons l'a priori simple

ci-dessus, l'hypothèse minimale cohérente est "toutes les puissances de 4", donc seuls 4 et 16 ont une probabilité non nulle d'être prédit.
- Ceci est bien sûr un exemple de surajustement.
- Étant donné $\mathcal{D} = \{16, 8, 2, 64\}$, l'hypothèse MAP est "toutes puissances de deux".
- Ainsi, la distribution prédictive du branchement s'élargit (ou reste la même) à mesure que nous voyons plus de données :
il commence étroit, mais est obligé de s'élargir à mesure qu'il voit plus de données.
- En revanche, dans l'approche bayésienne, nous commençons large puis nous rétrécissons au fur et à mesure que nous en apprenons plus, ce qui est plus intuitif.
- En particulier, étant donné $\mathcal{D} = \{16\}$, il existe de nombreuses hypothèses avec un appui a postériori non négligeable, la distribution prédictive est donc large.
- Cependant, lorsque ont voit $\mathcal{D} = \{16, 8, 2, 64\}$, l'a postériori concentre sa masse sur une hypothèse, donc la distribution prédictif devient plus étroite.
- Ainsi, les prédictions faites par une approche de branchement et une méthode bayésienne sont assez différentes dans le régime de petit échantillon, bien qu'elles convergent vers la même réponse lorque nous voyons plus de données .
- Le jeu de nombres impliquait de déduire une distribution sur une variable discrète tirée d'un espace d'hypothèse fini , $h \in \mathcal{H}$, étant donné une série d'observations discrètes.
- Cela a rendu les calculs particulièrement simple :
il suffisait de faire la somme, de multiplier et de diviser.
- Cependant, dans de nombreuses applications, les paramètres inconnus sont continus, donc l'espace d'hypothèse est (un sous-ensemble) de R^K , où K est le nombre de paramètres.
- Cela complique les mathématiques, puisqu'il faut remplacer sommes avec intégrales.
- Cependant, les idées de base sont les mêmes .

3.8 résumé

- On a vu un exemple d'utilisation du paradigme bayésien dans l'apprentissage automatique à savoir l'apprentissage du concept bayésien .
- L'utilisation du paradigme bayésien dans l'apprentissage automatique est beaucoup plus vaste ; en effet , on la retrouve dans :
le classificateur naive bayésien ,
les réseaux bayésiens .
- Mais ce qui est plus fondamental c'est que le paradigme bayésien est avant tout une philosophie qui peut s'appliquer dans beaucoup de cas , notamment en apprentissage automatique .

- Par exemple une variante de l'apprentissage approfondie est l'apprentissage approfondie bayésien ou on utilise des réseaux de neurones bayésiens , c'est à dire des réseaux de neurones ou les poids sont des variables aléatoires qui ont une distribution de probabilité a priori .
- On trouve aussi : la régression logistique bayésienne et la régression bayésienne ou le principe est le meme :
les paramétres inconnus du modéle sont des variables aléatoires .
- La théorie des probabilités peut être appliquée à tout problème impliquant une incertitude.
- En apprentissage automatique , l'incertitude se présente sous plusieurs formes : quelle est la meilleure prédiction (ou décision) compte tenu de données ? quel est le meilleur modèle compte tenu de certaines données ? quelle mesure dois-je effectuer ensuite ? etc.

- L'application systématique du raisonnement probabiliste à tous les problèmes inférentiels, y compris l'inférence des paramètres des modèles statistiques, est parfois appelée une approche bayésienne .

- D'autre part le paradigme bayésien a une histoire extrément liée aux méthodes mcmc , car bien qu'initialement la motivation et l'utilisation de ces méthodes était la physique statistique depuis relativement longtemps , les méthodes mcmc ne sont devenus populaires que récemment d'une part grace aux develeppement des ordinateurs et de la puissance de calcul et d'autre part au vu de leur application dans la statistique bayésienne .

- Parmi les exemples de méthodes d'apprentissage automatique traités dans cet exposé on a :
la régression logistique et la régression qui sont toutes les 2 traités d'un point de vue bayésien et auxquels on appliquera des méthodes mcmc .

4 RÉGRESSION LOGISTIQUE

4.1 Introduction à la régression :

-De façon informelle, un modéle explicatif est un modéle exprimant une variable $\mathcal{Y}$, appelée variable à expliquer (ou réponse), comme une fonction d'une ou de plusieurs variables dites variables explicatives ou prédicteurs .
- Toutefois si l'entité $\mathcal{Y}$ est considérée comme une variable aléatoire Y, un terme aléatoire, caractérisant l'incertitude de la prédiction, doit etre introduit d'une certaine façon dans l'équation du modéle .
- Dans un modéle de régression, on cherche essentiellement à déterminer la variation de l'espérance mathématique de Y en fonction des variables explicatives.
- En d'autres termes on étudie comment Y évolue "en moyenne" en fonction de ces variables explicatives.

- De plus cette entité explicative, que nous symboliserons par la lettre $\mathcal{X}$, sera une variable quantitative , pouvant prendre toute valeur dans un intervalle I de R .
- Aux différentes valeurs de $\mathcal{X}$ dans I correspondent, par hypothése, des v.a. distinctes et on est donc, en fait, en présence d'une famille de v.a. $\{Y(x)|x \in I\}$.
-Admettant que pour tout x l'espérance mathématique existe, alors $E(Y(x))$ est la fonction $g(x)$ qu'il s'agit de rechercher.
- Cette fonction mettant en évidence l'évolution moyenne de l'entité $\mathcal{Y}$ à expliquer en fonction de x est appelée fonction de régression .
- Dans cette approche on considére naturellement que l'incertitude de la prédiction de Y pour le "niveau" x de $\mathcal{X}$, se manifeste par une v.a. $\epsilon(x)$ venant s'ajouter à la composante déterministe $g(x)$.
- Dans sa forme la plus générale un modéle de régression simple s'écrit donc :

$$Y(x) = g(x) + \epsilon(x)$$

- Puisque $E(Y(x)) = g(x)$, on a nécessairement $E(\epsilon(x)) = 0$, quel que soit x .
- La V.a. $\epsilon(x)$ est appelée erreur ou aléa (d'ou la notation habituelle du "e" grec).
- Dans la plupart des modèles on suppose que l'erreur est de meme loi quel que soit x ce qui permet d'écrire $Y(x) = g(x) + \epsilon$.
(on écrit meme parfois simplement $Y = g(x) + \epsilon$ en omettant d'indiquer que la v.a. Y est assujette à la valeur x) .

4.2 La régression logistique :

-Une façon de construire un classificateur probabiliste est de créer un modèle conjoint de la forme $p(y, x)$ et puis de conditionner sur x , dérivant ainsi $p(y|x)$.
-C'est ce qu'on appelle l'approche générative .
-Une approche alternative consiste à ajuster directement un modèle de la forme $p(y|x)$.
-C'est ce qu'on appelle l'approche discriminente , et c'est l'approche qu'on va décrire .
-En particulier, nous supposerons que les modèles discriminatifs seront linéaires dans leurs paramètres.
-Cela simplifiera considérablement l'ajustement du modèle

4.2.1 le modéle de régression loigstique :

- Ce modéle est adapté au cas ou la variable à expliquer est binaire.
- En utilisant le codage 1/0 on la transforme en variable aléatoire de Bernoulli.
- Plus précisément, dans le formalisme conditionnel exposé précédemment , la loi de Y sachant $X = x$ est une loi $\mathcal{B}(p(x))$.
- La fonction de régression à estimer est donc :

$$E(Y|X = x) = p(x) \text{ ou } p(x) = P(Y = 1|X = x)$$

- Plus prosaiquement, le probléme est de déterminer comment la probabilité de «succés» évolue en fonction du niveau de la variable X .

- Ce modéle stipule que la probabilité conditionnelle de succés est de la forme :

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

- La fonction :

$$g(u) = \frac{e^u}{1 + e^u}$$

est appelée fonction logistique, elle est strictement croissante et prend ses valeurs dans l'intervalle $[0, 1]$.
- Sa fonction inverse est :

$$g^{-1}(u) = \log(\frac{u}{1-u})$$

et s'appelle fonction logit .

- Il existe d'autre fonctions dont le graphe présente une forme sigmoidale et qui sont

candidates pour modéliser le modéle , parmi eux :
la fonction probit : g^{-1} est alors la fonction inverse de la fonction de répartition d'une loi normale , et donc son expression n'est pas explicite .
-Plusieurs raison , tant théoriques que pratiques , font préférer la fonction logit .

- Pour une loi de Bernoulli $\mathcal{B}(p)$ le rapport $\frac{p}{1-p}$ a une certaine signification .
- On l'appelle parfois la chance ou la cote de succés (en anglais : odds).
- Dans le modéle logistique le logarithme de ce rapport est donc une fonction linéaire de la variable explicative :

$$\log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x$$

- Le modéle comporte donc deux paramétres inconnus β_0 et β_1 .
- On notera par β le couple (β_0, β_1) ou, indifféremment, le vecteur $(\beta_0, \beta_1)^t$.
- Contrairement à la régression classique il n'y a pas de variance de l'erreur à estimer puisqu'une loi de Bernoulli $\mathcal{B}(p(x))$ ne dépend que du paramétre $p(x)$.

- Supposons que nous observions indépendamment les v.a. binaires $Y_1, Y_2, ..., Y_n$ aux points $x_1, x_2, ..., x_n$ de la variable explicative .
- Pour tout i , $Y_i \sim \mathcal{B}(p(x_i))$ et la fonction de probabilité de Y_i est :

$$p(y) = p(x_i)^y (1 - p(x_i))^{1-y} ; y \in \{0, 1\}$$

- La fonction de vraisemblance de β associée à une réalisation $(y_1, y_2, ..., y_n)$ de $(Y_1, Y_2, ..., Y_n)$ est donc :

$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

avec :

$$p(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

- En régression logistique standard on cherche l'estmateur de maximum de vraisemblance
- Les équations qui déterminenent cet estimateur n'ont pas de solution explicite .
- Par contre vu que l'on traitera une régression logistique bayésienne , on aura pas besoin de résoudre ces équations , puisqu'on attribura une loi a priori aux paramétres β_0 et β_1 et on cherchera a simuler la loi a posteriori de ces paramétres qui sera à une constante multiplicative prés le produit de l'apriori et de la vraisemblance .

Flight	14	9	23	10	1	5	13	15	4	3	8	17	2	11	6	7	16	21	19	22	12	20	18
Failure	1	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
Temp.	53	57	58	63	66	67	67	67	68	69	70	70	70	70	72	73	75	75	76	76	78	79	81

FIGURE 4.1 – Température au moment du vol et défaillance des joints toriques

4.2.2 Exemple : régression logistique :

-Un modèle de régression utile pour les réponses binaire (0 -1) est le modèle logit, où la distribution de Y conditionnellement aux variables explicatives (ou dépendantes) $x \in R^p$ est modélisées par la relation
$P(Y = 1) = p = \frac{\exp(\alpha+x\beta)}{1+\exp(\alpha+x\beta)}$.
-De manière équivalente, la transformée logit de p, $logit(p) = \log(\frac{p}{1-p})$, satisfait la relation linéaire : $logit(p) = \alpha + x\beta$.
-En 1986, la navette spatiale Challenger explose au décollage, tuant les sept astronautes à bord.
- L'explosion a été le résultat d'une défaillance du joint torique, une rupture d'un anneau de caoutchouc qui scelle les parties de la navette ensemble.
- On pense que l'accident a été causé par le temps exceptionnellement froid (0 C) au moment du lancement, car il y a des raisons de croire que les probabilités de la défaillance du joint torique augmentent à mesure que la température diminue .
-Il est raisonnable d'ajuster une régression logistique,
avec p = probabilité d'une défaillance du joint torique et $x = temprature.$
- Nous avons ci-dessus une base de Données sur les lancements précédents de la navette spatiale et les défaillance du joint torique (1 en cas d'échec, 0 en cas de réussite) . -On observe (x_i, y_i), $i = 1, ..., n$ selon le modèle
$Y_i \sim Bernoulli(p(x_i))$, $p(x) = \frac{\exp(\alpha+x\beta)}{1+\exp(\alpha+x\beta)}$.
où $p(x)$ est la probabilité de défaillance d'un anneau torique à la température x.
-La vraisemblance est
$L(\alpha, \beta|y) \propto \prod_{i=1}^{n} (\frac{\exp(\alpha+x_i\beta)}{1+\exp(\alpha+x_i\beta)})^{y_i} (\frac{1}{1+\exp(\alpha+x_i\beta)})^{1-y_i}$
-Et on prend l'a priori $\pi_\alpha(\alpha|b)\pi_\beta(\beta) = \frac{1}{b}\exp(\alpha)\exp(-\frac{\exp(\alpha)}{b})d\alpha d\beta$.
qui met un a priori exponentiel sur $\log(\alpha)$ et un a priori plat sur β .
-Pour compléter la spécification de l'apriori, nous devons donner une valeur pour b, et nous choisissons la valeur dépendante des données qui fait $E(\alpha) = \hat{\alpha}$ où $\hat{\alpha}$ est l'estimateur de maximum de vraisemblance de α .
-on peut montrer que
$E(\alpha) = \int_0^\infty \frac{1}{b}\exp(\alpha)\exp(-\frac{\exp(\alpha)}{b})d\alpha = \int_0^\infty \log(w)\frac{1}{b}\exp(-\frac{w}{b})dw = \log(b) - \gamma$.
où γ est la constante d'Euler, égale à 0.577216. Ainsi nous prenons $\hat{b} = \exp(\hat{\alpha} + \gamma)$.

-La distribution a posteriori est proportionnelle à $L(\alpha,\beta|y)\pi(\alpha,\beta)$, et pour simuler cette distribution on prend un candidat indépendant
$g(\alpha,\beta) = \pi_\alpha(\alpha|\hat{b})\phi(\beta)$.
où $\phi(\beta)$ est une distribution normale de moyenne $\hat{\beta}$ et variance $\hat{\sigma}^2_\beta$.
-Générer une variable aléatoire à partir de $g(\alpha,\beta)$ est simple, car cela implique la génération d'une normale et d'une variable aléatoire exponentielle.
Si nous sommes au point (α_0,β_0) de la chaîne de Markov, et on génère (α',β') à partir de $g(\alpha,\beta)$, on accepte le candidat avec probabilité
$min\{\frac{L(\alpha',\beta'|y)}{L(\alpha_0,\beta_0|y)}\frac{\phi(\beta_0)}{\phi(\beta')},1\}$.
- On utilise la fonction glm de R pour trouver les valeur de $\hat{\alpha}$ et $\hat{\beta}$.
-voici le code R :

```
y=c(rep(1,4),rep(0,8),rep(1,2),rep(0,3),1,rep(0,5));
x=c(53,57,58,63,66,67,67,67,68,69,70,70,70,70,72,73,75,75,76,
76,78,79,81);n=length(y)
glm(y~x,family=binomial(link="logit"));
L=function(alpha,beta){return(
prod(((exp(alpha+beta*x)/(1+exp(alpha+beta*x)))^(y))*
((1/(1+exp(alpha+beta*x)))^(1-y))))};
apr=function(alpha,beta,b){return(
(1/b)*exp(alpha)*exp(-exp(alpha)/b))};alphachap=15.0429;
betachap=-0.2322;sigma2betachap=2.27;bchap=exp(alphachap+0.577216)
f=function(alpha,beta,b){return(L(alpha,beta)*apr(alpha,beta,b))}
g=function(alpha,beta){return(
apr(alpha,beta,bchap)*dnorm(beta,betachap,sqrt(sigma2betachap)))}
rho=function(A,B,U,V,b){return((f(U,V,b)/f(A,B,b))*(g(A,B)/g(U,V)))}
N=50000;U=NULL;V=NULL;A=NULL;B=NULL;A[1]=log(rexp(1,1/bchap));
B[1]=rnorm(1,betachap,sqrt(sigma2betachap))
for(i in 1:(N-1)){u=runif(1);U[i]=log(rexp(1,1/bchap))
V[i]=rnorm(1,betachap,sqrt(sigma2betachap));
A[i+1]=A[i]+(U[i]-A[i])*as.numeric(u<
rho(A[i],B[i],U[i],V[i],bchap));
B[i+1]=B[i]+(V[i]-B[i])*as.numeric(u<
rho(A[i],B[i],U[i],V[i],bchap))}
sum(y*log(exp(U[69]+V[69]*x)/(1+exp(U[69]+V[69]*x)))+
(1-y)*log(1/(1+exp(U[69]+V[69]*x))));hist(A,nclass=50)
windows();hist(B,,xlim=c(-0.3,-0.1),nclass=150)
MA=cumsum(A)/1:N;MB=cumsum(B)/1:N;plot(MA,type="s");
```

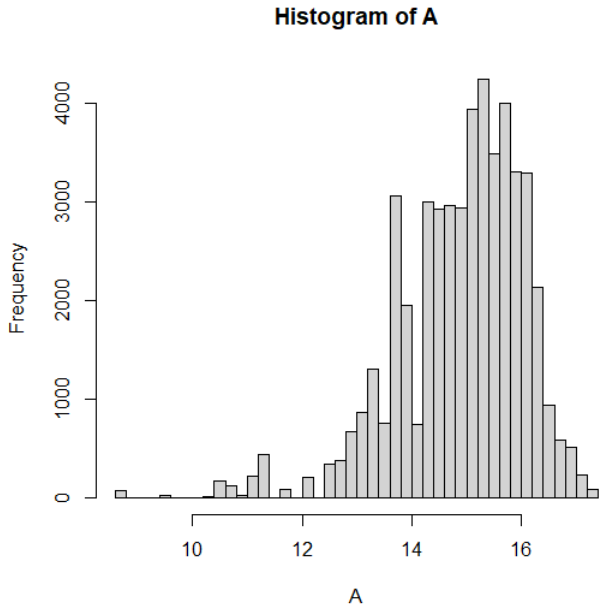

FIGURE 4.2 – densité de α

```
plot(MB,type="s")
```

- Nous avons ci-dessus et ci-dessous les résultats des distributions des paramétres α et β , - ainsi que l'illustration de la convergence des moyennes ci-dessous .

4.2.3 Régression probit :

- Cet exemple considère un modèle probit d'un point de vue bayésien.
- Rappelons que le modèle probit est un cas particulier de modèle linéaire généralisé où les y observés sont des variables binaires, prenant les valeurs 0 et 1, et les variables explicatives sont des vecteurs $x \in R^p$, pour une loi conditionnelle :

$$P(y = 1|x) = 1 - P(y = 0|x) = \phi(x^T\beta) \text{ , } \beta \in R^p$$

- Les données correspondant à ce modèle peuvent être facilement simulées, mais nous utilisons ici un jeu de données de R appelé Pima.tr qui est disponible dans la bibliothèque MASS.
- Ce jeu de données correspond à une étude de 200 femmes indiennes de l'éthnie Pima en termes de présence ou absence de diabète,

```
Pima.tr$type
```

(qui est donc la variable binaire y à expliquer) et différentes variables explicatives physiologiques.

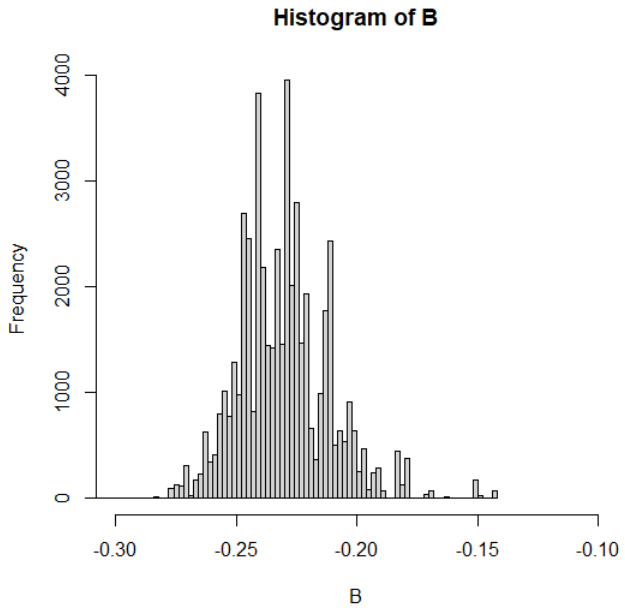

FIGURE 4.3 – densité de β

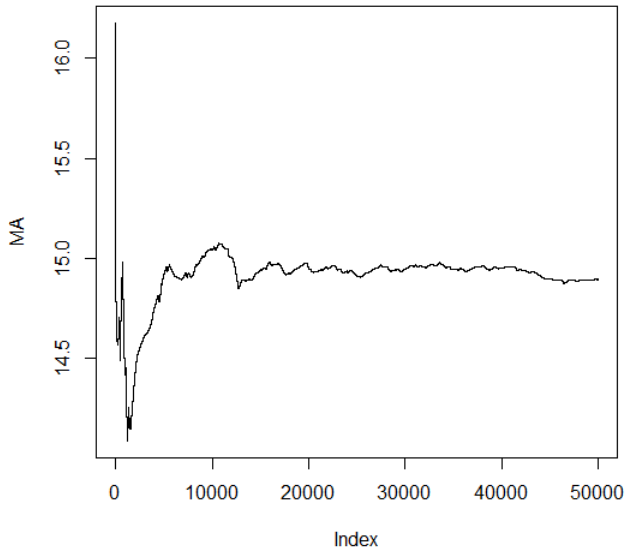

FIGURE 4.4 – moyenne de α

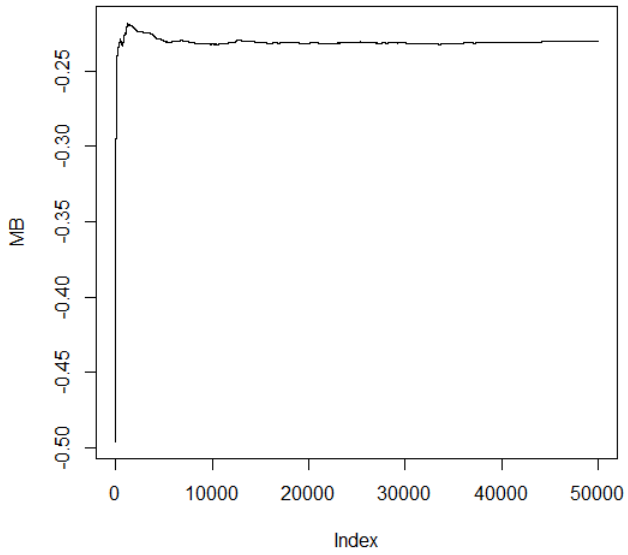

FIGURE 4.5 – moyenne de β

- Pour cet exemple, nous considérons seulement la variable de l'indice de masse corporelle,

```
Pima.tr$bmi
```

, avec une coordonnée à l'origine.
- Une estimation GLM standard du modèle est fournie par :

```
library("MASS");data(Pima.tr);X=Pima.tr;y=Pima.tr$type
x=Pima.tr$bmi;;y=as.numeric(y);x=as.vector(x);y=1-y
y=1+y;y=1-y;glm(y~x,family=binomial(link="probit"))
```

- On trouve : $\hat{\beta}_0 = -2.54$ et $\hat{\beta}_1 = 0.065$
- Dans une perspective bayésienne, nous introduisons une loi a priori sur $\beta = (\beta_0, \beta_1)$ qui est une distribution normale $\mathcal{N}(0, 100)$.
- La distribution a posteriori sur β est alors le produit de cette loi a priori presque plate avec la vraisemblance .
- Donc la densité a priori est :

$$\pi(\beta_0, \beta_1) \propto \exp\left(-1/200(\beta_0^2 + \beta_1^2)\right)$$

- On a la vraisemblance :

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \phi(\beta_0 + \beta_1 x_i)^{y_i} (1 - \phi(\beta_0 + \beta_1 x_i))^{1-y_i}$$

- Donc l'aposteriori est :

$$f(\beta_0, \beta_1) \propto \pi(\beta_0, \beta_1) L(\beta_0, \beta_1)$$

$$f(\beta_0, \beta_1) \propto \exp(-1/200(\beta_0^2 + \beta_1^2)) \prod_{i=1}^{n} \phi(\beta_0 + \beta_1 x_i)^{y_i} (1 - \phi(\beta_0 + \beta_1 x_i))^{1-y_i}$$

- On prend comme loi de proposition $\mathcal{N}(-2.54, 4^2)$ pour β_0 et $\mathcal{N}(0.065, 0.1^2)$ pour β_1 .
- Donc la loi de proposition est :

$$g \propto \exp(-1/2(\beta - m)' \Gamma^{-1} (\beta - m))$$

avec $m = \begin{pmatrix} -2.54 \\ 0.065 \end{pmatrix}$ et $\Gamma = \begin{pmatrix} 4^2 & 0 \\ 0 & 0.1^2 \end{pmatrix}$
- voici le code R :

```
library("MASS");data(Pima.tr);X=Pima.tr;y=Pima.tr$type
x=Pima.tr$bmi;y=as.numeric(y);y=1-y;y=1+y;y=1-y;
glm(y~x,family=binomial(link="probit"))
b0=-2.54;b1=0.065;
p=function(beta0,beta1,x){
return(pnorm(beta0+beta1*x,0,1))}
L=function(beta0,beta1,x,y){
return(prod(p(beta0,beta1,x)^y*(1-p(beta0,beta1,x))^(1-y)))}
apr=function(beta0,beta1){return(exp(-(1/200)*(beta0^2+beta1^2)))}
apo=function(beta0,beta1,x,y){
return(L(beta0,beta1,x,y)*apr(beta0,beta1))}
g=function(beta0,beta1){b=c(beta0,beta1)
m=c(-2.54,0.065);V=matrix(c(4^2,0,0,0.1^2),nrow=2)
return(exp(-(1/2)*t(b-m)%*%solve(V)%*%(b-m)))}
rho=function(beta0,beta1,y1,y2,x,y){
return(
min(1,(apo(y1,y2,x,y)/apo(beta0,beta1,x,y))*(g(beta0,beta1)/
g(y1,y2))))}
n=10^5;b0=NULL;b1=NULL;b0[1]=-2.54;b1[1]=0.065;y1=NULL;y2=NULL;
```

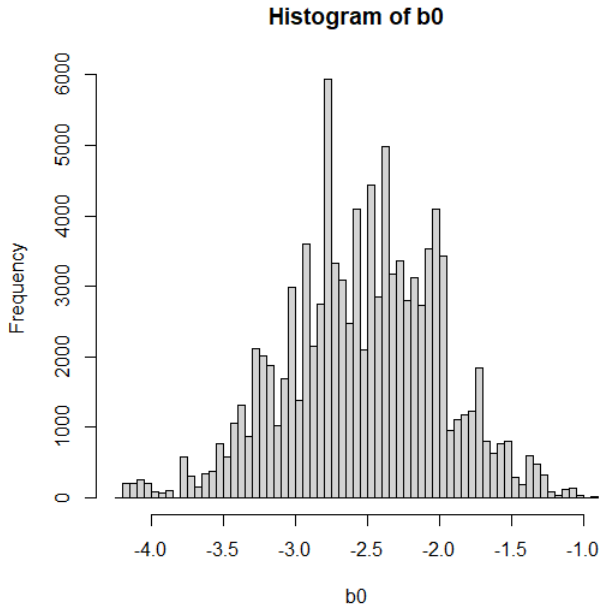

FIGURE 4.6 – densité de β_0

```
for(i in 1:(n-1)){y1[i]=rnorm(1,-2.54,4);y2[i]=rnorm(1,0.065,0.1);
u=runif(1);
b0[i+1]=b0[i]+(y1[i]-b0[i])*as.numeric(u<rho(b0[i],b1[i],
y1[i],y2[i],x,y))
b1[i+1]=b1[i]+(y2[i]-b1[i])*as.numeric(u<rho(b0[i],b1[i],
y1[i],y2[i],x,y))}
hist(b0,nclass=50);windows();hist(b1,nclass=50)
```

- Nous avons ci-dessus et ci-dessous les résultats des histogrammes des paramétres β_0 et β_1 .

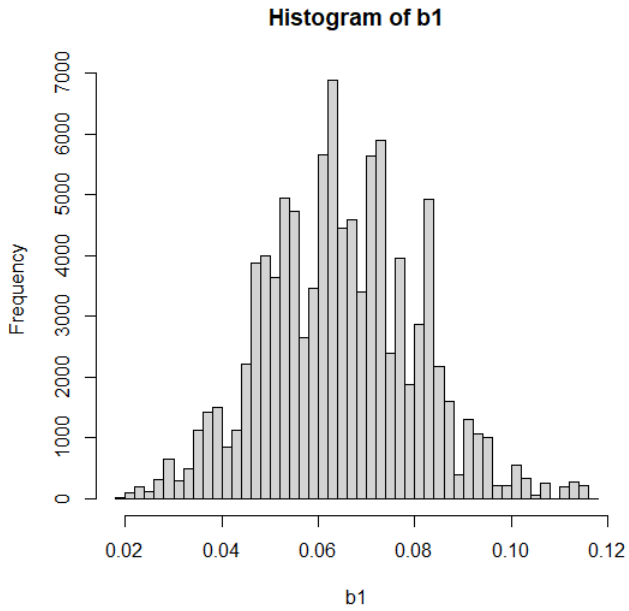

FIGURE 4.7 – densité de β_1

5 RÉGRESSION

5.1 Régression et apprentissage automatique :

- La régression linéaire est le «cheval de bataille» des statistiques et de l'apprentissage automatique (supervisé).
- Quand elle est augmenté de noyaux ou d'autres formes d'expansion de la fonction de base, elle peut également modéliser des relations non linéaire , Et quand la sortie gaussienne est remplacée par un Bernoulli ou une distribution multinomiale , elle peut être utilisé pour la classification .

5.2 Régression polynomiale :

- Nous allons traiter un exemple de réression polynomiale , c'est à dire un modéle de la forme :

$$y_{ij} = a + bx_i + cx_i^2 + \epsilon_{ij} \text{ , } i = 1, ..., k \text{ , } j = 1, ..., n_i;$$

ou on suppose que les v.a $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ sont indépendentes .

- Meme si nous allons faire une régression bayésienne , nous aurons besoin des estimateurs de maximum de vraisemblances pour avoir de bonnes lois candidates .

- Il suffit de simplement poser $z_i = x_i^2$ et de faire une régression linéaire multiple , à savoir pour :

$$\beta = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \text{ et } X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \dots & \dots & \dots \\ 1 & x_n & z_n \end{pmatrix}$$

on a : $y = X\beta + \epsilon$
et donc $\hat{\beta} = (X'X)^{-1}X'y$
- Sous R il suffira de faire la commande :

```
z=x^2
X=matrix(c(rep(1,n),x,x2),nrow=n)
summary(lm(y~x+z))
```

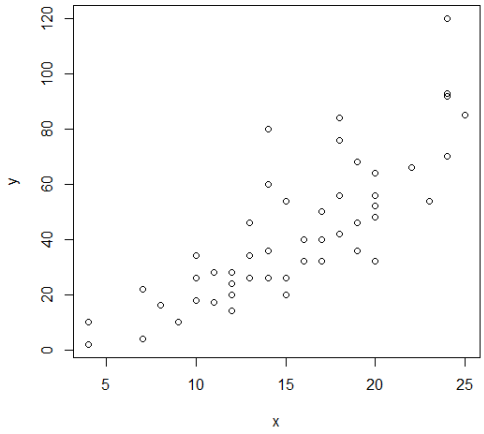

FIGURE 5.1 – courbe de y en fonction de x

5.2.1 Exemple : Metropolis–Hastings pour la régression :

- La base de données de R cars relie les distances de freinage (y) à la vitesse (x) pour un échantillon de voitures.
-On postule pour cette base de données un modèle quadratique :
$y_{i,j} = a + bx_i + cx_i^2 + \epsilon_{i,j}$, $i = 1, ..., k$, $j = 1, ..., n_i$,
où on suppose que les variables aléatoires $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ sont indépendantes.
- voici le code R pour génerer illustrer les données :

```
n=50;data(cars);X=cars;X=as.matrix(X);x=X[,1];y=as.vector(X[,2]);
plot(x,y)
```

- Nous avons le résultat ci-dessus ou on constate une courbe quadratique .
-La fonction de vraisemblance est donc proportionnelle à :

$$(\frac{1}{\sigma^2})^{\frac{N}{2}} \exp\,(-\frac{1}{\sigma^2} \sum_{i,j} (y_{i,j} - a - bx_i - cx_i^2)^2)$$

où $N = \sum_i n_i$ est le nombre total d'observations.
-On peut voir cette fonction de vraisemblance comme une loi a posteriori sur a, b, c et σ^2 (par exemple basée sur un a priori plat), et,on peut essayer de simuler dans cette loi avec un algorithme de Metropolis–Hastings .
-Pour commencer, on peut obtenir un candidat en générant des coefficients en fonction de leur loi d'échantillonnage obtenue par régression.
-Ainsi, on peut utiliser la commande R :

```
z=x^2
```

```
summary(lm(y~x+z))
```

-On peut utiliser comme lois candidates les lois normales centrées sur les EMV :
$a \sim \mathcal{N}(2.47, (14.8)^2)$,
$b \sim \mathcal{N}(0.91, (2.03)^2)$,
$c \sim \mathcal{N}(0.09, (0.065)^2)$,
$\sigma^{-2} \sim \mathcal{G}(\frac{n}{2}, (n-3)(15.17)^2)$,
dans un algorithme de Metropolis–Hastings afin de générer des échantillons $(a^{(i)}, b^{(i)}, c^{(i)})$ suivant la loi a posteriori .
-Voici le code R :

```
f=function(a,b,c){return(
exp((-1/(2*sigma2chap))*sum((y-a-b*x-c*x^2)^2)))}
m=c(2.47,0.91,0.09);u1=c(14.8^2,0,0);u2=c(0,2.03^2,0);
u3=c(0,0,0.065^2);V=matrix(c(u1,u2,u3),nrow=3);
g=function(y1,y2,y3){y=c(y1,y2,y3);
return(exp((-1/2)*t((y-m))%*%solve(V)%*%(y-m)))}
rho=function(a,b,c,y1,y2,y3){
return(min(1,(f(y1,y2,y3)/f(a,b,c))*(g(a,b,c)/g(y1,y2,y3))))}
sigma2=1/rgamma(5000,n/2,(n-3)*(15.17^2));sigma2chap=mean(sigma2);
A=NULL;B=NULL;C=NULL;A[1]=2.47;B[1]=0.91;C[1]=0.09;
for(i in 1:4999){y1=rnorm(1,2.47,14.8);y2=rnorm(1,0.91,2.03);
y3=rnorm(1,0.09,0.065);u=runif(1);
A[i+1]=A[i]+(y1-A[i])*
as.numeric(u<rho(A[i],B[i],C[i],y1,y2,y3));
B[i+1]=B[i]+(y2-B[i])*
as.numeric(u<rho(A[i],B[i],C[i],y1,y2,y3));
C[i+1]=C[i]+(y3-C[i])*
as.numeric(u<rho(A[i],B[i],C[i],y1,y2,y3))};
achap=A[(5000-500):5000];bchap=B[(5000-500):5000];
cchap=C[(5000-500):5000];a1=mean(achap);b1=mean(bchap);
c1=mean(cchap);y1=a1+b1*x+c1*x^2;plot(x,y1)
```

- Nous avons ci-dessous la courbe obtenue par régression .

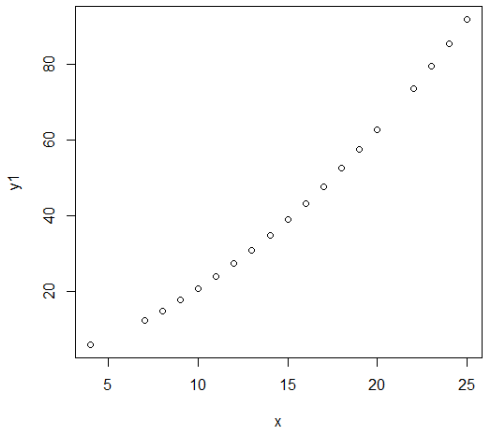

FIGURE 5.2 – courbe de régression de y en fonction de x

6 MODÉLE DE MARKOV CACHÉ

-

6.1 Qu'est-ce qu'un modèle de Markov caché ? :

-Un modèle de Markov caché (en abrégé HMM) est, grosso modo, une chaîne de Markov observée avec un bruit.
-En effet, le modèle comprend une chaîne de Markov, qu'on notera $\{X_k\}_{k\geq 0}$, où k est un indice entier.
-Cette chaîne de Markov est souvent supposé prendre des valeurs dans un ensemble fini, mais nous ne ferons pas cette restriction en général, permettant ainsi un espace d'état assez arbitraire.
- La chaîne de Markov est cachée, c'est-à-dire qu'elle n'est pas observable.
-Ce qui est disponible pour l'observateur est un autre processus stochastique $\{Y_k\}_{k\geq 0}$, lié à la chaîne de Markov telle que X_k régit la distribution de Y_k .
-Par exemple, Y_k peut avoir une distribution normale, dont la moyenne et la variance sont déterminées par X_k, ou Y_k peut avoir une distribution de Poisson dont la moyenne est déterminée par X_k.
-La chaîne de Markov sous-jacente $\{X_k\}$ est parfois appelée le régime, ou État.
-Toute inférence statistique, même sur la chaîne de Markov elle-même, doit être fait en termes de $\{Y_k\}$ uniquement, car $\{X_k\}$ n'est pas observé.
-Il y a aussi une autre hypothèse sur la relation entre la chaîne de Markov et le processus observable , en disant que X_k doit être la seule variable de la chaîne de Markov qui affecte la distribution de Y_k.
-Ceci est exprimé plus précisément dans la définition formelle qui suit .

6.1.1 Modéle de Markov caché :

-Un modèle de Markov caché est un processus bivarié en temps discret $\{X_k, Y_k\}_{k\geq 0}$, où $\{X_k\}$ est une chaîne de Markov et, conditionnellement à $\{X_k\}$, $\{Y_k\}$ est une suite de variables aléatoires indépendantes telles que la distribution conditionnelle de Y_k ne dépend que de X_k.

-La structure de dépendance d'un HMM peut être représentée par un modèle graphique .
-Les représentations de ce type utilisent un graphe orienté sans boucles pour décrire les structures de dépendance entre les variables aléatoires.

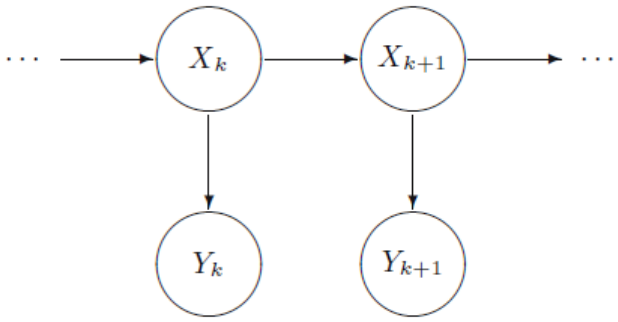

FIGURE 6.1 – Représentation graphique d'un HMM

- Les noeuds (cercles) du graphe correspondent aux variables aléatoires, et les arêtes (flèches) représentent la structure de la distribution de probabilité conjointe, avec les interprétations que ce dernier peut être factorisé comme un produit des distributions conditionnel de chaque noeud compte tenu de ses noeuds « parents » (ceux qui lui sont directement reliés par une flèche).
- Nous avons ci-dessus une représentation graphique de la structure de dépendance d'un modèle de Markov caché , où $\{Y_k\}$ est le processus observable et $\{X_k\}$ est la chaîne cachée.
-Nous supposerons que chaque HMM est homogène, c'est-à-dire que la chaîne de Markov $\{X_k\}$ est homogène (son noyau de transition ne dépend pas de l'indice de temps k), et que la loi conditionnelle de Y_k sachant X_k ne dépend pas non plus de k.

-Afin de garder cette discussion d'introduction simple, nous ne nous embarquons pas dans définitions mathématiques précises des concepts de chaîne de Markov tels que le noyaux transition par exemple.
-Comme mentionné ci-dessus, des deux processus $\{X_k\}$ et $\{Y_k\}$, seul $\{Y_k\}$ est effectivement observé, d'où l'inférence sur les paramètres du modèle doit être réalisé en utilisant $\{Y_k\}$ uniquement.
-L'autre sujet d'intérêt est bien sûr l'inférence sur le $\{X_k\}$ non observé : étant donné un modèle et quelques observations, peut-on estimer la suite inobservable d'états ?
-Ces deux grands objectifs statistiques sont en effet fortement liés.
-Des modèles qui comprennent des variables aléatoires non observées, comme le font les HMM, sont appelées des modèles à variables latentes, des modèles de données manquantes, ou encore des modèles avec des données incomplètes, où la variable latente fait référence aux quantités aléatoires non observables .

-Donnons déjà à ce stade un exemple simple et illustratif d'un HMM.

-Supposons que $\{X_k\}$ est une chaîne de Markov d'espace d'état $\{0, 1\}$ et que Y_k, conditionnellement à $X_k = i$, admet une distribution gaussienne $\mathcal{N}(\mu_i, \sigma_i^2)$.
- Autrement dit, la valeur du régime régit la moyenne et la variance de la distribution gaussienne à partir de laquelle nous tirons ensuite la sortie.
-Ce modèle illustre une caractéristique commune des HMM , à savoir que les distributions conditionnels de Y_k sachant X_k appartiennent toutes à une même famille paramétrique, avec des paramètres indexés par X_k.
-Dans ce cas, c'est la famille de distributions gaussienne , mais on peut bien sûr aussi considérer la famille Gamma, la Famille de Poisson, etc.

-Une observation significative, dans l'exemple actuel, est que la distribution marginale de $\{Y_k\}$ est celle d'un mélange de deux distributions gaussiennes.
-Par conséquent, nous pouvons également considérer les HMM comme une extension de modèles de mélange, y compris un certain degré de dépendance entre les observations.
-En effet, même si les variables Y sont conditionnellement indépendantes étant donné $\{X_k\}$, $\{Y_k\}$ n'est pas une suite indépendante à cause de la dépendance dans $\{X_k\}$.
-En fait, $\{Y_k\}$ n'est pas non plus une chaîne de Markov :
le processus joint $\{X_k, Y_k\}$ est bien sûr une chaîne de Markov, mais le processus observable $\{Y_k\}$ n'a pas la propriété de perte de mémoire des chaînes de Markov, dans le sens où la distribution conditionnel de Y_k sachant Y_0, . . . , Y_{k-1} dépend généralement conditionnelements de toutes les variables.
-Cependant, la dépendance dans la séquence $\{Y_k\}$ (définie dans un sens approprié) n'est pas plus forte que celle de $\{X_k\}$.

-Une autre vision consiste à considérer les HMM comme une extension des chaînes de Markov, dans le sens où l'observation $\{Y_k\}$ de l'état $\{X_k\}$ est déformée ou floue d'une certaine manière qui inclut un certain caractère aléatoire supplémentaire et indépendant.

-Dans le précédent exemple, la distorsion est simplement causée par un bruit gaussien additif, comme on peut écrire ce modèle sous la forme

$$Y_k = \mu_{X_k} + \sigma_{X_k} V_k$$

où $\{V_k\}_{k\geq 0}$ est une séquence aléatoire de variables gaussiennes standards i.i.d. (indépendante et identiquement distribuée) .

-Notre exemple simple n'est en aucun cas un cas singulier et, en grande généralité, tout

HMM peut être défini de manière équivalente par une fonction de représentation connue sous le nom de modèle d'espace d'états (général),

$$X_{k+1} = a(X_k, U_k)$$

,

$$Y_k = b(X_k, V_k)$$

, où $\{U_k\}_{k\geq 0}$et $\{V_k\}_{k\geq 0}$ sont des séquences de variables aléatoires i.i.d indépendantes de X_0 , mutuellement indépendants , et a et b sont des fonctions mesurables.

-La première équation est connue sous le nom d'équation d'état ou dynamique, tandis que la seconde est l'équation d'observation.
-Ces deux équations correspondent à une forme récursive et générative du modèle, par opposition à notre exposé initial, qui se concentrait sur la spécification de la distribution de probabilité conjointe des variables.
-La vue la plus naturelle et la plus fructueuse dépend généralement de ce que le HMM est censé modéliser et dans quel but il est utilisé .

6.2 Exemple :

-Les HMM et leurs généralisations sont aujourd'hui utilisés dans de nombreux domaines différents.
- Plusieurs livres spécialisés sont disponibles qui couvrent en grande partie les applications de HMM à certains domaines spécifiques tels que la reconnaissance vocale (Rabiner et Juang, 1993 ; Jelinek, 1997), l'économétrie (Hamilton, 1989 ; Kim et Nelson, 1999), la biologie computationnelle (Durbin et al., 1998 ; Koski, 2001), ou la vision par ordinateur (Bunke et Caelli, 2001).

-Il faut souligner que l'idée que l'on se fait de la nature La chaîne de Markov caché $\{X_k\}$ peut être assez différente d'un cas à l'autre.
-Dans certaines cas, elle a une signification physique bien définie, alors que dans d'autres cas, elle est conceptuellement plus diffuse, et dans d'autres cas encore, la chaîne de Markov peut être complètement fictif et la structure probabiliste du HMM est alors utilisé uniquement comme outil de modélisation de la dépendance des données.
- Les modéles de markov cachés ont l'avantage sur les chaines de Markov en ce qu'ils peuvent représenter des dépendances à long terme entre les observations, médiée par les variables latentes.

- En particulier, ils ne supposent pas que la propriété de Markov vaut pour les observations elles-mêmes.
- Ils peuvent également être utilisés pour définir des densités conditionnelles de classe à l'intérieur d'un classifieur génératif .

6.3 Modèle de Markov caché normale :

Par modèle de Markov caché normal, nous entendons un HMM dans lequel la distribution conditionnel de Y_k sachant X_k est gaussienne.
-Dans de nombreuses applications, l'espace d'état est fini, et nous supposerons alors qu'il est $\{1, 2, ..., r\}$.
-Dans ce cas, étant donné $X_k = i$, $Y_k \sim \mathcal{N}(\mu_i, \sigma_i^2)$,
de sorte que la distribution marginale de Y_k est un mélange fini de normales.

6.3.1 Reconnaissance vocale :

-Comme exemple de HMM normale, nous considérons des applications à la reconnaissance vocale, qui fut la première zone où les HMM ont été largement utilisés, à partir du début des années 1980. Le tâche de base est de, à partir d'un enregistrement de la voix d'une personne (ou en temps réel, en ligne), déterminer automatiquement ce qu'il a dit.
-Pour ce faire, le signal de parole enregistré et échantillonné est inséré dans de courts sections (également appelées trames), représentant généralement environ 20 millisecondes du signal d'origine.
-Chaque section est ensuite analysée séparément pour produire un ensemble de coefficients qui représentent la densité spectrale de puissance estimée du signal dans le cadre.
-Ce prétraitement aboutit à une série en temps multivarié de coefficients spectraux en temps discret .
-Pour qu'un mot donné soit reconnu (imaginez, pour simplicité, que les locuteurs ne prononcent que des mots simples), la longueur de la série de vecteurs résultant de ce prétraitement n'est pas déterminé au préalable mais dépend du temps mis par le locuteur pour prononcer le mot.
-Une exigence primaire sur le modèle est donc de faire face au problème d'alignement temporel afin de pouvoir comparer des séquences multivariées de longueurs inégales.
-Dans cette application, la chaîne de Markov cachée correspond aux sous-éléments de l'énoncé qui devraient avoir des caractéristiques spectrales comparables.
-En particulier, nous pouvons considérer chaque mot comme une suite de phonèmes (par exemple, red (rouge en anglais) : [r-e-d] ; classe : [$k - l - a$:-s]).

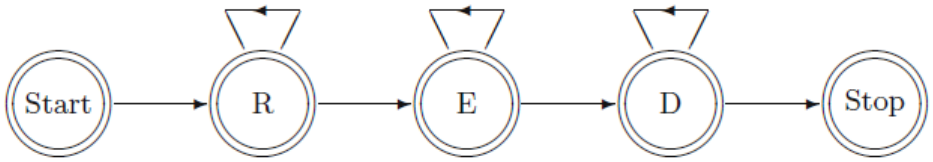

FIGURE 6.2 – Représentation en automate de la chaîne de Markov d'un HMM

-L'état de la chaîne de Markov est alors le phonème hypothétique qui est actuellement prononcé à un intervalle de temps donné.
- Ainsi, pour un mot à trois phonèmes, comme "red" par exemple, l'état de la chaîne de Markov peut évoluer selon la figure ci-dessus qui est une représentation en automate de la structure de la chaîne de Markov d'un HMM pour reconnaître le mot "red".
- La figure ci-dessus est une description d'automate d'une chaîne de Markov qui indique où la chaîne peut sauter compte tenu de son état actuel.
- Chaque flèche représente donc une transition possible qui est associée à une probabilité de transition non nul .
- On voit que chaque état correspondant à un phonème a une transition vers lui-même, c'est-à-dire une boucle ; c'est pour permettre au phonème de durer aussi longtemps que l'enregistrement de celui-ci .
- Les buts de l'état initial Start et l'état terminal Stop est simplement d'avoir des débuts et des fins bien définis de la chaîne de Markov ; l'état d'arrêt peut être considéré comme un état absorbant avec aucune observation associée .
- Les vecteurs d'observation associés à un état particulier (inobservable) sont supposés indépendants et se voient attribuer une distribution multivariée, le plus souvent un mélange de distributions gaussiennes.
- La variabilité induite par cette distribution est utilisée pour modéliser la variabilité spectrale au sein et entre les locuteurs.
- La reconnaissance vocale réelle est réalisée en exécutant le mot enregistré comme entrée à plusieurs HMM différents, chacun représentant un mot particulier, et en sélectionnant celui qui attribue la plus grande vraisemblance à la séquence observée.
- Dans une phase d'apprentissage préalable, les paramètres de chaque modèle du mot ont été estimés en utilisant un grand nombre d'énoncés enregistrés du mot.
- Notez que l'association des états de la chaîne cachée avec les phonèmes de la figure précédente est plus une vue conceptuelle qu'une description réelle de ce que fait le modèle.
- En pratique, les performances de la reconnaissance vocale basée sur HMM est bien meilleure que leur efficacité à segmenter les mots en phonèmes .

6.4 HMM à espace d'état général (continus) :

6.4.1 Exemple : Volatilité stochastique :

-Les propriétés distributives des prix spéculatifs ont des implications importantes pour plusieurs modèles financiers.
-Soit S_k le prix d'un actif financier, tel qu'un cours d'action, un indice boursier, ou taux de change — au temps k.
-Au lieu des prix, il est plus habituel de considérer les rendements relatifs $\frac{(S_k S_{k-1})}{S_{k-1}}$ ou les rendements logarithmique $\log\left(\frac{S_k}{S_{k-1}}\right)$, qui décrivent tous deux la variation relative dans le temps du processus du prix .
-Le célèbre modèle Black-Scholes, qui est un modèle en temps continu et qui postule un mouvement brownien géométrique pour le processus de prix, correspond aux rendements logarithmiques qui sont i.i.d. et avec une Distribution gaussienne $\mathcal{N}(\mu, \sigma^2)$, où σ est la volatilité (le mot volatilité est le mot utilisé en économétrie pour l'écart-type).
-Le modèle d'évaluation des options de Black et Scholes fournit la base de la théorie moderne de l'évaluation d'option .
-Dans les applications réelles, cependant, ce modèle a certains éléments de carences bien documentés .
-Les données des marchés financiers indiquent clairement que les distributions des retours ont généralement des queues plus lourdes que celles de la normale .
-De plus, même si les rendements sont environ non corrélé au fil du temps (comme prédit par le modèle Black et Scholes), ils ne sont pas indépendants.
-Ceci peut être facilement vérifié par le fait que les exemples d'autocorrélations des valeurs absolues (ou carrés) des rendements sont non nuls pour un grand nombre de retards
-La propriété indique que les rendements peuvent être modélisés par une séquence de bruit blanc (un processus stationnaire avec une autocorrélation nulle à tous les retards positifs), cette derniere propriété indique que les rendements sont dépendants et que la dépendance peut même s'étendre sur une assez longue période.
-La variance des rendements a tendance à changer avec le temps :
les grandes et les petites valeurs de l'échantillon se produisent en clusters.
Les grands changements ont tendance à être suivis de grands changements – de l'un ou l'autre signe –
et les petits changements ont tendance à être suivis de petits changements. changements, un phénomène souvent appelé regroupement de la volatilité.
-La plupart des modèles de données de rendements qui sont utilisés dans la pratique sont

d'ordre multiplicatif de la forme,

$$Y_k = \sigma_k V_k$$

, où $\{V_k\}_{k\geq 0}$ est une séquence i.i.d. et le processus de volatilité $\{\sigma_k\}_{k\geq 0}$ est un processus stochastique non négatif tel que σ_k et V_k soient indépendants pour tout k.

-La plupart du temps, $\{\sigma_k\}$ est supposé stationnaire au sens strict.
-On suppose souvent que V_k est symétrique ou, au moins, a une moyenne nulle.
-La justification de l'utilisation de ces modèles est assez simple.
-Tout d'abord, la direction des changements de prix est modélisé par le signe de V_k uniquement, indépendamment de l'ordre de grandeur de ce changement, qui est dirigé par la volatilité.
-Parce que σ_k et V_k sont indépendants et V_k est supposé avoir une variance unitaire, σ_k^2 est alors la variance conditionnel de X_k étant donné σ_k.
-La plupart des modèles supposent que σ_k est une fonction des valeurs passés.
-Le modèle le plus simple suppose que σ_k est une fonction des carrés des observations précédentes.

-Cela conduit au célèbre modèle autorégressif conditionnel d'hétéroscédasticité (ARCH) développé par Engle (1982) :

$$Y_k = \sqrt{X_k} V_k$$

$$X_k = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{k-i}^2$$

où $\alpha_0, ..., \alpha_p$ sont des constantes non négatives.
-Dans le modèle d'Engle (1982), $\{V_k\}$ Est normal ; par conséquent, la distribution d'erreur conditionnelle est normale, mais avec une variance conditionnel égale à une fonction linéaire des p dernières observations au carré.
-Les modèles ARCH sont ainsi capables de reproduire la tendance au valeurs extrême d'etre suivi par d'autres valeurs extrêmes, mais de signe imprévisible.
-La structure autorégressive peut être vue par l'argument suivant.
-En écrivant $v_k = Y_k^2 - X_k = X_k(V_k^2 - 1)$ on obtient
$Y_k^2 - \sum_{i=1}^{p} \alpha_i Y_{k-i}^2 = \alpha_0 + v_k$.
-Parce que $\{V_k\}$ est une suite i.i.d. de moyenne nulle et de variance unitaire, $\{v_k\}_{k>=0}$ est une séquence non corrélée.
-Parce que les processus ARCH(p) ne correspondent pas très bien aux rendements logarithmique sauf si l'ordre p est assez grand, plusieurs personnes ont pensé à des améliorations.

-Comme $Y_k^2 - \sum_{i=1}^{p} \alpha_i Y_{k-i}^2 = \alpha_0 + v_k$ a une certaine ressemblance avec une structure AR, une possible généralisation consiste à introduire une structure ARMA.
-Cette construction conduit au processus dit GARCH(p, q) (Bollerslev et al., 1994).
-Ce modèle affiche certaines similitudes frappantes avec les modèles autorégressifs avec le régime de Markov .
-Une alternative au cadre ARCH/GARCH est un modèle dans lequel le la variance est spécifiée pour suivre un processus stochastique latent.
-De tels modèles, appelés modèles à volatilité stochastique (SV), apparaissent dans la littérature théorique sur la tarification des options et la modélisation du taux de change.
-Contrairement au processus type GARCH , il n'y a pas de rétroaction directe des rendements passés sur le processus de volatilité , qui a été remis en question comme non naturel par certains auteurs.
- les versions Empiriques du modèle SV sont généralement formulées en temps discret, ce qui facilite le traitement des problèmes d'inférence.
- Le modèle canonique en SV pour les données en temps discret est (Hull et White, 1987 ; Jacquier et al., 1994),

$$X_{k+1} = \phi X_k + \sigma U_k \text{ , } U_k \sim \mathcal{N}(0,1)$$

$$Y_k = \beta \exp\left(\frac{X_k}{2}\right) V_k \text{ , } V_k \sim \mathcal{N}(0,1)$$

où les observations $\{Y_k\}_{k\geq 0}$ sont les rendements logarithmiques, $\{X_k\}_{k\geq 0}$ est la volatilité logarithmique, qui est supposée suivre une autorégression stationnaire d'ordre 1, et $\{U_k\}_{k\geq 0}$ et $\{V_k\}_{k\geq 0}$ sont des séquences indépendants i.i.d.
-Le paramètre β joue le rôle du facteur d'échelle constant, ϕ est la persistance (mémoire) dans la volatilité, et σ est la volatilité de la volatilité logarithmique.
-Malgré une représentation très parcimonieuse, ce modèle est capable d'exhiber un large éventail de comportements.
-Comme les modèles ARCH/GARCH, le modèle peut donner lieu à une forte persistance de la volatilité ("volatility clustering").
-Même avec $\phi = 0$, le modèle est un mélange à l'échelle gaussienne qui donnera lieu à un excès d'aplatissement dans le distribution marginale des données.
-Dans les modèles ARCH/GARCH avec des erreurs normales, le degré d'aplatissement est lié aux racines de l'équation de volatilité ; comme la volatilité devient plus corrélée, le degré d'aplatissement augmente également.
-Dans le modèle de volatilité stochastique, le paramètre σ régit le degré de mélange indépendamment du degré de lissage de l'évolution de la variance.
-Il est intéressant de noter que les modèles de volatilité stochastique sont liés aux modèles d'espace d'états linéaires conditionnellement gaussiens. En prenant les logarithmes de

rendements au carré, on obtient,

$$X_k = \phi X_{k-1} + \sigma U_{k-1}$$

$$\log(Y_k^2) = \log(\beta^2) + X_k + Z_k \text{ ou } Z_k = \log(V_k^2)$$

-Si V_k est normal, Z_k suit la distribution $\log \chi_1^2$.
-Cette répartition peut être approximée avec une précision arbitraire par un mélange fini de distributions gaussiennes , puis le modèle SV devient un modèle d'espace d'états linéaire gaussien conditionnellement (Sandmann et Koopman, 1998 ; Durbin et Koopman, 2000).
-Cette fois, la variable latente C_k est la composante du mélange et la modèle s'écrit

$$W_{k+1} = \phi W_k + U_k \text{ , } U_k \sim \mathcal{N}(0,1)$$

$$Y_k = W_k + (\mu(C_k) + \sigma_v(C_k)V_k) \text{ , } V_k \sim \mathcal{N}(0,1)$$

-Cette représentation du modèle de volatilité stochastique peut s'avérer utile lorsque on veut dériver des algorithmes numériques pour filtrer l'état caché ou estimer le modèle des paramètres.

6.4.2 Distribution conditionnelle de site unique en Modèle stochastique de volatilité :

-Pour illustrer l'échantillonneur de tranches, nous considérons le modèle stochastique de volatilité dont la forme d'espace d'état est la suivante :

$$X_{k+1} = \phi X_k + \sigma U_k$$

$$Y_k = \beta \exp(X_k/2) V_k$$

où $\{U_k\}_{k\geq 0}$ et $\{V_k\}_{k\geq 0}$ sont des bruits blancs gaussiens standards indépendants .
-Dans ce modèle, $\beta^2 \exp(X_k)$ est appelé la volatilité, et son estimation est l'un des objectifs de l'analyse .
-Nous considérons la distribution conditionnelle de X_k sachant X_{k-1}, X_{k+1} et Y_k, dont la fonction de densité de transition $\pi_k(x|x_{k-1}, x_k)$ est proportionnel à
$\pi_k(x|x_{k-1}, x_k) \propto \exp\left(-\left(\frac{(x_{k+1}-\phi x)^2}{2\sigma^2} + \frac{(x-\phi x_{k-1})^2}{2\sigma^2}\right)\right) \frac{1}{\beta \exp(x/2)} \exp\left(-\frac{y_k^2}{2\beta^2 \exp(x)}\right)$.
-En ignorant les constantes.
- En fait, les termes qui ne dépendent pas de x peuvent être ignorés , on peut compléter le carré (en x) pour obtenir :
$\pi_k(x|x_{k-1}, x_k) \propto \exp\left(-\frac{1+\phi^2}{2\sigma^2}\left((x-\mu_k)^2 + \frac{y_k^2\sigma^2}{(1+\phi^2)\beta^2} \exp(-x)\right)\right)$,

ou $\mu_k = \frac{\phi(x_{k+1}+x_{k-1})-\sigma^2/2}{1+\phi^2}$.
en définissant $\alpha_k = \frac{y_k^2\sigma^2\exp(-x)}{(1+\phi^2)\beta^2}$ et $\rho = \frac{1+\phi^2}{2\sigma^2}$, on a :
$\pi_k(x|x_{k-1}, x_k) \propto \exp\left(-\rho((x-\mu_k)^2 + \alpha_k \exp\left(-(x-\mu_k)\right))\right)$.
-Le paramètre μ_k correspond à un décalage simple qui ne pose aucun problème de simulation.
-Par conséquent, la forme générale de la fonction de densité de probabilité conditionnelle à partir de laquelle la simulation est requise est $\exp\left(-\rho(x^2 + \alpha\exp(-x))\right)$ pour les valeurs positives de ρ et α.
-La deuxième étape de l'algorithme de l'échantillonneur par tranches nécessite alors une simulation à partir de la distribution uniforme sur la ensemble
$A(u) = \{x : \exp\left(-\rho(x^2 + \alpha\exp(-x))\right) \geq u\} = \{x : x^2 + \alpha\exp(-x) \leq w\}$,
avec $w = -(1/\rho)\log(u)$.
-Or, alors que l'inversion de $x^2 + \alpha\exp(-x) = w$ est impossible analytiquement, le fait que cette fonction est convexe (pour $\alpha > 0$) et que la valeur précédente de x appartient à l'ensemble S(u) aide à résoudre cette équation par essais et erreurs numériques ou par des algorithmes de recherche de zéro plus élaborés.
-Il n'est pas non plus nécessaire de résoudre précisément ce problème d'équation, car la connaissance d'un intervalle contenant l'ensemble S(u) suffit à simuler à partir de la distribution uniforme sur S(u) :
il suffit alors de simuler des candidats e uniformément du plus grand ensemble et ne les accepter que si $e \in S(u)$.
-Voici le code R :

```
alpha=5;rho=1;f=function(x){return(exp(-rho*(x^2+alpha*exp(-x))))}
n=10^3;x=NULL;u=NULL;x[1]=runif(1);for(i in 1:(n-1)){
u[i+1]=runif(1,0,f(x[i]));y=runif(1,-10,10)
while((y^2+alpha*exp(-y))>((-1/rho)*log(u[i+1]))){
y=runif(1,-10,10)};x[i+1]=y};hist(x,nclass=50);acf(x)
```

- Nous avons ci-dessous les résultats de l'histogramme et du corrélogramme .
- Comme autre exemple d'application d'algorithme mcmc pour les modéles de markov cachés , nous allons considérer un modéle autorégressif d'ordre 1 .

6.4.3 Autorégression au carré et bruité :

-Considérons le modèle qui suit , où la chaîne de Markov cachée provient d'un modèle AR(1) régulier ,

$$x_{t+1} = \rho x_t + \epsilon_t \text{ , } (t \geq 1)$$

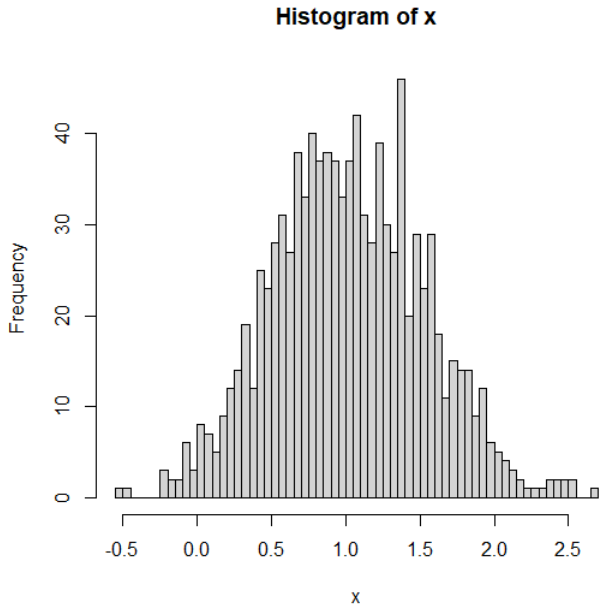

FIGURE 6.3 – densité pour $\alpha = 5$ et $\rho = 1$

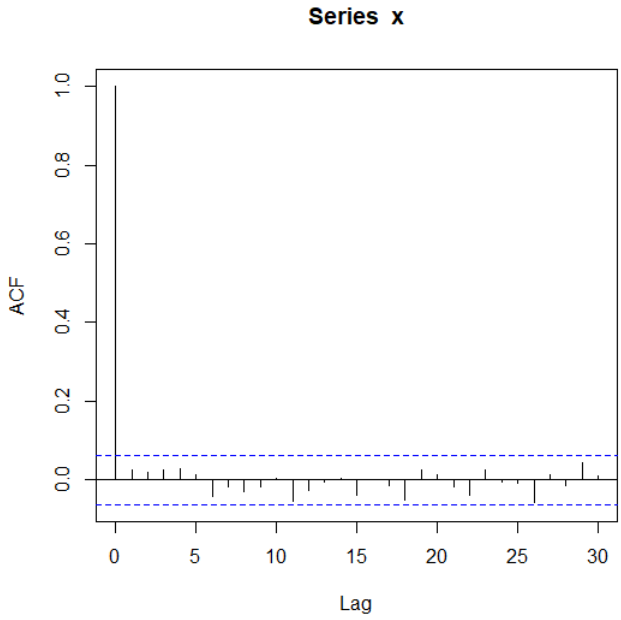

FIGURE 6.4 – corrélogramme

ou $\epsilon_t \sim \mathcal{N}(0,1)$ observé à travers :

$$y_t = x_t^2 + \zeta \text{ , ou } \zeta \sim \mathcal{N}(0,\tau^2)$$

- On s'intéresse à la distribution conditionnelle de x_t sachant (x_{t-1}, y_t, x_{t+1}) , $\pi(x_t|x_{t-1}, y_t, x_{t+1})$, proportionnelle à :

$$\exp\left(-(\tau^{-2}(y_t - x_t^2)^2 + (x_t - \rho x_{t-1})^2 + (x_{t+1} - \rho x_t)^2)/2\right)$$

- Cette densité n'est pas standard et, pour l'approcher, on peut lancer une chaîne de Metropolis–Hastings à marche aléatoire :

$$e^{(t)} \sim \mathcal{N}(x^{(t)}, \sigma^2)$$

- Voici le code R :

```
x0=-0.94;y=3.17;x1=-1.12;tau=0.2;r=0.85
pi=function(x){return(
exp(-1/2*(1/tau^2*(y-x^2)^2+(x-r*x0)^2+(x1-r*x)^2)))}
rho=function(x,e){return(min(1,pi(e)/pi(x)))}
n=10^5;sigma=0.9;x=NULL;e=NULL;x[1]=0;
for(i in 1:(n-1)){e[i]=rnorm(1,x[i],sigma);u=runif(1);
x[i+1]=(e[i]-x[i])*as.numeric(u<=rho(x[i],e[i]))+x[i]}
hist(x,nclass=100)
```

- pour $\sigma = 0.1$ la chaine reste bloquée sur un seul mode et n'arrive pas à explorer correctement la densité , mais pour $\sigma = 0.9$ la chaine explore correctement la densité .
- Nous avons ci-dessous les résultats des histogrammes pour $\rho = 0.85$, $\tau = 0.2$, $x_{t-1} = -0.94$, $y_t = 3.17$, $x_{t+1} = -1.12$.

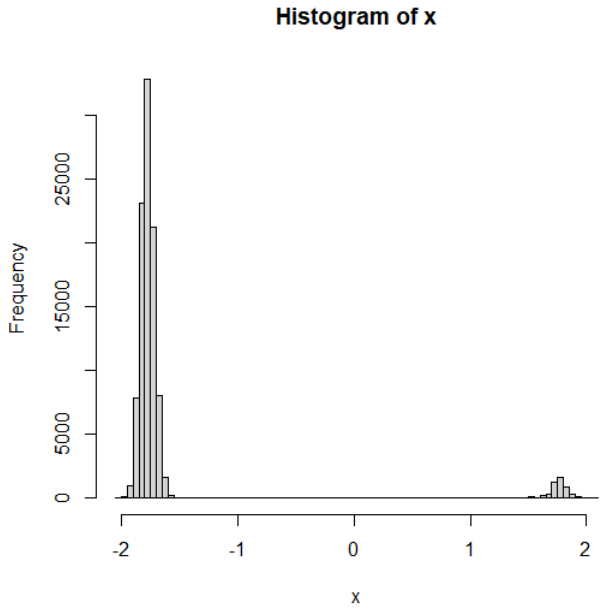

FIGURE 6.5 – densité pour $\sigma = 0.9$

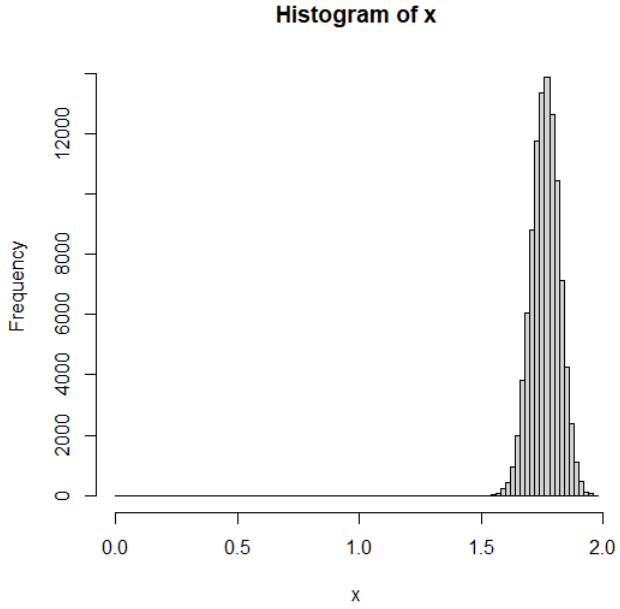

FIGURE 6.6 – densité pour $\sigma = 0.1$

7 MODÉLE À VARAIBLES LATENTES

- Les modéles à variables latentes jouent un role centrale en apprentissage automatique , on les rencontre par exemple en modéle de markov caché mais aussi dans les réseaux de neurones approfondies qui contiennent des couches cachés .

7.1 Modèles à données manquantes et démarginalisation :

-Les modèles à données manquantes sont des cas spéciaux de représentations $h(x) = E[H(x, Z)]$.
-Le mieux est de les voir comme des modèles où la loi des observations peut être exprimée comme :
$g(x|\theta) = \int_Z f(x, z|\theta)dz$.

7.1.1 Vraisemblance de données censurées :

-Des données censurées peuvent provenir d'expériences où des observations potentielles sont remplacées par une borne inférieure parce qu'elles prennent trop longtemps à être observées.
- Supposons qu'on observe $Y_1, ..., Y_m$, iid, tirées suivant $f(y - \theta)$ et que les $(n - m)$ observations restantes $(Y_{m+1}, ..., Y_n)$ soient censurées à un seuil a.
-La fonction de vraisemblance correspondante est alors :
$L(\theta|y) = (1 - F(a - \theta))^{n-m} \prod_{i=1}^{m} f(y_i - \theta)$
où F est la fonction de répartition associée à la densité f et où $y = (y_1, ..., y_m)$.
- Si nous avions observé les $n - m$ dernières valeurs (appelons les $z = (z_{m+1}, ..., z_n)$, avec $z_i \geq a$, $i = \overline{m + 1, ..., n}$, nous aurions construit la vraisemblance (des données complètes) :

$$L^c(\theta|y, z) = \prod_{i=1}^{m} f(y_i - \theta) \prod_{i=m+1}^{n} f(z_i - \theta)$$

-Notons que :

$$L(\theta|y) = E(L^c(\theta|y, z)) = \int_Z L^c(\theta|y, z) f(z|y, \theta)dz$$

où $f(z|y, \theta)$ est la loi des données manquantes conditionnellement aux données observées, c'est-à-dire le produit des $\frac{f(z_i-\theta)}{1-F(a-\theta)}$, ou encore $f(z-\theta)$ restreint au domaine $(a, +\infty)$.
-Lorsque $g(x|\theta) = \int_Z f(x, z|\theta)dz$ est vérifiée, le vecteur z ne sert qu'à simplifier les calculs.
- Il n'a pas forcément de sens particulier pour le problème statistique correspondant.

- On peut tout de même le voir comme un modèle à données manquantes au sens où Z peut être interprété comme manquant dans les observations.
-Nous appelons donc la fonction $L^c(\theta|x,z)) = f(x,z|\theta)$ vraisemblance du "modèle complet" ou des "données complètes".
- Elle correspond à la vraisemblance que nous obtiendrions si nous observions (x,z), qu'on appelle aussi données complètes (abusivement, puisque ce ne sont pas vraiment des données).

-Cette représentation est un cas particulier de démarginalisation, un contexte dans lequel une fonction d'intérêt est exprimée comme l'intégrale d'une quantité plus facile à manipuler sans contrainte supplémentaire.

7.1.2 L'algorithme EM :

-L'algorithme EM (pour Espérance–Maximisation) est une technique d'optimisation déterministe qui utilise la représentation

$$g(x|\theta) = \int_Z f(x,z|\theta)dz$$

pour construire une suite de problèmes de maximisation plus faciles, dont la limite est la réponse au problème initial.
-Supposons donc qu'on observe $X_1, ..., X_n$, de distribution jointe $g(x|\theta)$ vérifiant :
$g(x|\theta) = \int_Z f(x,z|\theta)dz$
et qu'on cherche à calculer $\hat{\theta} = arg_{max} L(\theta|x) = arg_{max} g(x|\theta)$.
-Puisque les données augmentées sont z, où $(X,Z) \sim f(x,z|\theta)$, la loi conditionnelle des données manquantes Z sachant les données observées x est :

$$k(z|\theta,x) = \frac{f(x,z|\theta)}{g(x|\theta)}$$

-On prend le logarithme de cette expression pour obtenir la relation suivante entre la vraisemblance des données complètes $L^c(\theta|x,z)$ et la vraisemblance des données observées $L(\theta|x)$:

$$\forall \theta_0 \ , \ \log(L(\theta|x)) = E_{\theta_0}(\log(L^c(\theta|x,Z)) - E_{\theta_0}(\log(k(Z|\theta,x))$$

où l'espérance est calculée par rapport à $k(z|\theta_0,x)$.

-Dans l'algorithme EM, on cherche à maximiser $\log(L(\theta|x))$, mais on ne considère que le terme $E_{\theta_0}(\log(L^c(\theta|x,Z))$.

-soit $Q(\theta|\theta_0, x) = E_{\theta_0}(\log(L^c(\theta|x, z))$,
l'algorithme EM procède de manière itérative en maximisant (en θ) $Q(\theta|\theta_0, x)$ à chaque itération.
-Si $\hat{\theta_1}$ est la valeur de θ qui maximise $Q(\theta|\theta_0, x)$,
on remplace θ_0 par la nouvelle valeur $\hat{\theta_1}$.
-On obtient ainsi une suite d'estimateurs $\{\hat{\theta_j}\}_j$ où $\hat{\theta_j}$ est la valeur qui maximise $Q(\theta|\hat{\theta}_{j-1}, x)$.
-Autrement dit : $Q(\hat{\theta}_j|\hat{\theta}_{j-1}, x)$=$max_\theta Q(\theta|\hat{\theta}_{j-1}, x)$.
-Ce processus itératif contient donc une étape de calcul d'espérance et une étape de maximisation, d'où son nom.

7.1.3 Algorithme 3 (L'algorithme EM) :

-Prendre une valeur initiale $\hat{\theta}_0$ (et donc m=0)
Répéter
1. Calculer (étape E)

$$Q(\theta|\hat{\theta}_m, x) = E_{\hat{\theta}_m}(\log(L^c(\theta|x, Z))$$

où l'espérance est calculée selon $k(z|\hat{\theta}_m, x)$.
2. Maximiser $Q(\theta|\hat{\theta}_m, x)$ en θ (étape M)
et prendre $\hat{\theta}_{m+1}$=$arg_{max} Q(\theta|\hat{\theta}_m, x)$
puis passer à m = m + 1, jusqu'à obtenir un point fixe,
c'est-à-dire jusqu'à ce que : $\hat{\theta}_{m+1} = \hat{\theta}_m$.

-L'inégalité de Jensen permet de démontrer aisément qu'à chaque itération de l'algorithme EM, la vraisemblance augmente :
$L(\hat{\theta}_{j+1}|x)(\hat{\theta}_j|x)$,
avec $L(\hat{\theta}_{j+1}|x) = L(\hat{\theta}_j|x) \Leftrightarrow Q(\hat{\theta}_{j+1}|\hat{\theta}_j, x)= Q(\hat{\theta}_j|\hat{\theta}_j, x)$.

-Cela signifie que, sous certaines conditions, tout point limite de la suite EM $\{\hat{\theta}_{(j)}\}$ est un point stationnaire de $L(\theta|x)$, même si ce n'est pas pour autant forcément un estimateur du maximum de vraisemblance ni même un maximum local.

-En pratique, il est donc recommandé de relancer l'algorithme EM plusieurs fois avec des points de départ différents choisis aléatoirement, si on veut éviter une mauvaise approximation du maximum global. (Ce point est le seul aspect aléatoire de l'algorithme EM :
si on commence avec la même valeur initiale $\hat{\theta}_0$, l'algorithme converge vers le même point fixe.)

-Pour implémenter l'algorithme EM, il faut donc être capable de :
(a) calculer la fonction $Q(\theta'|\theta, x)$.
(b) maximiser cette fonction.
-Beaucoup de modèles de données manquantes peuvent ainsi être analysés.

7.1.4 EM par Monte-Carlo :

-Une difficulté dans l'implantation de l'algorithme EM est que chaque étape E nécessite le calcul de l'espérance de la log-vraisemblance $Q(\theta|\theta_0, x)$.
-À l'exception de cas standards où la fonction Q est connue exactement, on peut approcher Q à l'aide de méthodes de Monte-Carlo puisque Q s'exprime naturellement comme une espérance.
-Ainsi, on peut par exemple simuler $Z_1, ..., Z_T$ à partir de la loi conditionnelle $k(z|x, \hat{\theta}_m)$ puis maximiser la log-vraisemblance approchée des données complètes :

$$\hat{Q}(\theta|\theta_0, x) = \frac{1}{T}\sum_{i=1}^{T} \log\left(L^c(\theta|x, z_i)\right)$$

comme le suggèrent Wei & Tanner (1990) sous le nom d'EM par Monte-Carlo (MCEM).

-Notons que l'algorithme MCEM ne vérifie pas la propriété de monotonie de l'algorithme EM. Il est donc important de vérifier que la suite des (β, σ) produite par l'algorithme MCEM converge vers une approximation du maximum de la vraisemblance du modèle, soit en évaluant numériquement la vraisemblance, soit en répétant l'algorithme avec différentes valeurs initiales.

-Contrairement aux méthodes de Monte-Carlo plus génériques, l'algorithme MCEM a néanmoins l'avantage d'approcher la suite EM convergente, plutôt que de maximiser une approximation de la vraisemblance.
-Contrairement à EM, MCEM est une méthode de Monte-Carlo et il est donc nécessaire d'évaluer l'erreur qui en résulte.
-Plutôt que la méthode grossière et coûteuse de la réplication des chemins, Booth Hobert (1999) donnent une approximation au premier ordre basée sur un développement limité de Q :
$var(\theta_1|\theta_0, x) \approx (\frac{\partial^2 Q}{\partial\theta\partial\theta^T}(\theta_1|\theta_0, x))^{-1} var(\frac{\partial Q}{\partial\theta}(\theta_1|\theta_0, x)(\frac{\partial^2 Q}{\partial\theta\partial\theta^T}(\theta_1|\theta_0, x))^{-1}$.
-Le terme de variance interne peut alors aisément être évalué à l'aide des variables manquantes simulées.

7.2 Données manquantes et variables latentes :

- Pour une distribution jointe cible $f(x,y)$, l'échantillonneur de Gibbs semble présenter une différence essentielle avec l'algorithme Metropolis–Hastings puisque ce dernier travaille sur une seule distribution , au sens où il génère toutes les composantes de (x,y) à la fois.

-Cette différence apparente du domaine d'application des deux algorithmes est en fait illusoire :

-$f(x,y)$ étant donné, on peut utiliser soit l'échantillonneur de Gibbs pertinent, soit un algorithme de Metropolis–Hastings générique.

-Inversement, si on dispose d'une densité marginale $f_X(x)$, on peut construire (ou compléter $f_X(x)$ en) une densité jointe correspondante $f(x,y)$ dans le seul but d'aider à la simulation, où la seconde variable Y est alors une variable auxiliaire qui peut ne pas être directement pertinente du point de vue statistique.

- Il existe de nombreux contextes où $f_X(x)$ peut être naturellement complétée en $f(x,y)$ et associée à un échantillonneur de Gibbs efficace.

-Ces considérations nous ramènent au cadre des modèles à données manquantes, où on a introduit la représentation :

$g(x|\theta) = \int_{D_Z} f(x,z|\theta)dz$,

ou $g(x|\theta)$ est la densité des observations (c'est-à dire la vraisemblance) et $f(x,z|\theta)$ représente la densité jointe complétée.

-Sous la contrainte $g(x|\theta) = \int_{D_Z} f(x,z|\theta)dz$, la densité f est arbitraire et elle peut être choisie de telle sorte que ses lois conditionnelles complètes soient faciles à simuler .

-L'algorithme de Gibbs peut alors être appliqué à f plutôt qu'à g et donc impliquer la loi conditionnelle correspondante de θ sachant (x,z).

-Selon le domaine d'application, ces représentations comme loi marginale portent des noms différents.

-Du point de vue mathématique, $g(x|\theta) = \int_{D_Z} f(x,z|\theta)dz$, est un modèle de mélange.

- En statistique, on parle le plus souvent de modèles à données manquantes, alors que les économètres préfèrent l'emploi de modèles à variable latente.

-En factorisant $f(x,z|\theta) = f(x|z,\theta)h(z|\theta)$ on obtient :

$$g(x|\theta) = \int_{D_Z} f(x|z,\theta)h(z|\theta)dz$$

et $h(z|\theta)$, la loi marginale des données manquantes z , est clairement une loi mélangeante.

-Dans le contexte général des données manquantes :

$$g(x) = \int_{D_Z} f(x,z)dz$$

pour $p \geq 2$, on écrit $y = (x,z) = (y_1,...,y_p)$ et on note les densités conditionnelles de $f(y) = f(y_1,...,y_p)$ sous la forme :
$Y_1|y_2,...,y_p \sim f_1(y_1|y_2,...,y_p)$,
$Y_2|y_1,y_3,...,y_p \sim f_2(y_2|y_1,y_3,...,y_p)$,
...
$Y_p|y_1,...,y_{p-1} \sim f_p(y_p|y_1,...,y_{p-1})$.
-Si on applique un échantillonneur de Gibbs à plusieurs étapes à ces lois conditionnelles complètes, en supposant qu'elles puissent toutes être simulées, on obtient une chaîne de Markov $(Y^{(t)})_t$ qui converge vers f et donc une sous-chaîne $(X^{(t)})_t$ qui converge vers g .

-Comme dernière illustration d'un échantillonneur de Gibbs sur un modèle à variable latente, nous avons l'échantillonneur par tranches (slice sampler en anglais), qui fonctionne comme une démarginalisation générique.

7.3 La connexion EM-Gibbs :

-Comme mentionné précédemment, l'algorithme EM peut être considéré comme un précurseur de l'échantillonneur de Gibbs à deux étapes dans les modèles de données manquantes , en ce qu'il exploite de même la distribution conditionnelle des variables manquantes.

-Si $X \sim g(x|\theta)$ représente les données observées , et si on augmente les données avec z , ou $A \sim f(x,z|\theta)$, alors on a les vraisemblances des données complétes et des données incomplétes :

$$L^C(\theta|x,z) = f(x,z|\theta) \text{ et } L(\theta|x) = g(x|\theta)$$

avec la densité de données manquantes : $K(z|x,\theta) = \frac{L^c(\theta|x,z)}{L(\theta|x)}$.

-Si nous pouvons normaliser la vraisemblance des données complètes en θ , c'est à dire si : $\int L^C(\theta|x,zd\theta < \infty$, alors on définit :

$$L^*(\theta|x,z) = \frac{L^C(\theta|x,z)}{\int L^C(\theta|x,z)d\theta}$$

et on crée l'échantillonneur de Gibbs à 2 étapes :

$$1.\ Z|\theta \sim k(z|x,\theta)$$

$$2.\ \theta|z \sim L^*(\theta|x,z)$$

-Notez la connexion directe à un algorithme EM basé sur L^C et k.
- L'étape "E" de l'algorithme EM calcule la valeur espérée du log-vraisemblance sur z, souvent en calculant $E(Z|x,\theta)$ et en substituant la log-vraisemblance.
-Dans l'échantillonneur de Gibbs, cette étape est remplacée par la génération d'une variable aléatoire de la densité k.
-L'étape "M" de l'algorithme EM prend alors comme valeur actuelle de θ le maximum de la log-vraisemblance espérée des données complètes.
-Dans l'échantillonneur de Gibbs, cette étape est remplacée par la génération d'une valeur de θ à partir de L^* , la vraisemblance normalisée des données complètes .

-La validité de l'échantillonneur "EM/Gibbs" suit de manière simple dès sa construction.
-Le noyau de transition de la chaîne de Markov est :

$$K(\theta,\theta'|x) = \int_Z K(z|x,\theta)L^*(\theta'|x,z)dz$$

et on peut montrer que la distribution invariante de la chaîne est la vraisemblance des données incomplètes, c'est-à-dire :

$$L(\theta'|x) = \int_\Theta K(\theta,\theta'|x)L(\theta|x)d\theta$$

Puisque $L(\theta'|x,z)$ est intégrable en θ, $L(\theta'|x)$ l'est aussi, et donc la distribution invariante est une densité propre ;
La chaîne de Markov est donc positive et la convergence s'en suit .
- Notons que l'algorithme EM est appliquée assez souvent en apprentissage automatique .

7.4 Réseaux de neurones :

- Les réseaux de neurones fournissent un autre type de modèle à données manquantes où les méthodes de simulation sont presque toujours nécessaires.
- Ces modèles sont fréquemment utilisés dans la classification et la reconnaissance de formes, ainsi qu'en robotique et en vision par ordinateur .

- En dehors du vocabulaire biologique et la connexion idéaliste avec les neurones réels, la théorie des réseaux de neurones couvre :
(1) la modélisation des relations non linéaires entre variables explicatif et dépendante (expliqué) ,
(2) l'estimation des paramètres de ces modèles à partir d'un échantillon (d'apprentissage)

- Bien que la littérature sur les réseaux de neurones évite généralement la modélisation probabiliste , ces modèles peuvent être analysés et estimés d'un point de vue statistique .
- Ils peuvent également être considérés comme un type particulier de problème d'estimation, où un enjeu majeur est alors l'identifiabilité.

- Un exemple classique simple d'un réseau de neurones est le modèle multicouche (également appelé le modèle de rétropropagation) qui relie les variables explicatives $x = (x_1, ..., x_n)$ aux variables dépendantes $y = (y_1, ..., y_n)$ à travers une "couche" cachée $h = (h_1, ..., h_p)$ ou $k = 1, ..., p$; $l = 1, ..., p$:

$$h_k = f(\alpha_{k_0} + \sum_j \alpha_{kj} x_j)$$

$$E(Y_l|h) = g(\beta_{l0} + \sum_{k=1}^{p} \beta_{lk} h_k)$$

et $var(Y_l) = \sigma^2$.
- Les fonctions f et g sont connues (ou choisies arbitrairement) à partir de catégories telles que : le seuil $f(t) = \mathbb{1}_{(t>0)}$, hyperbolique $f(t) = tan(t)$ ou sigmoid $f(t) = \frac{1}{1+e^{-t}}$.

- A titre d'exemple, considérons le problème de la reconnaissance des caractères, où l'écriture manuscrite est automatiquement déchiffrés.
- Les x peuvent correspondre à des caractéristiques géométriques d'un caractère numérisé, ou aux niveaux de gris des pixels, et les y sont les 26 lettres de l'alphabet, plus des symboles latéraux. (Voir Le Cun et al. 1989 pour une modélisation basée sur un échantillon de 7291 images 16×16 pixels, pour 9760 paramètres.)
- La vraisemblance du modèle multicouche comprend alors les paramètres $\alpha = (\alpha_{kj})$ et $\beta = (\beta_{lk})$ dans une structure non linéaire.
- En supposant la normalité, pour les observations (Y_t, X_t) , $t = 1, 2, ..., T$, la log-vraisemblance s'écrit :

$$l(\alpha, \beta|x, y) = -\sum_{t=1}^{T} \sum_{l=1}^{n} \frac{y_{tl} - E(y_{tl}|x_{tl})^2}{2\sigma^2}$$

- Une fonction objectif similaire peut être dérivée en utilisant un critère des moindres carrés.
- la maximisation de $l(\alpha, \beta|x, y)$ implique la détection et l'élimination de nombreux modes locaux .

7.4.1 Exemple : Gibbs sur des données censurées :

- On va traité un modèle à données censurées comme un modèle à données manquantes .
- On identifie $g(x|\theta)$ à la fonction de vraisemblance :

$$g(x|\theta) = L(\theta|x) \propto \prod_{i=1}^{m} e^{-(x_i-\theta)^2/2}$$

et

$$f(x, z|\theta) = L(\theta|x, z) \propto \prod_{i=1}^{m} e^{-(x_i-\theta)^2/2} \prod_{i=m+1}^{n} e^{-(z_i-\theta)^2/2}$$

est la vraisemblance des données complètes.

- Etant donnée une distribution a priori sur θ, $\pi(\theta)$, on peut alors créer un échantillonneur de Gibbs qui itère entre les distributions conditionnelles $\pi(\theta|x, z)$ et $f(z|x, \theta)$ et a pour distribution stationnaire $\pi(\theta|x, z)$, la distribution a posteriori de (θ, z) .
- Si on prend une loi a priori uniforme $\pi(\theta) = 1$, la distribution conditionnelle de $\theta|x, z$ est donnée par :

$$\theta|x, z \sim \mathcal{N}(\frac{m\overline{x} + (n-m)\overline{z}}{n}, \frac{1}{n})$$

et la distribution conditionnelle de $Z|x, \theta$ est le produit des lois normales tronquées :

$$Z_i|x, \theta \sim \varphi(z-\theta)\mathbb{1}(z > a)/(1-\phi(a-\theta))$$

où φ et ϕ sont respectivement les fonctions de masse et de répartition de la loi normale, puisque chaque Z_i doit être supérieur au point de troncation a .
- Voici le code R :

```
xbar=0;a=1;n=30;m=20;N=10^3;z=NULL;theta=NULL;theta[1]=rnorm(1,0,1)
for(i in 1:(N-1)){u=rnorm(1,theta[i],1)
while(u<=a){u=rnorm(1,theta[i],1)}
z[i]=(u/(1-pnorm(a-theta[i],0,1)))
theta[i+1]=rnorm(1,(m*xbar+(n-m)*mean(z))/n,sqrt(1/n))}
hist(theta,nclass=50);windows();hist(z,nclass=50)
```

- Nous avons ci-dessous le résultat des histogrammes .

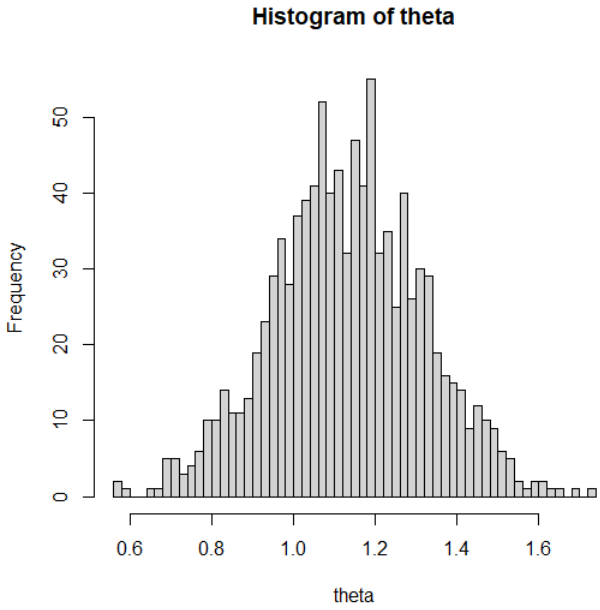

FIGURE 7.1 – densité de θ.

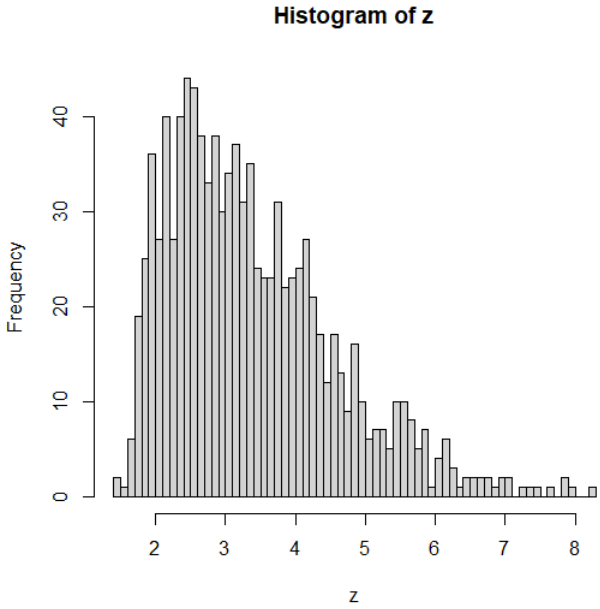

FIGURE 7.2 – densité de z.

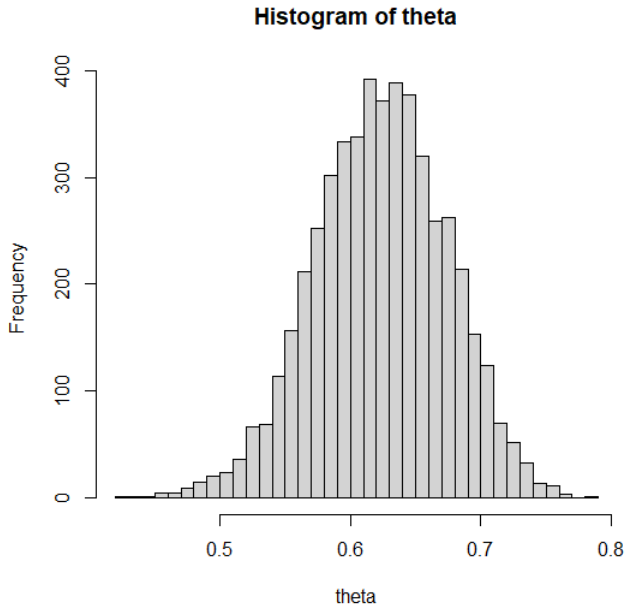

FIGURE 7.3 – densité de θ.

7.4.2 Exemple : Données multinomiales groupées :

-On considére le modèle multinomial
$\mathcal{M}(n, \frac{1}{2}+\frac{\theta}{4}; \frac{1}{4}(1-\theta); \frac{1}{4}(1-\theta); \frac{\theta}{4})$
en introduisant la variable latente Z avec la démarginalisation
$(z, x_1 - z, x_2, x_3, x_4) \sim \mathcal{M}(n, \frac{1}{2}+\frac{\theta}{4}; \frac{1}{4}(1-\theta); \frac{1}{4}(1-\theta); \frac{\theta}{4})$.
-Si on prend une distribution a priori uniforme sur les paramétres , les lois conditionnelles complètes peuvent être obtenues :
$\theta \sim \mathcal{B}e(z + x_4 + 1, x_2 + x_3 + 1)$ et $z \sim \mathcal{B}in(x_1, \frac{\theta}{2+\theta})$.
-Ce qui mène à l'échantillonneur de Gibbs codé sous R :

```
x1=125;x2=18;x3=20;x4=34;n=5000;theta=NULL;z=NULL
theta[1]=rbeta(1,x4+1,x2+x3+1);for(t in 1:(n-1)){
z[t]=rbinom(1,x1,theta[t]/(2+theta[t]))
theta[t+1]=rbeta(1,z[t]+x4+1,x2+x3+1)};hist(theta,nclass=50)
windows();hist(z,nclass=50)
```

pour certaines données x1 ;x2 ;x3 ;x4 .
-Cet exemple montre un cas où l'algorithme EM et l'échantillonnage de Gibbs s'appliquent tous deux.
- Nous avons ci-dessus et ci-dessous le résultat des histogrammes .

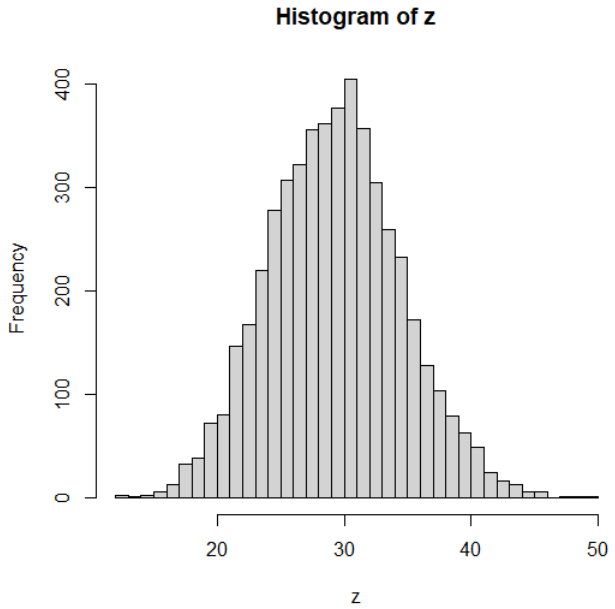

FIGURE 7.4 – densité de z.

CONCLUSION

- nous avons montré l'application de méthodes mcmc à quelques modéles d'apprentissage automatique à travers des exemples indirectes .
- nous avons aussi rappeler que les méthodes mcmc sont trés liées aux méthodes bayésiennes et que ces derniéres peuvent étre appliquées en apprentissage automatique .
- et pour finir nous avons aussi montré le lien entre l'échantillonneur de Gibbs et les modéles à varibales latentes qui jouent un role centrale en apprentissage automatique à travers entre autre les réseaux de neurones .
- En conclusion les méthodes mcmc représentent un outil puissant de simulation et d'approximation qui convient à beaucoup de problémes y compris l'apprentissage automatique .

BIBLIOGRAPHIE

[Collection Pratique R] Christian P. Robert, George Casella - Méthodes de Monte-Carlo avec R (2011, Springer Paris)

[Springer Texts in Statistics] Christian P. Robert - George Casella - Monte Carlo Statistical Methods (2004, Springer)

(Adaptive Computation and Machine Learning) Kevin P. Murphy - Machine Learning - A Probabilistic Perspective-The MIT Press (2012)

[Statistique et probabilités appliquées] Michel Lejeune - Statistique. La théorie et ses applications, Deuxième Édition (2010, Springer)

Philippe BESSE Data mining II. Modélisation Statistique Apprentissage - Université Paul Sabatier - www.lsp.ups-tlse.fr/Besse

[Springer series in statistics] Olivier Cappé, Eric Moulines, Tobias Ryden - Inference in Hidden Markov Models (2005, Springer)