

Méthodes de Monte Carlo par Chaînes de Markov

Gersende FORT

LTCI, CNRS - Telecom ParisTech

Méthodes MCMC

Introduction

Méthodes MCMC

Vers les procédures MCMC adaptatives

Algorithmes MCMC adaptatifs

Optimisation Stochastique

Bibliographie et Collaborations

Introduction

- Les méthodes **de Monte Carlo** permettent d'obtenir des réalisations de v.a. de loi $\pi \, d\nu$ spécifiée par l'utilisateur.
- Elles sont en particulier utilisées comme
 - méthodes numériques stochastiques pour l'approximation d'intégrales

$$\int \phi \pi \, d\nu = \mathbb{E}[\phi(X)] \quad X \sim \pi \, d\nu$$

- méthodes d'exploration d'une loi (recherche des maxima, de la médiane, des zones de plus haute densité, ...)

On peut regrouper les méthodes de Monte Carlo en deux grandes familles

1. **méthodes par acceptation-rejet** : se donner une loi instrumentale $g d\nu$,

- tirer des v.a. i.i.d. de loi $g d\nu$
- et accepter chaque tirage que si \dots - (critère faisant intervenir π et g) -
- approcher la loi cible par

$$\pi d\nu \approx \frac{1}{n} \sum_{k=1}^n \delta_{X_k} \quad \text{i.e. } \mathbb{E}[\phi(X)] \approx \frac{1}{n} \sum_{k=1}^n \phi(X_k)$$

On peut regrouper les méthodes de Monte Carlo en deux grandes familles

1. **méthodes par acceptation-rejet** : se donner une loi instrumentale $g \, d\nu$,

- tirer des v.a. i.i.d. de loi $g \, d\nu$
- et accepter chaque tirage que si \dots - (critère faisant intervenir π et g) -
- approcher la loi cible par

$$\pi \, d\nu \approx \frac{1}{n} \sum_{k=1}^n \delta_{X_k} \quad \text{i.e. } \mathbb{E}[\phi(X)] \approx \frac{1}{n} \sum_{k=1}^n \phi(X_k)$$

2. **méthodes par pondération** : se donner une loi instrumentale $g \, d\nu$,

- tirer des v.a. i.i.d. de loi g
- et approcher la loi cible par

$$\pi \, d\nu \approx \frac{1}{n} \sum_{k=1}^n \frac{\pi(X_k)}{g(X_k)} \delta_{X_k} \quad \text{i.e. } \mathbb{E}[\phi(X)] \approx \frac{1}{n} \sum_{k=1}^n \frac{\pi(X_k)}{g(X_k)} \phi(X_k)$$

Comment choisir la loi instrumentale g ?

C'est une condition d'autant plus contraignante que

- la dimension de l'espace de simulation X est grande.
- la loi cible est multimodale

Méthodes Monte Carlo par Chaînes de Markov

Etant donnée une loi cible $\pi d\nu$, il s'agit de construire une chaîne de Markov de sorte que

- $\pi d\nu$ soit la probabilité stationnaire de la chaîne.
- la chaîne vérifie des propriétés de convergence :
 - de type **ergodicité** :

$$\sup_{\{\phi, |\phi|_\infty \leq 1\}} \left| \mathbb{E}_x [\phi(X_n)] - \int \phi \pi d\nu \right| \rightarrow 0 \quad \forall x \in X$$

- ou de type **loi des grands nombres** :

$$\frac{1}{n} \sum_{k=1}^n \phi(X_k) \xrightarrow{\text{p.s.}} \int \phi \pi d\nu$$

- et éventuellement de type **TLC**

$$\frac{\sqrt{n}}{\sigma_\phi} \left(\frac{1}{n} \sum_{k=1}^n \phi(X_k) - \int \phi \pi d\nu \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Principalement, deux grandes familles d'algorithmes MCMC

④ algorithmes Hastings-Metropolis

- se donner un noyau de proposition $q(x, y) d\nu(y)$.
- Mise à jour $X_n \rightarrow X_{n+1}$: proposer un candidat $Y \sim q(X_n, \cdot) d\nu$ et

$$\begin{cases} X_{n+1} = Y & \text{avec probabilité } \alpha(X_n, Y) \\ X_{n+1} = X_n & \text{sinon} \end{cases}$$

où le ratio d'acceptation-rejet est donné par

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

Principalement, deux grandes familles d'algorithmes MCMC

① algorithmes Hastings-Metropolis

② algorithmes de Gibbs

- Décomposer l'espace d'état en sous-blocs $x = (x^{(1)}, \dots, x^{(d)})$
- Mise à jour $X_n \rightarrow X_{n+1}$: simuler le sous-bloc $X_{n+1}^{(i)}$ sous la loi conditionnelle du bloc i sachant les autres blocs

$$y \mapsto \frac{\pi \left(X_{n+1}^{(1)}, \dots, X_{n+1}^{(i-1)}, y, X_n^{(i+1)}, \dots, X_n^{(d)} \right)}{\int \pi \left(X_{n+1}^{(1)}, \dots, X_{n+1}^{(i-1)}, u, X_n^{(i+1)}, \dots, X_n^{(d)} \right) du}$$

Principalement, deux grandes familles d'algorithmes MCMC

① algorithmes Hastings-Metropolis

② algorithmes de Gibbs

③ et les versions hybrides : *Metropolis within Gibbs*

dans l'algorithme de Gibbs, remplacer la simulation sous la loi conditionnelle par une étape de Hastings-Metropolis dont la loi cible est cette loi conditionnelle.

Comportement des chaînes de Markov

- Par nature, la construction d'une chaîne de Markov $\{X_n, n \geq 0\}$ est itératif; la procédure algorithmique traduit une loi de transition

$$\mathbb{E}[\phi(X_{n+1})|\mathcal{F}_n] = \int \phi(y)P(X_n, dy)$$

- $(x, A) \mapsto P(x, A)$ est un noyau de transition. En particulier :

$$A \mapsto P(x, A)$$

est la "loi de X_{n+1} sachant $X_n = x$ ".

↪ [STA] A quelles conditions sur le noyau P , la chaîne (i) possède-t-elle une loi invariante, (ii) converge-t-elle vers cette loi et à quelle vitesse, (iii) contrôle de l'erreur L^p de l'approximation Monte Carlo, (iv) existence de théorèmes limites tq LGN, TCL, ...

Vers les procédures MCMC adaptatives

Dans la mise en oeuvre de ces algorithmes, différents “degrés de liberté” laissés à l'utilisateur : par exemple,

- Hastings-Metropolis : quelle loi de proposition choisir ?
- Gibbs : quelle partition de l'espace d'état X ? quel ordre de mise à jour des composantes et avec quelle fréquence ?

⇒ définition d'algorithmes **adaptatifs** construits pour

- que l'algorithme apprenne de son propre comportement passé, comment modifier les **paramètres d'implémentation**
- et les modifie au fur et à mesure de son déroulement : chaque itération comporte une étape de mise à jour de la chaîne, et de mise à jour des paramètres d'implémentation.

Méthodes MCMC

Algorithmes MCMC adaptatifs

Paramètres d'implémentation du SRW-HM

Choix des paramètres d'implémentation

Algorithmes MCMC adaptatifs

Optimisation Stochastique

Bibliographie et Collaborations

Dans la suite de cette section : illustrations à partir

- de l'algorithme de Hastings-Metropolis à marche aléatoire symétrique (SRW-HM)
- pour une densité cible π par rapport à la mesure de Lebesgue sur \mathbb{R}^d

Paramètres d'implémentation du SRW-HM (marche aléatoire symétrique Hastings-Metropolis)

Lorsque le noyau de proposition est de la forme

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

Exemple: $(x, y) \mapsto \frac{1}{\sqrt{2\pi}\sigma} \exp(-0.5(y - x)^2/\sigma^2)$

une itération $X_n \rightarrow X_{n+1}$ devient :

- tirer $Z \sim q(z)dz$

- Poser

$$\begin{cases} X_{n+1} = X_n + Z & \text{avec probabilité } 1 \wedge \frac{\pi(X_n + Z)}{\pi(X_n)} \\ X_{n+1} = X_n & \text{sinon} \end{cases}$$

Paramètres d'implémentation du SRW-HM (marche aléatoire symétrique Hastings-Metropolis)

Lorsque le noyau de proposition est de la forme

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

Exemple: $(x, y) \mapsto \frac{1}{\sqrt{2\pi}\sigma} \exp(-0.5(y - x)^2/\sigma^2)$

une itération $X_n \rightarrow X_{n+1}$ devient :

- tirer $Z \sim q(z)dz$

- Poser

$$\begin{cases} X_{n+1} = X_n + Z & \text{avec probabilité } 1 \wedge \frac{\pi(X_n + Z)}{\pi(X_n)} \\ X_{n+1} = X_n & \text{sinon} \end{cases}$$

En particulier, toute configuration $X_n + Z$ qui augmente la valeur de la densité cible est systématiquement acceptée.

Paramètres d'implémentation du SRW-HM (marche aléatoire symétrique Hastings-Metropolis)

Lorsque le noyau de proposition est de la forme

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

Exemple: $(x, y) \mapsto \frac{1}{\sqrt{2\pi}\sigma} \exp(-0.5(y - x)^2/\sigma^2)$

une itération $X_n \rightarrow X_{n+1}$ devient :

- tirer $Z \sim q(z)dz$

- Poser

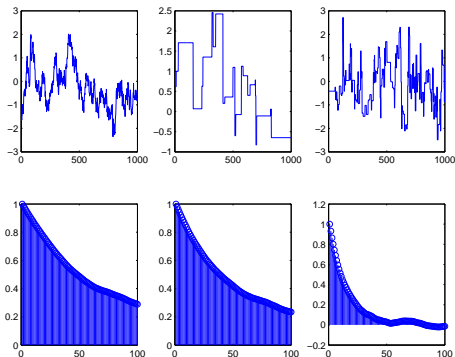
$$\begin{cases} X_{n+1} = X_n + Z & \text{avec probabilité } 1 \wedge \frac{\pi(X_n + Z)}{\pi(X_n)} \\ X_{n+1} = X_n & \text{sinon} \end{cases}$$

En particulier, toute configuration $X_n + Z$ qui augmente la valeur de la densité cible est systématiquement acceptée.

↪ En pratique, quel choix pour la loi de proposition q ?

Situation 1 : "Vitesse de déplacement"

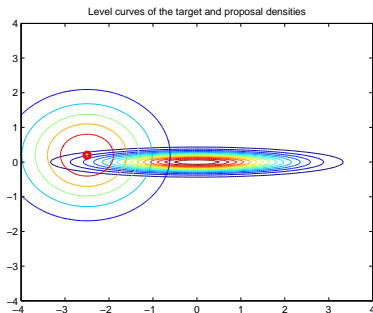
- Chaîne de Markov SRW-HM à valeur dans \mathbb{R}
- Avec loi de proposition $q \equiv \mathcal{N}(0, \sigma^2)$ pour différentes valeurs de σ^2 .



Pour différentes valeurs de σ : [haut] trajectoire de la chaîne. [bas] fonction d'auto-corrélation.

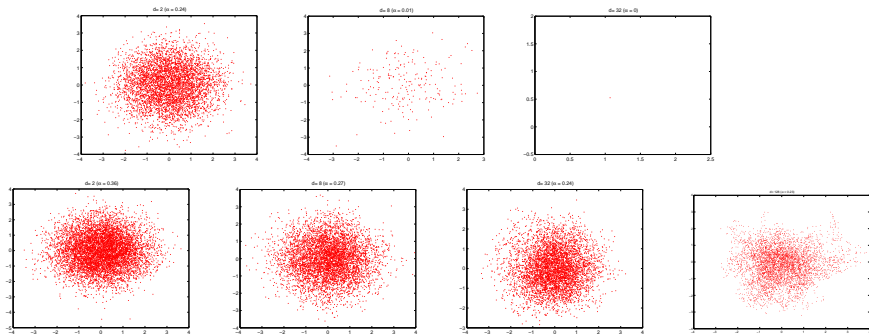
Situation 1bis : Structure de covariance de la densité cible

- Chaîne de Markov SRW-HM à valeur dans \mathbb{R}^2 de densité cible $\mathcal{N}_2(0, \Sigma)$
- Avec loi de proposition $\mathcal{N}_d(0, \text{Id})$



Situation 2 : Dimension de l'espace de simulation $X \subseteq \mathbb{R}^d$

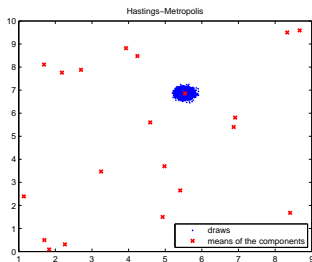
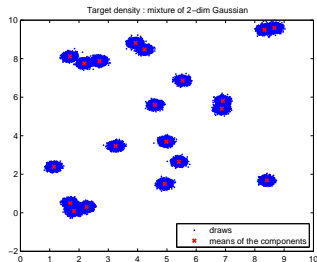
- Chaîne de Markov SRW-HM à valeur dans \mathbb{R}^d de loi cible $\mathcal{N}_d(0, \text{Id})$
- Avec loi de proposition $\mathcal{N}_d(0, \sigma^2 \text{Id})$



Pour différentes valeurs de d : chaîne projetée dans \mathbb{R}^2 [haut] σ est indépendant de d [bas] σ est de la forme c/\sqrt{d}

Situation 3 : Multimodalité de la densité cible

- Chaîne de Markov SRW-HM à valeur dans \mathbb{R}^2 de loi cible $\sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$
- Avec loi de proposition $\mathcal{N}_d(0, \sigma^2 \text{Id})$



[gauche] Tirages i.i.d. sous la loi cible. [droite] Chaîne SRW-HM

Choix des paramètres d'implémentation

Quelques réponses (aux Pbmes 1 et 2) dans le cas d'une loi de proposition

$\mathcal{N}_d(0, \Sigma)$:

$$\Sigma = \frac{2.38^2}{d} \Sigma_\pi$$

Choix des paramètres d'implémentation

Quelques réponses (aux Pbmes 1 et 2) dans le cas d'une loi de proposition

$\mathcal{N}_d(0, \Sigma)$:

$$\Sigma = \frac{2.38^2}{d} \Sigma_\pi$$

Réponses apportées

- sous des conditions plus ou moins restrictives sur la densité cible.
- pour différents algorithmes de type Hastings-Metropolis.
- basées sur la vitesse de convergence optimale de "limites diffusives" de la chaîne (renormalisée).

↪ [STA] *Méthodes de scaling pour l'étude fine d'échantillonneurs MCMC lorsque la dimension de l'espace de simulation $d \rightarrow \infty$ (et pour la dim infinie, Hilbert)*

Exemple d'algorithme adaptatif

En pratique, Σ_π est inconnue \implies coupler simulation et estimation

Notons

- θ_n l'estimation courante de la matrice Σ_π .
- P_θ le noyau de transition d'une chaîne SRW-HM de loi de proposition $\mathcal{N}_d(0, \theta)$

Une itération $(\theta_n, X_n) \rightarrow (\theta_{n+1}, X_{n+1})$ devient

- **Simuler** $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$
- **Mettre à jour l'estimation** de Σ_π par la formule

$$\theta_{n+1} = \theta_n + \frac{1}{n+1} \left((X_{n+1} - \mu_{n+1})(X_{n+1} - \mu_{n+1})^T - \theta_n \right)$$

Cas général: l'idée

- Famille de lois de proposition $\{q_\theta, \theta \in \Theta\}$ et donc, famille de noyaux de transition $\{P_\theta, \theta \in \Theta\}$ ayant tous la même loi invariante π .
- Il existe une valeur optimale θ_* solution d'un critère incalculable

$$\theta_* \text{ tel que } \int H(\theta_*, x)\pi(dx) = 0 \quad (1)$$

- Répéter:

Simulation $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$

Adaptation $\theta_{n+1} = \Xi(\theta_n, n, X_{n+1})$

où Ξ "oriente" $\{\theta_n, n \geq 0\}$ vers θ_* . On produit une suite aléatoire $\{X_n, n \geq 0\}$ qui (*espère-t-on*) approche π .

\hookrightarrow [STA] Convergence des schémas d'approximation stochastique pour la résolution de (1).

Mais danger !

- On a une famille de noyaux $\{P_\theta, \theta \in \Theta\}$ qui ont tous **même** loi invariante π . On les utilise dans un ordre **aléatoire**. Est-ce que l'on obtient une suite aléatoire de loi stationnaire π ?

Mais danger !

- On a une famille de noyaux $\{P_\theta, \theta \in \Theta\}$ qui ont tous **même** loi invariante π . On les utilise dans un ordre **aléatoire**. Est-ce que l'on obtient une suite aléatoire de loi stationnaire π ?
- Non, pas nécessairement.

Contre-exemple:

$$X_{t+1} \sim \begin{cases} P_0(X_t, \cdot) & \text{si } X_t = 0 \\ P_1(X_t, \cdot) & \text{si } X_t = 1 \end{cases} \quad \text{où } P_\ell = \begin{pmatrix} 1 - t_\ell & t_\ell \\ t_\ell & 1 - t_\ell \end{pmatrix}.$$

Les deux noyaux ont même loi invariante, égale à $\pi = (1/2, 1/2)$.

La matrice de transition de $(X_t)_t$ est

$$\tilde{P} = \begin{pmatrix} 1 - t_0 & t_0 \\ t_1 & 1 - t_1 \end{pmatrix} \quad \text{avec loi invariante } \tilde{\pi} \propto (t_1, t_0)$$

Bien-fondé des algorithmes adaptatifs

- On trouve dans la littérature des conditions suffisantes sur le schéma d'adaptation $\theta_n \rightarrow \theta_{n+1}$ et sur les noyaux de transition $\{P_\theta, \theta \in \Theta\}$ qui garantissent que la suite $\{X_n, n \geq 0\}$ *approche* π

$$\begin{aligned}\lim_n \mathbb{E}[\phi(X_n)] &= \int \phi(x) \pi \nu(dx) \\ \frac{1}{n} \sum_{k=1}^n \phi(X_k) &\xrightarrow{\text{p.s.}} \int \phi(x) \pi \nu(dx) \\ \sqrt{n} \left(\frac{1}{n} \sum_{k=1}^n \phi(X_k) - \int \phi(x) \pi \nu(dx) \right) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{\phi, \theta_\star}^2).\end{aligned}$$

↔ [STA] Conditions nécessaires pour la convergence des algorithmes MCMC adaptatifs

Méthodes MCMC

Algorithmes MCMC adaptatifs

Optimisation Stochastique

Introduction

Algorithme Gradient-Proximal

Algorithme Gradient-Proximal Stochastique

Bibliographie et Collaborations

Introduction (1/2)

- Un algorithme itératif exact

$$\theta_{n+1} = T(\theta_n)$$

dont on sait qu'il possède une *fonction de Lyapunov* V

$$V(T(\theta)) \leq V(\theta)$$

$$V(T(\theta)) < V(\theta) \quad \text{ssi} \quad \theta \notin \mathcal{L}$$

- Sous hyp adéquates, $\lim_n \text{dist}(\theta_n, \mathcal{L}) = 0$.

Introduction (1/2)

- Un algorithme itératif exact

$$\theta_{n+1} = T(\theta_n)$$

dont on sait qu'il possède une *fonction de Lyapunov* V

$$V(T(\theta)) \leq V(\theta)$$

$$V(T(\theta)) < V(\theta) \quad \text{ssi} \quad \theta \notin \mathcal{L}$$

- Sous hyp adéquates, $\lim_n \text{dist}(\theta_n, \mathcal{L}) = 0$.

- Si on perturbe à chaque itération l'opérateur T ,

$$\theta_{n+1} = F_n(\theta_n)$$

à quelles conditions sur cette perturbation a-t-on: $\lim_n \text{dist}(\theta_n, \mathcal{L}) = 0$?

Introduction (2/2)

Dans un contexte d'optimisation stochastique :

F_n est une perturbation aléatoire de T

④ Algorithme EM et ses versions stochastiques:

Objectif $\theta_\star \in \operatorname{argmax}_\theta \int p(z_{1:n}, y_{1:n}; \theta) d\mu(z_{1:n})$

mapping T $\theta_{n+1} = \operatorname{argmax}_\theta \mathbb{E} [\log p(Z_{1:n}, y_{1:n}; \theta) | y_{1:n}, \theta_n]$

mapping F_n $\theta_{n+1} = \operatorname{argmax}_\theta \frac{1}{n} \sum_{k=1}^n \log p(Z_{1:n}^{(k)}, y_{1:n}; \theta)$

Introduction (2/2)

Dans un contexte d'optimisation stochastique :

F_n est une perturbation aléatoire de T

- ① Algorithme EM et ses versions stochastiques:
- ② Algorithme gradient-proximal et ses versions stochastiques:

Objectif

$$\theta_\star \in \operatorname{argmin}_\theta \{g(\theta) + f(\theta)\} \quad \text{avec } g \text{ pénalité convexe, } \nabla f \text{ lipschitzien}$$

mapping T $\theta_{n+1} \in \operatorname{Prox}_{\gamma_{n+1}g} (\theta_n - \gamma_{n+1} \nabla f(\theta_n))$

mapping F_n $\theta_{n+1} \in \operatorname{Prox}_{\gamma_{n+1}g} \left(\theta_n - \gamma_{n+1} \widehat{\nabla f(\theta_n)}^{(n)} \right)$

Introduction (2/2)

Dans un contexte d'optimisation stochastique :

F_n est une perturbation aléatoire de T

- 1 Algorithme EM et ses versions stochastiques:
- 2 Algorithme gradient-proximal et ses versions stochastiques:

↔ [STA] Dans des contextes d'inférence statistique dans les modèles à données cachées, où la vraisemblance est définie comme une intégrale incalculable : quel outil d'optimisation et quelle approximation Monte Carlo ?

Algorithme Gradient Proximal (1/2)

$$(P) \quad \min_{\theta \in \Theta} V(\theta) \quad V(\theta) = f(\theta) + g(\theta),$$

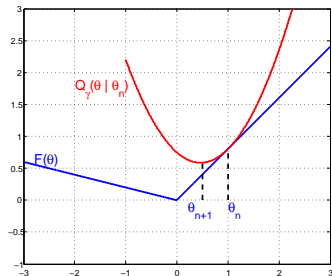
dans le cas où

- Θ espace euclidien de dimension finie, avec norme $\|\cdot\|$ et produit scalaire $\langle \cdot, \cdot \rangle$.
- $g : \Theta \rightarrow]-\infty, +\infty]$ convexe, non identiquement égale à $+\infty$, et semi-continue inférieurement.

La fonction $f : \Theta \rightarrow \mathbb{R}$ est C^1 et il existe L tq pour tout $\theta, \theta' \in \Theta$,

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\| .$$

Algorithme Gradient Proximal (2/2)



$$V(\vartheta) \leq Q_\gamma(\vartheta | \theta_n)$$

$$V(\theta_n) = Q_\gamma(\theta_n | \theta_n)$$

$$\theta_{n+1} = \min_{\vartheta \in \Theta} \left\{ g(\vartheta) + \frac{1}{2\gamma_{n+1}} \|\vartheta - (\theta_n - \gamma_{n+1} \nabla f(\theta_n))\|^2 \right\} \quad \gamma_{n+1} \in]0, 1/L[$$

Algorithme Gradient-Proximal Stochastique (1/2)

- En inférence statistique dans les modèles à données cachées

fonction g une pénalité de *sparsité* sur le paramètre

fonction f – la log-vraisemblance des observations la dépendance en les

observations est omise

$$\begin{aligned} f(\theta) &= -\log \int p(x; \theta) d\mu(x) \\ \nabla f(\theta) &= -\int \{\nabla \log p(x; \theta)\} \frac{p(x; \theta)}{\int p(u; \theta) d\mu(u)} d\mu(x) \\ &\approx -\frac{1}{m} \sum_{k=1}^m \nabla \log p(X^{(k)}; \theta) \end{aligned}$$

Algorithme Gradient-Proximal Stochastique (1/2)

- En inférence statistique dans les modèles à données cachées

fonction g une pénalité de *sparsité* sur le paramètre

fonction f – la log-vraisemblance des observations la dépendance en les

observations est omise

$$\begin{aligned} f(\theta) &= -\log \int p(x; \theta) d\mu(x) \\ \nabla f(\theta) &= -\int \{\nabla \log p(x; \theta)\} \frac{p(x; \theta)}{\int p(u; \theta) d\mu(u)} d\mu(x) \\ &\approx -\frac{1}{m} \sum_{k=1}^m \nabla \log p(X^{(k)}; \theta) \end{aligned}$$

- Plus généralement, tout contexte où ∇f a un coût de calcul prohibitif et/ou est incalculable,
MAIS peut être relu comme une espérance.

Algorithme Gradient-Proximal Stochastique (2/2)

$$\theta_{n+1} = \min_{\vartheta \in \Theta} \left\{ g(\vartheta) + \frac{1}{2\gamma_{n+1}} \|\vartheta - (\theta_n - \gamma_{n+1} H_{n+1})\|^2 \right\} \quad \gamma_{n+1} \in]0, 1/L[$$

- H_{n+1} est une approximation Monte Carlo de $\nabla f(\theta_n)$

$$\nabla f(\theta_n) = \int H_{\theta_n}(x) \, d\pi(x)$$

$$H_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{k,n})$$

- l'approximation Monte Carlo peut-être
 - non biaisée: cas de l'approximation i.i.d.
 - biaisée: cas de l'approximation MCMC.

Résultats de Convergence (1/2)

- 1 Peut-on assurer que cet algorithme perturbé ait le même comportement limite que l'algorithme exact :
 - convergence vers \mathcal{L}
 - vitesse de convergence

Résultats de Convergence (1/2)

- 1 Peut-on assurer que cet algorithme perturbé ait le même comportement limite que l'algorithme exact :
 - convergence vers \mathcal{L}
 - vitesse de convergence

Oui

- 2 et à quelles conditions sur
 - le choix des pas $\{\gamma_n, n \geq 0\}$ constant / décroissant, à quelle vitesse
 - la nature de la perturbation MCMC avec quelles vitesses de convergence des chaînes
 - la taille des blocs Monte Carlo $\{m_n, n \geq 0\}$ constant / croissant, à quelle vitesse

Résultats de Convergence (1/2)

- 1 Peut-on assurer que cet algorithme perturbé ait le même comportement limite que l'algorithme exact :
 - convergence vers \mathcal{L}
 - vitesse de convergence

Oui

- 2 et à quelles conditions sur
 - le choix des pas $\{\gamma_n, n \geq 0\}$ constant / décroissant, à quelle vitesse
 - la nature de la perturbation MCMC avec quelles vitesses de convergence des chaînes
 - la taille des blocs Monte Carlo $\{m_n, n \geq 0\}$ constant / croissant, à quelle vitesse

Il suffit

- Chaînes géométriquement ergodiques (extensions possibles).
- γ_n fixe et m_n croissant.
- OU : γ_n décroissant et m_n fixe.

Résultats de Convergence (2/2)

$$\theta_{n+1} = \min_{\vartheta \in \Theta} \left\{ g(\vartheta) + \frac{1}{2\gamma_{n+1}} \|\vartheta - (\theta_n - \gamma_{n+1} H_{n+1})\|^2 \right\} \quad \gamma_{n+1} \in]0, 1/L[$$

où H_{n+1} est une approximation de $\nabla f(\theta_n)$.

- ➊ Résultats pour n'importe quelle perturbation H_{n+1} (pas nécessairement Monte Carlo, pas nécessairement stochastique).
- ➋ Sur la convergence et le contrôle explicite de convergence de $\{V(\theta_n), n \geq 0\}$ mais aussi de $\{\theta_n, n \geq 0\}$.

↔ *[STA] Algorithmes itératifs perturbés, dont perturbation Monte Carlo biaisée*

Méthodes MCMC

Algorithmes MCMC adaptatifs

Optimisation Stochastique

Bibliographie et Collaborations

Collaborations

Eric Moulines, Pascal Bianchi, Olivier Cappé.

Pierre Priouret (Paris 6), Randal Douc (SudParis), Benjamin Jourdain (Pons), Tony Lelièvre (Pons), Gabriel Stoltz (Pons), C. Robert (Dauphine), Pierre Vandekerkhove (Marne).

Florence Forbes (Grenoble), J.F. Aujol (Bordeaux), C. Dossal (Bordeaux).

C. Andrieu (UK), Yves Atchadé (USA), Sean Meyn (USA), Gareth Roberts (UK), Jeff Rosenthal (Canada), Matti Vihola (Finlande).

Bibliographie

Y. Atchadé, G. Fort and E. Moulines. On stochastic proximal gradient algorithms, Submitted, February 2014. arXiv:1402:2365 math.ST. Revised in January 2015.

G. Fort. Central Limit Theorems for Stochastic Approximation with Controlled Markov Chain Dynamics. EsaimPS, 19:60-80, 2015. Accepted for publication in May 2014. arXiv math.PR 1309.3116

G. Fort, E. Moulines, P. Priouret and P. Vandekerkhove. A Central Limit Theorem for Adaptive and Interacting Markov Chains.arXiv math.ST 1107-2574 Supplement paper Bernoulli 20(2):457-485, 2014.

G. Fort, E. Moulines and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. Ann. Statist. 39(6):3262-3289, 2012. G. Fort, S. Meyn, E. Moulines and P. Priouret. The ODE method for the stability of skip-free

Markov Chains with applications to MCMC. Ann. Appl. Probab. 18(2) :664-707, 2008.

R. Douc, G. Fort, E. Moulines and P. Soulier. Practical drift conditions for subgeometric rates of convergence. Ann. Appl. Probab. 14(3) :1353-1377, 2004.

G. Fort and E. Moulines. Convergence of the Monte-Carlo EM for curved exponential families. Ann. Stat. 31(4):1220-1259, 2003.