Optimisation stochastique et méthodes MCMC : adaptation et convergence

Gersende Fort

Institut de Mathématiques de Toulouse
CNRS et Univ. de Toulouse

Optimisation des méthodes de Monte Carlo : exemple 1

Algorithme de Hastings-Metropolis

- Etant donné

  une loi cible $\pi \, \mathrm{d}\lambda$
  une loi de proposition $q(x, y)\mathrm{d}\lambda(y)$

- Algorithme : produit une suite de points $\{X_0, X_1, \cdots\}$ itérativement via

  - proposer un candidat $Y_{k+1} \sim q(X_k, \cdot)\mathrm{d}\lambda$
  - puis acceptation/rejet

$$X_{k+1} = \begin{cases} Y_{k+1} & \text{avec probabilité } \alpha(X_k, y) \stackrel{\text{def}}{=} 1 \wedge \left(\frac{\pi(y) q(y, X_k)}{\pi(X_k) q(X_k, y)} \right) \\ X_k & \text{sinon} \end{cases}$$

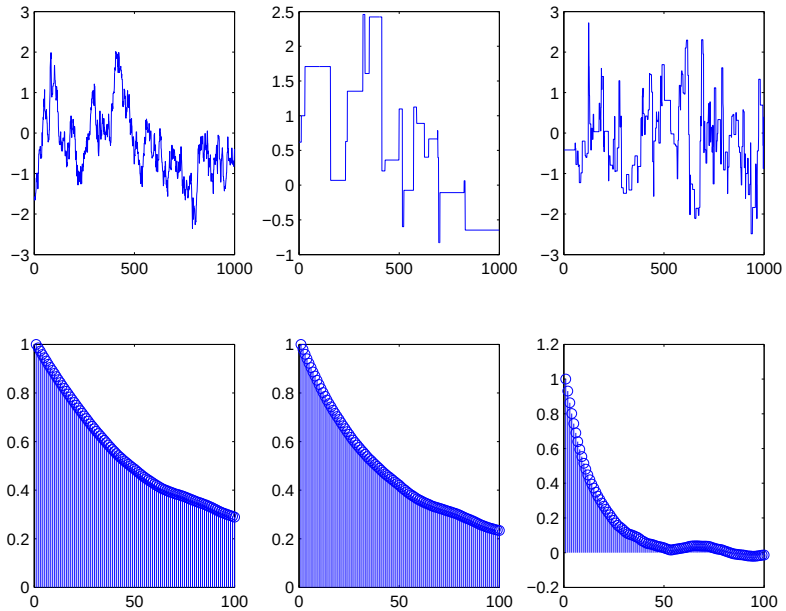
- Chaîne de Markov de noyau de transition

$$P(x, A) = \int_A \alpha(x, y) \; q(x, y)\mathrm{d}\lambda(y) + \delta_x(A) \int (1 - \alpha(x, y)) \; q(x, y)\mathrm{d}\lambda(y)$$

non stationnaire, mais ergodique.

Algorithme de Hastings-Metropolis : choix de la loi de proposition

- L'efficacité de l'échantillonneur dépend du mécanisme de proposition



- Résultats théoriques: parmi la famille de lois de proposition gaussiennes $q_\theta(x,\cdot) \sim \mathcal{N}_d(x,\theta)$, choix optimal Gelman, Roberts and Gilks (1996)

$$\theta_\star = \frac{2.38^2}{d}\ \mathbb{E}_\pi\left[(X - \mathbb{E}_\pi[X])\,(X - \mathbb{E}_\pi[X])^T\right]$$

inconnue !

Algorithme de Hastings-Metropolis : version adaptative Haario, Saksman and Taaminen (1999, 2001)

A l'itération $(n+1)$, étant donné le passé $\theta_{1:n}, X_{1:n}$,

- Tirer le nouveau point

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

selon un mécanisme dépendant du comportement passé de l'algorithme.

- Mise à jour du paramètre

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

selon un mécanisme d'approximation stochastique pour recherche du zero de la fonction

$$\theta \mapsto \mathrm{Cov}_\pi(X) - \theta$$

Plus généralement , MCMC adaptatif

- θ : pas nécessairement de dimension finie et peut dépendre du passé de processus auxiliaires (méthodes en *interactions*) Kou et al. (2006); Andrieu et al. (2007, 2008, 2011); Bercu et al. (2009); Brockwell et al. (2010)
- la procédure de mise à jour du paramètre peut ne pas être de type *approximation stochastique* Roberts and Rosenthal (2007); Andrieu and Thoms (2011); F., Jourdain, Lelièvre and Stoltz (2017)
- chaque noyau peut avoir sa propre mesure invariante π_θ Andrieu et al. (2007, 2008, 2011); Schreck, F. and Moulines (2013); F., Jourdain, Lelièvre and Stoltz (2015, 2016)

Mais le problème reste : si

$\star$ $\{(X_n, \theta_n), n \geq 0\}$ telle que

$$\mathcal{L}(X_{n+1} | \text{passé}_n) = P_{\theta_n}(X_n, \cdot)$$

$\{\theta_n, n \geq 0\}$ suite aléatoire dépendant du passé de l'algorithme

$\star$ chaque noyau P_θ a une unique loi invariante π_θ

à quelles conditions a-t-on convergence de

$$\frac{1}{n} \sum_{k=1}^{n} \phi(X_k), \qquad \mathcal{L}(X_n)$$

Maximum de Vraisemblance (pénalisé) dans des modèles à vraisemblance non explicite

Contexte

- N observations : $\mathbf{Y} = (Y_1, \cdots, Y_N)$
- Un modèle statistique (paramétrique) $\theta \in \Theta \subseteq \mathbb{R}^d$ la dépendance en $\mathbf{Y}$ est omise

$$\theta \mapsto L(\theta) \qquad \text{vraisemblance des observations}$$

- Eventuellement, une pénalité sur θ: $\quad \theta \mapsto g(\theta) \geq 0$

Objectif

$$\hat{\theta} \in \operatorname{argmax}_{\theta \in \Theta} \left(\frac{1}{N} \log L(\theta) - g(\theta) \right)$$

dans des modèles où L est non explicite.

Exemple 1 : Modèles à données cachées

- La vraisemblance est de la forme

$$\theta \mapsto L(\theta) = \int_{\mathsf{X}} p_\theta(x)\, \mu(\mathrm{d}x),$$

- Dans ces modèles

  - la densité *complète* $p_\theta(x)$ peut être évaluée explicitement

  - Inefficace d'approcher L par Monte Carlo; plus efficace de "restaurer" les donnés manquantes x par la loi *a posteriori*

  - Le gradient de L est de la forme

$$\nabla \log L(\theta) = \int_{\mathsf{X}} \partial_\theta \log p_\theta(x) \underbrace{\frac{p_\theta(x)\, \mu(\mathrm{d}x)}{\int p_\theta(z)\, \mu(\mathrm{d}z)}}_{\text{loi a posteriori}}$$

  i.e. une espérance, non explicite, par rapport à une loi connue à constante de normalisation près.

Exemple 2 : Champs de Gibbs

- N observations d'un graphe p noeuds, chaque noeud à valeur dans X fini : $Y_i \in \mathsf{X}^p$.
- Modèle statistique : i.i.d. de loi

$$y \mapsto \pi_\theta(y) \stackrel{\text{def}}{=} \frac{1}{Z_\theta} \exp\left(\langle \theta, B(y) \rangle\right)$$

- Dans ces modèles
  - la constante de normalisation Z_θ est incalculable.
  - la log-vraisemblance s'écrit

$$\frac{1}{N} \log L(\theta) = \left\langle \theta, \frac{1}{N} \sum_{i=1}^{N} B(Y_i) \right\rangle - \log Z_\theta$$

  - le gradient est de la forme

$$\nabla_\theta \left(\frac{1}{N} \log L(\theta) \right) = \frac{1}{N} \sum_{i=1}^{N} B(Y_i) - \int_{\mathsf{X}^p} B(y)\, \pi_\theta(y)\, \mu(\mathrm{d}y)$$

  i.e. une espérance, incalculable, par rapport à une loi connue à constante de normalisation près.

Résolution du problème d'estimation

$$\hat{\theta} \in \operatorname{argmax}_{\theta\in\Theta} \left(\ln L(\theta) - g(\theta)\right)$$

- $\ln L$ régulière, de gradient

$$\nabla(\ln L)(\theta) = \int H_\theta(x)\ \pi_\theta(x)\mathrm{d}\lambda(x)$$

- Approximation biaisée du gradient

$$\frac{1}{m}\sum_{k=1}^{m} H_\theta(X_{k,\theta}) \qquad \{X_{k,\theta}, k \geq 0\}\text{chaîne de Markov, cible } \pi_\theta \mathrm{d}\lambda$$

Le problème : si

* procédure d'optimisation itérative exploitant l'information de gradient
* dans laquelle, chaque gradient est approché par une somme de MC

à quelles conditions sur l'approximation, cet algorithme hérite-t-il des mêmes points limites que l'algorithme exact ?

Contexte

une famille de noyaux de transition markoviens : $\{P_\theta, \theta \in \Theta\}$ $\quad \Theta \subseteq \mathbb{R}^d$

chacun possédant sa propre probabilité invariante: π_θ

L'échantillonneur :

$$X_{k+1} | \text{passé}_k \sim P_{\theta_k}(X_k, \cdot)$$

où θ_k est aléatoire, déterminé par le passé de l'algorithme.

L'adaptation peut détruire la convergence

Même si tous les noyaux ont la même loi invariante π, l'échantillonneur peut ne pas converger vers π

- Soit $t_0 \neq t_1 \in]0,1[$ et les noyaux P_0, P_1

$$P_\ell = \begin{pmatrix} 1-t_\ell & t_\ell \\ t_\ell & 1-t_\ell \end{pmatrix}.$$

  Les deux noyaux ont même loi invariante : $\pi P_\ell = \pi$ avec $\pi = (1/2, 1/2)$.

- Algorithme adapté :

$$X_{k+1} \sim \begin{cases} P_0(X_k, \cdot) & \text{si } X_k = 0 \\ P_1(X_k, \cdot) & \text{si } X_k = 1 \end{cases}$$

- $\{X_k, k \geq 0\}$ est (ici) une chaîne de Markov; sa matrice de transition :

$$\begin{pmatrix} 1-t_0 & t_0 \\ t_1 & 1-t_1 \end{pmatrix} \qquad \text{de loi invariante } \tilde{\pi} \propto (t_1, t_0)$$

Dans la littérature

Contrôle explicite de convergence $\|P_\theta^n(x,\cdot) - \pi_\theta\|_V \leq C_\theta \ \mathsf{rate}_\theta(n) \ V(x)$

- Vitesses géométriques : Baxendale (1993), Meyn and Tweedie (1994), Rosenthal (1995), Lund and Tweedie (1996), Mengersen and Tweedie (1996); Roberts and Tweedie (1996)
- Vitesses sous-géométriques F.and Moulines (2003); Douc, Moulines and Soulier (2007); Andrieu, F., Vihola (2015)

Dans la littérature

Contrôle explicite de convergence $\|P_\theta^n(x,\cdot) - \pi_\theta\|_V \leq C_\theta \ \text{rate}_\theta(n) \ V(x)$

- Vitesses géométriques : Baxendale (1993), Meyn and Tweedie (1994), Rosenthal (1995), Lund and Tweedie (1996), Mengersen and Tweedie (1996); Roberts and Tweedie (1996)
- Vitesses sous-géométriques F.and Moulines (2003); Douc, Moulines and Soulier (2007); Andrieu, F., Vihola (2015)

Ergodicité Convergence de la loi de X_n

- Sous hyp d'ergodicité uniforme, vitesse géométrique Roberts and Rosenthal (2007)
- Sous hyp plus générales, vitesse géométrique et sous-géométrique Atchadé, F., Moulines and Priouret (2011): F., Moulines and Priouret (2012); Atchadé and F. (2009)

pour un algorithme particulier : Andrieu and Moulines (2006).

Dans la littérature

Contrôle explicite de convergence $\|P_\theta^n(x,\cdot) - \pi_\theta\|_V \leq C_\theta \ \text{rate}_\theta(n) \ V(x)$

- Vitesses géométriques : Baxendale (1993), Meyn and Tweedie (1994), Rosenthal (1995), Lund and Tweedie (1996), Mengersen and Tweedie (1996); Roberts and Tweedie (1996)
- Vitesses sous-géométriques F.and Moulines (2003); Douc, Moulines and Soulier (2007); Andrieu, F., Vihola (2015)

Ergodicité Convergence de la loi de X_n

- Sous hyp d'ergodicité uniforme, vitesse géométrique Roberts and Rosenthal (2007)
- Sous hyp plus générales, vitesse géométrique et sous-géométrique Atchadé, F., Moulines and Priouret (2011): F., Moulines and Priouret (2012); Atchadé and F. (2009)

pour un algorithme particulier : Andrieu and Moulines (2006).

LGN Convergence de moyenne empirique de fonctionnelles additives des tirages

- Loi faible et fonctionnelles bornées Roberts and Rosenthal (2007)
- Loi forte et fonctionnelles non bornées Andrieu and Moulines (2006); F., Moulines and Priouret (2012)
- Pour un algorithme particulier avec Θ compact et fonctions bornées Haario et al. (2001); Atchadé and Rosenthal (2005) - avec Θ non borné et fonctionnelles non bornées Andrieu and Moulines (2006), Saksman and Vihola (2010)

Dans la littérature

Contrôle explicite de convergence $\|P_\theta^n(x,\cdot) - \pi_\theta\|_V \leq C_\theta \ \text{rate}_\theta(n) \ V(x)$

- Vitesses géométriques : Baxendale (1993), Meyn and Tweedie (1994), Rosenthal (1995), Lund and Tweedie (1996), Mengersen and Tweedie (1996); Roberts and Tweedie (1996)
- Vitesses sous-géométriques F.and Moulines (2003); Douc, Moulines and Soulier (2007); Andrieu, F., Vihola (2015)

Ergodicité Convergence de la loi de X_n

- Sous hyp d'ergodicité uniforme, vitesse géométrique Roberts and Rosenthal (2007)
- Sous hyp plus générales, vitesse géométrique et sous-géométrique Atchadé, F., Moulines and Priouret (2011): F., Moulines and Priouret (2012); Atchadé and F. (2009)

pour un algorithme particulier : Andrieu and Moulines (2006).

LGN Convergence de moyenne empirique de fonctionnelles additives des tirages

- Loi faible et fonctionnelles bornées Roberts and Rosenthal (2007)
- Loi forte et fonctionnelles non bornées Andrieu and Moulines (2006); F., Moulines and Priouret (2012)
- Pour un algorithme particulier avec Θ compact et fonctions bornées Haario et al. (2001); Atchadé and Rosenthal (2005) - avec Θ non borné et fonctionnelles non bornées Andrieu and Moulines (2006), Saksman and Vihola (2010)

TCL

- Cas général F., Moulines, Priouret and Vandekerkhove (2014)
- Algorithme particulier Andrieu and Moulines (2006)

CS pour la convergence en loi : l'idée

- Pour simplifier : dans le cas où $\pi_\theta = \pi$ (tous les noyaux ont même loi invariante)
- Une inégalité triangulaire pour décomposer

$$\left| \mathbb{E}\left[f(X_n)\right] - \int f \, \pi \mathrm{d}\lambda \right|$$

Mécanisme de mise à jour

$$X_{n-r_n} \overset{P_{\theta_{n-r_n}}}{\longrightarrow} X_{n-r_n+1} \overset{P_{\theta_{n-r_n+1}}}{\longrightarrow} X_{n-r_n+2} \cdots \overset{P_{\theta_{n-1}}}{\longrightarrow} X_n$$

Comparé au schéma "adaptation gelée"

$$X_{n-r_n} \overset{P_{\theta_{n-r_n}}}{\longrightarrow} X_{n-r_n+1} \overset{P_{\theta_{n-r_n}}}{\longrightarrow} X_{n-r_n+2} \cdots \overset{P_{\theta_{n-r_n}}}{\longrightarrow} X_n$$

Comparé au schéma

$$X_n \sim \pi \, \mathrm{d}\lambda$$

CS pour la convergence en loi : "containment, diminishing adaption" (1/3)

Diminishing adaption condition

- Régularité des noyaux en θ : $\mathrm{dist}(P_\theta, P_{\theta'}) \leq \mathrm{dist}(\theta, \theta')$
- Mécanisme de mise à jour du paramètre tq $\lim_k(\theta_{k+1} - \theta_k) = 0$.

Containment condition

- Ergodicité de tous les noyaux, de façon analogue :

$$\|P_\theta^n(x, \cdot) - \pi\|_{\mathrm{VT}} \leq C_\theta \, V(x) \, \mathsf{rate}_\theta(n)$$

Conditions simples Roberts and Rosenthal, 2007

(dim) $\lim_n \sup_x \|P_{\theta_{n+1}}(x, \cdot) - P_{\theta_n}(x, \cdot)\|_{\mathrm{VT}} = 0$ en probabilité

(cont) $\forall \epsilon > 0, \{M_\epsilon(X_n, \theta_n), n \geq 0\}$ bornée en probabilité où

$$M_\epsilon(x, \theta) \stackrel{\mathrm{def}}{=} \inf\{n : \|P_\theta^n(x, \cdot) - \pi\|_{\mathrm{VT}} \leq \epsilon\}$$

Adaptées au cas

- chaîne sur espace fini/compact;
- ergodicité uniforme en x et en θ ($\rightarrow \Theta$ compact)

CS pour la convergence en loi : "containment, diminishing adaption" (2/3)

Conditions plus générales F., Moulines and Priouret (2012)

$\forall \epsilon > 0$, $\exists \{r(n), n \geq 0\}$ t.q. $\lim_n r(n)/n = 0$

(dim) $\lim_n \mathbb{E}\left[\| P^{r(n)}_{\theta_{n-r(n)}}(X_{n-r(n)}, \cdot) - \pi_{\theta_{n-r(n)}} \|_{\mathrm{VT}}\right] = 0$

(cont) $\sum_{j=0}^{r(n)} \mathbb{E}\left[\sup_x \frac{\| P_{\theta_{n-r(n)+j}}(x,\cdot) - P_{\theta_{n-r(n)}}(x,\cdot) \|_V}{V(x)}\right] \to 0$

Adaptées au cas

- chaîne à espace d'état quelconque
- ergodicité non simultanée, en x et θ
- $\{\theta_n, n \geq 0\}$ bornée

OU $\{\theta_n, n \geq 0\}$ à croissance contrôlée $\sup n^{-\tau} \|\theta_n\| < \infty$ p.s. - la vitesse à laquelle le paramètre est adapté, est liée à l'ergodicité des noyaux de transition. Vihola and Saksman (2010); F., Moulines and Priouret (2012)

CS pour la convergence en loi : "containment, diminishing adaption" (3/3)

En pratique, pour des échantillonneurs MCMC

- Conditions d'ergodicité géométrique, avec régularité en θ de la constante C_θ et du taux $\rho_\theta \in]0,1[$

$$\|P_\theta^n(x,\cdot) - \pi\|_V \leq C_\theta\, \rho_\theta^n\, V(x)$$

- Conditions de régularité en θ du noyau

$$\|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_V \leq C\, \|\theta - \theta'\|^a\, V(x) \qquad a \in]0,1]$$

- $\{\theta_n, n \geq 0\}$ bornée OU à croissance contrôlée

CS pour la convergence en loi : lorsque chaque noyau a sa propre loi invariante

$$\begin{aligned}\mathbb{E}\left[f(X_n)\right] &= \mathbb{E}\left[f(X_n) - P_{\theta_{n-r_n}}^{r_n} f(X_{n-r_n})\right] && \text{diminishing adaption}\\ &+ \mathbb{E}\left[P_{\theta_{n-r_n}}^{r_n} f(X_{n-r_n}) - \pi_{\theta_{n-r_n}}(f)\right] && \text{containment}\\ &+ \mathbb{E}\left[\pi_{\theta_{n-r_n}}(f)\right]\end{aligned}$$

Régularité en θ de π_θ

- Cas assez simple : on a l'expression de la loi invariante π_θ et on peut vérifier $\lim_n \pi_{\theta_n}(f)$ existe. Bardenet et al. (2015), F. et al. (2015,2016)
- Beaucoup plus technique : la régularité en θ de π_θ ne peut que se déduire de celle du noyau P_θ. F., Moulines and Priouret (2012)

Pour la résolution de

$$\text{argmin}_{\theta\in\Theta}\left(f(\theta)+g(\theta)\right)$$

- g convexe, non régulière, positive.
- f régulière, de gradient L-Lipschitz

- lorsque ∇f est non explicite et de la forme

$$\nabla f(\theta)=\int H_\theta(x)\,\pi_\theta(x)\mathrm{d}\lambda(x).$$

- par une approche **Gradient-Proximal** Combettes and Pesquet (2011)

$$\begin{aligned}\theta_{n+1} &= \text{argmin}_{\theta\in\Theta}\left(g(\theta)+\frac{1}{2\gamma_{n+1}}\|\theta-\theta_n+\gamma_{n+1}\nabla f(\tau_n)\|^2\right)\\ &= \text{Prox}_{\gamma_{n+1},g}\left(\theta_n-\gamma_{n+1}\nabla f(\theta_n)\right)\end{aligned}$$

les pas $\{\gamma_n, n\geq 0\}$ sont choisis par l'utilisateur.

Algorithme exact : itératif,

$$\tau_{n+1} = \text{argmin}_{\theta\in\Theta}\left(g(\theta) + \frac{1}{2\gamma_{n+1}}\|\theta - \tau_n + \gamma_{n+1}\nabla f(\tau_n)\|^2\right)$$

Algorithme perturbé : itératif,

- On supposera l'opérateur proximal explicite. Le gradient est approché

$$\theta_{n+1} = \text{argmin}_{\theta\in\Theta}\left(g(\theta) + \frac{1}{2\gamma_{n+1}}\|\theta - \theta_n + \gamma_{n+1}\widehat{\nabla f(\theta_n)}\|^2\right)$$

- Et plus précisément

$$\nabla f(\theta_n) = \int H_\theta(x)\,\pi_\theta(x)\mathrm{d}\lambda(x) \qquad \widehat{\nabla f(\theta_n)} = \frac{1}{m_{n+1}}\sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n})$$

- $\{X_{j,n}, j \geq 1\}$ chaîne de Markov non stationnaire de loi inv. $\pi_{\theta_n}\mathrm{d}\lambda$
- approximation biaisée, avec un biais "persistant" lorsque m_n constant

$$\mathbb{E}\left[\frac{1}{m_{n+1}}\sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}) \,\middle|\, \text{passé}_n\right] \neq \nabla f(\theta_n)$$

CS pour la convergence, cas convexe Atchadé, F. and Moulines (2017)

Résultats dans le cadre "convexe"

- g **convexe**, semi-cont inférieurement,
- f **convexe**, gradient L-lipschitz
- $\Theta \subset \mathbb{R}^d$ est le domaine de g; et $\operatorname{argmin}_\Theta(f+g)$ non vide.

pour prouver le comportement asymptotique : il existe $\theta_\infty \in \operatorname{argmin}_\Theta(f+g)$ t.q.

$$\lim_n \theta_n = \theta_\infty$$

- Conditions sur le pas $\sum_n \gamma_n = +\infty, \quad \gamma_n \in (0, 1/L]$
- Conditions sur l'approximation du gradient $\eta_{n+1} \stackrel{\text{def}}{=} \widehat{\nabla f(\theta_n)} - \nabla f(\theta_n)$

$$\sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^2, \qquad \sum_n \gamma_{n+1} \eta_{n+1}, \qquad \sum_n \gamma_{n+1} \langle \mathbf{A}_n(\theta_n), \eta_{n+1} \rangle$$

où $\mathbf{A}_n$ est l'opérateur gradient-proximal.

Approximation MC avec m_n croissant Atchadé, F. and Moulines (2017)

- Conditions sur le pas et le nombre de tirages MC par itération

$$\sum_n \gamma_n = +\infty, \qquad \sum_n \frac{\gamma_n^2}{m_n} < \infty; \qquad \sum_n \frac{\gamma_n}{m_n} < \infty \text{ (biased case)}$$

Ex. prendre $\gamma_n = \gamma$ et $m_n \sim n\,(\log n)^{1+\iota}$

- Conditions sur les chaînes de Markov
  - Il existe $p \geq 2$, $C > 0$, $\rho \in]0,1[$, et une fonction $W : \mathsf{X} \to [1, \infty[$ t.q.

$$\sup_{\theta \in \Theta} |H_\theta|_W < \infty, \qquad \sup_{\theta \in \Theta} \|P_\theta^n(x, \cdot) - \pi_\theta\|_{W^p} \leq C \rho^n W^p(x).$$

  - Θ est borné

Approximation MC avec m_n constant Atchadé, F. and Moulines (2017)

Régime "approximation stochastique", plus technique $\cdots$.

- Conditions sur le pas

$$\sum_n \gamma_n = +\infty \qquad \sum_n \gamma_n^2 < \infty \qquad \sum_n |\gamma_{n+1} - \gamma_n| < \infty$$

Ex. prendre $\gamma_n = O(1/n^\alpha)$ avec $\alpha \in]1/2, 1]$

- Conditions sur les chaînes de Markov
  - (les mêmes que dans le cas "m_n croissant")
  - Il existe C t.q. pour tout $\theta, \theta' \in \Theta$

$$|H_\theta - H_{\theta'}|_W + \sup_x \frac{\|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_W}{W(x)} + \|\pi_\theta - \pi_{\theta'}\|_W \leq C \|\theta - \theta'\|.$$

  - Θ est borné

Dans la littérature

- Résultats déterministes, couvrant tout type de perturbation du gradient
- Contrôles explicites de l'erreur $(f+g)(\theta_n) - \min(f+g)$
  Atchadé, F., Moulines (2017) combiné à de l'*averaging* : vitesse en $1/n$ pour des batchs croissants ($m_n \sim n(\log n)^{1+\iota}$), vitesse en $1/\sqrt{n}$ pour batch constant.

Dans la littérature

- Résultats déterministes, couvrant tout type de perturbation du gradient
- Contrôles explicites de l'erreur $(f+g)(\theta_n) - \min(f+g)$
  Atchadé, F., Moulines (2017) combiné à de l'*averaging* : vitesse en $1/n$ pour des batchs croissants ($m_n \sim n(\log n)^{1+\iota}$), vitesse en $1/\sqrt{n}$ pour batch constant.
- Appliqués au contexte "perturbation aléatoire" de type Monte Carlo, ces conditions ne supposent pas que
  - l'approximation du gradient non biaisée (i.e. Monte Carlo i.i.d.)
  - $\inf_n \gamma_n > 0$
  - le nombre de tirages Monte Carlo augmente avec les itérations

Combettes (2001); Combettes and Wajs (2005); Schmidt et al. (2011); Combettes and Pesquet (2015, 2016); Lin et al. (2015); Rosasco et al. (2014,2015)

Dans la littérature

- Résultats déterministes, couvrant tout type de perturbation du gradient
- Contrôles explicites de l'erreur $(f+g)(\theta_n) - \min(f+g)$
  Atchadé, F., Moulines (2017) combiné à de l'*averaging* : vitesse en $1/n$ pour des batchs croissants ($m_n \sim n(\log n)^{1+\iota}$), vitesse en $1/\sqrt{n}$ pour batch constant.
- Appliqués au contexte "perturbation aléatoire" de type Monte Carlo, ces conditions ne supposent pas que
  - l'approximation du gradient non biaisée (i.e. Monte Carlo i.i.d.)
  - $\inf_n \gamma_n > 0$
  - le nombre de tirages Monte Carlo augmente avec les itérations

  Combettes (2001); Combettes and Wajs (2005); Schmidt et al. (2011); Combettes and Pesquet (2015, 2016); Lin et al. (2015); Rosasco et al. (2014,2015)
- Résultats pour l'algorithme FISTA Aujol and Dossal (2016), Aujol, Dossal, F. and Moulines (2017)
- Calcul récursif de l'approximation du gradient F., Ollier, Samson (2017)

Dans la littérature

- Résultats déterministes, couvrant tout type de perturbation du gradient
- Contrôles explicites de l'erreur $(f+g)(\theta_n) - \min(f+g)$
  Atchadé, F., Moulines (2017) combiné à de l'*averaging* : vitesse en $1/n$ pour des batchs croissants ($m_n \sim n(\log n)^{1+\iota}$), vitesse en $1/\sqrt{n}$ pour batch constant.
- Appliqués au contexte "perturbation aléatoire" de type Monte Carlo, ces conditions ne supposent pas que
  - l'approximation du gradient non biaisée (i.e. Monte Carlo i.i.d.)
  - $\inf_n \gamma_n > 0$
  - le nombre de tirages Monte Carlo augmente avec les itérations

  Combettes (2001); Combettes and Wajs (2005); Schmidt et al. (2011); Combettes and Pesquet (2015, 2016); Lin et al. (2015); Rosasco et al. (2014,2015)
- Résultats pour l'algorithme FISTA Aujol and Dossal (2016), Aujol, Dossal, F. and Moulines (2017)
- Calcul récursif de l'approximation du gradient F., Ollier, Samson (2017)
- Analyse de convergence en non convexe, approximation sans biais Ghadimi et al. (2016)
- Techniques de preuve : algos itératifs perturbés possédant une fonction de Lyapunov (algos MM perturbés) Zangwill (1969), Meyer (1976), Robbins-Siegmund lemma, $\cdots$