# Échantillonnage de lois non-régulières en grande dimension par méthodes de Monte Carlo par Chaîne de Markov proximale

Alain Durmus[1], Éric Moulines[2], Marcelo Pereyra[3]

[2]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France

[1]LTCI, CNRS and Télécom ParisTech, 46 rue Barrault 75634 Paris Cedex 13, France

[3]Maxwell Institute for Mathematical Sciences & School of Mathematical and Computer Sciences
Heriot-Watt University, EH14 4AS, UK

alain.durmus@telecom-paristech.fr, eric.moulines@polytechnique.edu,
m.pereyra@hw.ac.uk

**Résumé** – L'échantillonnage de loi en grandes dimensions est devenu un enjeu fondamental pour pouvoir pleinement appliquer les outils de l'inférence statistique bayésienne. Dans de nombreux modèles la distribution à posteriori est log-concave, la log-vraissemblance étant même de plus gradient Lipschitz. Cependant, les distributions à priori induisant de la sparcité sont en général non-régulières, comme par exemple dans les cas des pénalités classiques associées au problème du LASSO ou de l'élastic net. Dans ce travail, nous introduisons une nouvelle méthode permettant l'échantillonnage de telles distributions. Cette dernière adapte les méthodes classiques d'échantillonnage grâce aux outils de l'optimisation proximale. Finalement nous appliquons cette nouvelle méthodologies à un problème de déconvolution et de reconstruction tomographique en traitement d'image.

**Abstract** – Sampling over high-dimensional space has become a prerequisite in the applications of Bayesian statistics to signal processing problems. In many situations of interest, the log-posterior distribution is concave. The likelihood part is generally smooth and gradient Lipshitz while the prior is concave but typically not smooth (the archetypical problem is the LASSO or the elastic-net penalty, but many other problems can be cast into this framework). In this paper, a new algorithm to sample from possibly non-smooth log-concave probability measures will be introduced. This algorithm uses Moreau-Yosida envelope combined with the Euler-Maruyama discretization of Langevin diffusions. Finally, this procedure is applied to a deconvolution problem and a tomographic reconstruction in image processing, which shows that they can be practically used in a high dimensional setting.

## 1 Introduction

La méthodologie bayésienne se caractérise par une modélisation probabiliste pour prendre en compte et quantifier l'incertitude sur les paramètres à inférer lors d'une analyse statistique. Plus précisément, à partir d'observations $y \in \mathbb{R}^p$, considérons un modèle probabiliste dominé par la mesure de Lebesgue et caractérisé à partir de sa vraisemblance $y \mapsto p(y|x)$ paramétrée par $x \in \mathbb{R}^d$. Le paramètre $x$ est alors supposé lui-même aléatoire et associé à une loi à priori de densité $p_\star$ par rapport à la mesure de Lebesgue. Cette approche permet alors de définir la loi jointe de $(y, x)$ et la loi conditionnelle de $x$ sachant $y$, appelée loi a posteriori et notée dans la suite $\pi$. Par application du théorème de Bayes, $\pi$ a pour densité la fonction définie pour tout $x \in \mathbb{R}^d$ par

$$\pi(x) = p(y|x)p_\star(x)/\mathcal{Z} \;,$$
$$\text{où } \mathcal{Z} = \int_{\mathbb{R}^d} p(y|x)p_\star(x)\mathrm{d}x \;.$$

Cependant dans la plupart des modèles d'intérêt, la constante de normalisation $\mathcal{Z}$ n'est pas calculable. Les algorithmes de Monte Carlo par chaîne de Markov (MCMC) sont alors devenus des outils fondamentaux pour échantillonner de telles lois. Mais avec l'augementation ces dernières années des capacités computationnelles et de la complexité des modèles, la dimension des paramètres à inférer et donc des lois à échantillonner, est devenu de plus en plus importante, ce qui impacte fortement la convergence des méthodes MCMC classiques. Nous nous concentrerons dans cet exposé à l'étude de la méthode MCMC basée sur la discrétisation de l'équation de Langevin associée à la distribution à posteriori.

Supposons que la densité de $\pi$ est positive sur $\mathbb{R}^d$ et est donc sous la forme $x \mapsto \mathrm{e}^{-U(x)}/\int_{\mathbb{R}^d} \mathrm{e}^{-U(y)}\mathrm{d}y$ avec un potentiel $U : \mathbb{R}^d \to \mathbb{R}$ continûment différentielle. L'équation de Langevin associée à $\pi$ est l'équation différentielle stochastique définie par :

$$\mathrm{d}\mathbf{X}_t = -\nabla U(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^d \;, \tag{1}$$

où $(B_t^d)_{t\geq 0}$ est un mouvement standard brownien de dimension $d$. Sous des hypothèses faibles sur $U$, cette équation possède une unique solution forte, et le semi groupe associé est réversible par rapport à $\pi$ et est ergodique. En outre lorsque $U$ est convexe, ce qui est le cas dans un certain nombre de modèles en statistique, des taux exponentiels en total variation et en distance de Wasserstein ont été établis. Le comportement en temps long du semi groupe associé à l'équation de Langevin est à la base de l'algorithme de Langevin non-ajusté (ULA). Cet algorithme est la méthode MCMC qui utilise la discrétisation de Euler-Maruyama de (1), pour échantillonner $\pi$. Cette discrétisation définie une chaîne de Markov $(X_k)_{k\geq 0}$ par :

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1} \tag{2}$$

où $(Z_k)_{k\geq 1}$ est une suite i.i.d. de variables aléatoires standard Gaussienne de dimension $d$ et $\gamma$ est la pas de discrétisation. Cette méthode a introduite en statistique computationnelle par [5]. Elle a déjà retenu l'attention de nombreux travaux, cf. [9]. En particulier, [7] montrent que sous des conditions appropriés sur $U$, la chaîne définie par (2) est $V$-géométriquement ergodique et converge vers une loi invariante $\pi_\gamma$ différente de $\pi$. Récemment sous l'hypothèse que le potentiel $U$ est fortement convexe ou convexe, [2, 3] ont obtenu des bornes non-asymptotiques en total variation entre la loi de la chaîne définie par (2) et $\pi$. Ces bornes sont accompagnées d'une étude précise de la convergence en fonction de la dimension.

Malheureusement, ULA n'est cependant pas bien défini lorsque la densité cible (ou le potentiel $U$) n'est pas lisse ce qui limite son application à certains problèmes d'inférence bayésienne. De plus, même lorsque le potentiel $U$ est sous-différentiable et Lipschitz, on peut observer empiriquement que ULA rencontre certaines difficultés de convergence. Dans cette contribution, nous proposons une nouvelle méthodologies MCMC pour échantillonner une loi $\pi$ log-concave mais non-régulière.

**Notations and conventions**. Soit $h : \mathbb{R}^d \to \mathbb{R}^m$. $h$ est dite $L$-Lipschitz pour $L \geq 0$, si pour tout $x, y \in \mathbb{R}^d$, $\|h(x) - h(y)\| \leq L \|x - y\|$. Pour deux densités de probabilité sur $\mathbb{R}^d$, $\mu, \nu$, nous considérons la distance en variation totale entre $\mu$ et $\nu$ définie par

$$\|\mu - \nu\|_{\mathrm{TV}} = \int_{\mathbb{R}^d} |\mu(x) - \nu(x)| \, \mathrm{d}x \; .$$

# 2 Échantillonnage de lois non-régulières par l'algorithme MYULA

## 2.1 Méthode proposée

Nous supposons que le potentiel $U$ vérifie la condition suivante.

**H1.** *$U = f + g$, où $f : \mathbb{R}^d \to \mathbb{R}$ et $g : \mathbb{R}^d \to (-\infty, +\infty]$ sont deux fonctions minorées satisfaisant :*

1. *$f$ est convexe, continûment différentiable, et tel que $\nabla f$ est $L_f$ Lipschitz.*
2. *$g$ est propre, convexe et semi-continue inférieurement.*

L'idée centrale de notre contribution est de remplacer $\pi$ par une approximation régulière bien choisie qui, par construction, satisfait les deux propriétés fondamentales suivantes : 1) l'approximation de Euler-Maruyama associé à cette approximation est toujours stable et possède de bonne propriété de convergence, et 2) cette approximation peut-être choisie arbitrairement proche de $\pi$ en ajustant le paramètre $\lambda > 0$. Pour cela, les objets fondamentaux que nous utilisons sont l'opérateur proximal et l'enveloppe de Moreau-Yosida associés à une fonction convexe.

Soit $\mathrm{g} : \mathbb{R}^d \to (-\infty, +\infty]$ une fonction convexe semi-continue inférieurement et $\lambda > 0$. L'enveloppe de Moreau-Yosida et l'opérateur proximal de paramètre $\lambda$ associée à g sont respectivement les fonction $\mathrm{g}^\lambda : \mathbb{R}^d \to \mathbb{R}$ et $\mathrm{prox}_{\mathrm{g}}^\lambda : \mathbb{R}^d \to \mathbb{R}^d$ (voir [8]) définies pour tout $x \in \mathbb{R}^d$ par

$$\mathrm{g}^\lambda(x) = \min_{y \in \mathbb{R}^d} \left\{ \mathrm{g}(y) + (2\lambda)^{-1} \|x - y\|^2 \right\} \; ,$$

$$\mathrm{prox}_{\mathrm{g}}^\lambda(x) = \arg\min_{y \in \mathbb{R}^d} \left\{ \mathrm{g}(y) + (2\lambda)^{-1} \|x - y\|^2 \right\} \; .$$

L'enveloppe de Moreau-Yosida $\mathrm{g}^\lambda$ est une version régularisée de g. En effet, $\mathrm{g}^\lambda$ est convexe et continûment différentiable.

Nous proposons alors d'approximer le potentiel $U$ par la fonction $U^\lambda : \mathbb{R}^d \to \mathbb{R}$ définie pour tout $x \in \mathbb{R}^d$ par

$$U^\lambda(x) = f(x) + g^\lambda(x) \; ,$$

où $g^\lambda$ est l'enveloppe de Moreau-Yosida de paramètre $\lambda > 0$ associé à $g$. Observons que sous **H** 1, par les propriétés fondamentales de l'enveloppe de Moreau-Yosida précédemment énoncées, $U^\lambda$ est convexe, continûment différentiable. Aussi, la proposition 1 ci-dessus montre que pour tout $\lambda > 0$, la densité de probabilité $\pi^\lambda : \mathbb{R}^d \to \mathbb{R}$ définie pour tout $x \in \mathbb{R}^d$ par

$$\pi^\lambda(x) = \frac{\mathrm{e}^{-U^\lambda(x)}}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(s)} \mathrm{d}s} \; ,$$

est bien définie et peut-être arbitrairement proche de $\pi$ en choisissant de manière appropriée $\lambda$.

**H2.** *Nous supposons que l'une de ces deux conditions est satisfaite :*

1. *$\mathrm{e}^{-g}$ est intégrable par rapport à la mesure de Lebesgue.*
2. *$g$ est Lipschitz.*

**Proposition 1.** *Supposons que les hypothèses **H**1 et **H**2 soient satisfaites.*

1. *Pour tout $\lambda > 0$, $\pi^\lambda$ définit une densité de probabilité propre sur $\mathbb{R}^d$, i.e.*

$$0 < \int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(y)} \mathrm{d}y < +\infty \; .$$

2. *Pour tout $\lambda > 0$, $\pi^\lambda$ est log-concave et continûment différentiable avec*

$$\nabla U^\lambda(x) = -\nabla \log \pi^\lambda(x) = \nabla f(x) + \lambda^{-1}(x - \mathrm{prox}_g^\lambda(x)) \; .$$

*De plus, $\nabla U^\lambda$ est Lipschitz de constante de Lipschitz $L \leq L_f + \lambda^{-1}$.*

3. *$(\pi^\lambda)_{\lambda \in \mathbb{R}^*}$ converge vers $\pi$ lorsque $\lambda \downarrow 0$ en variation totale, i.e.*

$$\lim_{\lambda \to 0} \|\pi^\lambda - \pi\|_{\mathrm{TV}} = 0 \; .$$

4. *Si de plus, **H**-2-2 est satisfaite alors pour tout $\lambda > 0$,*

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \leq \lambda \, \|g\|_{\mathrm{Lip}}^2 \; .$$

*Démonstration.* □

La méthode MCMC proposée dans cette contribution est alors une application de ULA pour échantillonner $\pi^\lambda$. Soit un paramètre de régularisation $\lambda > 0$ et un pas de discrétisation $\gamma > 0$ donnés, nous considérons la discrétisation de Euler-Maruyama de l'equation de Langevin associée à $\pi^\lambda$, définie pour tout for all $k \geq 0$ par

$$\begin{aligned} \text{MYULA} : \; X_{k+1}^{\mathrm{M}} =& (1 - \tfrac{\gamma}{\lambda}) X_k^{\mathrm{M}} - \gamma \nabla f(X_k^{\mathrm{M}}) \\ &+ \tfrac{\gamma}{\lambda} \, \mathrm{prox}_g^\lambda(X_k^{\mathrm{M}}) + \sqrt{2\gamma} Z_{k+1} \; . \end{aligned} \tag{3}$$

où $(Z_k)_{k \geq 1}$ est une suite i.i.d. de variables aléatoires standard Gaussienne de dimension $d$. Nous appellerons cette méthode l'algorithme de Langevin unajusté avec regularisation de Moreau-Yosida. Puisque $\nabla U^\lambda$ est gradient Lipschitz, la diffusion de Langevin associée à $U^\lambda$ converge vers $\pi^\lambda$ lorsque $t \to \infty$, et la chaîne de Markov associée à MYULA est par construction stable et peut être utiliser pour approcher $\pi^\lambda$. De plus, le paramètre $\lambda > 0$ contrôle l'erreur d'approximation induite par $\pi^\lambda$ comme substitut à $\pi$. Cette erreur peut être rendue arbitrairement petite et est bornée explicitement lorsque $g$ est Lipschitz (cette erreur d'approximation peut être aussi corrigée en utilisant un schéma d'échantillonnage d'importance, cependant nous ne considérons pas cette solution dans ce papier).

Finalement, la distribution stationnaire de (3) dépend des valeurs choisies pour $\lambda$ and $\gamma$. Nous montrons dans ... que pour un choix approprié de $\lambda$ et $\gamma$ cette distribution peut-être rendue arbitrairement proche de $\pi$. D'un point de vue plus pratique, afin de produire une chaîne possédant de bonnes propriétés de mélange, nous recommandons de choisir $\gamma \in [\lambda/5(L_f\lambda + 1), \lambda/2(L_f\lambda + 1)]$ et $\lambda$ de l'ordre de $L_f^{-1}$ (voir [4] pour une analyse détaillée).

# 3 Expériences numériques

Nous illustrons dans cette partie l'application de MYULA à l'inférence bayésienne de deux modèles rencontrés en traitement de l'image : un modèle de déconvolution et un modèle de reconstruction tomographique. Pour comparaison, nous prenons comme référence l'algorithme proximal de Langevin Metropolis ajusté (Px-MALA) [6] visant la densité a posteriori $\pi$ de façon exacte.

Dans notre premier exemple, nous considérons un modèle de déconvolution

$$\pi(x) \propto \exp\left[-(\|y - Hx\|^2/2\sigma^2) - \beta TV(x)\right] \tag{4}$$

où $TV : \mathbb{R}^d \to \mathbb{R}$ est la pseudo-norme de variation totale, $H$ est un filtre passe-bas de dimension $5 \times 5$, $\sigma = 0.47$ relatif à un rapport signal sur bruit de 40dB, $\beta = 0.03$, et où $y = Hx + w$ est la version floutée et bruitée de l'image de référence `Boat` de taille $d = 256 \times 256$ pixels. Nous avons pour ce modèle implémenté MYULA en choisissant pour $f$, la fonction $x \mapsto \|y - Hx\|^2/2\sigma^2$, et pour $g$, $x \mapsto \beta TV(x)$. De plus, nous avons choisi pour paramètres de $\lambda = L_f^{-1} = 0.45$ et $\gamma = L_f^{-1}/5 = 0.1$. Finalement, nous utilisons l'algorithme de Chambolle [1] pour évaluer l'opérateur proximal associé à la pseudo-norme $TV$. L'échantillonage d'une chaîne de longueur $10^5$ requière en moyenne un temps de calcul de l'ordre de 30 minutes.

Pour illustrer les performances de MYULA, la Figure 1(a) montre les régimes transients de deux chaînes associées respectivement à MYULA et Px-MALA à partir de l'évolution du potentiel a posteriori $U$, le long de leur trajectoire. Ces deux chaînes sont initialisées à la même position et pour une meilleurs visibilité, une échelle logarithmique est utilisée. Nous pouvons observer que MYULA ne requière que $10^2$ itérations pour atteindre une région "stable" pour le potentiel $U$, tandis que Px-MALA requière quant à lui $10^4$ itérations. La Figure 1(b) représente les fonctions d'autocorrélation de deux composantes des deux chaînes échantillonnées par MYULA et Px-MALA. Pour souligner l'efficacité de MYULA, nous considérons sur ce graphique la composante de la chaîne associée qui possède la plus grande variance empirique. Encore une fois, nous pouvons observer que MYULA converge bien plus rapidement que Px-MALA. D'un point de vue pratique, cette efficacité est accentuée également par le fait que une itération de MYULA possède une coût computationnel deux fois moins important que Px-MALA, car ce dernier inclus une étape d'acceptation/rejet. Finalement, une comparaison de statistiques (non reportés dans ce papier) indique que les estimations de MYULA ont une erreur d'approximation de l'ordre de 0.5% par rapport à Px-MALA (voir [4] pour les détails de cette analyse).

Dans notre seconde illustration, nous considérons un modèle de reconstruction tomographique :

$$\pi(x) \propto \exp\left[-\|y - AFx\|^2/2\sigma^2 - \beta TV(x)\right]. \tag{5}$$

de dimension $d = 128 \times 128$ pixels, où $F$ est la transformée de Fourier discrète, $A$ est un opérateur d'échentillonnage $\sigma = 7 \times 10^{-2}$, $\beta = 5$, et où $y = AFx + w$ est une mesure tomographique bruitée de l'image test `Shepp-Logan phantom`.

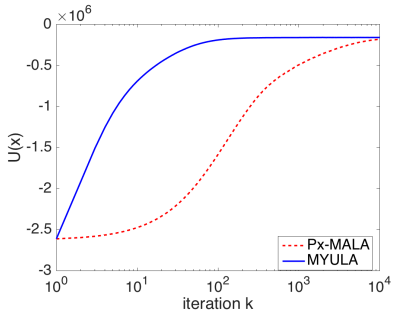
Encore une fois, MYULA est implémenté en choisissant pour $f$, $x \mapsto \|y - AFx\|^2/2\sigma^2$ et pour $g$ $x \mapsto \beta TV(x)$, avec

pour paramètres $\lambda = L_f^{-1} = 1 \times 10^{-4}$ et $\gamma_k = L_f^{-1}/10 = 10^{-5}$, et en utilisant [1] pour évaluer l'opérateur proximal de la norme $TV$. Le calcul de $10^5$ échantillons prend pour cette exemple de l'ordre 15 minutes.

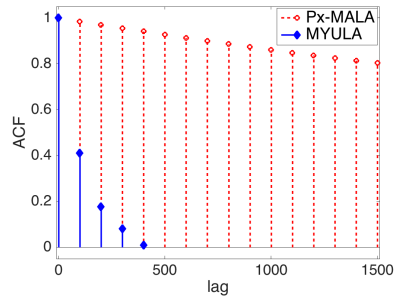
Pour analyser l'exactitude de MYULA, nous comparons en Figure 2(a) les seuils des régions de plus forte densité a posteriori (HPD), $\eta_\alpha$, estimée par MYULA et Px-MALA, définie pour tout $\alpha \in [0,1]$ par

$$\pi\{U \leq \eta_\alpha\} = 1 - \alpha \ .$$

Cette comparaison indique une erreur d'approximation de l'ordre de 3%. Concernant les performances computationnel, une analyse du temps d'autocorrélation intégrée (non reportée ici) révèle que ce modèle MYULA est approximativement deux fois plus efficace que Px-MALA. Pour illustration, la Figure 2(b) représente les fonctions d'autocorrelation de MYULA et Px-MALA pour la composante la plus lente de MYULA (voir également [4]).

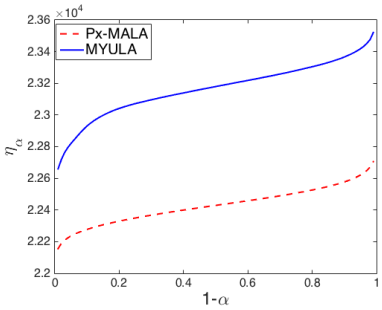



(a)

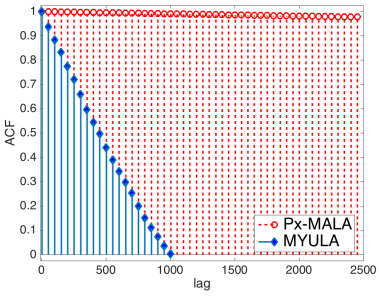



(b)

FIGURE 1 – Comparaison de MYULA et Px-MALA : (a) Convergence des deux chaînes vers une région typique de (4) (échelle logarithmique), (b) Fonction d'autocorrélation de deux chaînes (ACF).




(a)




(b)

FIGURE 2 – Reconstruction tomographique : (a) seuils des régions HDP $\eta_\alpha$, (b) fonctions d'autocorrélation.

# Références


[1] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1) :89–97, 2004.

[2] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, pages n/a–n/a, 2016.

[3] A. Durmus and É Moulines. Non-asymptotic convergence analysis for the unadjusted langevin algorithm. Accepted for publication in Ann. Appl. Probab. 1507.05021, arXiv, July 2015.

[4] A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo : when Langevin meets Moreau. *ArXiv e-prints*, December 2016.

[5] U. Grenander. Tutorial in pattern theory. Division of Applied Mathematics, Brown University, Providence, 1983.

[6] M. Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 2015. open access paper, http ://dx.doi.org/10.1007/s11222-015-9567-4.

[7] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4) :341–363, 1996.

[8] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[9] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4) :483–509 (1991), 1990.