Passty Langevin

Pascal Bianchi[1], Adil Salim[2], et Sholom Schechtman[3]

[1]LTCI, Télécom ParisTech, France
[2]VCC, KAUST, Saudi Arabia
[3]LIGM, Université Paris-Est Marne-la-Vallée, France

4 juin 2019

Résumé

Dans cet article nous proposons une nouvelle méthode pour simuler selon une densité log-concave lorsque la fonction convexe sous-jacente est une composée d'une fonction convexe et un opérateur affine. Ce type de problème est courant en machine learning ou imagerie computationnelle. Notre algorithme est une variante des algorithmes dits de Langevin, l'idée principale étant, par analogie avec les méthodes d'optimisation convexe, de poser un problème équivalent plus simple qu'on résout avec une légère modification de PGLA (Proximal Gradient Langevin Algorithm). Les analyses de convergence sont faites à l'aide de nouveaux liens établis entre ces méthodes et les algorithmes d'optimisation convexe. Des résultats expérimentaux sont présentés en dernière partie.

Mots-clef : Langevin, optimisation, MCMC

1 Introduction

Dans ce papier on s'intéresse à un cas particulier du problème général qui est de simuler un échantillon distribué selon une loi de densité log-concave $\pi \propto e^{-U}$, où $U : \mathbb{R}^d \to \mathbb{R}$ est une fonction convexe réelle.

La dimension de l'espace de départ d étant en général très grande pour simuler selon π on utilise des méthodes MCMC s'inspirant de l'équation différentielle stochastique de Langevin suivante :

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

où B_t est un mouvement brownien.

En effet, sous des conditions assez faibles sur U (cf. par exemple [RT96]) le processus solution de (1) aura comme mesure stationnaire π. Ainsi si on arrive à simuler une solution de (1) on aura asymptotiquement accès à l'échantillon voulu.
Les algorithmes de Langevin ciblant π sont alors une discrétisation sous une forme ou une autre de (1). Une des premières variantes est le ULA (Unajusted Langevin Algorithm) introduit dans [RT96] :

$$X_{k+1} = X_k - \gamma_k \nabla U(X_k) + \sqrt{2\gamma} W_{k+1} \qquad (1)$$

Avec W_{k+1} une gaussienne centrée réduite et $(\gamma_n)_{n\in\mathbb{N}}$ une suite de pas.

Des premiers résultats non-asymptotiques de la convergence (en norme TV) de la loi des itérés de ULA vers π ont été obtenus en 2013 par [Dal17]. Remarquons que la discrétisation (1) introduit un biais, et on n'a plus en général la convergence de la distribution des X_k vers la loi cible. Cependant ce biais peut être contrôlé en jouant sur la taille des pas $(\gamma_n)_{n\in\mathbb{N}}$.

Une autre remarque est le fait qu'une itération de (1) est pratiquement (à une perturbation gaussienne près) l'algorithme de descente de gradient visant à minimiser la fonction U. Ce premier lien entre l'optimisation et les méthodes de Langevin a été renforcé dans les trois papiers sortis en printemps 2018 ([DMM18], [Wib18], [Ber18]) les auteurs décomposent une itération de (1) en deux parties : en premier lieu la "descente de gradient" vise à minimiser U (ou ce qui est équivalent l'énergie potentielle associé), et le bruit gaussien cherche à minimiser l'entropie (l'entropie et l'énergie potentielle sont définis aux équations (10) et (11)). De ce fait, par analogie avec les méthodes d'optimisation des fonctions composées ULA tend à minimiser la somme des deux qui se trouve être (à constante près) la divergence de Kullback-Leibler par rapport à π.

Ce point de vue permet non seulement de mieux comprendre la nature des méthodes de Langevin, mais introduit aussi de nouvelles techniques de preuves et ouvre la voie à de nouveaux algorithmes. Ainsi dans [Ber18] l'étape de descente de gradient est remplacée

par un pas proximal, alors que dans [DMM18] le cas où U a une composante non lisse est étudié et la descente de gradient devient une descente de gradient proximale.

Le cas qui nous intéresse est le cas où U s'écrit sous une forme particulière :

$$U = f + g \circ \mathrm{M} \tag{2}$$

avec f et g deux fonctions convexes et M un opérateur affine. On supposera de plus que l'opérateur proximal de $g \circ M$ ne peut pas être calculé facilement. Des problèmes de ce type interviennent en machine learning et imagerie computationnelle dans un cadre bayésien, ici $g \circ \mathrm{M}$ représenterait la norme TV (cf. [Per16], [DMP18] pour plus de détails), une application au filtrage sur graphe sera donnée section 5

Si notre but était de trouver le minimum de U, un moyen de procéder serait alors d'effectuer le changement de variable $y = \mathrm{M}\, x, y \in \mathcal{Y}$ (ou $\mathcal{Y}$ est l'espace de départ de g) et remarquer que :

$$\inf_{x \in \mathbb{R}^d} U(x) = \inf_{y \in \mathrm{Im}\, \mathrm{M}} f(\mathrm{M}^{-1}\, y) + g(y) = \inf_{y \in \mathcal{Y}} F(y) + g(y)$$

où $F(y) = f(\mathrm{M}^{-1}\, y)$ si $y \in \mathrm{Im}\, \mathrm{M}$ et $+\infty$ sinon. Si l'opérateur proximal de F peut être calculé on peut alors trouver le minimum recherché par le schema de Passty [Pas79] :

$$y_{k+1} = \mathrm{prox}_F^{\gamma}(\mathrm{prox}_g^{\gamma}(y_k)) \tag{3}$$

$$x_{k+1} = \mathrm{M}^{-1}\, y_{k+1} \tag{4}$$

L'algorithme que nous proposons Passty-Langevin (PL) est alors analogue à ce schema. Nous minimisons dans E la fonction $f(M^{-1}\cdot)+g(\cdot)$ par la méthode qu'on vient de décrire et nous rajoutons un bruit gaussien dans l'espace E. Plus précisément en notant $(\gamma_k)_{k \in \mathbb{N}}$ une suite de pas l'algorithme que nous proposons est :

$$Y_{k+1} = \mathrm{prox}_F^{\gamma_{k+1}}(\mathrm{prox}_g^{\gamma_k}(Y_k)) + \sqrt{2\gamma_{k+1}}\Xi_{k+1} \tag{5}$$

$$X_{k+1} = \mathrm{M}^{-1}\, Y_{k+1} \tag{6}$$

où Ξ_{k+1} est une gaussienne standard dans $\mathrm{Im}\, M$. Nous montrerons que sous des conditions de régularité sur f et g, la convergence de la loi de X_{k+1} vers π.

Nous introduisons les concepts mathématiques nécessaires en section 2 puis posons rigoureusement le problème et présentons notre algorithme en section 3. En 4 l'analyse de convergence de l'algorithme est effectuée, la structure de cette partie, les résultats et les techniques des démonstrations suivent de près ceux de [DMM18]. Enfin en section 5 on compare notre algorithme à l'état de l'art sur le problème de filtrage sur graphe bayésien. C'est un cas où l'opérateur proximal de $g \circ \mathrm{M}$ de (2) ne peut qu'être approché par un algorithme itératif, les expériences numériques confirment alors l'avantage de notre méthode.

2 Notations

Pour $\gamma > 0$ et $\phi : \mathbb{R}^k \to \mathbb{R}$ l'opérateur proximal de pas γ pour ϕ est définit comme suit :

$$\mathrm{prox}_{\phi}^{\gamma}(x) = \arg\min_{y \in \mathbb{R}^k} \{\frac{1}{2\gamma} \|y - x\|^2 + \phi(y)\} \tag{7}$$

On sait que pour les fonctions convexes cet opérateur est bien défini, et vérifie la propriété suivante (voir par exemple [NPB14])

$$\mathrm{prox}_{\phi}^{\gamma}(x) \in x - \gamma \partial\phi(\mathrm{prox}_{\phi}^{\gamma}(x)) \tag{8}$$

où $\partial\phi(x) = \{v \in \mathbb{R}^k / \phi(y) \geqslant \phi(x) + \langle v, y - x\rangle \ \forall y \in \mathbb{R}^k\}$ est le sous-gradient de ϕ en x.

On note $\mathcal{P}_2(\mathbb{R}^d)$ l'espace des mesures de probabilités sur $\mathbb{R}^d$ admettant un moment d'ordre deux. Pour un sous-espace vectoriel $E \subset \mathbb{R}^d$, $\mathcal{P}_2(E) \subset \mathcal{P}_2(\mathbb{R}^d)$ sont les mesures de probabilités à support dans E.

Pour tout $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ on définit la distance de wasserstein d'ordre deux $\mathcal{W}_2$:

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{\substack{Y \sim \nu \\ X \sim \mu}} \mathbb{E}[(X - Y)^2] \tag{9}$$

$\mathcal{P}_2(\mathbb{R}^d)$ muni de cette distance est un espace métrique, séparable, complet et l'infimum dans (9) est atteint (cf. [Vil09] chapitre 6)

Pour E un sous-espace vectoriel de $\mathbb{R}^d$ nous définissons l'entropie relative à E comme une fonctionnelle $\mathcal{H}_E : \mathcal{P}_2(\mathbb{R}^d) \to]-\infty, +\infty]$:

$$\mathcal{H}_E(\mu) = \begin{cases} \int_E \frac{\mathrm{d}\mu}{\mathrm{d}\lambda_E} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda_E}\right) \mathrm{d}\lambda_E & \text{si } \mu \ll \lambda_E \\ +\infty \text{ sinon} \end{cases} \tag{10}$$

Pour U une fonction convexe nous définissons l'énergie potentielle associée à U, $\mathcal{E}_U(\mu) : \mathcal{P}_2(\mathbb{R}^d) \to]-\infty, +\infty]$ comme :

$$\mathcal{E}_U(\mu) = \int_x U(x)\mu(\mathrm{d}x) \tag{11}$$

La somme des deux est alors la fonctionnelle d'énergie libre associée à U.

$$\mathcal{F}(\mu) = \mathcal{H}_E(\mu) + \mathcal{E}_U(\mu) \tag{12}$$

Un calcul simple montre que si $\pi_E \in \mathcal{P}_2(\mathbb{R}^d)$ admet Ce^{-U} comme densité de probabilité par rapport à λ_E on a :

$$\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi) + \log(C) \tag{13}$$

où KL est la distance de Kullback-Leibler. (13) montre entre autre que π_E minimise $\mathcal{F}$. variable

3 Position du problème

Nous nous intéressons à simuler un échantillon selon une loi de probabilité absolument continue par rapport à $\lambda_{\mathbb{R}^d}$ de densité :

$$\pi(\mathrm{d}x) = \frac{e^{-U(x)}}{\int_{\mathbb{R}^d} e^{-U(x)}\lambda_{\mathbb{R}^d}(\mathrm{d}x)}\lambda_{\mathbb{R}^d}(\mathrm{d}x) \tag{14}$$

Nous supposons que la fonction $U : \mathbb{R}^d \to \mathbb{R}$ se décompose en $U = f + g \circ M$, avec $f : \mathbb{R}^d \to \mathbb{R}$ une fonction convexe différentiable de gradient L-Lipschitz, $g : \mathbb{R}^p \to \mathbb{R}$ une fonction convexe continue C-Lipschitz. Et $M : \mathbb{R}^d \to \mathbb{R}^p$ est une fonction affine injective.

Dans la suite on note le sous-espace vectoriel $E = \mathrm{Im}\,\mathrm{M}$, la dimension de E est donc d. Nous notons $F : \mathbb{R}^p \to \overline{\mathbb{R}}$ l'infimale post-composition de M par f ([BC11] section 12.5) devient par injectivité de M :

$$\begin{aligned} F(y) &= \inf_{y \in \mathrm{Im}\,\mathrm{M}} f(\mathrm{M}^{-1}\,y) \\ &= \begin{cases} f(\mathrm{M}^{-1}\,y) \text{ si } y \in \mathrm{Im}\,\mathrm{M} \\ +\infty \text{ sinon} \end{cases} \end{aligned} \tag{15}$$

F est alors convexe ([BC11] proposition 12.36) de domaine E.

Avec ces notations nous notons $\pi_E \in \mathcal{P}_2(E)$ la mesure de probabilité absolument continue par rapport à λ_E :

$$\pi_E(\mathrm{d}y) = \frac{e^{-F(y)+g(y)}}{\int e^{-F(y)+g(y)}\lambda_E(\mathrm{d}y)}\lambda_E(\mathrm{d}y) \tag{16}$$

par un simple changement de variable on a que si $Y \sim \pi_E$, $M^{-1}Y \sim \pi$

Une fois placé dans E notre algorithme est simplement une variante des algorithmes de Langevin usuels. Plus précisément partant d'une variable aléatoire $Y_0 \sim \mu_0 \in \mathcal{P}_2(E)$ et d'une suite de pas $(\gamma_k)_{k\in\mathbb{N}} \in \mathbb{R}_+$ l'algorithme que nous étudions (PL) est :

$$\begin{aligned} Y_{k+1/3} &= \mathrm{prox}_g^{\gamma_k}(Y_k) \\ Y_{k+2/3} &= \mathrm{prox}_F^{\gamma_{k+1}}(Y_{k+1/3}) \\ Y_{k+1} &= Y_{k+2/3} + \sqrt{2\gamma_{k+1}}\Xi_{k+1} \\ X_{k+1} &= \mathrm{M}^{-1}\,Y_{k+1} \end{aligned} \tag{17}$$

où Ξ_{k+1} est une gaussienne standard sur E.
Nous montrerons dans la section suivante qu'asymptotiquement la loi de Y_k sera proche de π_E en KL, par invariance par transformation affine de la KL on aura alors simplement la proximité de la loi de X_k à π

4 Analyse de la convergence

Une fois qu'on est sur l'espace E, l'algorithme (17) est pratiquement similaire au PGLD (Proximal Gradient Langevin Dynamics) de [DMM18] (section 4.2), nous remplaçons juste une descente de gradient par une descente proximale et prenons quelques précautions supplémentaire dû à la forme de F. Les théorèmes et les résultats présentés dans cette section suivent donc [DMM18].

Comme le montre l'équation (13) π_E est le minimiseur de $\mathcal{F}$. Pour prouver la convergence de l'algorithme nous contrôlerons à chaque itération la quantité $\mathcal{F}(\mu_{k+1}) - \mathcal{F}(\pi_E)$. Ainsi on introduit la décomposition suivante :

$$\begin{aligned} \mathcal{F}(\mu_{k+1}) - \mathcal{F}(\pi_E) = \mathcal{H}_E(\mu_{k+1}) - \mathcal{H}_E(\pi_E) \\ + \mathcal{E}_U(\mu_{k+1}) - \mathcal{E}_U(\mu_{k+2/3}) \\ + \mathcal{E}_U(\mu_{k+2/3}) - \mathcal{E}_U(\pi_E)) \end{aligned}$$

où μ_i est la distribution de probabilité de Y_i.
Le terme impliquant l'entropie est contrôlé par des techniques de flots de gradients comme dans [DMM18], les incréments d'énergie libre sont gérés à l'aide des propriétés de régularité et convexité de g et de F.

Lemme 1. *Si f est de gradient L-Lipschitz on a :*

$$\mathcal{E}_F(\mu_{k+1}) - \mathcal{E}_F(\mu_{k+2/3}) \leqslant Lp||\,\mathrm{M}^{-1}\,||_2^2\gamma_{k+1}$$

Démonstration. Soit x, y dans E. Comme f est convexe et de gradient L-Lipschitz on a :

$$\begin{aligned} F(x+y) - F(x) &= f(\mathrm{M}^{-1}(x+y)) - f(\mathrm{M}^{-1}\,x)) \\ &\leqslant \langle \nabla f(\mathrm{M}^{-1}\,x), \mathrm{M}^{-1}\,y\rangle + \frac{L}{2}||\,\mathrm{M}^{-1}\,y||^2 \end{aligned}$$

En prenant deux variables aléatoires Y, Z telles que $Y \sim \mu_{k+2/3}$ et Z une gaussienne standard dans $\mathrm{Im}\,E$ on a $Y + Z \sim \mu_{k+1}$. Ainsi en passant à l'espérance on obtient :

$$\begin{aligned} &\mathcal{E}_F(\mu_{k+1}) - \mathcal{E}_F(\mu_{k+2/3}) \\ &\leqslant \frac{\int_y\int_z \langle \nabla f(\mathrm{M}^{-1}\,y), \mathrm{M}^{-1}\,z\rangle e^{-\frac{||z-y||^2}{4\gamma}}\lambda_E(\mathrm{d}z)\lambda_E(\mathrm{d}y)}{(4\pi\gamma)^{p/2}} \\ &+ \frac{\int_y\int_z \frac{L}{2}||\,\mathrm{M}^{-1}\,z||^2 e^{-\frac{||z-y||^2}{4\gamma}}\lambda_E(\mathrm{d}z)\lambda_E(\mathrm{d}y)}{(4\pi\gamma)^{p/2}} \\ &\leqslant Lp||\,\mathrm{M}^{-1}\,||_2^2\gamma \end{aligned}$$

□

Lemme 2. *$\forall \nu \in \mathcal{P}_2(E)$ on a :*

$$\begin{aligned} &2\gamma_{k+1}(\mathcal{E}_F(\mu_{k+2/3}) - \mathcal{E}_F(\nu)) \\ &\quad \leqslant \mathcal{W}_2^2(\mu_{k+1/3}, \nu) - \mathcal{W}_2^2(\mu_{k+2/3}, \nu) \end{aligned} \tag{18}$$

Démonstration. Soit x, z dans $\mathbb{R}^d$, et $y = \text{prox}_F^{\gamma_{k+1}}(x)$. Alors on a :

$$\begin{aligned} ||y-z||^2 &= ||z-x||^2 + 2\langle y-x, x-z\rangle + ||y-x||^2 \\ &= ||z-x||^2 + 2\langle y-x, y-z\rangle - ||y-x||^2 \end{aligned}$$

On a d'après (8) $x - y \in \gamma_{k+1}\partial F(y)$ donc par convexité de F, on a $\langle y-x, y-z\rangle \leqslant \gamma_{k+1}(F(z) - F(y))$ et :

$$||y-z||^2 \leqslant ||z-x||^2 + 2\gamma_{k+1}(F(z) - F(y))$$

Nous obtenons alors notre inégalité en prenant le couple optimal $X \sim \mu_{k+1/3}, Z \sim \nu$ tel que $\mathbb{E}[||X - Z||^2] = \mathcal{W}_2^2(\mu_{k+1/3}, \nu)$ □

Nous pouvons maintenant énoncer le théorème suivant :

Théorème 1. *S f est de gradient L-Lipschitz, g est C-Lipschitz et μ_i est la densité de probabilité de Y_i on a :*

$$\begin{aligned} 2\gamma_{k+1}(\mathcal{F}(\mu_{k+1}) - \mathcal{F}(\pi_E)) \leqslant\ & 2\gamma_{k+1}^2(Lp||\,\mathrm{M}^{-1}\,||_2^2 + C^2) \\ &+ \mathcal{W}_2^2(\mu_{k+1/3}, \pi_E) \\ &- \mathcal{W}_2^2(\mu_{k+4/3}, \pi_E) \end{aligned} \tag{19}$$

Démonstration. En prenant $\nu = \pi_E$ les lemmes 1 et 2 permettent de contrôler l'incrément de l'énergie potentielle associée à F :

$$\begin{aligned} 2\gamma_{k+1}(\mathcal{E}_F(\mu_{k+2/3}) - \mathcal{E}_F(\pi_E)) = & 2Lp||\,\mathrm{M}^{-1}\,||_2^2\gamma_{k+1}^2 \\ &+ \mathcal{W}_2^2(\mu_{k+1/3}, \pi_E) \\ &- \mathcal{W}_2^2(\mu_{k+2/3}, \pi_E) \end{aligned}$$

Le lemme 29 de [DMM18] permet de contrôler l'incrément de l'énergie potentielle associée à g :

$$\begin{aligned} 2\gamma_{k+1}(\mathcal{E}_g(\mu_{k+1}) - \mathcal{E}_g(\pi_E)) \leqslant & \mathcal{W}_2^2(\mu_{k+1}, \pi_E) \\ &- \mathcal{W}_2^2(\mu_{k+4/3}, \pi_E) \\ &+ 2\gamma_{k+1}^2 C^2 \end{aligned}$$

Finalement le lemme 5 de [DMM18] nous donne une borne sur l'incrément de l'entropie relative

$$\begin{aligned} 2\gamma_{k+1}(\mathcal{H}_E(\mu_{k+1}) - \mathcal{H}(\pi_E)) \leqslant\ & \mathcal{W}_2^2(\mu_{k+2/3}, \pi_E) \\ &- \mathcal{W}_2^2(\mu_{k+1}, \pi_E) \end{aligned}$$

On obtient alors l'inégalité voulue en sommant les trois équations. □

Par convexité de la divergence de Kullback-Leibler nous obtenons un théorème du type théorème 6 de [DMM18] sur la convergence des moyennes des itérés :

Théorème 2. *Posons $\nu_{n+1} = \frac{1}{n+1}\sum_{i=0}^n \mu_i$, $C_3 = 2(Ld||\,\mathrm{M}^{-1}\,||_2^2 + C^2)$ et supposons $\mathcal{W}_2^2(\mu_{1/3}, \pi_E) \leqslant C_4$ où C_4 est une certaine constante alors en prenant un pas constant $\gamma = \frac{\epsilon}{2C_3}$ et $n+1 \geqslant \frac{4C_3C_4}{\epsilon^2}$ nous avons :*

$$\mathrm{KL}(\nu_n, \pi_E) \leqslant \epsilon \tag{20}$$

Démonstration. Par convexité de la KL on a :

$$\begin{aligned} &\mathrm{KL}(\nu_n, \pi_e) \leqslant \frac{1}{n+1}\sum_{k=0}^n \mathrm{KL}(\mu_k, \pi_E) \\ &\leqslant \sum_{k=0}^n \frac{(C_3\gamma + \frac{1}{\gamma}(\mathcal{W}_2^2(\mu_{k+1/3}, \pi_E) - \mathcal{W}_2^2(\mu_{k+4/3}, \pi_E)))}{n+1} \\ &\leqslant C_3\gamma + \frac{1}{\gamma(n+1)}(W_2^2(\mu_{1/3}, \pi_E) - \mathcal{W}_2^2(\mu_{n+4/3}, \pi_E)) \\ &\leqslant C_3\gamma + \frac{C_4}{\gamma(n+1)} \end{aligned} \tag{21}$$

L'inégalité voulue est obtenue en remplaçant γ et n par leurs valeurs respectives. □

Maintenant que la convergence dans l'espace E vers π_E est assurée, on obtient facilement la même borne entre les lois de $X_k = M^{-1}Y_k$ et π notre loi cible.

Théorème 3. *Notons $\tilde{\mu}_i$ la densité de $X_i = \mathrm{M}^{-1}\,Y_i$ et $\tilde{\nu}_n = \frac{1}{n}\sum_{i=1}^n \tilde{\mu}_i$ alors sous les mêmes hypothèse que le théorème 2 nous avons :*

$$\mathrm{KL}(\tilde{\nu}_n, \pi) = \mathrm{KL}(\nu_n, \pi_E) \leqslant \epsilon \tag{22}$$

Démonstration. Cela est simplement dû au fait que KL est invariante par transformation affine. □

5 Expériences numériques

Considérons un graphe non orienté $G = (V, E)$ où V est l'ensemble des nœuds et E l'ensemble des arêtes, une orientation arbitraire de ce graphe est la matrice d'incidence $\nabla : \mathbb{R}^V \to \mathbb{R}^E$ associée à cette orientation. Dans le contexte statistique bayésien décrit dans [WSST16], une réalisation d'un vecteur aléatoire $\theta \in \mathbb{R}^V$ est observée. La loi de ce vecteur aléatoire est paramétrée par un vecteur lui même aléatoire $x \in \mathbb{R}^V$ et de loi *a priori* $p(x) \propto \exp(-\lambda\|\nabla x\|_1)$. La vraisemblance du modèle est définie par $L(x|\theta) \propto \exp(-\|x - \theta\|_2^2)$, et l'estimateur du filtrage de tendance sur le

graphe G est l'estimateur du maximum a posteriori associé à ce modèle bayésien.

Dans cette partie, nous nous proposons d'échantillonner selon la loi *a posteriori* de ce modèle

$$\pi(x|\theta) \propto \exp(-U_\theta(x))$$

où $U_\theta(x) = \frac{1}{2}\|x-\theta\|_2^2 + \lambda\|\nabla x\|_1$. Pour cela, nous commençons par décomposer $\pi(\cdot|\theta)$ comme produit de deux lois indépendantes. Pour cela, notons P la projection orthogonale sur le noyau de la matrice ∇ et Q la projection orthogonale sur l'orthogonal du noyau. Notons également,

$$U_1(x) := \frac{1}{2}\|Q(\theta) - x\|_2^2 + \lambda\|\nabla Q(x)\|_1,$$

et

$$U_2(x) := \frac{1}{2}\|P(\theta) - x\|_2^2.$$

Alors, la relation de Pythagore donne $U_\theta(x) = U_1(Q(x)) + U_2(P(x))$. Pour échantillonner proportionnellement à $\exp(-U_2)$, il suffit d'échantillonner une loi gaussienne réduite centrée en $P(\theta)$ sur le noyau de ∇ (qui est une droite). Pour échantillonner proportionnellement à $\exp(-U_1)$ nous utilisons l'algorithme (PL) sur l'orthogonal du noyau avec $M = \nabla$, $g(y) = \lambda\|y\|_1$ et $f(x) = \frac{1}{2}\|Q(\theta) - x\|^2$. Le calcul de l'opérateur proximal de la post-composition F revient à la résolution d'un système linéaire Laplacien qui peut être réalisée efficacement [Spi10]. En utilisant le graphe "Facebook" de [LK14], nous comparons l'algorithme (PL) à l'algorithme

$$X_{k+1} = \mathrm{prox}_{U_\theta}^{\gamma_{k+1}}(X_k) + \sqrt{2\gamma_{k+1}}W_{k+1}$$

étudié dans [DMM18], que nous appelons "Proximal-Langevin". Le calcul de $\mathrm{prox}_{U_\theta}^{\gamma_{k+1}}$ se ramène à un calcul de l'opérateur proximal de $\|\nabla x\|_1$ qui est connu pour être difficile. Nous utilisons une sous-routine (l'algorithme du gradient projeté pour le problème dual) pour le calculer (voir par exemple [SBHJ16]). Nous représentons $\mathcal{F}(\tilde{\nu}_n)$ (nous estimons cette grandeur en utilisant 100 échantillons) en fonction du temps CPU (en utilisant un coeur d'un CPU de 2800 MHz et 256 GB de RAM).

L'algorithme Proximal-Langevin utilise une sous-routine à chaque étape, ce qui l'empêche de produire des itérés aussi souvent que Passty-Langevin. De plus, Proximal-Langevin est plus lent dans ce cas, ce qui est dû au calcul inexact de $\mathrm{prox}_{U_\theta}^{\gamma_{k+1}}$ à chaque étape.

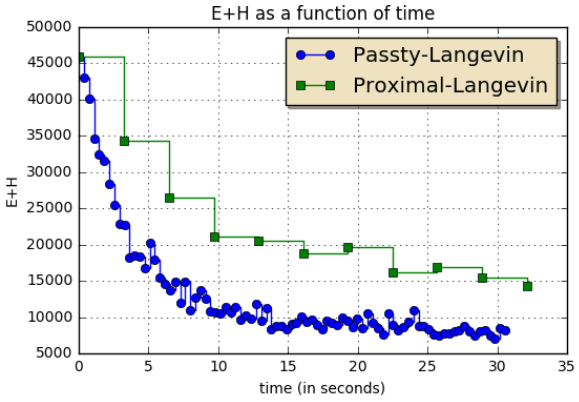

FIGURE 1 – $\mathcal{F}(\tilde{\nu}_n)$ en fonction du temps CPU pour les deux algorithmes

6 Conclusion

Nous présentons dans cet article un nouvel algorithme de type Langevin visant à simuler selon une densité log-concave lorsque la fonction convexe sous-jacente fait intervenir la composée d'une fonction convexe avec un opérateur affine et où l'opérateur proximal de cette composée ne peut pas être calculé facilement. L'exemple typique où ce problème intervient est un cadre bayésien où intervient la norme TV (cf. [DMP18] pour les problèmes d'imagerie computationnelle ou la section 5).

Notre méthode consiste à résoudre le problème dans un espace dual équivalent, mais où il n'est plus nécessaire de calculer l'opérateur proximal correspondant. On montre alors que la vitesse de convergence en nombre d'itérations reste alors la même ($\frac{1}{\epsilon^2}$ itérations pour une précision de ϵ), mais les expériences menées confirment l'avantage de notre méthode en temps CPU.

Ce projet a été soutenu par l'attribution d'une allocation de recherche de la Région Ile-de-France.

Références

[BC11] H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 1st edition, 2011.

[Ber18] E. Bernton. Langevin Monte Carlo and JKO splitting. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1777–1798. PMLR, 06–09 Jul 2018.

[Dal17] A. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. *J. R. Stat. Soc. B*, 79 :651–676, 2017.

[DMM18] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization, 2018.

[DMP18] A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo : When Langevin Meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1), 2018.

[LK14] Jure Leskovec and Andrej Krevl. SNAP Datasets : Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[NPB14] Neal N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3) :127–239, January 2014.

[Pas79] G. Passty. Ergodic convergence to a zero of the sum of monotone operators in hilbert space. 72 :383–390, 12 1979.

[Per16] M. Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4) :745–760, 2016.

[RT96] G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4) :341–363, 12 1996.

[SBHJ16] A. Salim, P. Bianchi, W. Hachem, and J. Jakubowicz. A stochastic proximal point algorithm for total variation regularization over large scale graphs. *IEEE CDC*, 2016.

[Spi10] Daniel A Spielman. Algorithms, graph theory, and linear equations in laplacian matrices. In *Proceedings of the ICM*, volume 4, pages 2698–2722, 2010.

[Vil09] C. Villani. *Optimal Transport : Old and New*. Grundlehren der mathematischen Wissenschaften 338. Springer-Verlag Berlin Heidelberg, 1 edition, 2009.

[Wib18] A. Wibisono. Sampling as optimization in the space of measures : The langevin dynamics as a composite optimization problem. In *COLT*, 2018.

[WSST16] Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1) :3651–3691, 2016.