Echantillonage de loi log-concave en grande dimension.

Alain Durmus [1]

[1] *LTCI, Telecom ParisTech & CNRS,*
46 rue Barrault, 75634 Paris Cedex 13, France.
alain.durmus@telecom-paristech.fr

Résumé. Echantilloner des lois en grandes dimensions est devenu un pré-requis dans les applications des statistiques Bayésiennes au machine learning. Dans de nombreuses cas 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 ne sont pas en général smooth. Les pénalités classiques sont celles associées au problème du LASSO ou de l'élastic net. Nous exposerons dans cette présentation des méthodes pour échantillonner de telles lois, qui proviennent de l'état de l'art des procédures d'optimisation en grande dimensions. Des bornes explicites en variation totale et en distance de Wasserstein seront aussi introduites. Une intention toute particulière sera d'expliciter la dépendence de ces bornes en la dimension. Ces méthodes seront appliqués à problème de déconvolution en image. Travaux en collaboration avec Eric Moulines et Marcelo Pereyra.

Mots-clés. diffusion de Langevin, méthode de Monte Carlo par chaîne de Markov, Metropolis Adjusted Langevin Algorithm, taux de convergence.

Abstract. Sampling over high-dimensional space has become a prerequisite in the applications of Bayesian statistics to machine learning problem. 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). We will describe methods to sample such distributions, which are adapted from the state-of-the-art optimization procedures which have been developed in this context. We will also provide convergence in Wasserstein distance to the equilibrium, showing explicitly the dependence in the dimension of the parameter space. Joint work with Éric Moulines et Marcelo Pereyra.

Keywords. Langevin diffusion, Markov Chain Monte Carlo, Metropolis Adjusted Langevin Algorithm, Rate of convergence.

1 Résumé long

Dans le cadre de l'inférence bayésienne, il est nécessaire d'échantilloner une loi connue à une constante près. En effet, à partir d'une loi de $N \in \mathbb{N}$ observations $\mathrm{y}_1, \cdots, \mathrm{y}_N$

paramétrée par un paramètre $\theta \in \mathbb{R}^d$, $p((\mathrm{y}_i)_{1\leq i\leq N}|\theta)$, et d'une loi à priori de densité $p_\star$ sur ce paramètre, une nouvelle loi à posteriori π est donnée par la formule de Bayes par la densité encore notée π :

$$\pi(\theta) = \frac{p((\mathrm{y}_i)_{1\leq i\leq N}|\theta)p_\star(\theta)}{\mathcal{Z}} \ , \ \text{où } \mathcal{Z} = \int_{\mathbb{R}^d} p((\mathrm{y}_i)_{1\leq i\leq N}|\theta)p_\star(\theta)\mathrm{d}\theta \ .$$

Cependant, sauf dans le cas de lois conjuguées, 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 échantilloner 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, deviennent de plus en plus importantes, ce qui impacte fortement la convergence des méthodes MCMC classiques, comme le random walk Metropolis ([10]-[16]). 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 π 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 π est l'équation différentielle stochastique définie par :

$$\mathrm{d}Y_t^{\mathrm{L}} = -\nabla U(Y_t^{\mathrm{L}})\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^d \ , \tag{1}$$

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 π et est ergodique [9]. 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. Ces taux dont la dépendance en la dimension d est explicite, ont été récemment obtenus en utilisant des inégalités fonctionnelles, tels que les inégalités de Poincaré et log-Sobolev (cf. [1, 3] [2]) ou par couplage [6]. 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. Cet algorithme est la méthode MCMC qui utilise la discrétisation de Euler-Maruyama de (1), pour échantillonner π. Cette discrétisation définie une chaîne de Markov $(X_k)_{k\geq 0}$ par :

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

où $(Z_k)_{k\geq 1}$ est une suite i.i.d. de variables aléatoires standard Gaussienne de dimension d et $(\gamma_k)_{k\geq 1}$ est une suite décroissante de pas, qui est soit constante ou tend vers 0.

Cette méthode a d'abord été proposé pour des applications en physique statistique par [17] et ensuite introduite en statistique computationnelle par [7] et [8]. Elle a déjà retenu l'attention de nombreux travaux. A pas constant, l'étude de la chaîne de Markov définie par (2) a été faite dans [18] et [11]. Dans ce cas, sous des conditions appropriés sur U, la chaîne est V-géométriquement ergodique et converge vers une loi invariante π_γ différente

de π. [14] et [13] ont aussi montré la convergence faible de la mesure empirique associée pour des fonctionnelles suffisamment régulières vers π, dans le cas des pas décroissants tendant vers 0. Récemment sous l'hypothèse que le potentiel U est fortement convexe, [4] a obtenu des bornes non-asymptotiques en total variation entre la loi de la chaîne définie par (2) et π. Ces bornes sont accompagnées d'une étude précise de la convergence en fonction de la dimension.

Nous présenterons une extension de ces différents résultats sous différentes hypothèses sur le potentiel U. En particulier, nous obtenons une étude de la convergence en fonction de la dimension similaire à [4] dans le cas où le potentiel U est simplement convexe. Dans le cadre de pas constants, nous exposerons aussi une borne explicite en V norme entre la loi invariante π_γ de (2) et π. Dans une seconde partie, nous proposerons un nouvel algorithme pour échantillonner une loi log-concave dont le potentiel n'est pas nécessairement continûment différentiable. Cet algorithme est construit en s'inspirant des nouvelles techniques d'optimisation en grande dimension. Ces techniques utilisent comme outil essentiel, les opérateurs proximaux. Pour conclure, des bornes non-asymptotiques pour cet algorithme seront aussi introduites.

Références

[1] D. Bakry, P. Cattiaux, and A. Guillin. Rate of convergence for ergodic continuous Markov processes : Lyapunov versus Poincaré. *J. Funct. Anal.*, 254(3) :727–759, 2008.

[2] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

[3] P. Cattiaux and A. Guillin. Trends to equilibrium in total variation distance. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1) :117–145, 2009.

[4] A. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. submitted 1412.7392, arXiv, December 2014.

[5] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5) :1423–1443, 2012.

[6] A. Eberle. Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, pages 1–36, 2015.

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

[8] U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *J. Roy. Statist. Soc. Ser. B*, 56(4) :549–603, 1994. With discussion and a reply by the authors.

[9] R. Z. Hasminskiĭ. *Stochastic stability of differential equations*, volume 7 of *Monographs and Textbooks on Mechanics of Solids and Fluids : Mechanics and Analysis*. Sijthoff & Noordhoff, Alphen aan den Rijn—Germantown, Md., 1980. Translated from the Russian by D. Louvish.

[10] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, April 1970.

[11] Mattingly J.C., Stuart A.M., and Higham D.J. Ergodicity for {SDEs} and approximations : locally lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2) :185 – 232, 2002.

[12] D. Lamberton and G. Pagès. Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3) :367–405, 2002.

[13] D. Lamberton and G. Pagès. Recursive computation of the invariant distribution of a diffusion : the case of a weakly mean reverting drift. *Stoch. Dyn.*, 3(4) :435–451, 2003.

[14] V. Lemaire. *Estimation de la mesure invariante d'un processus de diffusion*. PhD thesis, Université Paris-Est, 2005.

[15] V. Lemaire and S. Menozzi. On some non asymptotic bounds for the Euler scheme. *Electron. J. Probab.*, 15 :no. 53, 1645–1681, 2010.

[16] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6) :1087–1092, 1953.

[17] G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180 :378–384, 1981.

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