Log-concave sampling: Metropolis-Hastings algorithms are fast!

Raaz Dwivedi◇ RAAZ.RSK@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences, UC Berkeley

Yuansi Chen◇ YUANSI.CHEN@BERKELEY.EDU
Department of Statistics, UC Berkeley

Martin Wainwright WAINWRIG@BERKELEY.EDU
Bin Yu BINYU@BERKELEY.EDU
Department EECS and Department of Statistics, UC Berkeley

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We consider the problem of sampling from a strongly log-concave density in $\mathbb{R}^d$, and prove a non-asymptotic upper bound on the mixing time of the Metropolis-adjusted Langevin algorithm (MALA). The method draws samples by running a Markov chain obtained from the discretization of an appropriate Langevin diffusion, combined with an accept-reject step to ensure the correct stationary distribution. Relative to known guarantees for the unadjusted Langevin algorithm (ULA), our bounds reveal that the use of an accept-reject step in MALA leads to an exponentially improved dependence on the error-tolerance. Concretely, in order to obtain samples with TV error at most δ for a density with condition number κ, we show that MALA requires $\mathcal{O}\big(\kappa d \log(1/\delta)\big)$ steps, as compared to the $\mathcal{O}\big(\kappa^2 d/\delta^2\big)$ steps established in past work on ULA. We also demonstrate the gains of MALA over ULA for weakly log-concave densities. Furthermore, we derive mixing time bounds for a zeroth-order method Metropolized random walk (MRW) and show that it mixes $\mathcal{O}(\kappa d)$ slower than MALA.

Keywords: MCMC, sampling, random walk, Metropolis-adjusted Langevin algorithm, convergence

1. Main Results

Recent decades have witnessed great success of Markov Chain Monte Carlo (MCMC) algorithms suited for generating random samples; for instance, see the handbook (Brooks et al., 2011) and references therein. In a broad sense, these methods are based on two steps. The first step is to construct a Markov chain whose stationary distribution is either equal to the target distribution or close to it in a suitable metric. Given this chain, the second step is to draw samples by simulating the chain for a certain number of steps.

Many algorithms have been proposed and studied for sampling from probability distributions with a density on a continuous space. Two broad categories of these methods are *zeroth-order methods* and *first-order methods*. On one hand, a zeroth-order method is based on querying the density of the distribution (up to a proportionality constant) at a point in each iteration. By contrast, a first-order method also makes use of gradient information about the density. A few popular examples of ze-

. ◇Raaz Dwivedi and Yuansi Chen contributed equally to this work.
Extended abstract. Full version appears as [https://arxiv.org/abs/1801.02309, v2]

roth order algorithms include Metropolized random walk (MRW) (Mengersen et al., 1996; Roberts and Tweedie, 1996b), Ball Walk (Lovász and Simonovits, 1990; Dyer et al., 1991; Lovász and Simonovits, 1993) and the Hit-and-run algorithm (Bélisle et al., 1993; Kannan et al., 1995; Lovász, 1999; Lovász and Vempala, 2006, 2007). A number of first-order methods are based on the Langevin diffusion. Algorithms related to the Langevin diffusion include the Metropolis adjusted Langevin Algorithm (MALA) (Roberts and Tweedie, 1996a; Roberts and Stramer, 2002; Bou-Rabee and Hairer, 2012), the unadjusted Langevin algorithm (ULA) (Parisi, 1981; Grenander and Miller, 1994; Roberts and Tweedie, 1996a; Dalalyan, 2016), underdamped Langevin MCMC (Cheng et al., 2017), Riemannian MALA (Xifara et al., 2014), Proximal-MALA (Pereyra, 2016; Durmus et al., 2016), Metropolis adjusted Langevin truncated algorithm (Roberts and Tweedie, 1996a), Hamiltonian Monte carlo (Neal, 2011) and Projected ULA (Bubeck et al., 2015). There is now a rich body of work on these methods, and we do not attempt to provide a comprehensive summary in this paper. More details can be found in the survey (Roberts et al., 2004), which covers MCMC algorithms for general distributions, and the survey (Vempala, 2005), which focuses on random walks for compactly supported distributions.

In this paper, we study sampling algorithms for sampling from a log-concave distribution equipped with a density. A log-concave density takes the form

$$\pi\left(x\right)=\frac{e^{-f(x)}}{\displaystyle\int_{\mathbb{R}^d}e^{-f(y)}dy}\text{for all } x\in\mathbb{R}^d, \tag{1}$$

where f is a convex function on $\mathbb{R}^d$. Up to an additive constant, the function $-f$ corresponds to the log-likelihood defined by the density. Standard examples of log-concave distributions include the normal distribution, exponential distribution and Laplace distribution.

Some recent work has provided non-asymptotic bounds on the mixing times of Langevin type algorithms for sampling from a log-concave density. The mixing time corresponds to the number of steps, as function of both the problem dimension d and the error tolerance δ, to obtain a sample from a distribution that is δ-close to the target distribution in total variation distance. It is known that both the ULA updates (Dalalyan, 2016; Durmus and Moulines, 2016; Cheng and Bartlett, 2017) as well as underdamped Langevin MCMC (Cheng et al., 2017) have mixing times that scale polynomially in the dimension d, as well the inverse of the error tolerance $1/\delta$.

Both the ULA and underdamped-Langevin MCMC methods are based on evaluations of the gradient ∇f, along with the addition of Gaussian noise. Durmus and Moulines (2016) show that for an appropriate decaying step size schedule, the ULA algorithm converges to the right stationary distribution. However, their results, albeit non-asymptotic, are hard to quantify. In the sequel, we limit our discussion to Langevin algorithms based on constant step sizes, for which there are a number of explicit quantitative bounds on the mixing time. When one uses a fixed step size for these algorithms, an important issue is that the resulting random walks are asymptotically biased: due to the lack of Metropolis-Hastings correction step, the algorithms *will not* converge to the stationary distribution if run for a large number of steps. Furthermore, if the step size is not chosen carefully the chains may become transient (Roberts and Tweedie, 1996a). Thus, typical theory is based on running such a chain for a pre-specified number of steps, depending on the tolerance, dimension and other problem parameters.

In contrast, the Metropolis-Hastings step that underlies the MALA algorithm ensures that the resulting random walk has the correct stationary distribution. Roberts and Tweedie (1996a) derived sufficient conditions for exponential convergence of the Langevin diffusion and its discretizations, with and without Metropolis-adjustment. However, they considered the distributions with $f(x) = \|x\|_2^\alpha$ and proved geometric convergence of ULA and MALA under some specific conditions. In a more general setting, Bou-Rabee and Hairer (2012) and Eberle (2014) derived non-asymptotic mixing time bounds for MALA. However, all these bounds are non-explicit in the case of logconcave sampling, and so makes it difficult to extract an explicit dependence in terms of the dimension d and error tolerance δ. In particular, Eberle (2014) made significant contribution to establishing the accept-reject rate of MALA by assuming differentiability of the target function to fourth order, but its final mixing rate is only explicit when the sampling domain is contained in a ball with constant radius. A precise characterization of this dependence is needed if one wants to make quantitative comparisons with other algorithms, including ULA and other Langevin-type schemes. With this context, one of the main contributions of our paper is to provide an explicit upper bound on the mixing time of the MALA algorithm.

This work contains two main results, both having to do with the mixing times of MCMC methods for sampling. As described above, our first and primary contribution is an explicit analysis of the mixing time of Metropolis adjusted Langevin Algorithm (MALA). A second contribution is to use similar techniques to analyze a zeroth-order method called Metropolized random walk (MRW) and derive a explicit non-asymptotic mixing time bound for it. Unlike the ULA, these methods make use of the Metropolis-hastings accept-reject step and consequently converge to the target distributions in the limit of infinite steps. Here we provide explicit non-asymptotic mixing time bounds for MALA and MRW and show that MALA converges significantly faster than ULA. In particular, we show that if the density is strongly log-concave and smooth, the δ-mixing time for MALA scales as $\kappa d \log(1/\delta)$ which is significantly faster than ULA's convergence rate of order $\kappa^2 d/\delta^2$. We also show that MRW mixes $\mathcal{O}\left(\kappa d\right)$ slowly when compared to MALA. Furthermore, if the density is weakly log-concave, we show that MALA converges in $\mathcal{O}\left(d^2/\delta^{1.5}\right)$ time in comparison to the $\mathcal{O}\left(d^3/\delta^4\right)$ mixing time for ULA. These results are summarized in Table 1.

Acknowledgments

This work was supported by Office of Naval Research grant DOD ONR-N00014 to MJW, and by ARO W911NF1710005, NSF-DMS 1613002 and the Center for Science of Information (CSoI), US NSF Science and Technology Center, under grant agreement CCF-0939370 to BY. In addition, MJW was partially supported by National Science Foundation grant NSF-DMS-1612948, and RD was partially supported by the Berkeley Fellowship.

References

Claude JP Bélisle, H Edwin Romeijn, and Robert L Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993.

Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2012.

Steve Brooks, Andrew Gelman, Galin L Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

Random walk	**Strongly log-concave**	**Weakly log-concave**
ULA (Cheng and Bartlett, 2017)	$\mathcal{O}\left(\frac{d\kappa^2 \log((\log\beta)/\delta)}{\delta^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{dL^2}{\delta^6}\right)$
ULA (Dalalyan, 2016)	$\mathcal{O}\left(\frac{d\kappa^2 \log^2(\beta/\delta)}{\delta^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^3L^2}{\delta^4}\right)$
MRW	$\mathcal{O}\left(d^2\kappa^2 \log\left(\frac{\beta}{\delta}\right)\right)$	$\tilde{\mathcal{O}}\left(\frac{d^4\,L^{2.5}}{\delta^{1.5}}\right)$
MALA	$\mathcal{O}\left(\max\left\{d\kappa, d^{0.5}\kappa^{1.5}\right\} \log\left(\frac{\beta}{\delta}\right)\right)$	$\tilde{\mathcal{O}}\left(\frac{d^2\,L^{1.5}}{\delta^{1.5}}\right)$

Table 1. Scalings of upper bounds on δ-mixing time for different random walks in $\mathbb{R}^d$ with target $\pi \propto e^{-f}$. In the second column, we consider smooth and strongly log-concave densities, and report the bounds from a β-warm start for densities such that $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d$ for any $x \in \mathbb{R}^d$ and use $\kappa := L/m$ to denote the condition number of the density. The big-O notation hides universal constants. In the last column, we summarize the scaling for weakly log-concave smooth densities: $0 \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d$ for all $x \in \mathbb{R}^d$. For this case, the $\tilde{\mathcal{O}}$ notation is used to track scaling only with respect to d, δ and L and ignore dependence on the starting distribution and a few other parameters.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *arXiv preprint arXiv:1507.02564*, 2015.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *arXiv preprint arXiv:1612.07471*, 2016.

Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

Andreas Eberle. Error bounds for metropolis–hastings algorithms applied to perturbations of gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377, 2014.

Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 549–603, 1994.

Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(1):541–559, 1995.

László Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

László Lovász and Miklós Simonovits. The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of 31st Annual Symposium on Foundations of Computer Science, 1990*, pages 346–354. IEEE, 1990.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993.

László Lovász and Santosh Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4): 985–1005, 2006.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Kerrie L Mengersen, Richard L Tweedie, et al. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2 (11), 2011.

G Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4): 745–760, 2016.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996a.

Gareth O Roberts and Richard L Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996b.

Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

Santosh Vempala. Geometric random walks: a survey. *Combinatorial and Computational Geometry*, 52(573-612):2, 2005.

Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91: 14–19, 2014.