Non-asymptotic Analysis of Langevin Monte Carlo Algorithms: A Review of Three Influential Papers

Ghassen Jerfel

1 Introduction

Sampling from continuous distributions over high-dimensional state-spaces is a problem which has recently attracted a lot of research efforts in statistics, machine learning, and computational physics.
This boils down to sampling a target distribution π having a density with respect to the Lebesgue measure on $\mathbb{R}^d$, known up to a normalization factor; $\pi(\theta) = e^{-f(\theta)} / \int_{\mathbb{R}^d} e^{-f(y)} dy$ where f is continuously differentiable. In Bayesian inference, samples are used to construct statistical estimators from posterior summaries of interest such as expectations of desired quantities, credible intervals, and probabilities of rare events. In the frequentist framework, samples drawn from a suitable distribution can form confidence intervals for a point estimate.

1.1 Monte Carlo Markov Chain Methods

A popular class of methods are Monte Carlo Markov Chain methods where one constructs an irreducible and aperiodic discrete-time Markov chain whose stationary distribution is equal or close to π in total variation or some other distance. To obtain an ε-accurate sample, one needs to simulate the Markov chain for a certain number of steps n which is determined by a mixing time analysis.
Two broad categories of sampling methods are zeroth-order and first-order methods. On the one hand, zeroth-order methods are based on querying the density of the distribution (up to a normalizing constant). This includes Metropolized random walk, Ball Walk, and the Hit-and-run algorithms. However, vanilla random walk or Hit-and-run methods have been shown to scale poorly in higher dimensions. On the other hand, choosing an appropriate proposal distribution for Metropolis-Hastings (M-H) algorithms is a tricky subject. Furthermore, M-H methods typically require computations over the whole dataset. Therefore, these methods have gone out of fashion with the rise of large-scale datasets.
For this reason, it has been proposed to consider the discretization of continuous diffusion processes which leave the target distribution invariant. This includes the (overdamped) Langevin Monte Carlo methods which incorporate the gradient of the density to drive a random walk towards regions of high probability. LMC algorithms have also demonstrated faster convergence, in practice, for high-dimensional and large-scale applications.

1.2 Langevin Monte Carlo

Langevin-type algorithms are based on the Langevin diffusion, a stochastic process whose evolution is characterized by the stochastic differential equation (SDE):

$$dY_t = -\nabla f(Y_t)dt + \sqrt{2}dB_t \tag{1}$$

where B_t is the standard Brownian motion on $\mathbb{R}^d$.
Under smoothness assumptions on f (A1), this SDE admits a unique solution $(Y_t)_{t>0}$ which defines a strong Markov semigroup that converges to π in total variation (TV, 4) or Wasserstein distance (5).
However, simulating path solutions of such diffusion processes is not possible in most cases and discretizations of the SDE are used instead. We generally consider the Euler-Maruyama (forward Euler) discretization of the SDE (1) which defines a (possibly time-inhomogeneous) Markov chain X_k:

$$X_{t+1} = X_t - h\nabla f(X_t) + \sqrt{2h}\xi_{t+1}, \tag{2}$$

where $h > 0$ is a tuning parameter, the step size, and $\xi_1 ... \xi_t$ is a sequence of mutually independent (and independent of X_0) standard Gaussian vectors.
The use of this discretization to approximately sample from π is known as the Unadjusted Langevin Algorithm (ULA) or the Langevin Monte Carlo (LMC) algorithm:

Algorithm 1: Overdamped Langevin MCMC

Input : Step size $h < 1$, number of iterations n, initial point x_0, and gradient oracle $\nabla f(\cdot)$
for $i = 0, 1, \ldots, n-1$ **do**
  Sample $(x_{(i+1)}) \sim \mathcal{N}(x_{(i)} - h\nabla f(x_{(i)}), 2hI_{d\times d})$
end

1.3 Motivating the Analysis of Langevin Monte Carlo

Prior literature on MCMC algorithms has focused on establishing behavior and convergence of sampling algorithms in an asymptotic or a non-explicit sense, e.g., geometric and uniform ergodicity, asymptotic variance, and central limit theorems.
However, from such results, it is not easy to determine the computational complexity of various MCMC algorithms as a function of the problem dimension d, desired accuracy ε, and regularity of the potential $f(\cdot)$.
On the other hand, important question that practicioners face is how to choose a sampler for a particular problem, when to know to stop the algorithm, and how to select the step-size and other tuning parameters.
Therefore, we review three papers that analyze the non-asymptotic behavior and convergence of LMC in an explicit way that can be leveraged to guide the practical use of such sampling algorithms.

2 Notation

2.1 Markov Process Background

A (homogenous) Markov process $(Y_t)_{t\in\mathbb{R}_+}$ is a random process that satisfies the Markov property: for every bounded measurable function f and $s, t \in \mathbb{R}_+$ there is a bounded measurable function $P_s f$ such that

$$E[f(Y_{t+s}) \mid \{Y_r\}_{r\leq t}] = P_s f(Y_t)$$

$(P_t)_{t\geq 0}$ is the Markov semigroup associated with $(Y_t)_{t\geq 0}$: νP_t is the law of Y_t starting from $Y_0 \sim \nu$.
A probability measure π is said to be invariant or stationary if $\pi(P_t f) = \pi(f) \forall t \in \mathbb{R}_+$.
To interpret this notion, suppose that $Y_0 \sim \nu$ then $\mathbb{E}\left[g(Y_t)\right] = \mathbb{E}\left[\mathbb{E}\left[g(Y_t) \mid Y_0\right]\right] = \mathbb{E}\left[P_t g(Y_0)\right] = \nu(P_t g)$
We say that $(P_t)_{t\geq 0}$ is geometrically ergodic if there exists $\kappa \in [0, 1)$ such that for any initial distribution μ_0 and $t \geq 0$ we have for some constant $C(\mu_0) \geq 0$

$$\|\mu_0 P_t - \pi\|_{\text{TV}} \leq C(\mu_0)\kappa^t \tag{3}$$

Therefore, in the following analysis we will be interested in the mixing time which is the minimum number of steps, as function of both the problem dimension d and the error tolerance ε, to obtain a sample from a distribution that is ε-close to the target distribution in total variation (TV, 4) or other distances.

2.2 Relevant Distances

For any $d \in \mathcal{N}$ we write $\mathcal{B}(\mathbb{R}^d)$ for the σ-algebra of Borel sets of $\mathbb{R}^d$.
For two distributions ν and μ defined on the space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel-sigma algebra on $\mathbb{R}^d$, we use $\|\nu - \mu\|_{\text{TV}}$ to denote their total variation (TV) distance given by

$$\|\nu - \mu\|_{\text{TV}} = \sup_{A\in\mathcal{B}(\mathbb{R}^d)} |\nu(A) - \mu(A)|. \tag{4}$$

The Wasserstein-Monge-Kantorovich distance or order $p \geq 1$ W_p is defined by

$$W_p(\mu, \nu) = \Big(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\theta - \theta'\|^p \, d\gamma(\theta, \theta')\Big)^{1/p}, \tag{5}$$

where the $\inf$ is with respect to all joint distributions γ having μ and ν as marginal distributions.

2.3 Assumptions on f

A1 The function f is twice continuously-differentiable on $\mathbb{R}^d$ and has Lipschitz continuous gradients; that is, there exists a positive constant $L > 0$ such that for all $x, y \in \mathbb{R}^d$ we have

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{6}$$

A2 f is m-strongly convex, that is, there exists a positive constant $m > 0$ such that for all $x, y \in \mathbb{R}^d$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|x - y\|_2^2. \tag{7}$$

It is fairly easy to show that under these two assumptions the Hessian of f is positive definite throughout its domain, with $mI_{d\times d} \preceq \nabla^2 f(x) \preceq LI_{d\times d}$. We define $\kappa = L/m$ as the condition number.
Finally, we say that the density $\pi(\theta) \propto e^{-f(\theta)}$ is log-concave (resp. strongly log-concave) if the function f satisfies the inequality of A2 with $m = 0$ (resp. $m > 0$)

3 Theoretical Guarantees for Approximate Sampling from Smooth and Log-concave Densities (Dalalyan, 2017b)

Dalalyan (2017b) attempts to bridge the theoretical gap between sampling and optimization by developing explicit non-asymptotic convergence bounds for LMC under smoothness and log-concavity assumptions.
While it is known that gradient-based algorithms, under similar assumptions, are guaranteed logarithmic dependence on the error tolerance ε and independence of the dimension d, no prior convergence bounds for sampling algorithms explicitly analyzed the dependence on the dimension d and the precision ε.
The authors exploit the similarities between LMC and gradient descent algorithms to establish upper bounds on the TV distance between the target distribution π and its approximation by the distribution of the LMC iterates involving only explicit and easy-to-compute quantities.

3.1 Strength : A Novel Setting

Imposing smoothness and strong-convexity assumptions on the log of the density might not be too common in the the Bayesian inference literature. In fact, prior computable bounds involved in the geometric convergence of Markov chains are difficult to implement for getting tight bounds in high dimensions due to the great generality of considered processes.
On the other hand, smoothness and convexity assumptions are quite instrumental in the stochastic differential equations literature. In fact, under assumption (A1), the Langevin SDE (1) admits a unique strong solution which is a Markov process $\{Y_t\}_{t\geq 0}$. Furthermore, when f satisfies strong convexity, the process Y_t is geometrically ergodic in the sense of (3).
However, even when the diffusion is well behaved, the iterates defined by the discretization (2) have mixed behavior making them more difficult to study. For example, for sufficiently small fixed step sizes h, the distribution of iterates defined by (2) converges to a stationary distribution that is no longer equal to π.
Therefore, the authors depart from the classic approaches that directly study the spectral gap or conductance of the discrete Markov Chain convergence, under hard-to-quantify drift and recurrence conditions.
Instead, the authors analyze the discretization of the continuous process by decomposing the total variation distance into: (1) the error of approximating the Langevin diffusion Y_T by the discrete LMC process $X_{K,h}$, where $T = Kh$, and (2) the error of approximating the target distribution π by the distribution of the Langevin diffusion Y_T as follows:

$$\|\mathbf{P}_X^{K,h} - \pi\|_{\text{TV}} \leq \|\mathbf{P}_X^{K,h} - \mathbf{P}_Y^T\|_{\text{TV}} + \|\mathbf{P}_Y^T - \pi\|_{\text{TV}} \tag{8}$$

where $P_X^{K,h}$ is the law of $X_{K,h}$, P_Y^T is the law of Y_T.

3.2 Strength : Novel Analysis Techniques

To control the discretization error, the authors introduce a continuous-time diffusion process D_t as a continuous interpolation of the discrete process $(X_t)_{t\geq 0}$ such that the the distributions of the $X_{k,h}$ and D_{kh} random

vectors coincide at intervals of length h. An upper bound on the KL divergence between the interpolated process D_t and the original continuous process Y_t is obtained by a Girsanov-type change of measure. This KL bound controls the corresponding TV bound by Pinsker's inequality.
To bound the convergence error of the Langevin diffusion, the authors reference well-known results of geometric ergodicity (3) of the Langevin diffusion which implies an exponential convergence to the target distribution. The idea for LMC is then to approximate Y_t by $X_{k,h}$ where $t = kh$.

3.3 Strength : New Results

3.3.1 First Result on Polynomial dependency on d and ε^{-1}

This paper develops the very first theoretical result (Theorem 2) guaranteeing polynomial complexity in the dimension d as well as in inverse precision ε^{-1} for sampling from smooth and log-concave densities.
Let f: $\mathbb{R}^d \to \mathbb{R}$ be a function satisfying assumptions A1 and A2, and θ^* its global minimum point. Assume that for some $\alpha \geq 1$, we have $h \leq 1/(\alpha M)$ and $K \geq \alpha$. Then for any $T = Kh$, the total variation distance furnished by the LMC algorithm with the initial Gaussian distribution $\nu = \mathcal{N}_d(\theta^*, L^{-1}\mathbf{I_d})$ satisfies

$$\left\|\nu \mathbf{P}_X^{K,h} - \mathbf{P}_\pi\right\|_{\mathrm{TV}} \leq \frac{1}{2} \exp\left\{\frac{d}{4}\log\left(\frac{M}{m}\right) - \frac{Tm}{2}\right\} + \left\{\frac{dL^2Th\alpha}{4(2\alpha - 1)}\right\}^{1/2}. \tag{9}$$

Therefore, in order to get an error smaller than ε, it is sufficient to perform $K = O(d/\varepsilon^2)$ evaluations.

3.3.2 Rejecting Prior Beliefs that Metropolis-Hastings is Necessary

Based on the iterative formulation of the LMC algorithm (1), the authors establish a non-explosiveness guarantee on the iterates when $h \leq 1/L$ with an analysis that draws heavily on convex analysis techniques.
This implies that a Metropolis-Hastings correction step is not necessary to avoid transience in the Markov Chain. This allows for much easier parallelization of the unadjusted Langevin algorithm.
However, that the M-H step could still be beneficial for correcting the bias, especially in the case of constant step sizes h which would otherwise lead to a stationary distribution π_{ULA} that is different from π.

3.3.3 Simple and Extensible Analysis Techniques

The decomposition of the error into discretization error and continuous contraction error with the convex help of optimization techniques allows for straightforward extensions. This can be seen in the analysis of the warm-start initialization, the higher order discretization scheme, and the preconditioned LMC algorithm.
This type of analysis was also shown to easily extend to non-strongly log-concave densities although at the cost of deteriorated dependency on the dimension and precision ε.

3.3.4 Practical Guidance for Parameter Tuning and Early Stopping Rule

It is clear that the second error term is a decreasing function of $T = Kh$. On the other hand, the first error term vanishes when h tends to zero. However, for a fixed time-horizon T, a smaller h implies a larger number of iteration K. Therefore, a crucial trade-off between runtime and approximation error for a fixed time-horizon relies on the choice of K and h.
The explicit bounds established in this work (e.g. 9) provide concrete guidelines for choosing h and the stopping rule of the LMC algorithm to achieve a prescribed error rate for a variety of settings.

3.4 Weaknesses

3.4.1 Suitability of the Total Variation Distance

Total variation distance is difficult to compute for continuous measures thus requiring a proxy measure for convergence assessment. TV distance is also known to decay substantially only after a certain amount of time as manifested in the cut-off phenomena. Furthermore, TV does not directly provide the level of approximating the first order moment. This makes it unsuitable for convergence diagnostics.

Finally, the use of Pinsker's inequality in the analysis to upper bound the discretization error with a KL divergence term implies that a tighter bound could be achieved by controlling KL divergence directly. In fact, it was noted in Remark 1 that the bound can be vacuous for large T since the TV error cannot exceed 1 but one of the error terms in the upper bound might shoot to infinity.
Therefore, to quantify the approximation of a distribution by sampling, TV is not the most suitable distance.

3.4.2 Theoretical Gap between Sampling and Optimization

This paper endeavors to bridge the gap between sampling and optimization guarantees. However, the best result, achieved for a warm-start initialization, still leaves a big gap in mixing time: $O(d/\varepsilon^2 \log(d/\varepsilon))$ versus $O(\log(1/\varepsilon))$. While the authors indicate that sampling is intuitively a more difficult task than optimization, they fall short on justifying this huge gap.

3.4.3 Practicality of Warm-Start

The main result in (9) assumes a Gaussian start centered at the minimizer of f with a variance dependent on the smoothness condition. However, this result only guarantees an $O(d^3)$ dimensionality dependency.
The sharpest result, $O(d)$, can only be achieves with a warm-start which requires an initial distribution such that the χ^2 divergence to the target is bounded by a quantity independent of d.
However, the authors do not detail how difficult it is to find such an initialization. In fact, even for sampling from Gaussians, we cannot achieve a warm-start without some information about the variance of the target.
Finally, the authors do not make the dependence on L and m explicit for the warm-start result which further undermines its practical use.

3.4.4 Limitation of Constant Step-sizes

While the analysis of varying step-sizes is bound to be more complicated, the focus on constant step sizes of this work is another major weakness.
In fact, the authors do not emphasize, in enough detail, the asymptotic bias of using a constant step-size.
Furthermore, decaying step-sizes are more common in practice. Not incorporating these in the current manuscript limits the impact of the practical guidance that the authors strive for.

4 High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm (Durmus and Moulines, 2016)

Durmus and Moulines (2016) establish tighter non-asymptotic bounds for LMC under assumptions A1 and A2 in Wasserstein distances of order 2 (W_2) as well as in total variation distance based on a Euler discretization of the SDE (1) with both constant and varying step sizes h.

4.1 Strength : Reconsidering the Problem Setting

4.1.1 A Case for Wasserstein distances

The Wasserstein distance is often viewed as the intrinsic measure of closeness between two distributions. In fact, convergence with respect to W_p is equivalent to weak convergence of measures in addition to convergence of the first p moments. Therefore, bounds on W_p directly provide the level of approximating the first p moments. Consequently, W_2 is correctly deemed by the authors as more suitable for quantifying the quality of approximate sampling schemes than other distances such as the total variation.

4.1.2 Explicit Bounds for Non-increasing Step-sizes

The authors depart from the classical analysis of the discretization with fixed step-sizes and tackle the case of varying step-sizes such that $\lim_{k\to\infty} h_k = 0$ and $\sum_k h_k = \infty$.
The authors then develop explicit bounds for $h_k = h_1 k^{-\alpha}$ with $\alpha \in (0, 1]$ to investigate the different regimes for different choices of α. Finally, the authors establish the optimal choice of α for convergence in W_2 and TV for different smoothness assumptions.

4.1.3 Accurate Analysis of The Effect of the Initial Distribution

Dalalyan (2017b) showed that if the initial distribution is an appropriately chosen Gaussian or if warm-start is used, the number of steps required for an ε-close sample to π is of order $O(d^3\varepsilon^{-2})$ and $O(d\varepsilon^{-2})$ respectively. In this work, the authors more accurately study the impact of the initial distribution on the final precision. In fact, the authors establish various results, some of which match or outperform those reported for warm-start in (Dalalyan, 2017b) without any assumptions on the initial distribution. See Table 1 for a summary.

4.2 Strength : Simple Techniques for Analyzing Convergence in W_2

The authors first establish the geometric convergence of the continuous process, independently of the initial state, by straightforward synchronous coupling of the Langevin SDE solutions in Proposition 1.
The authors then investigate the convergence of the time-inhomogeneous discrete Markov chain $(X_k)_{k\geq 0}$ through the perspective of the step-size-dependent transition kernels R_h and the composition of kernels along the trajectory of the LMC iterates $Q_h^n = R_{h_1} \dots R_{h_n}$.
By the strong convexity and smoothness of f, the authors establish a Foster-Lyapunov condition on the kernel Q (in Proposition 2) which implies a strict contraction in W_2 which, in turn, implies geometric convergence of the sequence of discrete LMC iterates $\delta_x R_h^n$ to π_h in W_2.
Unlike Girsanov change-of-measure techniques for the analysis of the discretization error, this work proceeds by constructing a synchronous coupling between the Langevin diffusion and the linear interpolation of the Euler discretization $(Y_t, \bar{Y}_t)_{t\geq 0}$ defined for all $n \geq 0$ and $t \in [kh, (k+1)h)$:

$$\begin{cases} Y_t = Y_{kh} - \int_{kh}^{t} \nabla f(Y_s)ds + \sqrt{2}(B_t - B_{kh}) & \to dY_t = -\nabla f(Y_t)dt + \sqrt{2}dB_t \\ \bar{Y}_t = \bar{Y}_{kh} - \nabla f(\bar{Y}_{kh})(t-kh) + \sqrt{2}(B_t - B_{kh}) & \to d\bar{Y}_t = -\nabla f(\bar{Y}_{kh})dt + \sqrt{2}dB_t \end{cases} \tag{10}$$

Since π is invariant for P_t for all $t \geq 0$, it suffices to get some bounds on $W_2(\delta_x Q_h^n, \nu_0 P_{kh})$ and take $\nu_0 = \pi$. With $Y_0 \sim \pi$ and $\bar{Y}_0 = x$ we have that

$$W_2(\delta_x Q_h^k, \nu_0 P_{kh}) \leq E[\| Y_{kh} - \bar{Y}_{kh}\|_2] \tag{11}$$

Taking $\nu_0 = \pi$ we derive an explicit bound on the W_2 distance between $(\delta_x Q_h^k)_{k\geq 0}$ and the stationary measure π of the Langevin diffusion.
To improve their bounds, the authors introduce a Lipschitz-continuity assumption on the Hessian of f. Specifically, they assume that the f is three times continuously differentiable and there exists $\widetilde{L}$ such that for all $x, y \in \mathbb{R}^d$,

$$\| \nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq \widetilde{L}\| x - y\|_2 \tag{A3}$$

4.3 Strength : Improved TV Analysis

The manuscript first bounds the total variation distance of $\|\mu P_t - \nu P_t\|_{\text{TV}}$ by reflection coupling (and Lindvall's lemma) to establish the strict contraction of the semigroup P_t and geometric ergodicity of the associated Markov process. The propositions of Theorem 10 then lead to an upper bound on the TV distance of the continuous convergence in terms of W_1 : $\|\mu P_t - \nu P_t\|_{\text{TV}} \leq (4\pi t)^{-1/2} W_1(\mu, \nu)$
However, for the discrete process with non-increasing step-sizes, reflection coupling is no longer applicable. Instead, the authors propose a new coupling construction originally developed for Gaussian random walks. We will not delve into this technique and corresponding partial result, due to space constraints, but we would to emphasize that it seems to offer much tighter control on the discretization error than Girsanov-type change-of-measure inequalities.

4.4 Strength : New Results

The authors establish various explicit results, under different smoothness assumptions (with and without A3), both for a fixed step-size and for a non-increasing sequence of step-sizes $h_k = hk^{-\alpha}$.
The results cover both fixed-precision, where we care about the minimum number of iterations and optimal step-size to achieve a fixed target precision, and fixed-horizon where we care about the minimum distance in W_2 or TV for a fixed number of iterations.

4.4.1 Improved bounds in TV and W_2

Parameter	d	ε	L	m
Dalalyan (2017b) Gaussian Start	d^3	$\varepsilon^{-2}\|\log(\varepsilon)\|$	L^3	$m^{-2}\|\log(m)\|$
Dalalyan (2017b) Warm Start	d	$\varepsilon^{-2}\|\log(\varepsilon)\|$	-	-
Durmus and Moulines (2016)	$d\log(d)$	$\varepsilon^{-2}\|\log(\varepsilon)\|$	L^2	$m^{-3}\|\log(m)\|$
Durmus and Moulines (2016) under A3	$d\log(d)$	$\varepsilon^{-1}\|\log(\varepsilon)\|$	L^2	$m^{-2}\|\log(m)\|$

Table 1: Mixing Time Dependencies

The bounds derived in Theorem 5 hold for any initial distribution in $\mathcal{P}_2(\mathbb{R}^d)$ and match (up to logarithmic terms), with fixed step sizes, the dimension and precision dependency of the best bound derived in (Dalalyan, 2017b) which assumes warm-start for the initial distribution.
Furthermore, the dependency on L and m is made explicit in this result unlike in (Dalalyan, 2017b).
As we can see in Table 1, the result under A3 actually improves on existing bounds in terms of ε dependency.
If we further assume that $\widetilde{L} = 0$ (in (A3)), for the case of sampling from a d-dimensional Gaussian, the authors prove that the new bound $O(d^{1/2}\log(d))$ is sharp.

4.4.2 Explicit Fixed-Horizon Bounds for Varying Step-size: $h_k = h_1 k^{-\alpha}$

For varying step-sizes, the authors identify two regimes for the value of α.
If $\alpha \in (0, 1)$ then the novel explicit results state that the distance between the n^{th} iterate and the stationary distribution is $O((\frac{d\log(n)}{n^\alpha})^{1/2})$ under A1 and A2 or $O(\frac{d\log(n)}{n^\alpha})$ A1, A2, and A3.
If $\alpha = 1$, the authors extend this result to require $h_1 \geq 2\kappa^{-1}$ or $h_1 \geq 4\kappa^{-1}$ depending on the smoothness assumptions. The final bounds are similar to the above with the exception of α being set to 1.
While the analysis of varying step-sizes is quite rare, in this setting, explicit rates for a specific schedule are even rarer which makes this a major contribution of this paper.

4.5 Weaknesses

The upper bound $W_2(\nu_0, \pi)$ on the initial distance can be hard to compute since $\| \theta_0 - \hat{\theta}\|_2$ is not necessarily easy to evaluate. This undermines the practical usefulness of the main result.
Across most experiments, the Metropolis-adjusted Langevin Algorithm (MALA) seems to empirically outperform ULA. Some intuition as to why this is the case would be quite enriching to this work. In fact, the authors even claim that ULA is a substitute for MALA and Polya-Gamma even though the results are negative, in comparison to MALA.
While general results on the mixing time for varying step-sizes is provided in Theorem 5, explicit fixed-precision bounds for $h_k = h_1 k^{-\alpha}$ are surprisingly missing from the main paper. Without these results, it is difficult to assess the improvement such a decay schedule has on the mixing time.
The optimality with respect to sampling from a Gaussian for the special case of $\widetilde{L} = 0$ can mislead the reader into assuming that the bounds are sharp in general (close to a certain lower-bound). However, this specific sampling application is overly simplified and the result is thus far from surprising. Accordingly, we cannot see how this result translates to sharpness in most other scenarios.

5 Further and Stronger Analogy between Sampling and Optimization: Langevin Monte Carlo and Gradient Descent (Dalalyan, 2017a)

This paper revisits recent theoretical guarantees of LMC algorithms for sampling from a smooth and (strongly) log-concave density, which includes the reviewed papers (Dalalyan, 2017b, Durmus and Moulines, 2016).
While similarities to optimization have been exploited for improved theoretical guarantees of LMC, prior work has rarely justified the gap in the convergence rates between LMC and gradient descent (GD).

In this work, the authors establish sharper and simpler bounds on LMC convergence in terms of W_2 distances which look closer to that of GD than the existing LMC bounds.
The authors then provide further insights on the similarities between LMC and GD in an attempt to justify the disappointing gap in convergence rates.
Finally, this paper establishes convergence guarantees for the use of noisy gradients in LMC.

5.1 Strength : Simpler Analysis Techniques

This paper focuses on the Wasserstein distance as an the intrinsic metric for evaluating the closeness of two distributions while citing a few issues of total variation distances.
The convergence analysis in this manuscript does not leverage Girsanov-type change of measure or sophisticated coupling techniques. Instead, the authors introduce a continuous Langevin process Y_t, initialized at equilibrium ($Y_0 \sim \pi$) and driven by the same d-dimensional Brownian motion as the LMC iterates such that $W_{(k+1)h} - W_{kh} = \sqrt(h)\xi^{k+1}$.
The analysis then draws heavily on standard optimization techniques to inductively evaluate the approximation error at any given step k in the LMC procedure.
The authors finally establish a bound on $\mathbb{E}\left[\| Y_{(k+1)h} - X_{k+1,h}\|_2^2\right]$ and note that $Y_{(k+1)h} \sim \pi$ thus concluding that $W_2(\nu_{k+1}, \pi) \leq (\mathbb{E}\left[\| Y_{(k+1)h} - X_{k+1,h}\|_2^2\right])^{1/2}$.

5.2 Strength : Tighter and Simpler W_2 bounds

The authors demonstrate that, for LMC, if $h \leq 2/L$ then for $\rho = (1 - mh) \vee (Lh - 1)$

$$W_2(\nu_K, \pi) \leq \rho^K W_2(\nu_0, \pi) + \frac{Lh}{1-\rho}(5hd/3)^{1/2} \tag{12}$$

This bound holds under weaker conditions of $h \leq 2/M$ instead of $h \leq 1/(m + M)$ (the standard in prior work). It has a simpler remainder term than those appearing in the W_2 bounds of (Durmus and Moulines, 2016) and seems sharper, in terms of constants, than existing bounds as demonstrated in Figure 1.
In fact, this paper does a great job showcasing the improved sharpness in terms of the constants by plotting the minimum number of iterations for an ε-accurate sample over a wide range of dimensions d.
Finally, let's recall the standard result for the convergence analysis of gradient descent; if $h \leq 2/L$ then for $\rho = (1 - mh) \vee (Lh - 1)$ we have $\| X_K - X^*\|_2^2 \leq \rho^K \| X_0 - X^*\|_2^2$
Accordingly, one major contribution of this work is bridging the gap between the optimization algorithm (GD) and the sampling algorithm (LMC) as the tighter W_2 (12) closely mirrors that of GD and holds under the same conditions (on h) and for the same ρ constant.
Another contribution to the W_2 bounds is an easier-to-compute upper bound on the initial distance $W_2(\nu_0, \pi)^2$ which prior work such as (Durmus and Moulines, 2016) suggested to control with $\| \theta^0 - \hat{\theta}\|_2^2 + d/m$. On the other hand, this paper argues that it's often difficult to evaluate $\| \theta^0 - \hat{\theta}\|_2$ in practice and suggests an alternative based on the convexity of f as $W_2(\nu_0, \pi)^2 \leq \frac{2}{m}(f(\theta_0) - \int f(\theta) d\pi(\theta) + d)$.

5.3 Strength : Explaining the Gap between Sampling and Optimization

The authors attempt to explain the remaining gap between the optimization guarantees and the sampling guarantees, especially in terms of dimensionality dependence: $O(\log(1/\varepsilon))$ versus $O(d/\varepsilon^2 \log(d/\varepsilon))$.
First note that $f_\tau(\theta) = f(\theta)/\tau$ has the same optimum θ^* for any $\tau > 0$ whereas the density function $\pi_\tau \propto exp(-f_\tau(\theta))$ defines a different distribution for each choice of τ. This is standard intuition as to why optimization is a simpler task than sampling.
However, further note that the average value $\theta_\tau = \int \theta \pi_\tau(\theta) d\theta$ tends to θ^* as τ tends to 0. Simultaneously, $\pi_\tau(d\theta)$ tends to a Dirac measure at θ^*. Accordingly, the authors propose to analyze the limiting behavior of LMC as τ tends to 0 to analyze its convergence to θ^*. This should shed some light on the convergence of LMC in comparison to that GD when the target is the same.
The authors show that the resulting LMC updates (originally in 2) indeed mirrors those of GD. More interestingly, the limiting case of the W_2 error bound (12) is equivalent to the L_2 error bound for GD.

5.4 Strength : Noisy Gradient Analysis

LMC can be extended to the mini-batch setting by substituting the gradient with sub-sampled gradients.
To analyze this common setting, the authors assume independent zero-mean random vectors with a variance bounded by d. The noisy gradient can then be expressed as $Y_{k,h} = \nabla f(X_{k,h}) + \sigma\zeta^k$.
The authors find that using the sub-sampled gradient in the LMC algorithm does not cause a significant deterioration of the precision (an extra term $\sigma^2 \frac{(L+m)^2}{(L-m)^2}$) but considerably reduces the computational burden.

5.5 Weaknesses

One minor issue might be the potentially misleading conclusion that the convergence analysis of LMC is a natural counterpart to that of gradient descent. The finding that the LMC iterates converge to those of GD when the temperature $\tau \to 0$ is not surprising. As for the convergence analysis, taking τ to 0 would have eliminated the remainder term, no matter the magnitude of the constants in the term. Accordingly, we cannot claim that the bound is optimal just due to the fact that its limit when $\tau \to 0$ matches the GD bound.

6 Future Directions

The update rule of the LMC follows from replacing the gradient in the Langevin diffusion by its piecewise constant approximation. Therefore, the behavior of LMC is governed by two characteristics of the continuous-time process: the mixing rate and the smoothness of the sample paths.
For LMC, we know that the Langevin diffusion mixes exponentially fast with the precise rate $e^{-\kappa ht}$. In addition, all sample paths of Y are α-Holder-continuous. Combining these two properties, it has been shown that it suffices $O(d\varepsilon^{-2})$ iterations for the LMC algorithm to achieve an error smaller than ε.
However, several fundamental questions arise from this understanding.

6.1 Lower Bounds and Acceleration in MCMC

Can we can improve the dependence on the condition number, precision, and dimension, given the observed gap to optimization convergence guarantees?
However, without complexity lower bounds for MCMC, it is difficult to discern which inefficiency in the bound is due to an proof artifacts and which cannot be improved on, for a given class of sampling algorithms. Unfortunately, lower bounds in MCMC are largely unknown. Accordingly, it is difficult to estimate the gap between existing algorithms and optimal achievable rates. It is also difficult to discern when inefficiencies in the bounds are due to an artifact in the proof.
Lower bounds are well-known for convex optimization and have helped guide the design of accelerated algorithms such as Nesterov's Accelerated Gradient Descent.
This brings us to the next open question; is there an equivalent to Nesterov acceleration in sampling? We already know of accelerated or higher order variants of the Langevin dynamics such as the underdamped Langevin dynamics (equivalent to Hamiltonian dynamics under certain assumptions). The discretization of such algorithms has demonstrated faster mixing than LMC. However, the the achieved by underdamped Langevin dynamics $O(d^{1/2})$ are still worse than that proven for Hamiltonian Monte Carlo (HMC) $O(d^{1/4})$ which mostly only differs in its discretization scheme. In fact, while Euler discretization is a first-order method, the leapfrog integration of HMC is a second order integration method that can be applied multiple times per gradient evaluation. Accordingly, it is clear that the study of accelerated samplers and, if possible, optimal samplers, will require a combination of optimal diffusions and optimal discretization schemes.
Alternatively, one could investigate a variational (Lagrangian) formulation which would apply the acceleration principle directly in the space of measures.

6.2 Beyond Log-Concavity

Another interesting direction would be to explore if similar guarantees could be derived for sampling efficiently from non-log-concave distributions. Recent work has attempted to relax both the smoothness and the strong

log-concavity assumptions. This has resulted in a significant deterioration in the convergence rates such as exponential dependency on the dimension. On the other hand, it is well know that the Langevin diffusion is geometrically ergodic under weaker assumptions than strong log-concavity, such as the existence of a Log-Sobolev inequality. Therefore, it should be possible to investigate polynomial dependence on the dimension, with a tighter control on the discretization error, simply under smoothness and log-Sobolev assumptions.
One of the most challenging non-log-concave settings, however, is the multimodal setting. In fact, it is known that Langevin dynamics might take exponentially long to move from one mode to another as the gradient only captures local information about the nearest mode. Empirical progress has been made by exploring various schedules for the step-size to enable the crossing of wide valleys between modes. However, most of these approaches are ad-hoc such that they fail for certain general multimodal settings.
Furthermore, Langevin dynamics have been empirically shown to mix extremely slowly for sub-exponential distributions which are common for shrinkage priors such as the Cauchy and the Pareto distributions. However, little theoretical work has explicitly analyzed the mixing time for such target distributions.

6.3 Borrowing Ideas from Optimization

Another interesting direction is to borrow ideas from optimization that might accelerate or robustify existing samplers. For example, there has been a growing body of work on outlier-robustness for black-box stochastic optimization. This sort of approach can be considered orthogonally to the use of heavy-tail priors in order to detect outliers without suffering the slower mixing.
Another interesting perspective from optimization is black-box or universal algorithms that do not require knowledge of the smoothness parameter L. In fact, the optimal step-size is often a function of the strong convexity and smoothness constants m and L. However, it has been shown that it is easy to estimate m by backtracking whereas L remains hard to approximate which can lead to deterioration in the convergence speed. Accordingly, there has been much work on devising universal optimization algorithms that can estimate the smoothness and strong convexity parameters, on the fly, and adapt them across the state space. However, there has been no work on such universal and adaptive algorithms for sampling.
Several low-hanging fruit ideas from optimization that might be worth exploring include higher-order methods such as Newton's, conjugate gradients, adaptive preconditioning, and dual methods such as mirror descent.

6.4 Particle Methods

Interacting particle methods have found practical success thanks to the variance-reducing effect of the interaction term. One such recent method is Stein Variational Gradient Descent which has been observed to simulate gradient flow, just like Langevin diffusions, but on a different kernelized space of measures. Extending our analysis of diffusions to those involving (finitely-many) interacting particles could be a groundbreaking direction given the popularity of these methods in computational physics, statistics, machine learning, and molecular dynamics. However, there are still no known theoretical guarantees that explicitly quantify the convergence rate of such algorithms.

References

A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689, 2017a.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017b.

A. Durmus and E. Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.