

---

# Exponential ergodicity of mirror-Langevin diffusions

---

**Sinho Chewi**  
MIT  
schewi@mit.edu

**Thibaut Le Gouic**  
MIT  
tlegouic@mit.edu

**Chen Lu**  
MIT  
chenl819@mit.edu

**Tyler Maunu**  
MIT  
maunut@mit.edu

**Philippe Rigollet**  
MIT  
rigollet@mit.edu

**Austin Stromme**  
MIT  
astromme@mit.edu

## Abstract

Motivated by the problem of sampling from ill-conditioned log-concave distributions, we give a clean non-asymptotic convergence analysis of mirror-Langevin diffusions as introduced in [Zha+20]. As a special case of this framework, we propose a class of diffusions called Newton-Langevin diffusions and prove that they converge to stationarity exponentially fast with a rate which not only is dimension-free, but also has no dependence on the target distribution. We give an application of this result to the problem of sampling from the uniform distribution on a convex body using a strategy inspired by interior-point methods. Our general approach follows the recent trend of linking sampling and optimization and highlights the role of the chi-squared divergence. In particular, it yields new results on the convergence of the vanilla Langevin diffusion in Wasserstein distance.

## 1 Introduction

Sampling from a target distribution is a central task in statistics and machine learning with applications ranging from Bayesian inference [RC04; DM+19] to deep generative models [Goo+14]. Owing to a firm mathematical grounding in the theory of Markov processes [MT09], as well as its great versatility, Markov Chain Monte Carlo (MCMC) has emerged as a fundamental sampling paradigm. While traditional theoretical analyses are anchored in the asymptotic framework of ergodic theory, this work focuses on finite-time results that better witness the practical performance of MCMC for high-dimensional problems arising in machine learning.

This perspective parallels an earlier phenomenon in the much better understood field of optimization where convexity has played a preponderant role for both theoretical and methodological advances [Nes04; Bub15]. In fact, sampling and optimization share deep conceptual connections that have contributed to a renewed understanding of the theoretical properties of sampling algorithms [Dal17a; Wib18] building on the seminal work of Jordan, Kinderlehrer and Otto [JKO98].

We consider the following canonical sampling problem. Let  $\pi$  be a log-concave probability measure over  $\mathbb{R}^d$  so that  $\pi$  has density equal to  $e^{-V}$ , where the potential  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex. Throughout this paper, we also assume that  $V$  is twice continuously differentiable for convenience, though many of our results hold under weaker conditions.

Most MCMC algorithms designed for this problem are based on the *Langevin diffusion* (LD), that is the solution  $(X_t)_{t \geq 0}$  to the stochastic differential equation (SDE)

$$dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t, \tag{LD}$$

with  $(B_t)_{t \geq 0}$  a standard Brownian motion in  $\mathbb{R}^d$ . Indeed,  $\pi$  is the unique invariant distribution of (LD) and suitable discretizations result in algorithms that can be implemented when  $V$  is known only up to an additive constant, which is crucial for applications in Bayesian statistics and machine learning.

A first connection between sampling from log-concave measures and optimizing convex functions is easily seen from (LD): omitting the Brownian motion term yields the gradient flow  $\dot{x}_t = -\nabla V(x_t)$ , which results in the celebrated gradient descent algorithm when discretized in time [Dal17a; Dal17b]. There is, however, a much deeper connection involving the distribution of  $X_t$  rather than  $X_t$  itself, and this latter connection has been substantially more fruitful: the marginal distribution of a Langevin diffusion process  $(X_t)_{t \geq 0}$  evolves according to a *gradient flow*, over the Wasserstein space of probability measures, that minimizes the Kullback-Leibler (KL) divergence  $D_{\text{KL}}(\cdot \| \pi)$  [JKO98; AGS08; Vil09]. This point of view has led not only to a better theoretical understanding of the Langevin diffusion [Ber18; CB18; Wib18; DMM19; VW19] but it has also inspired new sampling algorithms based on classical optimization algorithms, such as proximal/splitting methods [Ber18; Wib18; Wib19; SKL20], mirror descent [Hsi+18; Zha+20], Nesterov’s accelerated gradient descent [Che+18; Ma+19; DR20], and Newton methods [Mar+12; Sim+16; WL20].

**Our contributions.** This paper further exploits the optimization perspective on sampling by establishing a theoretical framework for a large class of stochastic processes called *mirror-Langevin diffusions* (MLD) introduced in [Zha+20]. These processes correspond to alternative optimization schemes that minimize the KL divergence over the Wasserstein space by changing its geometry. They show better dependence in key parameters such as the condition number and the dimension.

Our theoretical analysis is streamlined by a technical device which is unexpected at first glance, yet proves to be elegant and effective: we track the progress of these schemes not by measuring the objective function itself, the KL divergence, but rather by measuring the chi-squared divergence to the target distribution  $\pi$  as a surrogate. This perspective highlights the central role of mirror Poincaré inequalities (MP) as sufficient conditions for exponentially fast convergence of the mirror-Langevin diffusion to stationarity in chi-squared divergence, which readily yields convergence in other well-known information divergences, such as the Kullback-Leibler divergence, the Hellinger distance, and the total variation distance [Tsy09, §2.4].

We also specialize our results to the case when the mirror map equals the potential  $V$ . This can be understood as the sampling analogue of Newton’s method, and we therefore call it the *Newton-Langevin diffusion* (NLD). In this case, the mirror Poincaré inequality translates into the Brascamp-Lieb inequality which automatically holds when  $V$  is twice-differentiable and *strictly* convex. In turn, it readily implies exponential convergence of the Newton-Langevin diffusion (Corollary 1) and can be used for approximate sampling even when the second derivative of  $V$  vanishes (Corollary 2). Strikingly, the rate of convergence *has no dependence on  $\pi$  or on the dimension  $d$*  and, in particular, is robust to cases where  $\nabla^2 V$  is arbitrarily close to zero. This *scale-invariant* convergence parallels that of Newton’s method in convex optimization and is the first result of this kind for sampling.

This invariance property is useful for approximately sampling from the uniform distribution over a convex body  $\mathcal{C}$ , which has been well-studied in the computer science literature [FKP94; KLS95; LV07]. By taking the target distribution  $\pi \propto \exp(-\beta V)$ , where  $V$  is any strictly convex *barrier function*, and  $\beta$ , the inverse temperature parameter, is taken to be small (depending on the target accuracy), we can use the Newton-Langevin diffusion, much in the spirit of interior point methods (as promoted by [LTV20]), to output a sample which is approximately uniformly distributed on  $\mathcal{C}$ ; see Corollary 3.

Throughout this paper, we work exclusively in the setting of continuous-time diffusions such as (LD). We refer to the works [DM15; Dal17a; Dal17b; RRT17; CB18; Wib18; DK19; DMM19; DRK19; Mou+19; VW19] for discretization error bounds, and leave this question open for future works.

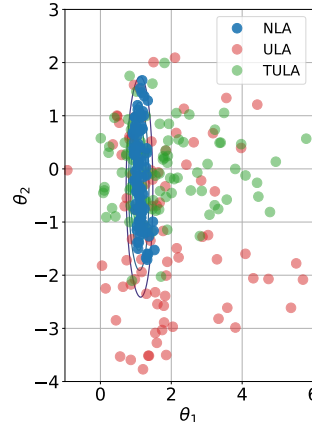


Figure 1: Samples from the posterior distribution of a 2D Bayesian logistic regression model using the Newton-Langevin Algorithm (NLA), the Unadjusted Langevin Algorithm (ULA), and the Tamed Unadjusted Langevin Algorithm (TULA) [Bro+19]. For details, see Section E.2.

**Related work.** The discretized Langevin algorithm, and the Metropolis-Hastings adjusted version, have been well-studied when used to sample from strongly log-concave distributions, or distributions satisfying a log-Sobolev inequality [Dal17b; DM17; CB18; Che+19; DK19; DM+19; Dwi+19; Mou+19; VW19]. Moreover, various ways of adapting Langevin diffusion to sample from bounded domains have been proposed [BEL18; Hsi+18; Zha+20]; in particular, [Zha+20] studied the discretized mirror-Langevin diffusion. Finally, we note that while our analysis and methods are inspired by the optimization perspective on sampling, it connects to a more traditional analysis based on coupling stochastic processes. Quantitative analysis of the continuous Langevin diffusion process associated to SDE (LD) has been performed with Poincaré and log-Sobolev inequalities [BGG12; BGL14; VW19], and with couplings of stochastic processes [CL89; Ebe16].

**Notation.** The Euclidean norm over  $\mathbb{R}^d$  is denoted by  $\|\cdot\|$ . Throughout, we simply write  $\int g$  to denote the integral with respect to the Lebesgue measure:  $\int g(x) dx$ . When the integral is with respect to a different measure  $\mu$ , we explicitly write  $\int g d\mu$ . The expectation and variance of  $g(X)$  when  $X \sim \mu$  are respectively denoted  $\mathbb{E}_\mu g = \int g d\mu$  and  $\text{var}_\mu g := \int (g - \mathbb{E}_\mu g)^2 d\mu$ . When clear from context, we sometimes abuse notation by identifying a measure  $\mu$  with its Lebesgue density.

## 2 Mirror-Langevin diffusions

Before introducing mirror-Langevin diffusions, our main objects of interest, we provide some intuition for their construction by drawing a parallel with convex optimization.

### 2.1 Gradient flows, mirror flows, and Newton’s method

We briefly recall some background on gradient flows and mirror flows; we refer readers to the monograph [Bub15] for the convergence analysis of the corresponding discrete-time algorithms.

Suppose we want to minimize a differentiable function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ . The *gradient flow* of  $f$  is the curve  $(x_t)_{t \geq 0}$  on  $\mathbb{R}^d$  solving  $\dot{x}_t = -\nabla f(x_t)$ . A suitable time discretization of this curve yields the well-known *gradient descent* (GD).

Although the gradient flow typically works well for optimization over Euclidean spaces, it may suffer from poor dimension scaling in more general cases such as Banach space optimization; a notable example is the case when  $f$  is defined over the probability simplex equipped with the  $\ell_1$  norm. This observation led Nemirovskii and Yudin [NJ79] to introduce the *mirror flow*, which is defined as follows. Let  $\phi : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  be a *mirror map*, that is a strictly convex twice continuously differentiable function of *Legendre type*<sup>1</sup>. The mirror flow  $(x_t)_{t \geq 0}$  satisfies  $\partial_t \nabla \phi(x_t) = -\nabla f(x_t)$ , or equivalently,  $\dot{x}_t = -[\nabla^2 \phi(x_t)]^{-1} \nabla f(x_t)$ . The corresponding discrete-time algorithms, called *mirror descent* (MD) algorithms, have been successfully employed in varied tasks of machine learning [Bub15] and online optimization [BC12] where the entropic mirror map plays an important role. In this work, we are primarily concerned with the following choices for the mirror map:

1. When  $\phi = \|\cdot\|^2/2$ , then the mirror flow reduces to the gradient flow.
2. Taking  $\phi = f$  and the discretization  $x_{k+1} = x_k - h_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$  yields another popular optimization algorithm known as (damped) *Newton’s method*. Newton’s method has the important property of being invariant under affine transformations of the problem, and its local convergence is known to be much faster than that of GD; see [Bub15, §5.3].

### 2.2 Mirror-Langevin diffusions

We now introduce the *mirror-Langevin diffusion* (MLD) of [Zha+20]. Just as LD corresponds to the gradient flow, the MLD is the sampling analogue of the mirror flow. To describe it, let  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$  be a mirror map as in the previous section. Then, the mirror-Langevin diffusion satisfies the SDE

$$X_t = \nabla \phi^*(Y_t), \quad dY_t = -\nabla V(X_t) dt + \sqrt{2} [\nabla^2 \phi(X_t)]^{1/2} dB_t, \quad (\text{MLD})$$

<sup>1</sup>This ensures that  $\nabla \phi$  is invertible, c.f. [Roc97, §26].

where  $\phi^*$  denotes the convex conjugate of  $\phi$  [BL06, §3.3]. In particular, if we choose the mirror map  $\phi$  to equal the potential  $V$ , then we arrive at a sampling analogue of *Newton’s method*, which we call the *Newton-Langevin diffusion* (NLD),

$$X_t = \nabla V^*(Y_t), \quad dY_t = -\nabla V(X_t) dt + \sqrt{2} [\nabla^2 V(X_t)]^{1/2} dB_t. \quad (\text{NLD})$$

From our intuition gained from optimization, we expect that NLD has special properties, such as affine invariance and faster convergence. We validate this intuition in Corollary 1 below by showing that, provided  $\pi$  is strictly log-concave, the NLD converges to stationarity exponentially fast, with no dependence on  $\pi$ . This should be contrasted with the vanilla Langevin diffusion (LD), for which the convergence rate depends on the Poincaré constant of  $\pi$ , as we discuss in the next section.

We end this section by comparing MLD and NLD with similar sampling algorithms proposed in the literature inspired by mirror descent and Newton’s method.

*Mirrored Langevin dynamics.* A variant of MLD, called “mirrored Langevin dynamics”, was introduced in [Hsi+18]. The mirrored Langevin dynamics is motivated by constrained sampling and corresponds to the vanilla Langevin algorithm applied to the new target measure  $(\nabla\phi)_{\#}\pi$ . In contrast, MLD can be understood as a Riemannian diffusion w.r.t. the Riemannian metric induced by the mirror map  $\phi$ . Thus, the motivations and properties of the two algorithms are different, and we refer to [Zha+20] for further comparison of the two algorithms.

An earlier draft of [Hsi+18] also introduced MLD, along with a continuous-time analysis of the diffusion. Their convergence analysis is based on the classical Bakry-Émery criterion (see [BGL14]), which is generally harder to check than the mirror Poincaré inequality (MP) that we introduce below; in particular, when  $\phi = V$ , we show that the mirror Poincaré inequality holds automatically.

*Quasi-Newton diffusion.* The paper [Sim+16] proposes a quasi-Newton sampling algorithm, based on L-BFGS, which is partly motivated by the desire to avoid computation of the third derivative  $\nabla^3 V$  while implementing the Newton-Langevin diffusion. We remark, however, that the form of NLD employed above, which treats  $V$  as a mirror map, does not in fact require the computation of  $\nabla^3 V$ , and thus can be implemented practically; see Section 5. Moreover, since we analyze the full NLD, rather than a quasi-Newton implementation, we are able to give a clean convergence result.

*Information Newton’s flow.* Inspired by the perspective of [JKO98], which views the Langevin diffusion as a gradient flow in the Wasserstein space of probability measures, the paper [WL20] proposes an approach termed “information Newton’s flow” that applies Newton’s method directly on the space of probability measures equipped with either the Fisher-Rao or the Wasserstein metric. However, unlike LD and NLD that both operate at the level of particles, information Newton’s flow faces significant challenges at the level of both implementation and analysis.

### 3 Convergence analysis

#### 3.1 Convergence of gradient flows and mirror flows

We provide a brief reminder about the convergence analysis of gradient flows and mirror flows defined in Section 2.1 to provide intuition for the next section. Throughout, let  $f$  be a differentiable function with minimizer  $x^*$ .

Consider first the gradient flow for  $f$ :  $\dot{x}_t = -\nabla f(x_t)$ . We get  $\partial_t[f(x_t) - f(x^*)] = -\|\nabla f(x_t)\|^2$  from a straightforward computation. From this identity, it is natural to assume a *Polyak-Łojasiewicz* (PL) *inequality*, which is well-known in the optimization literature [KNS16] and can be employed even when  $f$  is not convex [Che+20]. Indeed, if there exists a constant  $C_{\text{PL}} > 0$  with

$$f(x) - f(x^*) \leq \frac{C_{\text{PL}}}{2} \|\nabla f(x)\|^2 \quad \forall x \in \mathbb{R}^d, \quad (\text{PL})$$

then  $\partial_t[f(x_t) - f(x^*)] \leq -\frac{2}{C_{\text{PL}}}[f(x_t) - f(x^*)]$ . Together with Grönwall’s inequality, it readily yields exponentially fast convergence in objective value:  $f(x_t) \leq f(x_0) e^{-2t/C_{\text{PL}}}$ .

A similar analysis may be carried out for the mirror flow. Fix a mirror map  $\phi$  and consider the mirror flow:  $\dot{x}_t = -[\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)$ . It holds  $\partial_t[f(x_t) - f(x^*)] = -\langle \nabla f(x_t), [\nabla^2\phi(x_t)]^{-1}\nabla f(x_t) \rangle$ .

Therefore, the analogue of (PL) which guarantees exponential decay in the objective value is the following inequality, which we call a *mirror PL inequality*:

$$f(x) - f(x^*) \leq \frac{C_{\text{MPL}}}{2} \langle \nabla f(x), [\nabla^2 \phi(x)]^{-1} \nabla f(x) \rangle \quad \forall x \in \mathbb{R}^d. \quad (\text{MPL})$$

Next, we describe analogues of (PL) and (MPL) that guarantee convergence of LD and MLD.

### 3.2 Convergence of mirror-Langevin diffusions

The above analysis employs the objective function  $f$  to measure the progress of both the gradient and mirror flows. While this is the most natural choice, our approach below crucially relies on measuring progress via a *different functional*  $F$ . What should we use as  $F$ ? To answer this question, we first consider the simpler case of the vanilla Langevin diffusion (LD), which is a special case of MLD when the mirror map is  $\phi = \|\cdot\|^2/2$ . We keep this discussion informal and postpone rigorous arguments to Appendix A.

Since the work of [JKO98], it has been known that the marginal distribution  $\mu_t$  at time  $t \geq 0$  of LD evolves according to the *gradient flow* of the KL divergence  $D_{\text{KL}}(\cdot \parallel \pi)$  with respect to the 2-Wasserstein distance  $W_2$ ; we refer readers to [San17] for an overview of this work, and to [AGS08; Vil09] for comprehensive treatments. Therefore, the most natural choice for  $F$  is, as in Section 3.1, the objective function  $D_{\text{KL}}(\cdot \parallel \pi)$  itself. Following this approach, one can compute [Vil03, §9.1.5]

$$\partial_t D_{\text{KL}}(\mu_t \parallel \pi) = - \int \|\nabla \ln \frac{d\mu_t}{d\pi}\|^2 d\mu_t = -4 \int \|\nabla \sqrt{\frac{d\mu_t}{d\pi}}\|^2 d\pi.$$

In this setup, the role of the PL inequality (PL) is played by a *log-Sobolev inequality* of the form

$$\text{ent}_\pi(g^2) := \int g^2 \ln(g^2) d\pi - \left( \int g^2 d\pi \right) \ln \left( \int g^2 d\pi \right) \leq 2C_{\text{LSI}} \int \|\nabla g\|^2 d\pi. \quad (\text{LSI})$$

When  $g = \sqrt{d\mu_t/d\pi}$ , (LSI) reads  $D_{\text{KL}}(\mu_t \parallel \pi) \leq 2C_{\text{LSI}} \int \|\nabla \sqrt{d\mu_t/d\pi}\|^2 d\pi$ , which implies exponentially fast convergence:  $D_{\text{KL}}(\mu_t \parallel \pi) \leq D_{\text{KL}}(\mu_0 \parallel \pi) e^{-2t/C_{\text{LSI}}}$  by Grönwall's inequality.

A disadvantage of this approach, however, is that the log-Sobolev inequality (LSI) does not hold for any log-concave measure  $\pi$ , or it may hold with a poor constant  $C_{\text{LSI}}$ . For example, it is known that the log-Sobolev constant of an isotropic log-concave distribution must in general depend on the diameter of its support [LV18]. In contrast, we work below with a *Poincaré inequality*, which is conjecturally satisfied by such distributions with a *universal constant* [KLS95].

Motivated by [BCG08; CG09], we instead consider the *chi-squared divergence*

$$F(\mu) = \chi^2(\mu \parallel \pi) := \text{var}_\pi \frac{d\mu}{d\pi} = \int \left( \frac{d\mu}{d\pi} \right)^2 d\pi - 1, \quad \text{if } \mu \ll \pi,$$

and  $F(\mu) = \infty$  otherwise. It is well-known that the law  $(\mu_t)_{t \geq 0}$  of LD satisfies the Fokker-Planck equation in the weak sense [KS91, §5.7]:

$$\partial_t \mu_t = \text{div}(\mu_t \nabla \ln \frac{\mu_t}{\pi}).$$

Using this, we can compute the derivative of the chi-squared divergence:

$$\frac{1}{2} \partial_t F(\mu_t) = \int \frac{\mu_t}{\pi} \partial_t \mu_t = \int \frac{\mu_t}{\pi} \text{div}(\mu_t \nabla \ln \frac{\mu_t}{\pi}) = - \int \langle \nabla \ln \frac{\mu_t}{\pi}, \nabla \frac{\mu_t}{\pi} \rangle \mu_t = - \int \|\nabla \frac{\mu_t}{\pi}\|^2 \pi,$$

and exponential convergence of the chi-squared divergence follows if  $\pi$  satisfies a Poincaré inequality:

$$\text{var}_\pi g \leq C_{\text{P}} \mathbb{E}_\pi[\|\nabla g\|^2] \quad \text{for all locally Lipschitz } g \in L^2(\pi). \quad (\text{P})$$

Thus, when using the chi-squared divergence to track progress, the role of the PL inequality is played by a Poincaré inequality. As we discuss in Sections 4.1 and 4.3 below, the Poincaré inequality is significantly weaker than the log-Sobolev inequality.

A similar analysis may be carried out for MLD using an appropriate variation of Poincaré inequalities.

**Definition 1** (Mirror Poincaré inequality). Given a mirror map  $\phi$ , we say that the distribution  $\pi$  satisfies a *mirror Poincaré inequality* with constant  $C_{\text{MP}}$  if

$$\text{var}_\pi g \leq C_{\text{MP}} \mathbb{E}_\pi \langle \nabla g, (\nabla^2 \phi)^{-1} \nabla g \rangle \quad \text{for all locally Lipschitz } g \in L^2(\pi). \quad (\text{MP})$$

When  $\phi = \|\cdot\|^2/2$ , (MP) is simply called a *Poincaré inequality* and the smallest  $C_{\text{MP}}$  for which the inequality holds is the *Poincaré constant* of  $\pi$ , denoted  $C_{\text{P}}$ .

Using a similar argument as the one above, we show exponential convergence of MLD in  $\chi^2(\cdot \parallel \pi)$  under (MP). Together with standard comparison inequalities between information divergences [Tsy09, §2.4], it implies exponential convergence in a variety of commonly used divergences, including the total variation (TV) distance  $\|\cdot - \pi\|_{\text{TV}}$ , the Hellinger distance  $H(\cdot, \pi)$ , and the KL divergence  $D_{\text{KL}}(\cdot \parallel \pi)$ .

**Theorem 1.** For each  $t \geq 0$ , let  $\mu_t$  be the marginal distribution of MLD with target distribution  $\pi$  at time  $t$ . Then if  $\pi$  satisfies the mirror Poincaré inequality (MP) with constant  $C_{\text{MP}}$ , it holds

$$2\|\mu_t - \pi\|_{\text{TV}}^2, H^2(\mu_t, \pi), D_{\text{KL}}(\mu_t \parallel \pi), \chi^2(\mu_t \parallel \pi) \leq e^{-\frac{2t}{C_{\text{MP}}}} \chi^2(\mu_0 \parallel \pi), \quad \forall t \geq 0.$$

We give two proofs of this result in Appendix A.

Recall that LD can be understood as a gradient flow for the KL divergence on the 2-Wasserstein space. In light of this interpretation, the above bound for the KL divergence yields a convergence rate *in objective value*, and it is natural to wonder whether a similar rate holds for the iterates themselves in terms of 2-Wasserstein distance. From the works [Din15; Led18; Liu20], it is known that a Poincaré inequality (P) implies the transportation-cost inequality

$$W_2^2(\mu, \pi) \leq 2C_{\text{P}} \chi^2(\mu \parallel \pi), \quad \forall \mu \ll \pi. \quad (1)$$

Initially unaware of these works, we proved that a Poincaré inequality implies a suboptimal chi-squared transportation inequality. Since the suboptimal inequality already suffices for our purposes, we state and prove it in Appendix B. We thank Jon Niles-Weed for bringing this to our attention.

The inequality (1) implies that if  $\pi$  has a finite Poincaré constant  $C_{\text{P}}$  then Theorem 1 also yields exponential convergence in Wasserstein distance. In the rest of the paper, we write this result as

$$\frac{1}{2C_{\text{P}}} W_2^2(\mu_t, \pi) \leq e^{-\frac{2t}{C_{\text{MP}}}} \chi^2(\mu_0 \parallel \pi),$$

for *any* target measure  $\pi$  that satisfies a mirror Poincaré inequality, with the convention that  $C_{\text{P}} = \infty$  when  $\pi$  fails to satisfy a Poincaré inequality. In this case, the above inequality is simply vacuous.

## 4 Applications

We specialize Theorem 1 to the following important applications.

### 4.1 Newton-Langevin diffusion

For NLD, we assume that  $V$  is strictly convex and twice continuously differentiable; take  $\phi = V$ . In this case, the mirror Poincaré inequality (MP) reduces to the *Brascamp-Lieb inequality*, which is known to hold with constant  $C_{\text{MP}} = 1$  for any strictly log-concave distribution  $\pi$  [BL76; BL00; Gen08]. It yields the following remarkable result where the exponential contraction rate has no dependence on  $\pi$  nor on the dimension  $d$ .

**Corollary 1.** Suppose that  $V$  is strictly convex and twice continuously differentiable. Then, the law  $(\mu_t)_{t \geq 0}$  of NLD satisfies

$$2\|\mu_t - \pi\|_{\text{TV}}^2, H^2(\mu_t, \pi), D_{\text{KL}}(\mu_t \parallel \pi), \chi^2(\mu_t \parallel \mu), \frac{1}{2C_{\text{P}}} W_2^2(\mu_t, \pi) \leq e^{-2t} \chi^2(\mu_0 \parallel \pi).$$



If  $\pi$  is log-concave, then it satisfies a Poincaré inequality [AB15; LV17] so that the result in Wasserstein distance holds. In fact, contingent on the famous *Kannan-Lovász-Simonovitz* (KLS) conjecture ([KLS95]), the Poincaré constant of any log-concave distribution  $\pi$  is upper bounded by a constant, independent of the dimension, times the largest eigenvalue of the covariance matrix of  $\pi$ .

At this point, one may wonder, under the same assumptions as the Brascamp-Lieb inequality, whether a mirror version of the log-Sobolev inequality (LSI) holds. This question was answered negatively in [BL00], thus reinforcing our use of the chi-squared divergence as a surrogate for the KL divergence.

If the potential  $V$  is convex, but degenerate (i.e., not strictly convex) we cannot use NLD directly with  $\pi$  as the target distribution. Instead, we perturb  $\pi$  slightly to a new measure  $\pi_\beta$ , which is strongly log-concave, and for which we can use NLD. Crucially, due to the scale invariance of NLD, the time it takes for NLD to mix does not depend on  $\beta$ , the parameter which governs the approximation error.

**Corollary 2.** Fix a target accuracy  $\varepsilon > 0$ . Suppose  $\pi = e^{-V}$  is log-concave and set  $\pi_\beta \propto e^{-V-\beta\|\cdot\|^2}$ , where  $\beta \leq \varepsilon^2/(2 \int \|\cdot\|^2 d\pi)$ . Then, the law  $(\mu_t)_{t \geq 0}$  of NLD with target distribution  $\pi_\beta$  satisfies  $\|\mu_t - \pi\|_{\text{TV}} \leq \varepsilon$  by time  $t = \frac{1}{2} \ln[2\chi^2(\mu_0 \parallel \pi_\beta)] + \ln(1/\varepsilon)$ .

*Proof.* From our assumption, it holds

$$D_{\text{KL}}(\pi \parallel \pi_\beta) = \int \ln \frac{d\pi}{d\pi_\beta} d\pi = \beta \int \|\cdot\|^2 d\pi + \ln \int e^{-\beta\|\cdot\|^2} d\pi \leq \beta \int \|\cdot\|^2 d\pi \leq \frac{\varepsilon^2}{2}.$$

Moreover, Theorem 1 with the above choice of  $t$  yields  $D_{\text{KL}}(\mu_t \parallel \pi_\beta) \leq \varepsilon^2/2$ . To conclude, we use Pinsker's inequality and the triangle inequality for  $\|\cdot\|_{\text{TV}}$ .  $\square$

Convergence guarantees for other cases where  $\phi$  is only a *proxy* for  $V$  are presented in Appendix C.

## 4.2 Sampling from the uniform distribution on a convex body

Next, we consider an application of NLD to the problem of sampling from the uniform distribution  $\pi$  on a convex body  $\mathcal{C}$ . A natural method of outputting an approximate sample from  $\pi$  is to take a strictly convex function  $\tilde{V} : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  such that  $\text{dom } \tilde{V} = \mathcal{C}$  and  $\tilde{V}(x) \rightarrow \infty$  as  $x \rightarrow \partial\mathcal{C}$ , and to run NLD with target distribution  $\pi_\beta \propto e^{-\beta\tilde{V}}$ , where the inverse temperature  $\beta$  is taken to be small (so that  $\pi_\beta \approx \pi$ ). The function  $\tilde{V}$  is known as a *barrier function*.

Although we can take any choice of barrier function  $\tilde{V}$ , we obtain a clean theoretical result if we assume that  $\tilde{V}$  is  $\nu^{-1}$ -exp-concave, that is, the mapping  $\exp(-\nu^{-1}\tilde{V})$  is concave. Interestingly, this assumption further deepens the rich analogy between sampling and optimization, since such barriers are widely studied in the optimization literature. There, the property of exp-concavity is typically paired with the property of *self-concordance*, and barrier functions satisfying these two properties are a cornerstone of the theory of *interior point algorithms* (see [Bub15, §5.3] and [Nes04, §4]).

We now formulate our sampling result. In our continuous framework, it does not require self-concordance of the barrier function.

**Corollary 3.** Fix a target accuracy  $\varepsilon > 0$ . Let  $\pi$  be the uniform distribution over a convex body  $\mathcal{C}$  and let  $\tilde{V}$  be a  $\nu^{-1}$ -exp-concave barrier for  $\mathcal{C}$ . Then, the law  $(\mu_t)_{t \geq 0}$  of NLD with target density  $\pi_\beta \propto e^{-\beta\tilde{V}}$  for  $\beta \leq \varepsilon^2/(2\nu)$  satisfies  $\|\mu_t - \pi\|_{\text{TV}} \leq \varepsilon$  by time  $t = \frac{1}{2} \ln[2\chi^2(\mu_0 \parallel \pi_\beta)] + \ln(1/\varepsilon)$ .

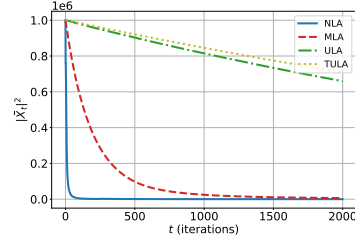


Figure 2: Approximately sampling from  $\pi \propto e^{-\|\cdot\|}$  by sampling from  $\pi_\beta \propto e^{-\|\cdot\|-\beta\|\cdot\|^2}$  ( $\beta = .0005$ ). Algorithms are initialized at a random  $X_0$  with  $\|X_0\| = 1000$ . The plot shows the squared distance of the running means to 0.

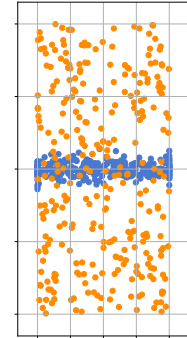


Figure 3: Uniform sampling from the set  $[-0.01, 0.01] \times [-1, 1]$ : PLA (blue) vs. NLA (orange). See Section E.3.

*Proof.* Lemma 1 in Appendix D ensures that  $D_{\text{KL}}(\pi_\beta \parallel \pi) \leq \varepsilon^2/2$ . We conclude as in the proof of Corollary 2, by using Theorem 1, Pinsker’s inequality, and the triangle inequality for  $\|\cdot\|_{\text{TV}}$ .  $\square$

We demonstrate the efficacy of NLD in a simple simulation: sampling uniformly from the ill-conditioned rectangle  $[-a, a] \times [-1, 1]$  with  $a = 0.01$  (Figure 3). We compare NLA with the Projected Langevin Algorithm (PLA) [BEL18], both with 200 iterations and  $h = 10^{-4}$ . For NLA, we take  $\tilde{V}(x) = -\log(1 - x_1^2) - \log(a^2 - x_2^2)$  and  $\beta = 10^{-4}$ .

### 4.3 Langevin diffusion under a Poincaré inequality

We conclude this section by giving some implications of Theorem 1 to the classical Langevin diffusion (LD) when  $\phi = \|\cdot\|^2/2$ . In this case, the mirror Poincaré inequality (MP) reduces to the classical Poincaré inequality (P) as in Section 3.2.

**Corollary 4.** *Suppose that  $\pi$  satisfies a Poincaré inequality (P) with constant  $C_P > 0$ . Then, the law  $(\mu_t)_{t \geq 0}$  of the Langevin diffusion (LD) satisfies*

$$2\|\mu_t - \pi\|_{\text{TV}}^2, H^2(\mu_t, \pi), D_{\text{KL}}(\mu_t \parallel \pi), \chi^2(\mu_t \parallel \mu), \frac{1}{2C_P} W_2^2(\mu_t, \pi) \leq e^{-\frac{2t}{C_P}} \chi^2(\mu_0 \parallel \pi).$$

The convergence in TV distance recovers results of [Dal17b; DM17]. Bounds for the stronger error metric  $\chi^2(\cdot \parallel \pi)$  have appeared explicitly in [CLL19; VW19] and is implicit in the work of [BCG08; CG09] on which the TV bound of [DM17] is based.

Moreover, it is classical that if  $\pi$  satisfies a log-Sobolev inequality (LSI) with constant  $C_{\text{LSI}}$  then it has Poincaré constant  $C_P \leq C_{\text{LSI}}$ . Thus, the choice of the chi-squared divergence as a surrogate for the KL divergence when tracking progress indeed requires weaker assumptions on  $\pi$ .

## 5 Numerical experiments

In this section, we examine the numerical performance of the *Newton-Langevin Algorithm* (NLA), which is given by the following Euler discretization of NLD:

$$\nabla V(X_{k+1}) = (1 - h)\nabla V(X_k) + \sqrt{2h} [\nabla^2 V(X_k)]^{1/2} \xi_k, \quad (\text{NLA})$$

where  $(\xi_k)_{k \in \mathbb{N}}$  is a sequence of i.i.d.  $\mathcal{N}(0, I_d)$  variables. In cases where  $\nabla V$  does not have a closed-form inverse, such as the logistic regression case of Section E.2, we invert it numerically by solving the convex optimization problem  $\nabla V^*(y) = \operatorname{argmax}_{x \in \mathbb{R}^d} \{ \langle x, y \rangle - V(x) \}$ .

We focus here on sampling from an ill-conditioned generalized Gaussian distribution on  $\mathbb{R}^{100}$  with  $V(x) = \langle x, \Sigma^{-1}x \rangle^\gamma/2$  for  $\gamma = 3/4$  to demonstrate the scale invariance of NLD established in Corollary 1. Additional experiments, including the Gaussian case  $\gamma = 1$ , are given in Appendix E.

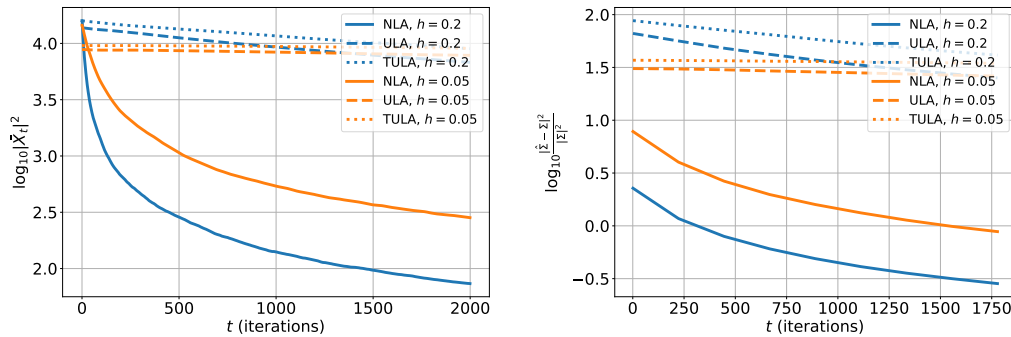


Figure 4:  $V(x) = \langle x, \Sigma^{-1}x \rangle^{3/4}/2$ ,  $\Sigma = \operatorname{diag}(1, 2, \dots, 100)$ . Left: absolute squared error of the mean 0. Right: relative squared error for the scatter matrix  $\Sigma$ .

Figure 4 compares the performance of NLA to that of the Unadjusted Langevin Algorithm (ULA) [DM+19] and of the Tamed Unadjusted Langevin Algorithm (TULA) [Bro+19]. We run the



algorithms 50 times and compute running estimates for the mean and scatter matrix of the family following [ZWG13]. Convergence is measured in terms of squared distance between means and relative squared distance between scatter matrices,  $\|\hat{\Sigma} - \Sigma\|^2 / \|\Sigma\|^2$ . NLA generates samples that rapidly approximate the true distribution and also displays stability to the choice of the step size  $h$ .

## 6 Open questions

We conclude this paper by discussing several intriguing directions for future research. In this paper, we focused on giving clean convergence results for the continuous-time diffusions MLD and NLD, and we leave open the problem of obtaining discretization error bounds. In discrete time, Newton’s method can be unstable, and one uses methods such as damped Newton, Levenburg-Marquardt, or cubic-regularized Newton [CGT00; NP06]; it is an interesting question to develop sampling analogues of these optimization methods. In a different direction, we ask the following question: are there appropriate variants of other popular sampling methods, such as accelerated Langevin [Ma+19] or Hamiltonian Monte Carlo [Nea12], which also enjoy the scale invariance of NLD?

## Broader impact

The sampling algorithms designed in this paper have the potential to improve a wide variety of Bayesian methods and therefore have an indirect impact on various domains such as health and medicine where such methods are pervasive. Sampling algorithms are also used for the generation of automated spam messages, which have potentially negative effects on society. Since this paper is primarily focused on theory, these questions are not addressed here.

## Acknowledgments

Philippe Rigollet was supported by NSF awards IIS-1838071, DMS-1712596, DMS-TRIPODS-1740751. Sinho Chewi and Austin Stromme were supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. Thibaut Le Gouic was supported by ONR grant N00014-17-1-2147 and NSF IIS-1838071.

We thank the reviewers for very helpful suggestions regarding the presentation of the paper.

## References

- [AB15] D. Alonso-Gutiérrez and J. Bastero. *Approaching the Kannan-Lovász-Simonovits and variance conjectures*. Vol. 2131. Lecture Notes in Mathematics. Springer, Cham, 2015, pp. x+148.
- [AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
- [BC12] S. Bubeck and N. Cesa-Bianchi. “Regret analysis of stochastic and nonstochastic multi-armed bandit problems”. In: *Foundations and Trends in Machine Learning* 5.1 (2012), pp. 1–122.
- [BCG08] D. Bakry, P. Cattiaux, and A. Guillin. “Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré”. In: *Journal of Functional Analysis* 254.3 (2008), pp. 727–759.
- [BEL18] S. Bubeck, R. Eldan, and J. Lehec. “Sampling from a log-concave distribution with projected Langevin Monte Carlo”. In: *Discrete & Computational Geometry* 59.4 (2018), pp. 757–783.
- [Ber18] E. Bernton. “Langevin Monte Carlo and JKO splitting”. In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1777–1798.
- [BGG12] F. Bolley, I. Gentil, and A. Guillin. “Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations”. In: *Journal of Functional Analysis* 263.8 (2012), pp. 2430–2457.

- [BGL14] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552.
- [BL00] S. G. Bobkov and M. Ledoux. “From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities”. In: *Geom. Funct. Anal.* 10.5 (2000), pp. 1028–1052.
- [BL06] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization*. Second. Vol. 3. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Theory and examples. Springer, New York, 2006, pp. xii+310.
- [BL76] H. J. Brascamp and E. H. Lieb. “On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation”. In: *J. Functional Analysis* 22.4 (1976), pp. 366–389.
- [Bro+19] N. Brosse, A. Durmus, É. Moulines, and S. Sabanis. “The tamed unadjusted Langevin algorithm”. In: *Stochastic Processes and their Applications* 129.10 (2019), pp. 3638–3663.
- [Bub15] S. Bubeck. “Convex optimization: algorithms and complexity”. In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.
- [CB18] X. Cheng and P. Bartlett. “Convergence of Langevin MCMC in KL-divergence”. In: *Algorithmic Learning Theory 2018*. Vol. 83. Proc. Mach. Learn. Res. (PMLR). Proceedings of Machine Learning Research PMLR, 2018, p. 26.
- [CG09] P. Cattiaux and A. Guillin. “Trends to equilibrium in total variation distance”. In: *Ann. Inst. Henri Poincaré Probab. Stat.* 45.1 (2009), pp. 117–145.
- [CGT00] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods*. Vol. 1. SIAM, 2000.
- [Che+18] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. “Underdamped Langevin MCMC: A non-asymptotic analysis”. In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 300–323.
- [Che+19] X. Cheng, D. Yin, P. L. Bartlett, and M. I. Jordan. “Quantitative  $W_1$  convergence of Langevin-like stochastic processes with non-convex potential and state-dependent noise”. In: *arXiv e-prints*, arXiv:1907.03215 (July 2019).
- [Che+20] S. Chewi, T. Maunu, P. Rigollet, and A. Stromme. “Gradient descent algorithms for Bures-Wasserstein barycenters”. In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 1276–1304.
- [CL89] M.-F. Chen and S.-F. Li. “Coupling methods for multidimensional diffusion processes”. In: *The Annals of Probability* (1989), pp. 151–177.
- [CLL19] Y. Cao, J. Lu, and Y. Lu. “Exponential decay of Rényi divergence under Fokker-Planck equations”. In: *J. Stat. Phys.* 176.5 (2019), pp. 1172–1184.
- [Dal17a] A. Dalalyan. “Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent”. In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, 2017, pp. 678–689.
- [Dal17b] A. S. Dalalyan. “Theoretical guarantees for approximate sampling from smooth and log-concave densities”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.
- [Din15] Y. Ding. “A note on quadratic transportation and divergence inequality”. In: *Statist. Probab. Lett.* 100 (2015), pp. 115–123.
- [DK19] A. S. Dalalyan and A. Karagulyan. “User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient”. In: *Stoch. Proc. Appl.* 129.12 (2019), pp. 5278–5311.
- [DM+19] A. Durmus, É. Moulines, et al. “High-dimensional Bayesian inference via the unadjusted Langevin algorithm”. In: *Bernoulli* 25.4A (2019), pp. 2854–2882.

- [DM15] A. Durmus and É. Moulines. “Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm”. In: *Statistics and Computing* 25.1 (Jan. 2015), pp. 5–19.
- [DM17] A. Durmus and É. Moulines. “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm”. In: *Ann. Appl. Probab.* 27.3 (2017), pp. 1551–1587.
- [DMM19] A. Durmus, S. Majewski, and B. Miasojedow. “Analysis of Langevin Monte Carlo via convex optimization”. In: *J. Mach. Learn. Res.* 20 (2019), Paper No. 73, 46.
- [DR20] A. S. Dalalyan and L. Riou-Durand. “On sampling from a log-concave density using kinetic Langevin diffusions”. In: *Bernoulli* 26.3 (2020), pp. 1956–1988.
- [DRK19] A. S. Dalalyan, L. Riou-Durand, and A. Karagulyan. “Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets”. In: *arXiv e-prints*, arxiv:1906.08530 (June 2019).
- [Dwi+19] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. “Log-concave sampling: Metropolis-Hastings algorithms are fast”. In: *Journal of Machine Learning Research* 20.183 (2019), pp. 1–42.
- [Ebe16] A. Eberle. “Reflection couplings and contraction rates for diffusions”. In: *Probability Theory and Related Fields* 166.3-4 (2016), pp. 851–886.
- [FKP94] A. Frieze, R. Kannan, and N. Polson. “Sampling from log-concave distributions”. In: *The Annals of Applied Probability* 4.3 (1994), pp. 812–837.
- [Gen08] I. Gentil. “From the Prékopa-Leindler inequality to modified logarithmic Sobolev inequality”. In: *Ann. Fac. Sci. Toulouse Math. (6)* 17.2 (2008), pp. 291–308.
- [Goo+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets”. In: *Advances in Neural Information Processing Systems* 27. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. 2014, pp. 2672–2680.
- [Hsi+18] Y.-P. Hsieh, A. Kavis, P. Rolland, and V. Cevher. “Mirrored Langevin dynamics”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 2878–2887.
- [JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. “The variational formulation of the Fokker-Planck equation”. In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17.
- [KLS95] R. Kannan, L. Lovász, and M. Simonovits. “Isoperimetric problems for convex bodies and a localization lemma”. In: *Discrete Comput. Geom.* 13.3-4 (1995), pp. 541–559.
- [KNS16] H. Karimi, J. Nutini, and M. Schmidt. “Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 795–811.
- [KS91] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*. Second. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xxiv+470.
- [Led18] M. Ledoux. *Remarks on some transportation cost inequalities*. 2018.
- [Liu20] Y. Liu. “The Poincaré inequality and quadratic transportation-variance inequalities”. In: *Electron. J. Probab.* 25 (2020), Paper No. 1, 16.
- [LTV20] A. Laddha, Y. Tat Lee, and S. Vempala. “Strong self-concordance and sampling”. In: *STOC* (2020).
- [LV07] L. Lovász and S. Vempala. “The geometry of logconcave functions and sampling algorithms”. In: *Random Structures & Algorithms* 30.3 (2007), pp. 307–358.
- [LV17] Y. T. Lee and S. S. Vempala. “Eldan’s stochastic localization and the KLS hyperplane conjecture: an improved lower bound for expansion”. In: *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*. IEEE Computer Soc., Los Alamitos, CA, 2017, pp. 998–1007.

- [LV18] Y. T. Lee and S. S. Vempala. “Stochastic localization + Stieltjes barrier = tight bound for log-Sobolev”. In: *STOC’18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2018, pp. 1122–1129.
- [Ma+19] Y.-A. Ma, N. Chatterji, X. Cheng, N. Flammarion, P. Bartlett, and M. I. Jordan. “Is there an analog of Nesterov acceleration for MCMC?” In: *arXiv e-prints*, arXiv:1902.00996 (Feb. 2019).
- [Mar+12] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. “A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion”. In: *SIAM J. Sci. Comput.* 34.3 (2012), A1460–A1487.
- [Mou+19] W. Mou, N. Flammarion, M. J. Wainwright, and P. L. Bartlett. “Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity”. In: *arXiv e-prints*, arXiv:1907.11331 (July 2019).
- [MT09] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. 2nd. USA: Cambridge University Press, 2009.
- [Nea12] R. Neal. “MCMC using Hamiltonian dynamics”. In: *Handbook of Markov Chain Monte Carlo* (June 2012).
- [Nes04] Y. Nesterov. *Introductory lectures on convex optimization*. Vol. 87. Applied Optimization. A basic course. Kluwer Academic Publishers, Boston, MA, 2004, pp. xviii+236.
- [NJ79] A. S. Nemirovskii and D. B. Judin. *Complexity of problems and efficiency of optimization methods*. 1979.
- [NP06] Y. Nesterov and B. T. Polyak. “Cubic regularization of Newton method and its global performance”. In: *Mathematical Programming* 108.1 (2006), pp. 177–205.
- [RC04] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.
- [Roc97] R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Reprint of the 1970 original, Princeton Paperbacks. Princeton University Press, Princeton, NJ, 1997, pp. xviii+451.
- [RRT17] M. Raginsky, A. Rakhlin, and M. Telgarsky. “Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis”. In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, July 2017, pp. 1674–1703.
- [San17] F. Santambrogio. “{Euclidean, metric, and Wasserstein} gradient flows: an overview”. In: *Bulletin of Mathematical Sciences* 7.1 (2017), pp. 87–154.
- [Sim+16] U. Simsekli, R. Badeau, A. T. Cemgil, and G. Richard. “Stochastic quasi-Newton Langevin Monte Carlo”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML’16. New York, NY, USA: JMLR.org, 2016, pp. 642–651.
- [SKL20] A. Salim, A. Korba, and G. Luise. “Wasserstein proximal gradient”. In: *arXiv e-prints*, arxiv:2002.03035 (Feb. 2020).
- [Tsy09] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. Springer, New York, 2009, pp. xii+214.
- [Vil03] C. Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370.
- [Vil09] C. Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973.
- [VW19] S. Vempala and A. Wibisono. “Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8094–8106.

- [Wib18] A. Wibisono. “Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem”. In: *Conference on Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2093–3027.
- [Wib19] A. Wibisono. “Proximal Langevin algorithm: rapid convergence under isoperimetry”. In: *arXiv e-prints*, arxiv:1911.01469 (Nov. 2019).
- [WL20] Y. Wang and W. Li. “Information Newton’s flow: second-order optimization method in probability space”. In: *arXiv e-prints*, arxiv:2001.04341 (Jan. 2020).
- [Zha+20] K. S. Zhang, G. Peyré, J. Fadili, and M. Pereyra. “Wasserstein control of mirror Langevin Monte Carlo”. In: *arXiv e-prints*, arxiv:2002.04363 (Feb. 2020).
- [ZWG13] T. Zhang, A. Wiesel, and M. S. Greco. “Multivariate generalized Gaussian distribution: convexity and graphical models”. In: *IEEE Transactions on Signal Processing* 61.16 (2013), pp. 4141–4148.