Langevin Monte Carlo without log-concavity

Leello Dadi

Abstract

Langevin Monte Carlo (LMC) is a Markov Chain Monte Carlo sampling method that generates samples by adding gradient information to Gaussian noise. Since it queries only local information, we might expect it to have all the limitations of local methods like Gradient Descent. The presence of noise, however, makes LMC capable of provably sampling from a large class of distributions for which global information cannot be inferred from local queries.

This statement was proved by Vempala and Wibisono in [VW19] under two mild assumptions. This will be the starting point of our report. We will then see that, although their work shows the region of tractability for LMC extends beyond strong log-concavity, there is a potential exponential blow-up of constants in the convergence bound that ensures that the computational hardness of non-convexity is never violated. Tzen et al's trajectory-wise analysis of LMC [TLR18] will shed some light on the causes of this exponential blow-up of convergence time. With this understanding, we will naturally ask whether or not this blow-up occurs when trying to sample from real world distributions. By discussing the success of Song and Ermon's work [SE19] using LMC to generate natural images, we will finish by arguing that the real world may well be within the tractable region.

1 Introduction

In keeping with the tradition of secretive Monte Carlo research initiated by Ulam and Von Neuman [Eck], we will consider in this report the problem of sampling from a probability distribution π while keeping secret our motivations for doing so.

The considered distribution π, our target measure, is a probability measure over $\mathbb{R}^d$ that admits a density p_π that can be expressed as

$$p_\pi(x) = \frac{\exp\left(-f(x)\right)}{Z},$$

where $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a continuously differentiable function, referred to as the *potential*, and $Z \in \mathbb{R}$ is a normalization constant. We assume that we can query f and ∇f at any point $x \in \mathbb{R}^d$, but Z will remain unknown.

Our goal is to generate samples approximately distributed according to π. One approach for doing so is to initialize a random walk at some point $\mathbf{X}_0 \in \mathbb{R}^d$ and make Gaussian steps exploring the space hoping to land in the typical regions of π [SFR10]. Before committing to a step, we can check if the walk is leading us towards a higher probability regions by querying f then decide to accept or reject this step. Eventually our random walk will output samples from π at the risk of having to reject many steps.

Here, however, since we have access to the gradients, we can inform our random walk by using this first-order information: the gradient can guide us towards high probability regions.

This is the idea of Langevin Monte Carlo (LMC). The algorithm consists of generating the following Markov Chain

$$\mathbf{X}_{k+1} = \mathbf{X}_k - \eta \nabla f(\mathbf{X}_k) + \sqrt{2\eta}\mathbf{Z}_{k+1}, \tag{LMC}$$

where $\eta > 0$ is referred to as the step-size and $(\mathbf{Z}_k)_k$ is an sequence of independent identically distributed $\mathcal{N}(0,1)$ variables.

LMC is a gradient augmented random walk and using it sample was first suggested in the chemical physics literature [RDF78]. The asymptotic convergence of the scheme for a fixed step-size was discussed in [RT96] and the authors argued for the necessity of an Metropolis accept/reject step to ensure that the iterates would eventually sample from π.

Indeed, without such a step, the distributions of the iterates $(\mathbf{X}_k)_{k\in\mathbb{N}}$ tend to a biased limit π_η, that is in most cases different from π. Dalalyan [Dal16], through a non asymptotic analysis, offered a possible way of side stepping this need for a Metropolis filter. He showed that given a precision ϵ, it was possible to pick η small enough, then run the (LMC) chain for long enough to generate samples that were ϵ close π.

The non-asymptotic convergence guarantees were first derived in the strongly-log-concave case [Dal16], i.e when f is strongly convex. In this report our focus will be on guarantees of convergence that do not require strong-convexity or even convexity of f. This is of particular interest because (LMC) only sees f through local queries of the gradient and one may believe some form of convexity, or conditions on stationary points, would be necessary to obtain strong guarantees. The presence of noise seems to dispense with the need for such strong conditions when sampling is the goal.

We will see that it is possible to establish convergence of (LMC) under two mild assumptions that, in effect, only put restrictions on the tails of f. The proof we will study is Vempala and Wibisono's in [VW19] where they show convergence of the iterates as measured by the KL divergence from π (Section 3).

The convergence bound will, at first sight, appear to always be polynomial in dimension. But a closer inspection, informed by the work of Tzen, Liang and Raginsky [TLR18], will reveal that the convergence can be extremely slow, and in fact be exponential in dimension. The divide between the tractable realm of polynomial convergence and the intractable is captured by a single constant appearing in the bound : the log-Sobolev constant.

It seems that this log-Sobolev constant is a good measure of benign non-convexity [CCAY+20, Li21, PVBL+20]: a sequence involving the sum of a gradient term and a noise term seems to be able to tackle non-convexity as long as the log-Sobolev constant does not have a poor dependence on dimension.

Naturally, we will ask whether or not the real world exhibits this benign non-convexity. Distributions like images, are at the very least multi-modal and therefore cannot be log-concave let alone strongly log-concave. They could however have log-Sobolev constants that have a good polynomial dependence on dimension. Song and Ermon propose a method for applying (LMC) to real-world distributions. By studying their work, we will discuss whether or not their results provide sufficient evidence for the claim that the real-world admits well-behaved log-Sobolev constants.

We will conclude by discussing open questions regarding convergence of (LMC) and reviewing the possible ways in which the tools we understood here to analyze (LMC) can help us analyze the closely related iterates of Stochastic Gradient Descent, an algorithm in heavy use today, that also mixes gradients with noise, albeit with a different weighing.

2 Notation and setting

The distributions considered in this report are measures over $(\mathbb{R}^d, B(\mathbb{R}^d))$, where $B(\mathbb{R}^d)$ denotes the Borel sigma field. The Euclidean norm on $\mathbb{R}^d$ is denoted $\|\cdot\|$. For a probability distribution ν, we write p_ν for its Lebesgue density. The expectation with respect to ν is denoted $\mathbb{E}_\nu$. The gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$ is denoted ∇f, if it is multivariate, the Jacobian is denoted J_f. A function is said to be *L-smooth* if its gradient is L-Lipschitz. We will consider Ito stochastic differential equations of the form

$$dX_t = b(X_t)dt + \sigma dB_t,$$

where $\sigma > 0$, $(B_t)_{t\in[0,+\infty[}$ is a standard Brownian motion on $\mathbb{R}^d$, and $b : \mathbb{R}^d \mapsto \mathbb{R}^d$ is Lipschitz. For appropriate initializations X_0, this equation has a unique solution (Theorem 5.2.1 [Øk03]) and for each $t \in [0, +\infty[$, the law of X_t admits a density denoted p_t with respect to the Lebesgue measure, moreover, $t \mapsto p_t$ is continuously differentiable (Thm 4.1 [Pav14]).

The KL divergence from π of a measure ν and the relative Fisher information I of ν with respect to π are defined as

$$\mathrm{KL}(\nu||\pi) = \int_{\mathbb{R}^d} p_\nu(x) \log \frac{p_\nu(x)}{p_\pi(x)} dx \quad \text{and} \quad I(\nu||\pi) = \int_{\mathbb{R}^d} p_\nu(x) \left\| \nabla \log \frac{p_\nu(x)}{p_\pi(x)} \right\|^2 dx.$$

3 Convergence under isoperimetry and smoothness

We begin our study with the proof that the distribution of iterates $(\mathbf{X}_k)_k$, for an appropriate fixed step-size η, will indeed approach the target distribution π.

As Vempala and Wibisono's show, this convergence will hold even without convexity of the potential f. A convergence analysis for two divergence measures is provided in their paper[VW19], the KL and the Renyi divergence. Our focus will be on the KL divergence as it the more complete contribution of the paper.

The main result of [VW19] is a bound on the decrease of KL divergence at each single step of (LMC), when going from $\mathbf{X}_k$ to $\mathbf{X}_{k+1}$. The convergence of the whole sequence is then obtained by induction. Two assumptions are needed for this result. The first is smoothness of the target potential. The second is a functional inequality called the log-Sobolev inequality (LSI), which we will examine shortly. Using only these two, Vempala and Wibisono prove the following theorem.

Theorem 3.1. Let π be a distribution satisfying the (LSI) with constant $\alpha > 0$. If the potential is L-smooth, we have for any $\eta < \frac{\alpha}{4L^2}$, the following decrease at each step:

$$\mathrm{KL}(\rho_{k+1}||\pi) \leq e^{-\alpha\eta}\mathrm{KL}(\rho_k||\pi) + 6\eta^2 dL^2,$$

where ρ_k denotes the distribution of $\mathbf{X}_k$.

In this section, we will first discuss the (LSI) assumption to get a better grasp of the scope of the result. We will then discuss the proof of Theorem 3.1. After briefly mentioning the Renyi result, we will discuss the implications of the convergence bound.

3.1 The log-Sobolev inequality

The main assumptions the authors of [VW19] make is one they refer to as an *isoperimetric* assumption. In particular, they assume that π verifies the log-Sobolev inequality. This inequality relates the KL divergence from π to the relative Fisher information with respect to π. It is defined as follows.

Definition 3.1. There exists $\alpha > 0$ such that for any probability measure ν over $\mathbb{R}^d$ such that $\mathrm{KL}(\nu||\pi) < \infty$, we have

$$\mathrm{KL}(\nu||\pi) \leq \frac{1}{2\alpha} I(\nu||\pi). \tag{LSI}$$

The biggest α for which the inequality above holds is called the log-Sobolev constant of π. This inequality was first introduced by Gross [Gro75] and we can view it as merely an analytic tool to show sub-gaussian concentration through a line of reasoning called the Herbst argument.

The Herbst argument Given a zero mean, real random variable X, a concentration inequality is a bound on the tails of the law of X. The random variable is said to be sub-gaussian if it verifies the following inequality:

$$\mathbb{P}(|X| > t) \leq C \exp(-\frac{t^2}{c}),$$

for some absolute constants $C, c > 0$. The Chernoff method, which is a simple application of Markov's inequality to $e^{\lambda X}$ for $\lambda > 0$ tells us that to establish such a bound, we only need to control the moment generating function $\lambda \mapsto \mathbb{E}[e^{\lambda X}]$. The *Herbst argument* is a basic sequence of steps showing that the log-Sobolev inequality is enough to upper bound the moment generating function and obtain sub-gaussian tails [BLM13, ABC+00].

In other words, assuming the (LSI) is saying that we have an easy way of showing sub-gaussian tails or, correspondingly, of showing the quadratic growth of f. This will be our understanding of (LSI) : it is a convenient guarantee of sub-gaussian concentration. It is however important to note that (LSI) is strictly stronger than sub-gaussianity. Here, in [VW19], the (LSI) is referred to with the term *isoperimetric* and this comes from the geometric view of measure concentration which quantifies how much measures concentrate around bodies of a certain volume.

The log-Sobolev assumption has been used in prior publications, namely [Dal14], [RRT17] and [MCJ+19]. In those papers, the (LSI) is derived from stronger assumptions like strong-log-concavity or dissipativity (see Assumption 3). For the first two, [Dal14] and [RRT17], the stronger assumptions are used again elsewhere in their results, ensuring that Vempala and Wibisono's result cannot be deduced from theirs. The last paper [MCJ+19], however, does an almost identical analysis albeit with a slightly more complicated proof for essentially the same Lemma (Lemma 12 in [VW19] and Lemma 5 in [MCJ+19]). A non-exhaustive diagram of the commonly used assumptions in proving convergence of (LMC) is provided in Figure 1.

The (LSI) is a well-behaved inequality in the sense that it is preserved under bounded perturbations of the potential f [ABC+00]. It is also stable through differentiable Lipschitz transformations. This shows that a broad class of distributions verify the (LSI) and from Figure 1 we can see that it is a relatively weak assumptions. Proving convergence without making any stronger additional assumption is therefore a valuable result.

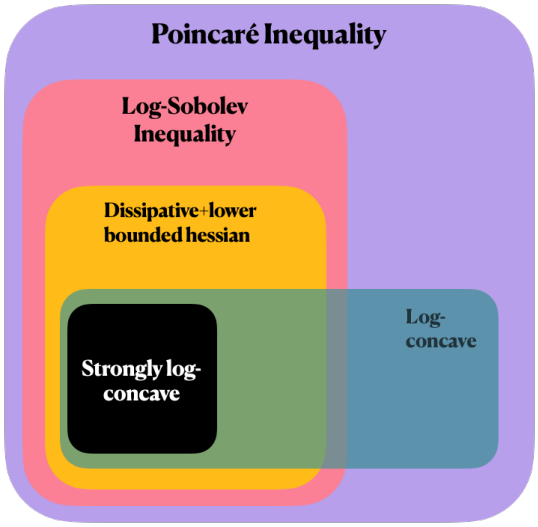

Figure 1. A diagram showing the order between commonly used assumptions. A proof that strongly-log concave densities are dissipative can be derived from the quadratic lower bound at 0. Corollary 2.1.(2) of [CGW10] shows that dissipativity and lower bounded hessians imply LSI. The fact that Poincaré inequalities hold for log-concave measures is shown in Corollary 1.9 of [BBCG08]. The proof that LSI implies Poincaré can be found in [BGL14](Proposition 5.1.3).

3.2 A proof through conditional Fokker-Planck

As mentionned earlier, Vempala and Wibisono prove convergence of (LMC) by analyzing how the KL divergence evolves during a single step of (LMC). We briefly discuss their proof.

To establish Theorem 3.1, the first step is to observe that a single iteration of (LMC) corresponds to the *exact* solution of the continous-time diffusion

$$\begin{cases} dX_t = -\nabla f(\mathbf{X}_k)dt + \sqrt{2}dB_t \\ X_0 = \mathbf{X}_k \end{cases} . \tag{1}$$

Notice that the drift term is independent of X_t and the exact solution of (1) evaluated at time η is $X_\eta = \mathbf{X}_{k+1}$. By writing the discrete iterates in this interpolated way, we gain access to the vast toolbox of stochastic differential equations.

Now, our goal is to know how the KL divergence from π evolves when going from $\mathbf{X}_k$ to $\mathbf{X}_{k+1}$. Using our interpolated process this means we want to know how the continuous object $t \mapsto \mathrm{KL}(p_t||\pi)$ evolves between $t = 0$ and $t = \eta$. To do so we compute its time derivative :

$$\frac{d}{dt}\mathrm{KL}(p_t||\pi) = \int_{\mathbb{R}^d} \frac{\partial p_t}{\partial t}(x) \ln \frac{p_t(x)}{\pi(x)} dx. \tag{2}$$

From (2), we see that in order to know the behavior of the KL divergence, we need to know how the density of p_t of $\hat{X}_t$ evolves, i.e, we need to know $\frac{\partial p_t}{\partial t}(x)$.

Luckily for us, since (1) is a constant drift diffusion, when conditionned on $\mathbf{X}_k$, the time derivative $\frac{\partial}{\partial t} p_{t|\mathbf{X}_k}$ of the conditional density is given to us by a standard formula : the Fokker-Planck equation [Pav14]. Once we have the evolution of the conditionned density, we can integrate with respect to $\mathbf{X}_k$ to determine the evolution of the unconditional density $\frac{\partial}{\partial t} p_t$. This requires a few interchanges between derivatives and expectations.

On the interchange of derivatives and integrals: Since we are not integrating over compact domains, the interchange of derivatives and integrals requires justifications. A closely related paper [MFWB19] cares about this interchange and provides lemmas justifying it while [VW19] does not mention any concerns. A careful read, however, shows that steps resembling the induction argument of [MFWB19] are also done in [VW19](Lemma 16, 17) where it is shown that the (LSI) is preserved when applying the gradient step and adding noise. Since (LSI) implies sub-gaussian tails, the fast decay required to justify the interchange is obtained as long as the initial distribution verifies the (LSI) (Theorem 4.1 [Pav14]).

Plugging in (2) the formula for the unconditional density determined through this interchange, we find that

$$\frac{d}{dt}\mathrm{KL}(\rho_t||\pi) \leq -\frac{3}{4}I(\rho_t||\pi) + \mathbb{E}\left[\|\nabla f(X_t) - \nabla f(\mathbf{X}_k)\|_2^2\right]. \tag{3}$$

The first term, ignoring the multiplicative factor, is what we would have obtained if the drift term in (2) was not frozen in time. The second term is the price we pay for this frozen drift. In other words, it is the price of discretization.

The authors of [VW19] show that the control of the right hand side of (3) can be achieved using only (LSI) and smoothness.

3.3 Getting to Grönwall only using LSI and smoothness

It is clear that the first term in equation (3) can immediately be upperbounded using the (LSI) assumption. An important contribution of this paper is a simple proof showing that the second term can be uniformely bounded for $t \in [0, \eta]$ (Lemma 12, [VW19]).

The problem boils down to upperbounding the expected squared norm of the gradients of f. Vempala and Wibisono show that, as long as ∇f is Lipschitz smooth, a *transport inequality*, which is an inequality relating a Wassertein distance to the KL divergence, called Talagrand's T_2 inequality, is sufficient to control this expected square norm.

Otto and Villani [OV00] proved that (LSI) is (strictly) stronger than Talagrand's T_2 inequality, consequently, the (LSI) assumption is sufficient to control the expected squared norms.

With this, the authors are able to derive an inequality roughly of the form

$$\frac{d}{dt}\mathrm{KL}(\rho_t||\pi) \leq -C\mathrm{KL}(\rho_t||\pi) + D,$$

where $C, D \geq 0$ are constants. A straightforward application of Grönwall's lemma, which is a common tool to establish exponential decay, yields the result in Theorem 3.1.

3.4 The Renyi Result

In addition to analyzing the behavior of the KL divergence along the iterates of (LMC), the paper also provides, in a second part, analysis of the evolution of the Renyi divergence.

The Renyi divergence is defined as follows. For a given $q > 0$, the Renyi divergence of order q of ν from π, denoted $R_q(\nu||\pi)$, is given by

$$R_q(\nu||\pi) = \frac{1}{q-1}\log \mathbb{E}_\pi\left[\left(\frac{p_\nu(X)}{p_\pi(X)}\right)^q\right].$$

The Renyi divergence is increasing with respect to q, and for $q = 1$, we can recover the KL divergence as a limit. It is therefore a generalisation of KL for which a generalized version of Pinsker's inequality also holds.

In [VW19], it is shown that for a given step size η, the iterates of (LMC) converge to the *biased* limit π_η in Renyi divergence at a fast rate when the biased limit verifies the (LSI). A missing result however is how far this biased limit is from the actual target measure π.

This glaring gap prevents their analysis from being a complete one. They only provide a Gaussian example showing that the assumption that (LSI) holds for the biased limit and that the distance to the true target is controllable by scaling down η. For this reason, their analysis is only a first step in proving convergence in Renyi divergence.

Here is how they manage to show their result. The biased limit π_η is stationnary for the dynamics of (LMC) : starting from a variable distributed according to π_η, applying a gradient step and then adding gaussian noise returns us back to a variable distributed according to π_η. Observing this, Vempala and Wibisono show that: first, applying a gradient step does not change the Renyi divergence between two distributions and second, adding Gaussian noise, under the (LSI), decreases the divergence between distributions by a multiplicative factor.

Putting these two facts together, and exploiting the fact that π_η is stationnary, [VW19] show that, at each step of (LMC), the Renyi divergence is decreased by a multiplicative factor, thus establishing the fast geometric rate. Without the control of the divergence of this biased limit from the true target, this result remains incomplete.

3.5 Discussion

Vempala and Wibisono provide simple and straightforward analysis showing convergence of (LMC) to the target measure. Indeed a simple induction on Theorem 3.1, shows that in order to be ϵ-close in KL divergence to the target measure π, we can pick $\eta = O(\frac{\alpha\epsilon}{dL^2})$, initialize from an appropriate distribution and run (LMC) for

$$O\left(\frac{L^2 d}{\alpha^2 \epsilon}\right)$$

iterations. This convergence time has a polynomial dependence on d only if none of the constants involved are hiding possible dimension dependencies. Since the log-Sobolev inequality encapsulates a very broad class of distributions and is in effect a restriction only on the tails of the potential, it is unreasonable to expect that this convergence time will always be polynomial in dimension. Indeed, any hard non-convex optimization problem can be cast as a sampling problem, as we will see in the next section. And a convergence bound that is always polynomial in dimension would risk violating the assumptions of computational hardness.

Consequently, some constants in the bound must be hiding possible dimension dependencies and explode exponentially in the dimension for some targets π. The log-Sobolev constant is the exploding constant. In fact, it is believed that for most generic cases, this constant is exponentially small in dimension [RRT17]. In the next section we will give a concrete meaning to the slow convergence time predicted by the bound when LSI constants are very small.

4 Analyzing the trajectories of LMC

To better understand how exponentially small LSI constants can affect the dynamics, we turn our attention to the work of Tzen, Liang and Raginsky [TLR18].

The goal of their paper is to describe the behavior of the trajectories of LMC. Here, LMC is viewed as a tool to optimize the potential f. This introduces a parameter β, referred to as the *inverse temperature* in our LMC iteration :

$$\mathbf{X}_{k+1} = X_k - \eta \nabla f(X_k) + \sqrt{\frac{2\eta}{\beta}} Z_{k+1} \tag{LMC-β}$$

which is a discretization of the diffusion

$$dX_t = -\nabla f(X_t)dt + \sqrt{\frac{2}{\beta}} dB_t. \tag{LD-β}$$

The diffusion above has a stationary measure π_β whose density is proportional to $e^{-\beta f}$. When $\beta \to \infty$, this stationary measure concentrates around the minimizers of f (Section 3.5, [RRT17]). Consequently, by choosing high enough values of β, we can optimize f using (LMC-β).

Tzen et al study the behavior of the iterates of this algorithm around local minimizers of f and they find that with an appropriate choice of η and β, the iterates can be made to stay within the neighborhood of some local minimzer for an arbitrarely long time with high probability.

This indicates that the possibly slow mixing of LMC can be caused by the long times spent around local minimizers.

4.1 Setting

In [TLR18], the potential $f : \mathbb{R}^d \to \mathbb{R}$ on which LMC is applied is an empirical risk approximation of a population risk. Indeed, taking the population risk to be of the form $F(x) := \mathbb{E}_\nu[\ell(x, Z)]$, where $Z \sim \nu$ is some unknown probability measure over some set $\mathcal{Z}$, we can attempt to optimize F by considering the empirical risk

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(x, Z_i), \tag{4}$$

where $Z_1, \dots, Z_n$ are independent, identically distributed samples from ν. The following assumptions are made on the functions ℓ which translate to properties of f.

Assumption 1. For any $z \in \mathcal{Z}$, the function $x \mapsto \ell(x, z)$ is continuously twice differentiable and there exists B such that $\|\nabla \ell(0, z)\| \leq B$ for all $z \in \mathcal{Z}$.

Assumption 2. There exist $L > 0$ and $M > 0$ such that for any $z \in \mathcal{Z}$,

$$\|\nabla \ell(x, z) - \nabla \ell(y, z)\| \leq L\|x - y\| \quad \text{and} \quad \|\nabla^2 \ell(x, z) - \nabla^2 \ell(y, z)\|_2 \leq M\|x - y\|.$$

This implies that f is L-Lipschitz gradient and M-Lipschitz hessian.

Assumption 3. The function f is (m, b)-dissipative :

$$\exists m > 0,\ b \geq 0, \quad \forall x \in \mathbb{R}^d, \quad \langle x, \nabla f(x) \rangle \geq m\|w\|^2 - b.$$

We note in passing that these assumptions imply that f verifies the LSI (see Figure 1), but the problem of convergence to stationarity here is not considered here. Rather, Tzen et al focus on proving the following result on time spent around local minima when using (LMC-β) to optimize (4).

4.2 The main result

We now state the main result of [TLR18]. First, pick a nondegenerate local minimum $\bar{x}$ of f where $H = \nabla^2 f(\bar{x})$ is positive definite and initialize (LMC-β) within a distance of at most $r > 0$ of $\bar{x}$.

> *For any $\delta \in [0, 1]$, for any small $\epsilon > 0$, we set $T_{rec} = \frac{2}{m}\log(\frac{8r}{\epsilon})$. For any escape time $T_{\text{esc}} > T_{\text{rec}}$ of our choice, there exist an η small enough, scaling as $O(\frac{1}{T_{\text{esc}}})$, and a β big enough, scaling as $O(\log(T_{\text{esc}}))$, such that, for the iterates of (LMC-β), we have*
>
> $$\mathbb{P}(\text{Escape from } \epsilon\text{-neighborhood of } w_H \text{ in } [T_{\text{rec}}, T_{\text{esc}}]) \leq \delta.$$

This result tells us that we can make the iterates of (LMC-β) stay, with high probability, within the neighborhood of a local minimum for as long as we desire with an appropriate choice of η and β. In other words, there is a choice of η and β such that (LMC-β) is trapped close to a local minimum for a long time.

4.3 A summary of the proof

We can briefly outline the main ideas underlying the proof. The goal is to control the escape of the discrete iterates from a neighborhood of $\bar{x}$. To do so, we will control the probability that the continuous diffusion (LD-β) escapes, then we will relate this control to the discretization (LMC-β).

Controlling the escape of diffusion : The first step is to linearize the gradient around the local minimum $\bar{x}$: we can write $\nabla f(x) = H(x - \bar{x}) - \rho(x - \bar{x})$. This allows us to express the diffusion (LD-β) as a sum of a well behaved process coming from the linear term and a remainder process coming from the error of the linearization ρ. Controlling the escape therefore boils down to controlling the sum of these two processes. At this point, Tzen et al exploit the following seemingly trivial result on the control of a sum.

Lemma 4.1. Let A and B be two real random variables, then the following inclusion of events holds

$$\{A + B \geq c\} \subseteq \{A \geq c_1\} \cup \{B \geq c_2\}$$

for **any** c_1, c_2 such that $c_1 + c_2 = c$.

The result above tells us that controlling the sum of two random variables can be achieved by controlling the terms individually with the added benefit of having some freedom to choose the thresholds c_1 and c_2. In particular if we have some knowledge of B such that we can find a choice of c_2 that makes $\mathbb{P}(B > c_2) = 0$, *we can shift the entire burden of controlling the sum onto A.*

This is precisely what is done in [TLR18]. The remainder process, because of the Lipschitzness of the Hessian (Assumption 2), can be upperbounded. Consequently, we can find a choice of threshold that puts the burden of controlling the escape entirely on the well behaved process coming from the linear term. By doing so and by exploiting standard results for the control of the well behaved process, [TLR18] are able to derive an upper bound on the probability of escape of the diffusion.

Escape of the discrete process With the diffusion handled, the goal now is to relate the probability of escape of (LMC-β) to that (LD-β). Previous work, namely the result of Dalalyan [Dal16] as well as [RRT17], has shown that for η small enough, the behavior of the discrete process is close to the diffusion *evaluated on a grid*. We would like to simply upper bound the probability of escape of the discrete process with the probability of escape of the diffusion. But, since escape for the discrete process can only be related to escape of the diffusion on a grid, it is *weaker* than controlling the escape on a continuous interval. Indeed, saying that a process does not escape when evaluated on a grid does not exclude the possibility of escape *in between* the grid points. Consequently, a further control of the process in between the grid points is necessary. Adding further conditions on η and β, [TLR18] are able to do so to prove their result.

4.4 Is this a compelling theorem ?

In reading the theorem, we might take issue with the freedom we are given in setting the escape time T_{esc}. The theorem allows us to *choose* how long the process remains trapped. It is less of a theorem on how (LMC-β) behaves and more of a result on what we can make (LMC-β) do.

A stronger claim would be one that has β chosen independently of T_{esc}. Indeed, we can first choose β such that the stationary measure of (LD-β) concentrates sufficiently around the minimizers of f. As shown by [RRT17], β only needs to scale polynomially in d. And then ask how long the process with take to explore the modes.

Tzen, Liang and Raginsky allude to such a result in Remark 2 of [TLR18]. The Eyring-Kramers law they mention relates the mean exit time of a particle following the Langevin diffusion from the neighborhood of a local minimum to the *height of the barrier* it has to cross, i.e the difference in function value between the minimum and the saddle point it has to cross to exit. This exit time was first related to the log-Sobolev constant by [MS14]. Their result informally states the following. If the stationary measure admits a log-sobelev constant equal to α, then there is a local minimum for which the average of the exit time τ from the neighborhood of the minimum verifies:

$$\frac{1}{\alpha} \lesssim \mathbb{E}[\tau]$$

We can clearly from this result how small log-Sobolev constants affect the dynamics. For a given β, the LSI constant α_β of the stationary distribution π_β tells us that there is a minimum around which the iterates will be trapped on average for a time longer than $\frac{1}{\alpha_\beta}$. If α_β is exponentially small in the dimension, this implies an exponentially long wait time to transition out of the basin of that local minimum.

5 LMC in the real world

We have seen in the previous section that the price to pay for a small log-Sobolev constant is time exponentially long in the dimension spent around local minimizers. The question now is to figure out where real world distributions lie. Do they live in the intractable realm of exponentially small LSI constants ?

To try to answer this question, we study in this section the work of Song and Ermon [SE19]. Their paper proposes the following method for sampling from real world distributions: first learn the *score*, i.e $\nabla \log(p_\pi)$ from available data using a neural network then sample from

the learnt score using Langevin Monte Carlo. A naive application of this idea yields samples of poor quality and the authors show that, for this approach to work, the perturbation of the data by noise is instrumental.

We discuss their method in detail by first explaining how the score is learnt, then we describe the proposed LMC scheme and finish by arguing that the demonstrated success of [SE19] provides strong support for the view that the real world is tractable.

5.1 Score matching

Score matching consists of learning the score of a distribution π when only having access to independent samples.

A vector field $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is parametrized with a neural network, where θ denotes the network parameters included in some set Θ. This network is then trained to learn the score of the data distribution by minimizing the relative Fisher information

$$\min_{\theta\in\Theta} \mathbb{E}_{X\sim\pi}\left[\|s_\theta(X) - \nabla \log p_\pi(X)\|_2^2\right]. \tag{5}$$

If the minimum 0 is attained for some parameter $\theta^\star$, the equality $s_{\theta^\star} = \nabla \log p_\pi(x)$ will hold π-almost everywhere. The loss (5) is therefore a sensible one. But since it involves the unknown score $\nabla \log p_\pi(x)$, it is unusable in practice.

Hyvärinen [Hyv05] noticed that if p_π is decays sufficiently fast at infinity, such that for any $\theta \in \Theta$, $\lim_{\|x\|_2\to\infty} p(x)s_\theta(x) = 0$, then, a simple integration by parts will yield the equivalent optimization problem

$$\min_{\theta\in\Theta} \mathbb{E}_{X\sim\pi}\left[\mathrm{tr}(J_{s_\theta}(X)) + \|s_\theta(X)\|_2^2\right], \tag{6}$$

where $J_{s_\theta}(x)$ denotes the Jacobian of s_θ at x. Although this is now implementable, the trace of the Jacobian is an expensive quantity to compute. Two methods to alleviate this computational burden are discussed in [SE19].

The first is *sliced score matching*, proposed by Song and his collaborators, replaces the trace by an unbiased estimator $v^T J s_\theta(X) v$ where v is an isotropic random vector, in order to leverage the fast implementation of Jacobian-vector products available in automatic differentiation packages. The second method, better suited for the scores involved in this paper, is *Denoising Score Matching (DSM)* [Vin11]. Unlike sliced score matching, DSM does not estimate the score of the data distribution π but rather the score of data distribution perturbed by Gaussian noise $\pi * \mathcal{N}(0, \sigma^2 I)$. An integration by parts shows that learning scores of perturbed distributions amounts to minimizing the implementable loss

$$\min_{\theta\in\Theta} \mathbb{E}_{X\sim\pi, Z\sim\mathcal{N}(0,\sigma^2 I)}\left[\|s_\theta(X+Z) - \frac{Z}{2\sigma^2}\|_2^2\right].$$

Learning the perturbed score is well suited here because, as we will see next, perturbation is necessary for score matching to work well.

5.2 The need for noise

Given a set of samples from some real world distribution, attempting to minimize (6) is likely to fail.

First, (6) assumes that a score exists. This requires that the data distribution admit a density with respect to the Lebesgue measure that is differentiable and positive over the

entire space. There is no reason to believe this is the case for the distribution of natural images. In fact, a commonly held belief is that the support of this natural distribution is some lower dimensional manifold to which the Lebesgue measure assigns no mass. Therefore, it is unrealistic to expect that this distribution admits density let alone a differentiable one.

Moreover, the loss in (6) is a weighted L_2 loss that enforces score matching in regions of high probability and discounts mismatches in low probability regions. This causes the learnt score to be innacurate in low probability regions. The aim being to apply a random walk algorithm that traverses the space, going from one high probability region to another, low accuracy scores are likely to induce it in error.

Gaussian smoothing can address both these issues. Indeed, convolution by a Gaussian confers all the necessary regularity : first, convolution with a absolutely continuous distribution like the Gaussian guarantees the existence of a density with respect to the Lebesgue measure, second, since the Gaussian is infinitely supported, the density is positive everywhere, and finally the continuous differentiability of the Gaussian density is inherited by the density. Moreover, perturbation with a Gaussian with high enough variance increases the likelihood that samples will land in regions that were originally of low probability. This improves the score matching in those regions.

The addition of noise to the samples is therefore crucial to successfully match the score of the data distribution. It ensures the well-posedness of the problem and improves the matching in low data density regions.

5.3 Learning the score in practice

We have previously established that perturbing the data is necessary. Instead of perturbing the data once, the authors propose to perturb the data with a decreasing sequence of noise scales and learning, jointly, the scores of the decreasingly perturbed distributions. The proposed method, inspired by simulated annealing, consists of first setting a geometrically decreasing sequence of noise scales $\sigma_1 > \cdots > \sigma_L$. A conditional[1] neural network $s_\theta(x, \sigma)$, referred to as a *Noise Conditional Score Network*, is chosen to parametrize the L vector fields that will be matched to the gradient fields of the perturbed distributions.

For each of the noise scales, the loss defined in (6) is denoted $\ell(\theta, \sigma)$. The network is then trained to jointly minimize all those losses by enforcing the minimization of

$$\frac{1}{L}\sum_{i=1}^{L}\lambda(\sigma_i)\ell(\theta, \sigma_i),$$

where $\lambda(\sigma_i)$ is a weighing coefficient set to σ_i^2 to ensure that the products $\lambda(\sigma_i)\ell(\theta, \sigma_i)$ are roughly of the same order for each noise scale σ_i.

To learn the score of an image distribution, the authors recommend using an architecture that is well suited for image classification. In the experiments, a U-Net architecture is chosen to parametrize the vector fields, it is trained with Adam.

The once unknown score of the distribution π is now approximated by a sequence of score networks.

[1]Here, conditional is borrowed from the deep learning literature, and simply means that the input variable is augmented with an additional input to *condition* the output.

5.4 Annealed Langevin Dynamics

With the trained score network in hand, the next step is to sample from the distribution it represents using LMC. Although our ultimate goal is to sample from the data distribution, we know, from the discussion in 5.2, that the best we can do is to sample from the least perturbed distribution.

Following the simulated annealing inspiration, the authors propose that instead of directly attempting to sample from the least perturbed distribution, it is better to use the sequence of scores to progressively warm start LMC at each round. The algorithm proceeds as follows : an initial sample is drawn form a uniform distribution over the hypercube, then LMC is run for T steps to sample from π_{σ_0}, the output is then used to as the initial point for sampling from the next noise level. The pseudo-code is provided in 1. This scheme is a chained sequence of (LMC) iterations.

Algorithm 1 Annealed Langevin dynamics.

Require: $\{\sigma_i\}_{i=1}^{L}, \eta, T$.
1: Initialize $\tilde{\mathbf{x}}_0$
2: **for** $i \leftarrow 1$ to L **do**
3: $\quad \eta_i \leftarrow \eta \cdot \sigma_i^2/\sigma_L^2$ {η_i is the step size for noise level σ_i}
4: $\quad$ **for** $t \leftarrow 1$ to T **do**
5: $\quad\quad$ Draw $z_t \sim \mathcal{N}(0, I)$
6: $\quad\quad \tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \eta_i s_\theta(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{2\eta_i}\, z_t$
7: $\quad$ **end for**
8: $\quad \tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
9: **end for**
10: **return** $\tilde{\mathbf{x}}_T$ =0

5.5 Evaluation metrics

We discuss the two metrics reported in the paper to evaluate the results of the scheme we have just described. The Inception score and the FID score are measures of quality of the generated samples. We briefly describe them to understand the properties they are evaluating.

Inception score [SGZ+16] For a given generative model G, we write $X \sim G$, for the samples generated by G. We define the conditional class distribution $\mathbb{P}_{Y|X} :=$ `Inceptionv3(X)`, where `Inceptionv3` is a convolutional neural network trained for classification on the same dataset G was trained on. It outputs a distribution over the class labels. The Inception score $\mathrm{IS}(G)$ is then defined as

$$\mathrm{IS}(G) = \exp\left(\mathbb{E}_{X\sim G}[\mathrm{KL}(\mathbb{P}_{Y|X}||\mathbb{P}_Y)]\right).$$

FID score An improvement over the Inception score was proposed by [HRU+17]. The *Frechet Inception Distance*(FID) also makes use of a trained `Inceptionv3`. Two statistics, the mean and the variance, of the intermediate features extracted by the network are computed. These two quantities are then used to define a Gaussian random variable. The FID score is taken to be the Wassertein distance between this Gaussian and the one obtained from real data.

According to these measures of success, the score based scheme is a successful one. At the time of publication, they achieved the best inception score on the CIFAR-10 dataset and report FID scores they deem competitive. In addition to these quantitative metrics, the authors provide uncurated images that they have sampled using their method. The images look compelling (to me). A nearest neighbor analysis of a few samples, which consists of looking for the closest image in ℓ_2 norm in the training dataset, reveals that the samples are not merely memorized.

From this we conclude that it is reasonable to state that this scheme is successful in sampling from real image distributions.

5.6 Discussion

The success of the Annealed Langevin dynamics scheme, which is merely a chained sequence of (LMC) iterations, is particularly striking when we take a look at the chosen hyper-parameters. It is with a choice of $T = 100$ and $\eta = 10^{-5}$ that Song and Ermon generated the convincing images in [SE19]. This is a very small number of iterations in view of the convergence bounds of Section 3. In fact, forgetting about the discretization and considering the Langevin diffusion in continuous time, the effective time $t = T \times \eta = 10^{-3}$ is extremely small to yield meaningful convergence towards the target distribution even for the diffusion unless the constants involved are very friendly.

This leads us to the following two options. Either the measures of success of practical image generation schemes, as well as our visual evaluation sample diversity, are deeply flawed and correlate poorly with statistical measures like the KL divergence. Or the constants characterizing natural distributions do not have poor dimension dependencies and we can be sample from them with few iterations.

A sceptical reading would argue that the empirical estimation of the score does not even guarantee that the vector field learnt through score matching is a gradient field and that therefore, we might not be performing proper (LMC) iterations. Moreover, without evaluating the scheme on synthetic data for which the statistical measures of convergence can be tested, the heuristic measures of success can not be fully trusted.

We argue that there is an optimistic view that sees that since approaches that share strong similarities with [SE19], i.e score-based diffusion models, are leading the benchmarks for image generation [noa], there is some indication that constants characterizing distributions in the real world must be well behaved. This view can be bolstered by applying the annealed scheme to synthetic data and studying simple well known settings.

6 Conclusion and open questions

We began with the analysis of (LMC) that relied on the existence of a log-Sobolev constant to derive a convergence bound. We then saw, through inspection of the behavior of (LMC) around local minima, that good dependence on dimension of this constant was necessary for the convergence bound to be meaningful, otherwise, we would only guarantee convergence in a number of steps that rivals the complexity of naive grid based algorithm. Finally, we argued that there may be evidence that the log-Sobolev constants in the real world are well behaved.

The apparent success of score-based generative methods has already generated interest in providing explanations. The paper of [BMRR20] claims that the manifold hypothesis can help explain the success of (LMC) in sampling from images. They argue that dimensionality dependence involved in the dynamics is the dimension of the data manifold and not the

dimension of the ambient space. Since this data manifold is believed to be low dimensional, the fast convergence of (LMC) would follow. Another recent line of work [CCNW21], more theoretical, attempts to derive bounds for (LSI) constants of measures convolved with Gaussian noise. They show that for a probability measure contained inside a ball of radius R, the (LSI) constant will be at least $O(e^{-4R^2/t\sigma^2})$ when convolved with noise of variance σ. Sadly, gray-scale images are contained in a hypercube, and the ball containing it has radius $R = \sqrt{d}$. Therefore their results only yields an exponentially small lower bound, so it is unsatisfactory for our purposes. Justifications for the (perceived) success of score-based generative models have therefore yet to emerge.

In addition to this, there are still a number of gaps left to understand the convergence of (LMC). Of particular interest is whether or not adaptively preconditioned versions of (LMC) converge. For instance, the default implementation of Stochastic Gradient Langevin Dynamics, a variant of (LMC), of TensorFlow is preconditioned with RMSProp. There is to date no convincing proof that this algorithm, or any other adaptively preconditioned scheme (as used in deep learning optimization) works.

A final line of inquiry that we can pursue from here are the questions surrounding the behavior of Stochastic Gradient Descent. The iterates of SGD loosly ressemble those of (LMC) except that they weigh the gradient more heavily than the noise contrary to (LMC). The same tools we reviewed in this report, like using a continuous time diffusion, invoking standard results like the Fokker-Planck equation and linearizing around a minimum, have been used to prove, for instance, that the empirical observation that parameter-dependent noise is better than isotropic noise for finding flat minima in [XSS20].

The danger here in wanting to use these methods to prove further results on SGD is that the diffusion approximation and the abstraction afforded by theory can lead to spurious results. The very recent paper [ZLU21], providing a nice set of counter examples to beliefs about SGD, claims to show that on the real line, a diffusion approximation of SGD can be made to have *any* stationary measure, for *any* loss function f as long as we get to choose the noise covariance. They further conjecture that this is true in higher dimensions as well. Consequently, unless the noise is accurately modelled, it is possible to obtain misleading results. Establishing which types of questions on SGD the continuous time methods reviewed in this report can answer and which ones they cannot is an interesting direction for future research.

References

[ABC+00] Cécile Ané, Sébastien Blachère, Djalil Chafai, Pierre Fougères, Ivan Gentil, Florent Malrieu, Cyril Roberto, and Grégory Scheffer. *Sur les inégalités de Sobolev logarithmiques*. Société Mathématique de France, 2000.

[BBCG08] Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13(none):60–66, January 2008. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. Markov Semigroups. In Dominique Bakry, Ivan Gentil, and Michel Ledoux, editors, *Analysis and Geometry of Markov Diffusion Operators*, Grundlehren der mathematischen Wissenschaften, pages 3–75. Springer International Publishing, Cham, 2014.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[BMRR20] Adam Block, Youssef Mroueh, Alexander Rakhlin, and Jerret Ross. Fast Mixing of Multi-Scale Langevin Dynamics under the Manifold Hypothesis. *arXiv:2006.11166 [cs, stat]*, June 2020. arXiv: 2006.11166.

[CCAY+20] Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv:1805.01648 [cs, math, stat]*, July 2020. arXiv: 1805.01648.

[CCNW21] Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-Sobolev inequalities for mixture distributions. *arXiv:2102.11476 [math]*, March 2021. arXiv: 2102.11476 version: 2.

[CGW10] Patrick Cattiaux, Arnaud Guillin, and Li-Ming Wu. A note on Talagrand's transportation inequality and logarithmic Sobolev inequality. *Probability Theory and Related Fields*, 148(1):285–304, September 2010.

[Dal14] Arnak Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. Working Paper 2014-45, Center for Research in Economics and Statistics, December 2014.

[Dal16] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *arXiv:1412.7392 [math, stat]*, December 2016. arXiv: 1412.7392.

[Eck] Roger Eckhardt. Stan Ulam, John von Neuman and the Monte Carlo method. *Monte Carlo*, page 11.

[Gro75] Leonard Gross. Logarithmic Sobolev Inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975. Publisher: Johns Hopkins University Press.

[HRU+17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Hyv05] Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[Li21] Mufan (Bill) Li. On Escape Time, Lyapunov Function, Poincaré Inequality, and the KLS Conjecture Beyond Convexity, January 2021.

[MCJ+19] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, October 2019. ISBN: 9781820003112 Publisher: National Academy of Sciences Section: Physical Sciences.

[MFWB19] Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. Improved Bounds for Discretization of Langevin Diffusions: Near-Optimal Rates without Convexity. *arXiv:1907.11331 [math, stat]*, November 2019. arXiv: 1907.11331.

[MS14] Georg Menz and André Schlichting. POINCARÉ AND LOGARITHMIC SOBOLEV INEQUALITIES BY DECOMPOSITION OF THE ENERGY LANDSCAPE. *The Annals of Probability*, 42(5):1809–1884, 2014. Publisher: Institute of Mathematical Statistics.

[noa] Papers with Code - Image Generation.

[OV00] F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400, June 2000.

[Pav14] Grigorios A. Pavliotis. The Fokker–Planck Equation. In Grigorios A. Pavliotis, editor, *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, Texts in Applied Mathematics, pages 87–137. Springer, New York, NY, 2014.

[PVBL+20] Loucas Pillaud-Vivien, Francis Bach, Tony Lelièvre, Alessandro Rudi, and Gabriel Stoltz. Statistical Estimation of the Poincaré constant and Application to Sampling Multimodal Distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 2753–2763. PMLR, June 2020. ISSN: 2640-3498.

[RDF78] P. J. Rossky, J. D. Doll, and H. L. Friedman. Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, November 1978. Publisher: American Institute of Physics.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, June 2017. ISSN: 2640-3498.

[RT96] Gareth O. Roberts and Richard L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability.

[SE19] Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[SFR10] Chris Sherlock, Paul Fearnhead, and Gareth O. Roberts. The Random Walk Metropolis: Linking Theory and Practice Through a Case Study. *Statistical Science*, 25(2):172–190, May 2010. Publisher: Institute of Mathematical Statistics.

[SGZ+16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[TLR18] Belinda Tzen, Tengyuan Liang, and Maxim Raginsky. Local Optimality and Generalization Guarantees for the Langevin Algorithm via Empirical Metastability. In *Conference On Learning Theory*, pages 857–875. PMLR, July 2018. ISSN: 2640-3498.

[Vin11] Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, July 2011. Conference Name: Neural Computation.

[VW19] Santosh Vempala and Andre Wibisono. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[XSS20] Zeke Xie, Issei Sato, and Masashi Sugiyama. A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima. September 2020.

[ZLU21] Liu Ziyin, Botao Li, and Masahito Ueda. SGD May Never Escape Saddle Points. *arXiv:2107.11774 [cs, math, stat]*, July 2021. arXiv: 2107.11774.

[Øk03] Bernt Øksendal. Stochastic Differential Equations. In Bernt Øksendal, editor, *Stochastic Differential Equations: An Introduction with Applications*, Universitext, pages 65–84. Springer, Berlin, Heidelberg, 2003.