Non-Smooth Optimization via Sampling

Nahom Seyoum nahom.seyoum@yale.edu
Department of Statistics and Data Science
Yale University

Haoxiang You haoxiang.you@yale.edu
Department of Mechanical Engineering and Material Science
Yale University

Editor: Seyoum, You

Keywords: Non-Smooth Optimization, Langevin Dynamics, Stochastic Opt, Sampling

1 Introduction

Many applications in science, engineering, finance, etc necessitate solving non-convex optimization problems of the form:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x), \tag{1}$$

where $f(x)$ is non-convex and non-smooth. The goal is to identify its global minimizer or, at the very least, obtain a high-quality local one.

We aim to solve the non-convex optimization via sampling algorithms. The idea is inspired by recent progress in the sampling algorithms and the relationship between sampling and optimization (Wibisono, 2018; Lee et al., 2021; Chen et al., 2022), where these works provide insights for analysis sampling algorithm from an optimization perspective. We look at the problem the other way around, where we view the optimization problem as the problem of drawing samples from the distribution $\nu(x) \propto e^{-f(x)}$ induced by optimization problem (1). Intuitively, $\nu(x)$ assigns high probability density to regions where $f(x)$ is small, making samples drawn from $\nu(x)$ likely to minimize $f(x)$ (See Figure 1).

To concretize how non-convex optimization may arise in practice, we provide a sequential decision-making problem in example 1. Then, in the upcoming section, we'll outline initial ideas for convergence analysis on these sampling algorithms that apply to (non-convex) optimization problems. Finally, we will close this article with a short discussion section.

Example 1 *We consider a trajectory optimization problem*

$$\begin{array}{ll} \underset{s_i,u_i}{\textit{minimize}} & \sum_{i=0}^{T-1} l(s_i,u_i) + l_T(s_T) \\ \textit{subject to} & s_{i+1} = g(s_i,u_i) \\ & s_i = a_0, \end{array} \tag{2}$$

where $s_k \in \mathbb{R}^n$ represents the states, $u_k \in \mathbb{R}^m$ denotes the control inputs, a_0 is the initial state, $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ represents the dynamics, and $l : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$ along with

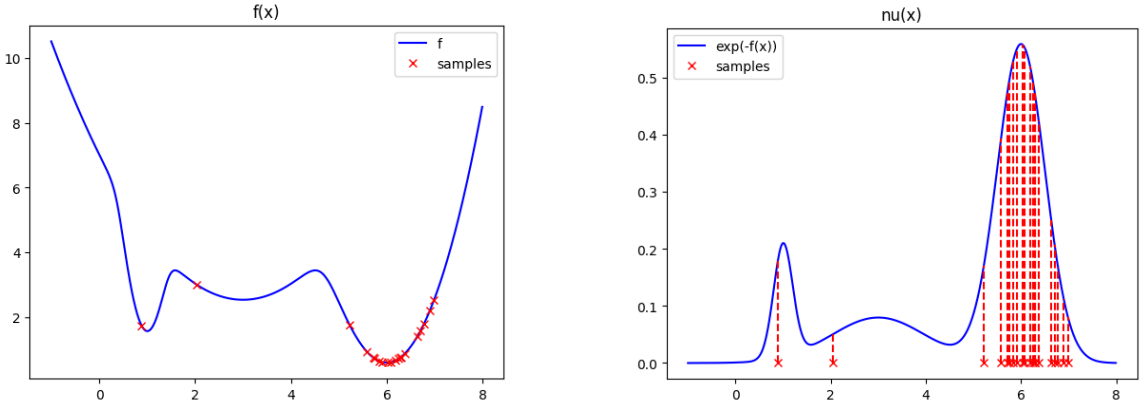

Figure 1: Samples draw from ν is likely to around global minimizer of f(x)

$l_T : \mathbb{R}^n \to \mathbb{R}_+$ *define our objective. The problem is non-convex when the dynamics* g *is non-linear.*

One approach to tackle problem (2) *is through shooting, where the problem is reformulated as an optimization problem solely in the sequence of control inputs, denoted as* $U = [u_0^\top, \ldots, u_{T-1}^\top]^\top \in \mathbb{R}^{m(T-1)}$. *In this method, the states* s_k *are computed explicitly as* $s_k = g(g(g(\ldots, u_{k-3}), u_{k-2}), u_{k-1})$, *and the objective function is defined as* $J(U) = \sum_{i=0}^{T-1} l(s_i, u_i) + l_T(s_T)$, *which give us the non-convex optimization problem*

$$\underset{U \in \mathbb{R}^{m(T-1)}}{\textit{minimize}} \quad J(U) \tag{3}$$

2 Theoretical Analysis

In this section, we aim to conduct a theoretical analysis of optimization algorithms within the sampling framework. Our analysis will be divided into two parts:

1. Determining the number of iterations needed to run a sampling algorithm to obtain a sample X with distribution $\nu \propto e^{-f(x)}$.

2. Assessing the probability that a sample X drawn from ν is close to the global minimizer of $f(x)$.

For simplicity, we'll assume that the optimization problem (1) has a unique global minimizer x^* and that the objective function f is smooth. We'll focus exclusively on the Markov chain Monte Carlo (MCMC) sampling algorithm within the scope of this discussion.

2.1 Background on Langevin Dynamics

Here we introduce one MCMC sampling algorithm we considered in this article: Langevin dynamics(LD). It's worth noting that several other sampling algorithms, such as Hamiltonian Monte Carlo, Gibbs sampling, and Metropolis-Hastings, could also be suitable. We selected LD as a concrete example for our analysis, but the good choice of sampling algorithm for a particular nonconvex optimization problem may vary in practice.

The LD is defined by following the stochastic differential equation(SDE)

$$dX_t = \underbrace{-\nabla f(X_t)dt}_{\text{gradient flow}} + \underbrace{\sqrt{2}dW_t}_{\text{Brownian motion}} \quad . \tag{4}$$

Since LD can be viewed as "a gradient flow + noise", it is no wonder why we can potentially use it for optimization.

To sample from the target distribution $\nu(x)$ using LD, we initialize a random variable from an arbitrary distribution μ_0 (often chosen as the standard Gaussian distribution): $X_0 \sim \mu_0$. We then simulate the LD (4) forward to obtain X_t with distribution μ_t. As we progress in time, the distribution μ_t approximates the target distribution $\nu(x)$.

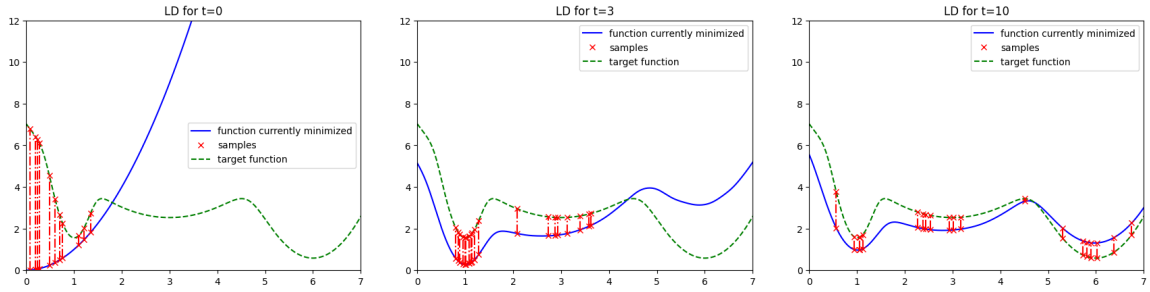

Figure 2: The objective function to minimize involves along LD: Initially, the samples X_0 are drawn from a standard Gaussian distribution $\mu_0 \propto e^{-x^2}$, implying that with high probability, X_0 minimizes the quadratic function $h_0(x) = x^2$. Then, as the LD progresses, the distribution of X_t, denoted as $\mu_t \propto e^{-h_t(x)}$, approaches the target distribution ν. Equivalently, we can say that the objective function $h_t(x)$, which is minimized by X_t with high probability, gradually approaches the target objective $f(x)$.

We can interpret the process of LD as follows: we start from samples $X_0 \sim \mu_0 \propto e^{-h_0(x)}$ where the $h_0(x)$ is relatively easy to optimize, and as we progress, $X_t \sim \mu_t \propto e^{-h_t(x)}$ evolves to minimize $h_t(x)$, which gradually becomes closer to our target objective $f(x)$. Eventually, X_t converges to a point where it is highly likely to minimize $f(x)$. Figure 2 illustrates this idea of the objective function shifting along LD.

We gauge the progress of LD by measuring the Kullback-Leibler (KL) divergence between the induced density function $\mathsf{KL}(\mu_t|\nu)$. In the next subsection, we demonstrate that if we allow LD to run for a sufficiently long time, the samples will essentially follow the distribution $\nu(x)$, i.e., $X_t \sim \nu(x)$, regardless of the form of $f(x)$. Additionally, we show that when the target distribution ν satisfies logarithmic Sobolev inequalities (LSI), the converge rate is linear in terms of KL divergence.

2.2 Convergence of LD

Here we provide a convergence analysis of the LD algorithm, we show the KL divergence between the distribution of $X_t \sim \mu_t$ and the target distribution $\nu \propto e^{-f(x)}$ is always decreasing along LD (with some minor regularity conditions). Moreover, when our target ν satisfies α-LSI condition, the KL divergence will decrease linearly over time.

Below we outline several important lemmas that are instrumental in proving our main results.

Lemma 1 *Let $\mu_t \in \mu(\mathbb{R}^d)$ be the distribution of $X_t \in \mathbb{R}^d$. If X_t envolves following the LD (4), then the density μ_t envolves following the Fokker-Planck equation, which is the following partial differential equations:*

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t \nabla f) + \Delta \mu_t \tag{5}$$

Here, $\nabla \cdot (\mu_t \nabla f)$ is the divergence of the vector fields $\mu_t \nabla f$ and $\Delta \mu_t$ is the Laplacian of distribution μ_t.

Recall for any vector field $b(x) = [b_1(x), \ldots, b_d(x)]^\top \in \mathbb{R}^d$, the divergence of b is given by

$$(\nabla \cdot b) = \sum_{i=1}^{d} \frac{\partial b_i}{\partial x_i}. \tag{6}$$

And for any scalar function $\phi(x) \in \mathbb{R}$, the Laplacian is given by

$$\Delta \phi(x) = \sum_{i=1}^{d} \frac{\partial^2 \phi(x)}{\partial x^2} = \nabla \cdot (\nabla \phi). \tag{7}$$

We won't delve into proving the Fokker-Planck equation in this article, as there are numerous online resources available for that purpose. Instead, we provide the following Corollary from the Lemma 1.

Corollary 2 *The target distribution $\nu \propto e^{-f(x)}$ is the stationary distribution of LD.*

Proof The Fokker-Planck equation (5) can be written as

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot \left(\mu_t \nabla \log \frac{\mu_t}{\nu} \right). \tag{8}$$

To see so, recall we can write $\nabla f = -\nabla \log \nu$ and $\Delta \mu_t = \nabla \cdot (\nabla \mu_t) = \nabla \cdot (\mu_t \nabla \log \mu_t)$. Then we can develop Fokker-Planck equation (5) as

$$\begin{aligned} \frac{\partial \mu_t}{\partial t} &= -\nabla \cdot (\mu_t \nabla \log \nu) + \nabla \cdot (\mu_t \nabla \log \mu_t) \\ &= \nabla \cdot (\mu_t (\nabla \log \mu_t - \nabla \log \nu)) \\ &= \nabla \cdot \left(\mu_t \nabla \log \frac{\mu_t}{\nu} \right). \end{aligned} \tag{9}$$

Then for any x, if we substitute $\mu_t(x) = \nu(x)$, we have $\frac{\partial \mu_t}{\partial t}(\nu(x)) = \nabla \cdot \left(\nu(x) \nabla \log \frac{\nu(x)}{\nu(x)} \right) = 0$. ■

Now, we want to show for any $\mu_t \neq \nu$, along the LD, we have the KL divergence always decreasing. This is given the de Bruijn's Identity:

Theorem 3 *(de Bruijn's Identity) If for any μ_t, we have that μ_t and ν are differentiable and vanish at infinity: $\mathbb{E}_{\mu_t}[|\nabla \log \mu_t(x)|] < \infty$, $\mathbb{E}_{\nu}[|\nabla \log \nu_t(x)|] < \infty$, and $\lim_{|x|\to\infty}[\mu_t(x)\nabla\nu(x)] = 0$, then along the LD, we have the KL divergence between the distribution μ_t and the target ν involve as the negative relative Fisher information:*

$$\frac{d\mathsf{KL}(\mu_t\|\nu)}{dt} = -\mathsf{FI}(\mu_t\|\nu) \tag{10}$$

Proof Recall the KL divergence between μ_t and ν is given by

$$\mathsf{KL}(\mu_t\|\nu) = \mathbb{E}_{\mu_t}[\log \frac{\mu_t}{\nu}] = \int_{\mathbb{R}^d} \mu_t(x) \log \frac{\mu_t(x)}{\nu(x)} dx \tag{11}$$

We prove de Bruijn's Identity by the Fokker-Planck Equation and integration by parts:

$$\begin{aligned}
\frac{d\mathsf{KL}(\mu_t\|\nu)}{dt} &= \frac{\partial \int_{\mathbb{R}^d} \mu_t(x) \log \frac{\mu_t(x)}{\nu(x)} dx}{\partial t} \\
&= \int_{\mathbb{R}^d} \frac{\partial \mu_t}{\partial x}(x) \log \frac{\mu_t(x)}{\nu(x)} dx + \underbrace{\int_{\mathbb{R}^d} \mu_t(x) \frac{1}{\mu_t(x)} \frac{\partial \mu_t}{\partial t}(x) dx}_{=\frac{\partial \int_{\mathbb{R}^d} \mu_t(x) dx}{\partial t} = \frac{\partial}{\partial t} 1 = 0} \text{ (By chain rule)} \\
&= \int_{\mathbb{R}^d} \nabla \cdot (\mu_t \nabla \log \frac{\mu_t}{\nu}) \log \frac{\mu_t}{\nu} dx \text{ (By Fokker-Planck equation)} \\
&= -\int_{\mathbb{R}^d} \mu_t \langle \nabla \log \frac{\mu_t}{\nu}, \nabla \log \frac{\mu_t}{\nu} \rangle dx \text{ (By integration by parts)} \qquad (12) \\
&= -\mathsf{FI}(\mu_t\|\nu) \text{ (By definition of relative Fisher information).} \qquad (13)
\end{aligned}$$

■

Given that the relative Fisher information is always positive when $\mu_t \neq \nu$, the KL between μ_t and ν consistently decreases along LD trajectory. As ν is the only stationary distribution, running LD for an adequate duration ensures the convergence of the X_t distribution towards ν. We will delve into the speed of this convergence, or the LD algorithm's mixing time, focusing particularly on the target distribution ν that α-LSI.

Definition 4 *We say ν satisfies the Log-Sobolev Inequality(LSI) with constant $\alpha > 0$ if for any distribution μ the following holds:*

$$\mathsf{FI}(\mu\|\nu) \geq 2\alpha \mathsf{KL}(\mu\|\nu). \tag{14}$$

If our target distribution ν is α-log concave, or equivalently if $f(x)$ is α-strongly convex, it satisfies the α-LSI condition with the same constant alpha. However, the α-LSI family encompasses a broader range of distributions than just the alpha-log concave ones. For instance, the mixture of Gaussian distribution depicted in Figure 1 also falls within the LSI family, despite the fact that the $f(x)$ function that induces this distribution is clearly non-convex. We won't delve into the conditions under which the distribution induced by $\nu(x)$ satisfies α-LSI or how to determine the constant α. However, we'll outline the convergence rate under LD if the target distribution ν does satisfy the α-LSI condition.

Theorem 5 *If ν satisfying α-LSI, then along LD for $X_t \sim \mu_t$, for all $t \geq 0$:*

$$\mathsf{KL}(\mu_t \| \nu) \leq e^{-2\alpha t} \mathsf{KL}(\mu_0 \| \nu). \tag{15}$$

Proof By Theorem 3 and α-LSI, we have

$$\frac{d}{dt}\mathsf{KL}(\mu_t \| \nu) = -\mathsf{FI}(\mu_t \| \nu) \leq -2\alpha \mathsf{KL}(\mu_t \| \nu). \tag{16}$$

Then by Gronwall's lemma, we conclude:

$$\mathsf{KL}(\mu_t \| \nu) \leq e^{-2\alpha t} \mathsf{KL}(\mu_0 \| \nu). \tag{17}$$

■

Under the KL Divergence, it is not awfully difficult to think of the iteration needed to approximate the target distribution. We simply need to set a threshold ϵ for the KL divergence:

$$KL(\mu_t \parallel \nu) \leq \epsilon$$

Using the exponential decay result:

$$KL(\mu_0 \parallel \nu) e^{-2\alpha t} \leq \epsilon$$

Solving for t:

$$t \geq \frac{1}{2\alpha} \log\left(\frac{KL(\mu_0 \parallel \nu)}{\epsilon}\right)$$

2.3 Smoothness and Convexity

As could have been observed, the above results make certain assumptions about the smoothness and convexity of the functions concerned. The goal in the next section is to relax some of the assumptions and find the iterations required analytically. One way this assumption can be relaxed is use the Total Variation as a measure of distance as opposed to the Kullback–Leibler divergence.

Therefore, in this section, we aim to show that the LD Markov chain converges to the target distribution ν in TV distance. The proof involves demonstrating that the transition kernel is contractive in TV distance and then using this property to show convergence.

Definition of TV Distance

The TV distance between two probability measures μ and ν on a measurable space $(\mathbb{R}^d, \mathcal{B})$ is defined as:

$$d_{TV}(\mu, \nu) = \frac{1}{2} \int_{\mathbb{R}^d} |\mu(x) - \nu(x)| \, dx$$

Contractivity in TV Distance

For a Markov chain with transition kernel P, the TV distance after one step is given by:

$$d_{TV}(\mu P, \nu P) \leq \gamma d_{TV}(\mu, \nu)$$

where $\gamma < 1$ is a contraction coefficient.

The transition kernel P for the Langevin Dynamics with step size η and drift term $-\nabla f(x)$ is:

$$P(x, dy) = N(x - \eta \nabla f(x), 2\eta I)\, dy$$

Here, $N(m, \Sigma)$ denotes the normal distribution with mean m and covariance Σ.

Contractivity of the Transition Kernel

To show that the transition kernel P is contractive in TV distance, we use the fact that Gaussian smoothing via the transition kernel reduces the TV distance between two distributions. Specifically, if we have two initial distributions μ and ν, the application of the transition kernel can be viewed as convolving these distributions with a Gaussian kernel.

Consider the TV distance between the distributions after one step of the Markov chain:

$$d_{TV}(\mu P, \nu P) \leq \int_{\mathbb{R}^d} d_{TV}(N(x - \eta \nabla f(x), 2\eta I), N(y - \eta \nabla f(y), 2\eta I))\, d_{TV}(\mu(dx), \nu(dy))$$

Using properties of the Gaussian distribution, we can bound this distance by:

$$d_{TV}(N(x - \eta \nabla f(x), 2\eta I), N(y - \eta \nabla f(y), 2\eta I)) \leq \exp\left(-\frac{\|x - y\|^2}{8\eta}\right)$$

Thus, the TV distance after one step satisfies:

$$d_{TV}(\mu P, \nu P) \leq \int_{\mathbb{R}^d} \exp\left(-\frac{\|x - y\|^2}{8\eta}\right) d_{TV}(\mu(dx), \nu(dy))$$

Since the exponential term is less than 1, we can set $\gamma = \exp\left(-\frac{\|x-y\|^2}{8\eta}\right)$ and obtain:

$$d_{TV}(\mu P, \nu P) \leq \gamma d_{TV}(\mu, \nu)$$

where $\gamma < 1$.

Convergence and Iteration

By iterating the contractive property, we have:

$$d_{TV}(\mu_{k+1}, \nu) \leq \gamma d_{TV}(\mu_k, \nu)$$

Applying this recursively, we get:

$$d_{TV}(\mu_t, \nu) \leq \gamma^t d_{TV}(\mu_0, \nu)$$

To achieve $d_{TV}(\mu_t, \nu) \leq \epsilon$, we need:

$$\gamma^t d_{TV}(\mu_0, \nu) \leq \epsilon$$

Solving for t, we obtain:

$$t \geq \frac{\log(\epsilon) - \log(d_{TV}(\mu_0, \nu))}{\log(\gamma)}$$

Since $\gamma < 1$, $\log(\gamma)$ is negative, and taking the negative reciprocal gives a positive bound:

$$t \geq \frac{\log(d_{TV}(\mu_0, \nu)) - \log(\epsilon)}{-\log(\gamma)}$$

This demonstrates that the Langevin Dynamics algorithm converges to the target distribution ν in TV distance, with the number of iterations t depending logarithmically on the initial TV distance and the desired accuracy ϵ.

2.4 Probability of Samples to be Global Minimum

In the previous section, we discussed how long we need to run a sampling algorithm to obtain a sample X that follows the distribution $\nu(x) \propto e^{-f(x)}$. In this section, we'll explore what happens if we indeed obtain a sample from $\nu(x)$ and what the probability is that it's close to the global minimizer, denoted as $\Pr(\|X - x^*\|_2^2 \leq \epsilon^2)$. This analysis will help us determine the number of i.i.d samples we need to draw from $\nu(x)$ to ensure that there exists a sample in our set of i.i.d samples that lie within an ϵ-ball of the global minimizer with high probability defined by users. For simplicity, in this subsection, we focus on one-dimensional cases, i.e. $X \in \mathbb{R}$. However, the techniques and analysis used can naturally scale to the general case, i.e. $X \in \mathbb{R}^d$.

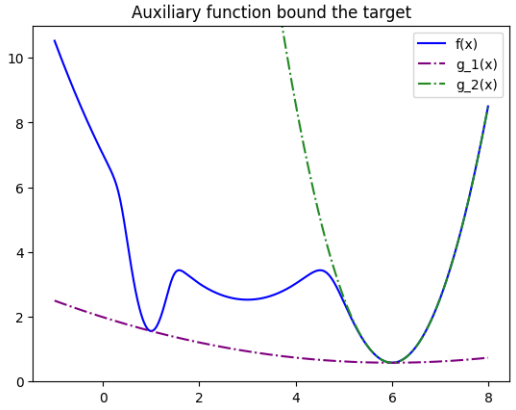

Figure 3: Auxiliary function upper/lower bound the f(x)

Our primary approach for establishing the upper and lower bound of $\Pr(|X - x^*|_2^2 \leq \epsilon^2)$ involves identifying an auxiliary function $g_1(x)$ and $g_2(x)$ such that $f(x) \geq g_1(x)$ and $f(x) \leq g_2(x)$ for all x, and the distributions $\rho_1(x) \propto e^{-g_1(x)}$, $\rho_2(x) \propto e^{-g_2(x)}$ has easy to compute cumulative density function (CDF). Figure 3 provides an example of the quadratic

auxiliary functions that lower/upper bounds the target $f(x)$. We denote $Z_\nu = \int_{\mathbb{R}} e^{-f(x)}dx$, $Z_{\rho_1} = \int_{\mathbb{R}} e^{-g_1(x)}dx$ and $Z_{\rho_2} = \int_{\mathbb{R}} e^{-g_2(x)}dx$. By construction, we have $e^{-g_2(x)} \leq e^{-f(x)} \leq e^{-g_1(x)}$ and $Z_{\rho_2} \leq Z_\nu \leq Z_{\rho_1}$.

Theorem 6 *Assume there exists a constant $\sigma > 0$ such that $f(x) \geq f(x^*) + \frac{\sigma}{2}\|x - x^*\|_2^2$. Then, we have the probability of the random variable X drawn from the distribution $\nu \propto e^{-f(x)}$ within the ϵ-ball of the global minimizer x^* upper bounded by:*

$$\Pr(\|X - x^*\| \leq \epsilon) \leq \frac{Z_{\rho_1}}{Z_\nu}(\phi(\sqrt{\sigma}\epsilon) - \phi(-\sqrt{\sigma}\epsilon)), \tag{18}$$

where $Z_\nu = \int_{\mathbb{R}} e^{-f(x)}dx$, $Z_{\rho_1} = \int_{\mathbb{R}} e^{-g_1(x)}dx$, with $g_1(x) = f(x^*) + \frac{\sigma}{2}\|x - x^*\|_2^2$ and $\phi(x)$ is the CDF of standard Gaussian distribution.

Note our condition does not require $f(x)$ to be convex, although it automatically satisfies the condition when $f(x)$ is σ-strongly convex.

Proof Noticing the distribution $\rho_1 \propto e^{-g_1(x)}$ is a Gaussian distribution with mean x^* and variance $1/\sigma$. Therefore, its density function can be explicitly written as

$$\frac{e^{-g_1(x)}}{Z_{\rho_1}} = \frac{\sqrt{\sigma}}{\sqrt{2\pi}} e^{-\frac{\sigma\|x-x^*\|^2}{2}} \tag{19}$$

We can upper bound the probability of the random variable $X \sim \nu(x) \propto e^{-f(x)}$ be ϵ closed to the global minimizer by assuming the random variable drawn from ρ_1:

$$\begin{aligned}
\Pr(\|X - x^*\| \leq \epsilon) =& \frac{1}{Z_\nu}\int_{x^*-\epsilon}^{x^*+\epsilon} e^{-f(x)}dx \\
\leq& \frac{1}{Z_\nu}\int_{x^*-\epsilon}^{x^*+\epsilon} e^{-g_1(x)}dx \\
=& \frac{Z_{\rho_1}}{Z_\nu}\int_{x^*-\epsilon}^{x^*+\epsilon} \frac{e^{-g_1(x)}}{Z_{\rho_1}}dx \\
=& \frac{Z_{\rho_1}}{Z_\nu}\int_{x^*-\epsilon}^{x^*+\epsilon} \frac{\sqrt{\sigma}}{\sqrt{2\pi}} e^{-\frac{\sigma\|x-x^*\|^2}{2}}dx \\
=& \frac{Z_{\rho_1}}{Z_\nu}\int_{-\sqrt{\sigma}\epsilon}^{\sqrt{\sigma}\epsilon} \frac{\sqrt{\sigma}}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} d(\frac{y}{\sqrt{\sigma}} + x^*) \\
=& \frac{Z_{\rho_1}}{Z_\nu}(\int_{-\infty}^{\sqrt{\sigma}\epsilon} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}dy - \int_{-\infty}^{-\sqrt{\sigma}\epsilon} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}dy) \\
=& \frac{Z_{\rho_1}}{Z_\nu}(\phi(\sqrt{\sigma}\epsilon) - \phi(-\sqrt{\sigma}\epsilon)),
\end{aligned} \tag{20}$$

■

Theorem 7 *Assume the $f(x)$ is L-smooth, then we have the probability of the random variable X drawn from the distribution $\nu \propto e^{-f(x)}$ within the ϵ-ball of the global minimizer x^* lower bounded by:*

$$\Pr(\|X - x^*\| \leq \epsilon) \geq \frac{Z_{\rho_2}}{Z_\nu}(\phi(\sqrt{L}\epsilon) - \phi(-\sqrt{L}\epsilon)), \tag{21}$$

where $Z_{\rho_2} = \int_{\mathbb{R}} e^{-g_2(x)}dx$, with $g_2(x) = f(x^*) + \frac{L}{2}\|x - x^*\|_2^2$.

By smoothness, we have $f(x) \leq f(x^*) + \frac{L}{2}\|x - x^*\|_2^2, \forall x$. Then, by similar techniques, we applied to prove Theorem 6 we can obtain the results.

With Theorem 6 and Theorem 7, we can compute the number of samples needed from the distribution ν to ensure that the probability of at least one sample lying within an ϵ-ball of x^* exceeds a given threshold δ.

Theorem 8 *Assume we have i.i.d a set samples $\mathcal{X} = \{X_1, X_2, \dots\}$ drawn from the distribution $\nu \propto e^{-f(x)}$ and our target objective $f(x)$ satisfies the conditions of Theorems 6 and 7. Then, to guarantee that the probability of at least one sample lying within an ϵ-ball of x^* is greater than δ, i.e.,*

$$\Pr(\exists X_i \in \mathcal{X}, \text{ such that } \|X_i - x^*\| \leq \epsilon) \geq \delta, \tag{22}$$

the minimum number of samples we need to draw, denoted as N, is within the range

$$\frac{\ln(1-\delta)}{\ln(1 - \frac{Z_{\rho_2}}{Z_\nu}(\phi(\sqrt{L}\epsilon) - \phi(-\sqrt{L}\epsilon))} \leq N \leq \frac{\ln(1-\delta)}{\ln(1 - \frac{Z_{\rho_1}}{Z_\nu}(\phi(\sqrt{\sigma}\epsilon) - \phi(-\sqrt{\sigma}\epsilon))} \tag{23}$$

Proof To simplify our analysis, we assume there exists an integer number N, such that $\Pr(\exists X_i \in \mathcal{X}, \text{ such that } \|X_i - x^*\| \leq \epsilon) = \delta$, when $\mathcal{X}$ contains N samples.

The probability of at least one sample lying within an ϵ-ball of x^* equals to the probability of all samples are outside the ϵ-ball of x^*, then we have

$$\begin{gathered}
\Pr(\forall X_i \in \mathcal{X}, \|X_i - x^*\| > \epsilon) = 1 - \Pr(\exists X_i \in \mathcal{X}, \text{ such that } \|X_i - x^*\| \leq \epsilon) \\
\Pr(\forall X_i \in \mathcal{X}, \|X_i - x^*\| > \epsilon) = 1 - \delta \\
(1 - \Pr(\|X - x^*\| \leq \epsilon))^N = 1 - \delta \quad \text{(By i.i.d samples)} \\
N = \frac{\ln(1-\delta)}{\ln(1 - \Pr(\|X - x^*\| \leq \epsilon))}
\end{gathered} \tag{24}$$

Substituting the bounds for $\Pr(\|X - x^*\| \leq \epsilon))$ from Theorem 6 and Theorem 7 completes the proof. ■

2.5 Extension to Non-smooth functions

To extend the analysis for non-smooth cases, we need to address the challenge of non-smoothness in the objective function $f(x)$. We will use auxiliary functions that approximate the non-smooth function from above and below, providing bounds for the probability that a sample drawn from the distribution $\nu \propto e^{-f(x)}$ is close to the global minimizer x^*.

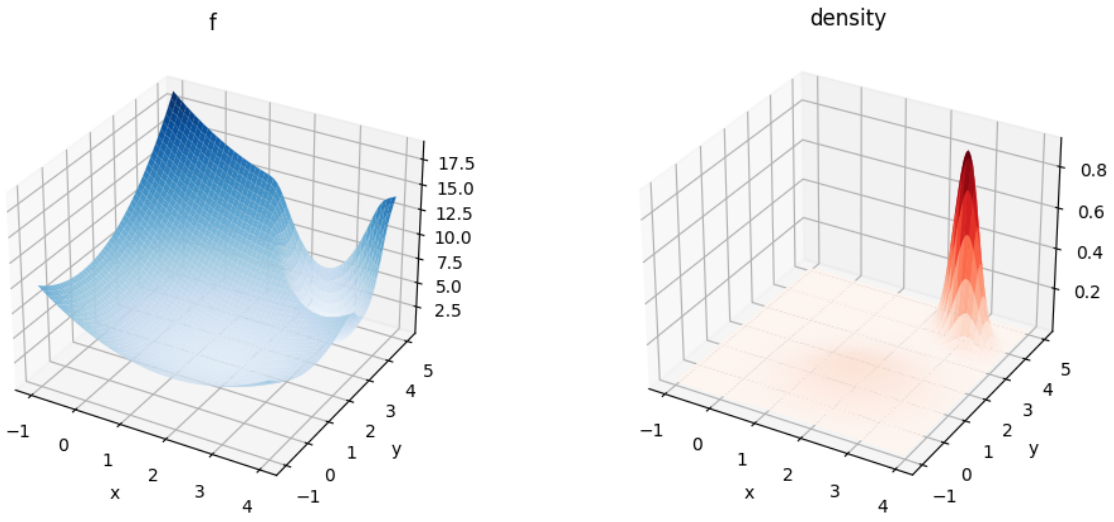

Figure 4: Objective function with one local minimal at (2,1) and one global minimal at (3,4)

Setup

Given the non-smooth function $f(x)$, we assume there exist auxiliary functions $g_1(x)$ and $g_2(x)$ such that:

$$f(x) \geq g_1(x) \quad \text{and} \quad f(x) \leq g_2(x)$$

for all x. These auxiliary functions are chosen to be smooth, making the resulting distributions $\rho_1(x) \propto e^{-g_1(x)}$ and $\rho_2(x) \propto e^{-g_2(x)}$ easier to work with analytically. The idea is to bound the probability that a sample drawn from $\nu \propto e^{-f(x)}$ is close to the global minimizer x^* using these smoother approximations.

2.5.1 Existence of Quadratic Bounds

One crucial assumption is the existence of quadratic bounds on the function $f(x)$. Specifically, we assume there exist constants $\sigma > 0$ and $L > 0$ such that:

$$f(x) \geq f(x^*) + \frac{\sigma}{2}\|x - x^*\|^2$$

$$f(x) \leq f(x^*) + \frac{L}{2}\|x - x^*\|^2$$

for all x. These conditions imply that the function $f(x)$ is bounded below by a strongly convex function and above by a smooth quadratic function. The lower bound ensures that $f(x)$ has a steep enough slope near the global minimizer, while the upper bound ensures that $f(x)$ does not grow too rapidly, providing control over the behavior of the function.

Theorem 6 for Non-Smooth Case

Assume there exists a constant $\sigma > 0$ such that:

$$f(x) \geq f(x^*) + \frac{\sigma}{2}\|x - x^*\|^2$$

Then, we have the upper bound for the probability that a sample from $\nu \propto e^{-f(x)}$ lies within an ϵ-ball of the global minimizer x^*:

$$\Pr(\|X - x^*\| \leq \epsilon) \leq \frac{Z_{\rho_1}}{Z_\nu}(\Phi(\sqrt{\sigma}\epsilon) - \Phi(-\sqrt{\sigma}\epsilon))$$

where $Z_\nu = \int e^{-f(x)}\,dx$, $Z_{\rho_1} = \int e^{-g_1(x)}\,dx$, and $g_1(x) = f(x^*) + \frac{\sigma}{2}\|x - x^*\|^2$. $\Phi(x)$ is the CDF of the standard Gaussian distribution.

Theorem 7 for Non-Smooth Case

Assume $f(x)$ is L-smooth in the sense that:

$$f(x) \leq f(x^*) + \frac{L}{2}\|x - x^*\|^2$$

Then, we have the lower bound for the probability that a sample from $\nu \propto e^{-f(x)}$ lies within an ϵ-ball of the global minimizer x^*:

$$\Pr(\|X - x^*\| \leq \epsilon) \geq \frac{Z_{\rho_2}}{Z_\nu}(\Phi(\sqrt{L}\epsilon) - \Phi(-\sqrt{L}\epsilon))$$

where $Z_{\rho_2} = \int e^{-g_2(x)}\,dx$, and $g_2(x) = f(x^*) + \frac{L}{2}\|x - x^*\|^2$.

Theorem 8 for Non-Smooth Case

Assume we have a set of i.i.d. samples $X = \{X_1, X_2, \ldots\}$ drawn from $\nu \propto e^{-f(x)}$, and $f(x)$ satisfies the conditions of Theorems 6 and 7. Then, the minimum number of samples N required to guarantee that the probability of at least one sample lying within an ϵ-ball of x^* is greater than δ is:

$$\ln(1 - \delta) \leq N \ln(1 - \frac{Z_{\rho_2}}{Z_\nu}(\Phi(\sqrt{L}\epsilon) - \Phi(-\sqrt{L}\epsilon)))$$

$$\ln(1 - \delta) \geq N \ln(1 - \frac{Z_{\rho_1}}{Z_\nu}(\Phi(\sqrt{\sigma}\epsilon) - \Phi(-\sqrt{\sigma}\epsilon)))$$

Solving for N:

$$\frac{\ln(1-\delta)}{\ln(1 - \frac{Z_{\rho_2}}{Z_\nu}(\Phi(\sqrt{L}\epsilon) - \Phi(-\sqrt{L}\epsilon)))} \leq N \leq \frac{\ln(1-\delta)}{\ln(1 - \frac{Z_{\rho_1}}{Z_\nu}(\Phi(\sqrt{\sigma}\epsilon) - \Phi(-\sqrt{\sigma}\epsilon)))}$$

2.5.2 Proof for Theorem 6

Given that $f(x) \geq f(x^*) + \frac{\sigma}{2}\|x - x^*\|^2$, the auxiliary function $g_1(x) = f(x^*) + \frac{\sigma}{2}\|x - x^*\|^2$ is used. The distribution $\rho_1 \propto e^{-g_1(x)}$ is Gaussian with mean x^* and variance $1/\sigma$:

$$e^{-g_1(x)} = e^{-\left(f(x^*) + \frac{\sigma}{2}\|x - x^*\|^2\right)} = e^{-f(x^*)} e^{-\frac{\sigma}{2}\|x - x^*\|^2}$$

$$\rho_1(x) = \frac{\sqrt{\sigma}}{\sqrt{2\pi}} e^{-\frac{\sigma}{2}\|x - x^*\|^2}$$

Thus, the probability $X \sim \nu$ lies within ϵ-ball of x^* can be upper-bounded by assuming $X \sim \rho_1$:

$$\Pr(\|X - x^*\| \leq \epsilon) \leq \frac{Z_{\rho_1}}{Z_\nu} \int_{x^* - \epsilon}^{x^* + \epsilon} \rho_1(x)\, dx$$

Since $\rho_1(x)$ is Gaussian:

$$\int_{x^* - \epsilon}^{x^* + \epsilon} \rho_1(x)\, dx = \Phi(\sqrt{\sigma}\epsilon) - \Phi(-\sqrt{\sigma}\epsilon)$$

Therefore:

$$\Pr(\|X - x^*\| \leq \epsilon) \leq \frac{Z_{\rho_1}}{Z_\nu}(\Phi(\sqrt{\sigma}\epsilon) - \Phi(-\sqrt{\sigma}\epsilon))$$

2.5.3 Proof for Theorem 7

Given that $f(x) \leq f(x^*) + \frac{L}{2}\|x - x^*\|^2$, the auxiliary function $g_2(x) = f(x^*) + \frac{L}{2}\|x - x^*\|^2$ is used. The distribution $\rho_2 \propto e^{-g_2(x)}$ is Gaussian with mean x^* and variance $1/L$:

$$e^{-g_2(x)} = e^{-\left(f(x^*) + \frac{L}{2}\|x - x^*\|^2\right)} = e^{-f(x^*)} e^{-\frac{L}{2}\|x - x^*\|^2}$$

$$\rho_2(x) = \frac{\sqrt{L}}{\sqrt{2\pi}} e^{-\frac{L}{2}\|x - x^*\|^2}$$

Thus, the probability $X \sim \nu$ lies within ϵ-ball of x^* can be lower-bounded by assuming $X \sim \rho_2$:

$$\Pr(\|X - x^*\| \leq \epsilon) \geq \frac{Z_{\rho_2}}{Z_\nu} \int_{x^* - \epsilon}^{x^* + \epsilon} \rho_2(x)\, dx$$

Since $\rho_2(x)$ is Gaussian:

$$\int_{x^* - \epsilon}^{x^* + \epsilon} \rho_2(x)\, dx = \Phi(\sqrt{L}\epsilon) - \Phi(-\sqrt{L}\epsilon)$$

Therefore:

$$\Pr(\|X - x^*\| \leq \epsilon) \geq \frac{Z_{\rho_2}}{Z_\nu}(\Phi(\sqrt{L}\epsilon) - \Phi(-\sqrt{L}\epsilon))$$

Number of Samples Needed

To ensure that with high probability (at least δ) at least one sample from the i.i.d. set $X = \{X_1, X_2, \ldots, X_N\}$ lies within an ϵ-ball of x^*, we derive N:
Using the bounds from Theorems 6 and 7:

$$\Pr(\exists X_i \in X \text{ such that } \|X_i - x^*\| \leq \epsilon) = 1 - \Pr(\forall X_i \in X, \|X_i - x^*\| > \epsilon)$$

$$= 1 - (1 - \Pr(\|X - x^*\| \leq \epsilon))^N$$

To ensure this probability is at least δ:

$$1 - (1 - \Pr(\|X - x^*\| \leq \epsilon))^N \geq \delta$$

$$(1 - \Pr(\|X - x^*\| \leq \epsilon))^N \leq 1 - \delta$$

$$N \geq \frac{\ln(1-\delta)}{\ln(1 - \Pr(\|X - x^*\| \leq \epsilon))}$$

Using the bounds from Theorems 6 and 7:

$$\frac{\ln(1-\delta)}{\ln(1 - \frac{Z_{\rho_2}}{Z_\nu}(\Phi(\sqrt{L}\epsilon) - \Phi(-\sqrt{L}\epsilon)))} \leq N \leq \frac{\ln(1-\delta)}{\ln(1 - \frac{Z_{\rho_1}}{Z_\nu}(\Phi(\sqrt{\sigma}\epsilon) - \Phi(-\sqrt{\sigma}\epsilon)))}$$

2.6 TV Distance and Probability Bound

Ultimately, we want to show that this measure of distance does converge to the ϵ of the global minimizer. That is what we will try to establish in this section.

First, recall that the Total Variation (TV) distance between two probability measures μ_t and ν on a measurable space $(\mathbb{R}^d, \mathcal{B})$ is defined as:

$$d_{\text{TV}}(\mu_t, \nu) = \frac{1}{2} \int_{\mathbb{R}^d} |\mu_t(x) - \nu(x)| \, dx$$

For any measurable set $A \in \mathcal{B}$, the TV distance provides an upper bound on the difference in the probabilities assigned to A by μ_t and ν:

$$|\mu_t(A) - \nu(A)| \leq d_{\text{TV}}(\mu_t, \nu)$$

Now, consider the specific set $A_\epsilon = \{x \in \mathbb{R}^d : \|x - x^*\| \leq \epsilon\}$, which is the ϵ-ball around the global minimizer x^*. The probability that a sample $X_t \sim \mu_t$ lies within this ϵ-ball is:

$$\mu_t(A_\epsilon) = \int_{A_\epsilon} \mu_t(x) \, dx$$

The TV distance gives us the following relationship between $\mu_t(A_\epsilon)$ and $\nu(A_\epsilon)$:

$$|\mu_t(A_\epsilon) - \nu(A_\epsilon)| \leq d_{\text{TV}}(\mu_t, \nu)$$

This inequality can be rewritten to express $\mu_t(A_\epsilon)$ in terms of $\nu(A_\epsilon)$ and the TV distance:

$$\nu(A_\epsilon) - d_{\text{TV}}(\mu_t, \nu) \leq \mu_t(A_\epsilon) \leq \nu(A_\epsilon) + d_{\text{TV}}(\mu_t, \nu)$$

2.6.1 Evolution of the TV Distance

Langevin Dynamics (LD) induces a Markov chain with a transition kernel P, and under the assumptions that allow for contractivity, we have:

$$d_{\mathrm{TV}}(\mu_{t+1}, \nu) \leq \gamma d_{\mathrm{TV}}(\mu_t, \nu)$$

where $0 < \gamma < 1$ is the contraction coefficient.
Iterating this contractive property gives:

$$d_{\mathrm{TV}}(\mu_t, \nu) \leq \gamma^t d_{\mathrm{TV}}(\mu_0, \nu)$$

This implies that as t increases, $d_{\mathrm{TV}}(\mu_t, \nu)$ decreases exponentially, meaning μ_t becomes closer to ν in the TV distance.

2.6.2 Bounding the Probability Over Time

Using the relationship between $\mu_t(A_\epsilon)$ and $\nu(A_\epsilon)$, and substituting the expression for $d_{\mathrm{TV}}(\mu_t, \nu)$, we have:

$$\nu(A_\epsilon) - \gamma^t d_{\mathrm{TV}}(\mu_0, \nu) \leq \mu_t(A_\epsilon) \leq \nu(A_\epsilon) + \gamma^t d_{\mathrm{TV}}(\mu_0, \nu)$$

This inequality provides bounds on the probability $\mu_t(A_\epsilon)$ that the sample X_t lies within the ϵ-ball around x^*. As $t \to \infty$, $\gamma^t d_{\mathrm{TV}}(\mu_0, \nu) \to 0$, so the probability $\mu_t(A_\epsilon)$ converges to $\nu(A_\epsilon)$:

$$\lim_{t \to \infty} \mu_t(A_\epsilon) = \nu(A_\epsilon)$$

For finite t, the probability can be expressed as:

$$\mu_t(A_\epsilon) \geq \nu(A_\epsilon) - \gamma^t d_{\mathrm{TV}}(\mu_0, \nu)$$

This shows that as t increases, the probability that X_t lies within the ϵ-ball around x^* becomes increasingly close to $\nu(A_\epsilon)$, which is the probability under the target distribution ν.

3 Discussion

In this section, we'll discuss why approaching optimization in a stochastic manner could offer benefits compared to deterministic methods like gradient descent. Additionally, we'll highlight some valuable references and potential avenues for further research to refine the algorithm and enhance its practical utility.

3.1 Stochastic VS Deterministic Optimization

As mentioned before, the LD in (4) can be viewed as deterministic gradient flow plus some random noise. This raises the question: what are the advantages of solving optimization problems in a stochastic manner compared to using deterministic gradient flow? We believe that sampling algorithms can potentially offer three key advantages for optimization:

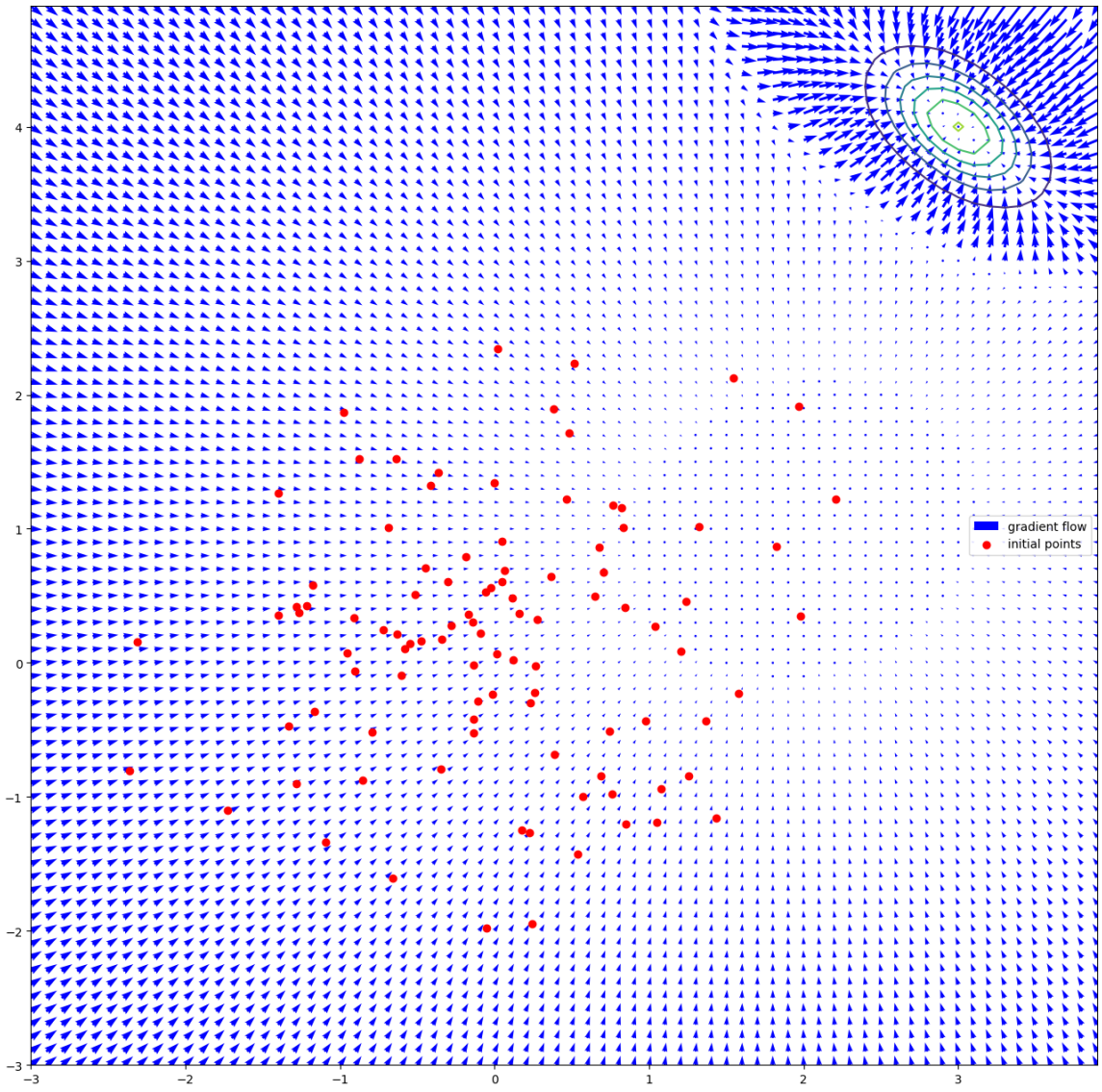

Figure 5: Gradient vector fields and initial points

1. Enable an exploit-explore strategy to discover better local optima.
2. Enable parallel computation for acceleration.
3. Faster convergence in terms of total iterations for certain nonconvex optimization.

Exploit-Explore Strategy: Even though we apply deterministic gradient flow since the initial condition is randomly assigned, the resulting trajectory when running gradient descent can still be viewed as a stochastic process. Moreover, since the algorithm is deterministic, the last iteration X_T will always contract to one or several local minimal, and where it is contracted to is heavily dependent on the distribution of initial guess X_0. On the other hand, LD or another sampling algorithm naturally enables an exploit-explore

strategy where the Brownian motion explores the surroundings of the current iteration and the gradient follows to exploit the gradient information of the target distribution. This automatic exploration makes the sampling algorithm less sensitive to the initial condition, resulting in finding better local minima for nonconvex objective functions.

We illustrate our idea with a concrete example involving the minimization of a 2-variable nonconvex objective function exhibiting one local minimum and one global minimum (Figure 4). We initialize our decision variables randomly from a standard Gaussian distribution (Figure 5). We then apply both a sampling algorithm and a deterministic algorithm to each initialized point and obtain their respective last iteration distributions. In Figure 6, we observe that along the gradient flow, X_t contracts to the local minimum. In contrast, when running the Unadjusted Langevin Algorithm (ULA), a discrete algorithm for simulating LD, the distribution at the last iteration approaches $\nu(x)$, allowing us to escape the local minimal by selecting the solution that has the minimum target value.

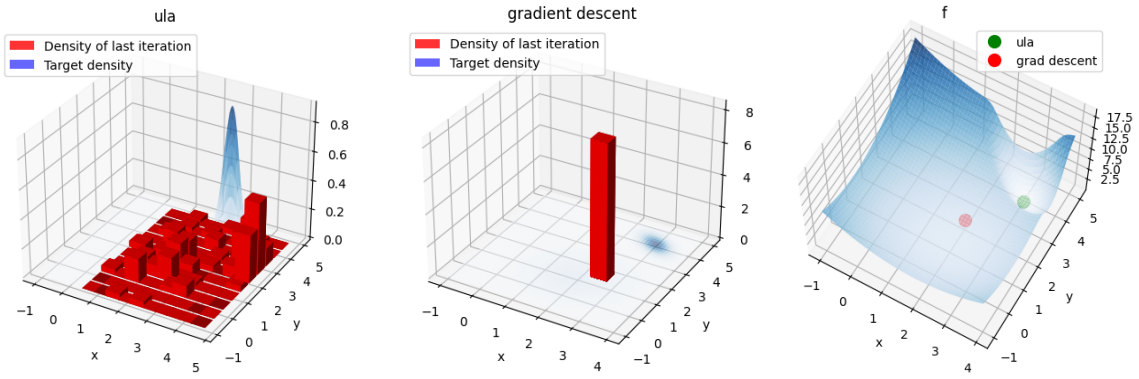

Figure 6: Last iteration distribution of stochastic and deterministic algorithms: In the ULA (Unadjusted Langevin Algorithm), a discrete algorithm for simulating LD, we observe the last iteration distribution approaching the target ν, while in gradient descent, it converges to a local minimum. In the final figure, for each algorithm, we select the samples with the lowest function value as the output. We observe that ULA successfully finds the global minimum, while gradient descent becomes trapped in a local minimum.

Parallel Computing: As illustrated in Figure 6, when employing a stochastic algorithm, the distribution of the last iteration aligns with the induced distribution ν. Conversely, when applying deterministic gradient flow, the last iteration tends to converge to the same local minimum. This phenomenon underscores the significance of parallel computing with GPU for running stochastic algorithms. With stochastic algorithms, we can initialize numerous different starting points and run them to obtain various last iterations, allowing us to select the best one. In contrast, for deterministic algorithms, regardless of the initial starting point, the last iteration is likely to be the same, rendering parallel computing meaningless in such cases.

Faster convergence: In general, finding the global minimum of a nonconvex function can be an NP-hard problem. In a study by (Ma et al., 2019), the authors investigated a class of nonconvex optimization problems and observed that the number of iterations

required to find the global minimum scales exponentially with the dimensionality of the decision variables when using a certain class of optimization algorithms. However, when employing MCMC sampling algorithms, the number of iterations scales linearly instead. This highlights the potential advantage of sampling algorithms in addressing nonconvex optimization problems, particularly in high-dimensional spaces. This finding suggests that sampling algorithms may even be faster in terms of the total number of iterations required for convergence.

3.2 Related Works

The concept of optimization via sampling has been around for some time. An early algorithm, simulated annealing, precisely embodies this idea for finding global optima in nonconvex optimization problems. In simulated annealing, the target distribution $e^{-\beta f(x)}$ is sequentially addressed by introducing a positive "temperature" coefficient $\beta > 0$, which controls how concentrated the last iteration is around the global minimum x^*. Equation (23) indicates that the minimum number required to achieve a certain probability threshold for the last iteration to be close to the global minimizer scales with the smoothness coefficient, assuming $\frac{Z_{\rho_2}}{Z_\nu}$ remains roughly the same.

A closely related work to ours may be (Xu et al., 2017), where they also analyze the global convergence of Langevin dynamics-based algorithms for nonconvex optimization.

While our convergence analysis primarily emphasizes the mixing time, other studies, such as those by (Zhang et al., 2017; Raginsky et al., 2017)delve into the hitting time of sampling algorithms, which directly measures the time it takes for a sampling algorithm to reach the minimum of the objective function.

Yet, this article serves as a preliminary exploration into solving optimization via sampling. Moving forward, an important question to address is which specific sampling algorithms are most suitable for practical nonconvex optimization tasks, such as those encountered in robotics. With advancements in both sampling algorithms and hardware capabilities, there's potential for sampling-based approaches to offer significant advantages in optimizing complex systems.

4 Acknowledgement

We conceived the concept of "Optimization via Sampling" during our study of Probabilistic Machine Learning, where the emphasis is on the converse direction, "Sampling via Optimization". As a result, certain proofs in the "Convergence of LD" section draw heavily from the lecture notes of Probabilistic Machine Learning. We extend our gratitude to Professor Andre Wibisono for his invaluable insights and suggestions.

As this article represents a preliminary exploration into solving optimization via sampling, we realize that a significant majority of our proofs and initial thoughts are predicated on a number of assumptions about the nature of the functions we are interrogating. Throughout this paper, we try to relax notions of smoothness to try and generalize our findings but we definitely see potential in relaxing other assumptions about differentiability and convexity. Therefore, we appreciate the readers who bear to finish reading our article. We welcome anyone interested in contributing to the development of sampling algorithms

for optimization to contact us and collaborate on creating practical and theoretically sound algorithms.

References

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

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, 2019.

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, 2017.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *arXiv preprint arXiv:1707.06618*, 2017.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022. PMLR, 2017.