Discretized Diffusions and their Applications in Global Non-Convex Optimization

by

Andrew Toulis

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

Discretized Diffusions and their Applications
in Global Non-Convex Optimization

Andrew Toulis
Master of Science
Graduate Department of Computer Science
University of Toronto
2019

In certain optimization problems where gradient descent fails, specially designed diffusions can converge to the global minimizers. A popular stochastic analogue of gradient descent is the Langevin algorithm. In this work, we study both this classical method and consider more general diffusions that can be used for optimization even in the non-convex setting. We extend results for discretization error, making it applicable as the number of steps is extended to infinity. This provides new optimization and discretization error bounds for the Langevin diffusion and a broad class of generalized diffusions that will be introduced. The conditions that must be met by these diffusions to achieve global optimization are broad and verifiable, thus allowing practitioners to utilize these methods. Furthermore, to aid practitioners, a recipe for analysis of the approximation error due to discretization is provided. We provide upper bounds for the Euler discretization, the most popular discretization that is employed in the Langevin Algorithm.

Acknowledgements

First of all, I would like to thank my supervisor, Murat Erdogdu. Thank you for taking me on as your first student under such short notice, and for helping me enter a completely new field of research. I am very grateful for the many in-depth discussions we had together, and for the countless proofs we worked out in your office. It was truly a privilege to be given so much one-on-one time with an expert, and you always managed to make the mathematics seem approachable. I cannot thank you enough for teaching me and giving me access to new mathematical ideas and tools. For many years, I have dreamed of doing theoretical research, and you have made this possible. Thank you for your patience, kindness and for providing me with motivation throughout the stages of my master's.

Thank you to my family. To my parents, growing up you provided me with everything I needed to succeed and you fostered the strong curiosity in me. To my grandparents, thank you for inspiring me and allowing our family to have opportunities like this. To my brothers, thank you for your support and friendship, and especially for driving me back and forth many times.

To Daniel Roy, thank you for agreeing to read my thesis and providing me with helpful feedback.

I am also grateful to my peers. To Rasa Hosseinzadeh, thank you for your curiosity, teamwork and detailed feedback on my thesis. Also thank you for the many times you caught mistakes early before I spent time going in the wrong direction!

Lastly, to my girlfriend, Minae Nemoto, thank you for your support and for believing in me. Thank you for motivating me and encouraging me to focus on my studies. Thank you for reminding me of how much I have worked to have this opportunity, and how much I have grown through this experience.

Contents

Chapter 1

Introduction

This thesis is a theoretical work that aims to develop the applicability of using diffusions to perform optimization. We will design diffusions that can be appropriately discretized to closely approximate solutions to global optimization problems. The design of these diffusions will be such that their convergence rate is rapid. Moreover, under the right assumptions, we establish optimization guarantees that apply even in the non-convex setting where traditional methods such as gradient descent can fail to work.

1.1 Optimization in Machine Learning

Many applications in statistics and machine learning can be formulated as a minimization problem. Several optimization algorithms have been proposed to solve minimization problems by relying on the gradient of the function to be minimized. In practice, this gradient is non-linear. As a result, iterative optimization methods are used to solve these problems [Erd17]. Recent advances in machine learning, due to increased data set sizes and computational resources, have inspired many types of iterative methods, such as the Stochastic Langevin Gradient Descent algorithm [VZ$^+$15]. We will be developing theoretical guarantees for a method that outperforms common methods currently employed in practice.

In this work, we are interested in applying optimization in real-world settings. Therefore, we move beyond traditionally studied settings such as convex optimization in order to make this theoretical work applicable. We study non-convex problems that have many local minimas and saddle points. A majority of machine learning applications fall under this category.

Optimal theoretical guarantees have been established in convex optimization. However, the non-convex setting is rich with opportunities for improved guarantees. Strong convexity is not assumed in this work. Instead, other restrictions are made on the objective functions. We note that some restrictions made in the literature can be overly specific and hard to verify. Part of this work will be to determine a set of realistic, broad and verifiable conditions that will aid practitioners in designing an optimization method for their application.

1.2 Diffusions and Optimization

In general, we consider the following unconstrained optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is a non-convex function.

A familiar optimization algorithm which is widely applied in machine learning applications is gradient descent:

$$\theta_{m+1}=\theta_m-\eta\nabla f(\theta_m).$$

In this work, we will study modifications of the gradient descent method, where carefully crafted noise is added to each step:

$$\theta_{m+1}=\theta_m-\eta\nabla f(\theta_m)+\sqrt{\frac{2\eta}{\gamma}}\bar{W}_m,$$

where $\gamma>0$ controls how much noise is added to the system and $\bar{W}_m$ is an isotropic Gaussian vector with co-variance matrix equal to the identity matrix: $\bar{W}_m\sim N(0,I)$.

To inspire the tools that will be used in this work, we will show how this simple modification of gradient descent has an analogue in the continuous setting which acts as a diffusion. We first manipulate the above equation and make a change notation to emphasize time dependence:

$$\begin{aligned}\theta_{t+\eta}=&\theta_t-\eta\nabla f(\theta_t)+\sqrt{\frac{2\eta}{\gamma}}N(0,I),\\ =&\theta_t-\eta\nabla f(\theta_t)+\sqrt{\frac{2}{\gamma}}N(0,\eta I).\end{aligned}$$

Since increments of a Wiener process W_t are Gaussian with variance proportional to the time elapsed, we can re-write $N(0,\eta I)$ as the difference between the process after time has elapsed: $W_{t+\eta}-W_t\sim N(0,\eta I)$. Re-arranging terms and dividing by the step-size:

$$\frac{\theta_{t+\eta}-\theta_t}{\eta}=-\nabla f(\theta_t)+\sqrt{\frac{2}{\gamma}}\frac{W_{t+\eta}-W_t}{\eta}.$$

Letting $\eta\to 0$, we have:

$$d\theta_t=-\nabla f(\theta_t)dt+\sqrt{\frac{2}{\gamma}}dW_t.$$

This last equation is a stochastic differential equation, known as the Langevin diffusion. We are interested in understanding which functions this method and other diffusions can minimize, as well as evaluating the performance of the proposed methods such as their convergence rates. In the literature, non-convex applications have been studied when $f(x)$ has strongly convex tails [CCAY+18, RRT17]. We extend on this work, using the tools presented in the follow-up work of [EMS18].

1.3 Summary of Contributions

In this work, we review the literature and make the following contributions to the literature:

1. (Chapter 3) Establish upper bounds for the weak error of the Euler discretization, avoiding assuming uniformly bounded functions or a torus domain as typically done in [CDC15, VZ$^+$15], and references therein.

2. (Chapter 4) Develop a strong convergence result that generalizes the results of [CCBJ17] for general diffusions and discretizations.

3. (Chapters 5–6) Develop an extension of Milstein's theorem on inferring the weak error of a discretization [Mil94] from finite time intervals to infinite time.

4. (Chapters 7–8) Provide a new break-down of the expected optimization error under our proposed diffusion method, without the need for averaging as in [EMS18, CDC15].

1.4 Thesis Layout

To accomplish our goals, this paper is structured as follows:

1. Chapter 1 is the introduction to the thesis.

2. In Chapter 2, we introduce the notation used in this work. We introduce the main methodology used for optimization, namely Ito Diffusions. We review some important preliminary results and tools for analysis that will be applied in the following chapters.

3. In Chapter 3, we study the error introduced due to discretization. We first start with weak error, and analyze it for the Euler discretization in order to obtain upper bounds. We briefly study higher-order integrators that a similar analysis could also be applied to.

4. In Chapter 4, we move on to strong error. We present a motivating result due to strong error and a type of convergence behaviour analyzed in the following chapter.

5. In Chapter 5, we study the Wasserstein rate of a diffusion. We highlight important sufficient conditions. Finally, we present useful applications which are used in the next chapter.

6. In Chapter 6, we apply tools from prior chapters to yield a new result for the infinite time weak error of a discretization.

7. In Chapter 7, we study how diffusions can be used for optimization. We review results on the long-term behaviour and the expected sub-optimality of diffusions. We show how previously discussed conditions allow us to control these properties, and hence ensure the applicability of our proposed methods to optimization problems.

8. In Chapter 8, we combine and apply our results from previous chapters to obtain bounds on the optimization error.

9. Finally, we review the major contributions of this thesis in the concluding chapter, and outline some possible directions for future work.

Chapter 2

Preliminaries

2.1 Ito Diffusions

In this work we will study how a diffusion obeying a stochastic differential equation (SDE) can be used to perform optimization. In this section, we review some important concepts relating to Ito Diffusions which obey a particular type of SDE. In general, we denote Z_t as the Ito Diffusion obeying the following SDE:

$$dZ_t = b(Z_t)dt + \sigma(Z_t)dW_t,$$

where W_t is a d-dimensional Wiener process independent of Z_t and previous times. Let the initial conditions of the diffusion be x_0. For simplicity, we will omit the initial condition $Z_t^{x_0}$ from notation, and simply use Z_t. When calculating expectations, the initial condition will be made clear using the notation $E^{x_0}[Z_t]$.

2.2 Existence and Uniqueness of a Solution

Let $\|A\|_F$ denote the Frobenius norm of a matrix, and $\|A\|_{op}$ the operator norm of a matrix.

We will assume standard conditions on the coefficients $b(x)$ and $\sigma(x)$ to ensure the existence and uniqueness of a solution to Z_t [Øks03]:

Condition 1. *Linearity and Lipschitz continuity of coefficients*

- *Linear growth of coefficients:*

$$\begin{aligned} \|b(x)\|_2 \leq& \lambda_b[1 + \|x\|_2], \\ \|\sigma(x)\|_F \leq& \lambda_\sigma[1 + \|x\|_2], \\ \left\|\sigma(x)\sigma(x)^T\right\|_{op} \leq& \lambda_a[1 + \|x\|_2^r], \end{aligned}$$

for $r \in 1, 2$.

- *Lipschitz continuity of coefficients:*

$$\begin{aligned}\|b(x)-b(y)\|_2 \leq & D(d)\|x-y\|_2,\\ \|\sigma(x)-\sigma(y)\|_F \leq & D(d)\|x-y\|_2,\end{aligned}$$

  where $D(d)$ is dimension-dependent.

We note that Lipschitz continuity and boundedness of the coefficients at $x = 0$ would be sufficient to imply linear growth. The condition on the operator norm has been added to the conditions in [Øks03] in order to control moment bounds of the diffusion in later sections, as per the work of [EMS18].

2.3 The Euler Discretization

For positive integers m, the Euler discretization X_m corresponding to the diffusion Z_t is the Markov chain with update equation:

$$X_{m+1} = X_m + \eta b(X_m) + \sqrt{\eta}\sigma(X_m)\bar{W}_m,$$

where η is the step-size and $\bar{W}_m$ is an isotropic Gaussian vector, independent of X_m and previous time-steps. Again, we have omitted initial conditions: $X_0 = x_0$.

One can think of this discretization as the stochastic analogue to the classical Euler discretization for solving a differential equation. The classic method simply arises from a Taylor expansion of the function of interest, where terms are matched up to first order. In a later section, we will show how in the stochastic setting, the Euler discretization has an analogous derivation where it also matches stochastic Taylor series terms up to first order.

2.4 Ito's Lemma

One of the most important results in stochastic calculus is Ito's Lemma, which states:

$$df(Z_t) = \left(\langle b(Z_t), \nabla f(Z_t)\rangle + \frac{1}{2}\langle \sigma(Z_t)\sigma(Z_t)^T, \nabla^2 f(Z_t)\rangle_F\right)dt + \langle \nabla f(Z_t), \sigma(Z_t)dW_t\rangle,$$

where $\langle A, B\rangle_F \triangleq \mathrm{tr}(A^T B)$ is the Frobenius inner product between matrices.

There are several important theorems that apply to an Ito diffusion. The ones we will briefly introduce here are Dynkin's formula, the Kolmogorov's Backward Equation, and the Fokker–Planck equation. For a more in-depth treatment, we refer readers to [Øks03].

2.5 Smooth Statistics

Instead of working directly with sample-paths, it is often useful to study how the expected value of any suitably smooth statistic of Z_t evolves in time. We define:

$$u(x_0, t) \triangleq E^{x_0}[f(Z_t)],$$

where t is the time of evaluation of the expectation and the initial position is x_0. $f(x)$ is typically taken to be twice continuously differentiable with compact support. The operator $E^{(\cdot)}[f(Z_t)]$ is called the semi-group operator and is well-defined for such functions.

This motivates the introduction of an infinitesimal generator which tracks how the smooth statistic evolves in time:

$$\mathcal{A}f(Z_t) \triangleq \lim_{h \to 0^+} \frac{E^{x_0}[f(Z_{t+h}) - f(Z_t)]}{h}.$$

Dynkin's formula is an integral form involving the generator, and can be thought of as a stochastic generalization of the second fundamental theorem of calculus:

$$E^{x_0}[f(Z_t)] = f(x_0) + E^{x_0}\left[\int_0^t \mathcal{A}f(Z_s)ds\right].$$

There exists an explicit form of the generator. To obtain it, we apply expectations to the statement of Ito's Lemma:

$$dE^{x_0}[f(Z_t)] = E^{x_0}\left[\langle b(Z_t), \nabla f(Z_t)\rangle + \frac{1}{2}\langle \sigma(Z_t)\sigma(Z_t)^T, \nabla^2 f(Z_t)\rangle_F\right]dt.$$

Matching differential quantities, we can conclude that the generator is an operator of the following form:

$$\mathcal{A}f(x) \triangleq \langle b(x), \nabla f(x)\rangle + \frac{1}{2}\langle \sigma(x)\sigma(x)^T, \nabla^2 f(x)\rangle_F.$$

The generator is used to formulate Kolmogorov's backward equation, which tells us how a smooth statistic varies in time. The backward equation states that $u(x, t)$ is time-differentiable and obeys the following differential equation:

$$\begin{aligned} u(x, 0) =& f(x), \\ \frac{\partial}{\partial t} u(x, t) =& \mathcal{A}u(x, t). \end{aligned}$$

Under conditions to be discussed, $E^x[f(Z_t)] = e^{t\mathcal{A}} f(x)$ is a solution to this differential equation [CDC15]. Here, we make use of the exponential map applied to the generator, which returns an operator that acts on functions:

$$e^{t\mathcal{A}} \triangleq I + \sum_{i=1}^{\infty} \frac{t^i}{i!}\mathcal{A}^i.$$

By performing a Taylor series expansion, one can see how the exponential map operator arises naturally as a solution [CDC15]:

$$\begin{aligned}
u(x,t) =& u(x,0) + \sum_{i=1}^{\infty} \frac{t^i}{i!} \frac{\partial^i}{\partial s^i} u(x,s)|_{s=0}, \\
=& u(x,0) + \sum_{i=1}^{\infty} \frac{t^i}{i!} \frac{\partial^{i-1}}{\partial s^{i-1}} \frac{\partial}{\partial s} u(x,s)|_{s=0}, \\
=& u(x,0) + \sum_{i=1}^{\infty} \frac{t^i}{i!} \mathcal{A} \frac{\partial^{i-1}}{\partial s^{i-1}} u(x,s)|_{s=0}, \\
=& f(x) + \sum_{i=1}^{\infty} \frac{t^i}{i!} A^i f(x),
\end{aligned}$$

where the interchange of time and spatial derivatives is permitted and in the last step we evaluate $u(x,0) = f(x)$.

As we will see in the following sections, by bounding derivatives of sufficient order, we can expand the above sum up to order k such that the remainder term is bounded by $O(t^{k+1})$. Hence we can approximate the value of smooth statistics up to some order.

2.6 Invariant Measures

The Fokker–Planck equation is another major result in the analysis of Ito Diffusions. It describes how the probability density of Z_t changes through time:

$$\frac{\partial}{\partial t} p(x,t) = -\frac{\partial}{\partial x}[b(x)p(x,t)] + \frac{\partial^2}{\partial x^2}[\frac{1}{2}a(x)p(x,t)].$$

This differential equation offers a way to find an invariant measure of a diffusion such that:

$$\frac{\partial}{\partial t} p(x,t) = 0.$$

In other words, the probability density of Z_t through time is unchanging. This is also called a stationary distribution. This will be applied in later sections to the diffusions that we design.

2.7 Dissipativity

We now introduce a standard condition that is sufficient for the existence of a unique invariant measure for Z_t.

Condition 2. *α, β Dissipativity*

$$\mathcal{A}\|x\|^2 \leq -\alpha\|x\|^2 + \beta \quad \alpha, \beta > 0.$$

This condition ensures that the diffusion does not drift but rather travels inwards when too far from the origin.

In later sections, more specialized forms of dissipativity will be discussed. Uniform dissipativity and distant dissipativity both will imply standard dissipativity. Therefore we will often omit explicit reference to this condition. One should note that dissipativity of the second moment carries over to higher moments when combined with Condition 1 [EMS18].

2.8 Bounded Moments

In the following, we show how moments of the diffusion can be bounded via application of dissipativity. Dissipativity ensures the long-term moments of the diffusion are finite, which would not be the case if we only assume linearity.

Lemma 1. *[EMS18] Suppose the diffusion Z_t satisfies Conditions 1 and 2.*

Then for some fixed $\beta_n > 0$:

$$\begin{aligned} E^{x_0}[\|Z_t\|_2^n] \leq & e^{-\alpha t}\|x_0\|_2^n + \frac{\beta_n}{\alpha}(1 - e^{-\alpha t}), \\ \leq & \|x_0\|_2^n + \frac{\beta_n}{\alpha}. \end{aligned}$$

Proof We begin by using the fact that dissipativity of the second moment carries over to higher moments:

$$\mathcal{A}\|x\|_2^n \leq -\alpha\|x\|_2^n + \beta_n,$$

where β_n depends on r, λ_a from Condition 1 [EMS18].

We now apply Dynkin's formula to the function $e^{\alpha t}\|Z_t\|_2^n$:

$$\begin{aligned} E^{x_0}[e^{\alpha t}\|Z_t\|_2^n] = & \|x_0\|_2^n + \int_0^t E^{x_0}[\mathcal{A}(e^{\alpha t}\|Z_s\|_2^n)]ds, \\ = & \|x_0\|_2^n + \int_0^t E^{x_0}[\alpha e^{\alpha t}\|Z_s\|_2^n + e^{\alpha t}\mathcal{A}\|Z_s\|_2^n]ds, \\ \leq & \|x_0\|_2^n + \int_0^t E^{x_0}[\alpha e^{\alpha t}\|Z_s\|_2^n - e^{\alpha t}(\alpha\|Z_t\|_2^n - \beta_n)]ds, \\ = & \|x_0\|_2^n + \frac{1}{\alpha}(1 - e^{\alpha t})\beta_n. \end{aligned}$$

Dividing both sides by $e^{\alpha t}$ gives the desired result. □

For the discretized diffusion, a similar lemma applies:

Lemma 2. *[EMS18] Suppose the diffusion Z_t satisfies Conditions 1 and 2, and that we have selected the Euler discretization as our discretization method.*

Furthermore, assume the discretization is conducted with sufficiently small step-size:

$$\eta < 1 \wedge \frac{\alpha}{2(n_e - 1)!!(1 + \lambda_\sigma + \lambda_b)^{n_e}}.$$

Then:

$$E^{x_0}[\|X_m\|_2^n] \leq \|x_0\|_2^{n_e} + 1 + 2\frac{\beta_{n_e}}{\alpha},$$

where $n_e \geq n$ is an even integer.

In general we will assume that dissipativity is met such that the moments of both the continuous diffusion and discretization are bounded. Define M to be the maximum of both moment bounds, where in general M depends on the dimension of the problem d.

2.9 Weak Taylor series

We now will inspire the stochastic analogue to the Taylor series expansion of a function, often called a weak Taylor series or Ito-Taylor expansion [Mil94].

To start, let us consider the case of an ordinary differential equation:

$$\frac{dX_t}{dt} = a(X_t).$$

We assume that a is sufficiently smooth and obeys linear growth such that a solution X_t exists. Moreover, let f be a sufficiently smooth function. By the chain rule, we have the familiar statement:

$$\frac{d}{dt}f(X_t) = \frac{df(X_t)}{dx}\frac{dX_t}{dt} = a(X_t)\frac{d}{dx}f(X_t).$$

Define the operator $L = a(X)\frac{d}{dx}$. This implies:

$$f(X_t) = f(x_0) + \int_0^t Lf(X_s)ds.$$

This argument can then be applied repeatedly inside the integral, to obtain the familiar Taylor expansion, a summation with powers of L:

$$\begin{aligned} f(X_t) =& f(x_0) + \int_0^t L\big[f(x_0) + \int_0^s Lf(X_r)dr\big]ds, \\ =& f(x_0) + Lf(x_0)t + \int_0^t \int_0^s L^2 f(X_r)drds, \\ =& f(x_0) + Lf(x_0)t + \frac{L^2}{2}f(x_0)t^2 + ... \quad . \end{aligned}$$

The same method follows for the stochastic case, where the operator L is now defined by Ito's lemma. To eliminate Martingale terms, expectations must be taken, as in Dynkin's formula. The operator L under the expectation is therefore simply the generator $\mathcal{A}$. This can be iterated in the same way to obtain the weak Taylor series introduced earlier:

$$\begin{aligned} E^{x_0}[f(Z_t)] =& f(x_0) + E^{x_0}\left[\int_0^t \mathcal{A}f(Z_s)ds\right], \\ =& f(x_0) + \mathcal{A}f(x_0)t + E^{x_0}\left[\int_0^t\int_0^s \mathcal{A}^2 f(Z_r)drds\right], \\ =& f(x_0) + \mathcal{A}f(x_0)t + \frac{\mathcal{A}^2}{2}f(x_0) + ... \quad . \end{aligned}$$

2.10 Euler Discretization Taylor Series Analysis

We present one useful application of the weak Taylor expansion. This is a popular method for analyzing the approximation error of a discretization. Similar arguments apply to other discretizations [Zyg11].

Lemma 3. *The Euler discretization approximates expectations after one-step to second-order accuracy:*

$$|\mathbb{E}^{X_m}[f(Z_\eta)] - \mathbb{E}^{X_m}[f(X_{m+1})]| = O(\eta^2).$$

Proof First let us establish some useful facts. Define the increment:

$$\delta_{X_{m+1}} \triangleq X_{m+1} - X_m = \eta b(X_m) + \sqrt{\eta}\sigma(X_m)\bar{W}_m.$$

We have the following identities:

$$\begin{aligned} \mathbb{E}^{X_m}[\delta_{X_{m+1}}] =& \eta b(X_m), \\ \mathbb{E}^{X_m}[\delta_{X_{m+1}}\delta_{X_{m+1}}^T] =& \eta\sigma(X_m)\sigma(X_m)^T + \eta^2 b(X_m)b^T(X_m). \end{aligned}$$

We begin the proof by performing a Taylor series expansion of f around the point X_m:

$$f(X_{m+1}) = f(X_m) + \langle\delta_{X_{m+1}}, \nabla f(X_m)\rangle + \frac{1}{2}\langle\delta_{X_{m+1}}\delta_{X_{m+1}}^T, \nabla^2 f(X_m)\rangle_F + O(\|\delta_{X_{m+1}}\|_2^3).$$

Now take expectations, and use the definition of $\delta_{X_{m+1}}$ and the identities above:

$$\begin{aligned} \mathbb{E}^{X_m} f(X_{m+1}) =& f(X_m) + \langle\eta b(X_m), \nabla f(X_m)\rangle \\ &+ \frac{1}{2}\langle\eta\sigma(X_m)\sigma(X_m)^T + \eta^2 b(X_m)b^T(X_m), \nabla^2 f(X_m)\rangle_F + O(\eta^2), \\ =& (I + \eta\mathcal{A})f(X_m) + O(\eta^2). \end{aligned}$$

We notice that this matches the exact Taylor series expansion up to order one, which proves the result:

$$\begin{aligned} \mathbb{E}^{X_m} f(Z_\eta) =& e^{\eta\mathcal{A}} f(X_m), \\ =& (I + \eta\mathcal{A})f(X_m) + O(\eta^2). \end{aligned}$$

□

2.11 Langevin Dynamics

Langevin Dynamics is the special case of an Ito Diffusion when $b(x) = -\nabla f(x)$, where $f(x)$ is the objective function of interest. The dynamics can be thought of as gradient descent with carefully scaled noise. The discretization of Langevin dynamics is known as the Langevin algorithm. The Langevin algorithm has been used to practically sample from densities of interest in order to estimate intractable expectations [VZ+15].

In this work, we will study the Langevin algorithm's relationship to optimization. As will be demonstrated in later sections, Langevin Dynamics converges to an invariant measure which is equal to the Gibbs measure of the function $f(x)$. In turn, expectations under this measure are shown to closely approximate the minimum of the function of interest. In particular, when the diffusion is dissipative and the gradient of interest is Lipschitz continuous, then Langevin dynamics is, in expectation, an approximate global optimizer even for non-convex and multi-modal objective functions [EMS18].

However, the assumption that the diffusion is dissipative is not always satisfied. We note that the dissipativity condition has a simple form for Langevin dynamics where $b(x) = -\nabla f(x)$ and $\sigma(x) = \sqrt{\frac{2}{\gamma}}I$:

$$\mathcal{A}\|x\|^2 = 2\langle -\nabla f(x), x\rangle + \frac{2d}{\gamma} \leq -\alpha\|x\|^2 + \beta,$$

where d is the dimension. Hence one can quickly verify for a choice of $f(x)$ whether Langevin dynamics will be dissipative and hence inherit a unique invariant measure. For example, the functions $\sqrt{1 + \|x\|^2}$ and $\log(1 + \|x\|^2)$ have obvious minima, yet would fail to converge to a unique invariant measure. In future sections, we will consider generalized diffusions that optimize these functions and for which dissipative behaviour is satisfied.

Recent works have decomposed the optimization error of the Langevin algorithm and other generalized methods [EMS18]. We build on this work by decomposing the error in a new way that does not involve averaging samples.

There are two main types of Langevin Dynamics, first-order and second-order. Second-order dynamics includes a momentum term. While most of this work applies to the first-order case, it is an area of future work to perform similar analysis for the second-order case. Second-order dynamics is of interest to practitioners due to its empirically better performance [CCBJ17]. In the strongly convex setting, recent analysis has confirmed this observation. A practical MCMC algorithm, based on a discretization of second-order Langevin dynamics, was proposed [CCBJ17]. It was shown to achieve ϵ error (in L^2-Wasserstein distance) in $O(\frac{\sqrt{d}}{\epsilon}\log(\frac{1}{\epsilon}))$ steps for sufficiently low step-size.

There is one final practical point of interest that is an important area of future work. Due to increasing sizes of data-sets, often one must sample a subset of data when performing gradient descent. As a result, stochastic gradient descent methods have been studied and used. We note that data-stochastic analogues of the algorithms we study here exist and have been applied in practice, such as the stochastic version of Langevin dynamics. However, in this theoretical work, we will focus on the case where the full gradient is known so that we can quantify the error accumulated purely from discretization.

Chapter 3

Weak Error

3.1 Introduction to Weak Error

In machine learning problems, we are often in a situation where it is intractable to compute an exact expectation under some density: $E^{x \sim p(x)}[f(x)]$. For example, the target density p might arise as the posterior in a Bayesian inference problem:

$$p(\theta) = p_0(\theta) \prod_{i=0}^{N} p(X_i|\theta),$$

where $p_0(\theta)$ is the prior distribution, X_i are i.i.d. observations, and θ is the model parameters to be optimized.

A standard approach to approximating expectations is to design a diffusion that on average or in the long-term gives us samples that are close to the density of interest. In later sections, we will explore how to design such a diffusion.

While it is often possible to find a diffusion that tracks the density of interest, we need to be able to simulate the diffusion. In practice, samples are taken from a Markov chain [VZ+15]. One approach is to directly design the Markov chain to have an equilibrium that is near the density of interest, which can involve strategies such as an accept-reject step. However, this can be very costly due to high rejection rates. Another approach, which we will consider here, is to design a Markov chain that closely matches the dynamics of the diffusion.

Therefore, it is of interest to quantify the error due to sampling from a Markov chain compared to an exact diffusion. In the literature, the local order characterizes how well a given discretization (sometimes called an integrator) approximates an exact process after one-step. We call it a local error since we are assuming only a single step has occurred with some small step-size η. We provide the definition below.

Definition 1. *Weak Local Error of Order k A discretization is said to be a k^{th}-order local integrator of f if, for any starting position X_m, the following is satisfied:*

$$|\mathbb{E}^{X_m}[f(Z_\eta)] - \mathbb{E}^{X_m}[f(X_{m+1})]| = C[1 + \|X_m\|_2]^s \eta^k,$$

where in general C, s and k depend on the underlying diffusion and discretization.

As we already saw, the Euler discretization has a weak local error of order two. Integrators with higher-order weak local error were analyzed in [CDC15], and we will briefly introduce the symmetric scheme from their work.

In later sections, we will demonstrate how one can apply weak error to optimization problems. We will design a continuous diffusion that appropriately optimizes a function of interest, and then admit discretization errors when attempting to perform this optimization in practice.

3.2 Preliminary Analysis

In this chapter, we will obtain upper bounds for the approximation error of the Euler discretization after a single time-step $\Delta t = \eta$.

To do so, we must begin with some preliminaries relating to the generator. We then construct weak Taylor series for the continuous process and discretization, subtracting like-terms and bounding higher-order remainder terms using a convenient representation.

Recall, the generator of the SDE Z_t with drift $b(x)$ and diffusion coefficient $a(x) \triangleq \sigma(x)\sigma(x)^T$ is an operator of the following form:

$$\begin{aligned}\mathcal{A}f(x) \triangleq& \langle b(x), \nabla f(x)\rangle + \frac{1}{2}\langle a(x), \nabla^2 f(x)\rangle_F,\\ \triangleq& \sum_i b_i \partial_i f + \frac{1}{2}\sum_{i,j} a_{ij}\partial_{ij} f,\end{aligned}$$

where we have used the shorthand notation ∂_i in place of $\frac{\partial}{\partial_i}$ and the summation indices are from 1 to d, where d is the dimension.

With some work, one can find the form of the generator squared:

$$\begin{aligned}\mathcal{A}^2 f =& \sum_s b_s\partial_s(b_i\partial_i f + \frac{1}{2}\sum_{i,j} a_{ij}\partial_{ij} f) + \frac{1}{2}\sum_{s,t} a_{st}\partial_{st}(b_i\partial_i f + \frac{1}{2}\sum_{i,j} a_{ij}\partial_{ij} f),\\ =& \sum_{s,i} b_s(\partial_s b_i)(\partial_i f) + b_s b_i \partial_{is} f + \frac{1}{2}\sum_{s,i,j} b_s(\partial_s a_{ij})(\partial_{ij} f) + b_s a_{ij}\partial_{sij} f\\ &+ \frac{1}{2}\sum_{s,t,i} a_{st}(\partial_{st} b_i)(\partial_i f) + 2a_{st}(\partial_s b_i)(\partial_{ti} f) + a_{st} b_i \partial_{sti} f\\ &+ \frac{1}{4}\sum_{s,t,i,j} a_{st}(\partial_{st} a_{ij})(\partial_{ij} f) + 2a_{st}(\partial_s a_{ij})(\partial_{tij} f) + a_{st} a_{ij}\partial_{stij} f,\end{aligned}$$

where we have used that $a = a^T$ and the product rule for second derivatives. Note that the term $b_s a_{ij}\partial_{sij} f$ can be grouped with $a_{st} b_i \partial_{sti} f$ due to symmetry of derivatives, as done below.

In general, $\mathcal{A}^n f$ contains the first $2n$ derivatives of f and the first $2(n-1)$ derivatives of b and a. This is since $\mathcal{A}$ is a second-order differential operator and the first application of $\mathcal{A}$ is what introduces b and a.

For the Euler discretization, the corresponding generator evaluated at a point x_0 is:

$$\tilde{\mathcal{A}}_{x_0} f(x) \triangleq \langle b(x_0), \nabla f(x)\rangle + \frac{1}{2}\langle a(x_0), \nabla^2 f(x)\rangle.$$

One can see this either by performing a weak Taylor expansion of $E^{x_0}[f(Z_t)]$ at x_0, or by considering the continuous diffusion $\tilde{X}_t$ which behaves like the Euler discretization (having constant coefficients for each time interval corresponding to a discrete time-step). See [CCBJ17] for an example.

$\tilde{\mathcal{A}}^2_{x_0}$ simply contains all terms from $\mathcal{A}^2$ that do not take derivatives with respect to the drift and diffusion coefficients, which are now constants due to the evaluation of the functions at x_0:

$$\tilde{\mathcal{A}}^2_{x_0} f = \sum_{s,i} b_s b_i \partial_{is} f + \sum_{s,i,j} b_s a_{ij} \partial_{sij} f + \frac{1}{4} \sum_{s,t,i,j} a_{st} a_{ij} \partial_{stij} f.$$

As will be done in the later part of this section, these summations can be bounded by applying maximums:

$$|\tilde{\mathcal{A}}^2_{x_0} f| \leq d^2 \max_i |b_i|^2 \max_{i,j} |\partial_{ij} f| + d^3 \max_i |b_i| \max_{i,j} |a_{ij}| \max_{s,i,j} |\partial_{sij} f| + d^4 \max_{i,j} |a_{ij}|^2 \max_{s,t,i,j} |\partial_{stij} f|.$$

For simplicity, we notate maximum norms as follows:

$$\text{For a vector:} \quad \|b(x)\|_\infty \triangleq \max_i |b_i(x)|.$$
$$\text{For a matrix:} \quad \|\sigma(x)\|_\infty \triangleq \max_{i,j} |\sigma_{ij}(x)|.$$

and so on for tensors of higher order. The previous expression simplifies to:

$$|\tilde{\mathcal{A}}^2_{x_0} f| \leq d^2 \|b\|^2_\infty \left\|\nabla^2 f\right\|_\infty + d^3 \|b\|_\infty \|a\|_\infty \left\|\nabla^3 f\right\|_\infty + d^4 \|a\|^2_\infty \left\|\nabla^4 f\right\|_\infty.$$

3.3 Weak Taylor Analysis

We will compare expectations under the continuous diffusion and the Euler discretization using Taylor series expansions.

For each diffusion, expectations can be expanded in terms of powers of their generators:

$$E^{x_0}[f(Z_\eta)] = (I + \eta \mathcal{A} + \frac{1}{2}\eta^2 \mathcal{A}^2 + ...) f(x_0),$$
$$E^{x_0}[f(\tilde{X}_\eta)] = (I + \eta \tilde{\mathcal{A}}_{x_0} + \frac{1}{2}\eta^2 \tilde{\mathcal{A}}^2_{x_0} + ...) f(x_0).$$

Recall the definition $u(x_0, t) \triangleq E^{x_0}[f(Z_t)]$. As shown in the introductory chapter, this expansion follows from a weak Taylor expansion of the expectation $E^{x_0}[f(Z_t)]$. The Taylor expansion is performed through time $t = 0$ to $t = \eta$, and time derivatives are converted to expressions in terms of the generator via application of the backward equation $\frac{\partial}{\partial s} u(x_0, s) = \mathcal{A} u(x_0, s)$ [CDC15]:

$$\begin{aligned} u(x_0, \eta) &= u(x_0, 0) + \sum_{i=1}^{\infty} \frac{\eta^i}{i!} \frac{\partial^i}{\partial s^i} u(x_0, s)|_{s=0}, \\ &= u(x_0, 0) + \sum_{i=1}^{\infty} \frac{\eta^i}{i!} \mathcal{A}^i u(x_0, s)|_{s=0}, \\ &= f(x_0) + + \sum_{i=1}^{\infty} \frac{\eta^i}{i!} \mathcal{A}^i f(x_0). \end{aligned}$$

Expanding the Taylor series up to first order, and applying Taylor's theorem with the mean-value forms of the remainder, we have for some $t \in [0, \eta]$:

$$u(x_0, \eta) = (I + \eta \mathcal{A}) f(x_0) + \frac{1}{2}\eta^2 \mathcal{A}^2 u(x_0, s)|_{s=t}.$$

Re-arranging terms, we have the following estimate of the remainder error by approximating the expectation up to first order:

$$\begin{aligned} |u(x_0, \eta) - (I + \eta\mathcal{A}) f(x_0)| \leq & \frac{1}{2}\eta^2 \max_{0 \leq t \leq \eta} |\mathcal{A}^2 u(x_0, t)|, \\ = & \frac{1}{2}\eta^2 \max_{0 \leq t \leq \eta} |E^{x_0}[\mathcal{A}^2 f(Z_t)]|, \end{aligned}$$

where in the last line we have used the definition of $u(x, t)$ and utilized the commutativity of $\mathcal{A}$ and $E^{(\cdot)}[f(Z_t)]$ [Øks03].

Applying this same procedure to each diffusion, we have the following two inequalities:

$$\begin{aligned} |E^{x_0}[f(Z_\eta)] - (I + \eta\mathcal{A}) f(x_0)| \leq & \frac{1}{2}\eta^2 \max_{0 \leq t \leq \eta} |E^{x_0}[\mathcal{A}^2 f(Z_t)]|, \\ |E^{x_0}[f(\tilde{X}_\eta)] - (I + \eta\tilde{\mathcal{A}}_{x_0}) f(x_0)| \leq & \frac{1}{2}\eta^2 \max_{0 \leq t \leq \eta} |E^{x_0}[\tilde{\mathcal{A}}^2_{x_0} f(\tilde{X}_t)]|. \end{aligned}$$

3.4 Upper Bounds for the Euler Discretization

While in general $\mathcal{A}$ and $\tilde{\mathcal{A}}_{x_0}$ are not the same, when they are both evaluated at x_0, they are equal. Hence $\mathcal{A} f(x_0) - \tilde{\mathcal{A}}_{x_0} f(x_0) = 0$ and we have by the triangle inequality:

$$\begin{aligned} \frac{2}{\eta^2}\left| E^{x_0}[f(Z_\eta) - f(X_1)] \right| & \leq \max_{0 \leq t \leq \eta} \left| E^{x_0}[\mathcal{A}^2 f(Z_t)] \right| + \max_{0 \leq t \leq \eta} \left| E^{x_0}[\tilde{\mathcal{A}}^2_{x_0} f(\tilde{X}_t)] \right|, \\ & \leq \max_{0 \leq t \leq \eta} E^{x_0} \left[|\mathcal{A}^2 f(Z_t)| \right] + \max_{0 \leq t \leq \eta} E^{x_0} \left[|\tilde{\mathcal{A}}^2_{x_0} f(\tilde{X}_t)| \right], \\ & \leq \max_{0 \leq t \leq \eta} E^{x_0} \Big[d^2 \|b(Z_t)\|_\infty \|\nabla b(Z_t)\|_\infty \|\nabla f(Z_t)\|_\infty + d^2 \|b(Z_t)\|^2_\infty \left\|\nabla^2 f(Z_t)\right\|_\infty \\ & + d^3 \|b(Z_t)\|_\infty \|\nabla a(Z_t))\|_\infty \left\|\nabla^2 f(Z_t)\right\|_\infty + d^3 \|b(Z_t)\|_\infty \|a(Z_t)\|_\infty \left\|\nabla^3 f(Z_t)\right\|_\infty \\ & + d^3 \|a(Z_t)\|_\infty \left\|\nabla^2 b(Z_t)\right\|_\infty \|\nabla f(Z_t)\|_\infty + d^3 \|a(Z_t)\|_\infty \|\nabla b(Z_t)\|_\infty \left\|\nabla^2 f(Z_t)\right\|_\infty \\ & + d^4 \|a(Z_t)\|_\infty \left\|\nabla^2 a(Z_t)\right\|_\infty \left\|\nabla^2 f(Z_t)\right\|_\infty + d^4 \|a(Z_t)\|_\infty \|\nabla a(Z_t)\|_\infty \left\|\nabla^3 f(Z_t))\right\|_\infty \\ & + d^4 \|a(Z_t)\|^2_\infty \left\|\nabla^4 f(Z_t)\right\|_\infty \Big] \\ & + \max_{0 \leq t \leq \eta} E^{x_0} \left[d^2 \left\| b(\tilde{X}_t) \right\|^2_\infty \left\| \nabla^2 f(\tilde{X}_t) \right\|_\infty + d^3 \left\| a(\tilde{X}_t) \right\|_\infty \left\| b(\tilde{X}_t) \right\|_\infty \left\|\nabla^3 f(Z_t)\right\|_\infty \right. \\ & \left. + d^4 \left\| a(\tilde{X}_t) \right\|^2_\infty \left\| \nabla^4 f(\tilde{X}_t) \right\|_\infty \right], \end{aligned}$$

where d is the dimension. Here we used the bounding method from the preliminary subsection of this chapter, which takes advantage of maximum norms to control the summations involved in each generator squared.

To bound this entire expression, we will assume the following bound relating to each term:

Condition 3. *Polynomial growth of derivatives*

1. *Derivatives up to order two of the drift and diffusion coefficients are bounded by a polynomial, where we have shared the maximum of parameters across derivatives for simplicity:*

$$\begin{aligned}\|\nabla b(x)\|_\infty &\triangleq \max_{s,i} |\partial_s b_i(x)| \le \lambda_b[1+\|x\|_2^{s_b}],\\ \left\|\nabla^2 b(x)\right\|_\infty &\triangleq \max_{s,t,i} |\partial_{st} b_i(x)| \le \lambda_b[1+\|x\|_2^{s_b}],\\ \|\nabla \sigma(x)\|_\infty &\triangleq \max_{s,i,j} |\partial_s \sigma_{ij}(x)| \le \lambda_\sigma[1+\|x\|_2^{s_\sigma}],\\ \left\|\nabla^2 \sigma(x)\right\|_\infty &\triangleq \max_{s,t,i,j} |\partial_{st} \sigma_{ij}(x)| \le \lambda_\sigma[1+\|x\|_2^{s_\sigma}].\end{aligned}$$

2. *Moreover, the first four derivatives of f are bounded by a polynomial of a similar form, with λ_f, s_f defined similarly.*

Condition 3 implies the following when combined with Condition 1 (linear growth of coefficients). Defining: $s = \max\{1, s_b, s_\sigma, s_f\}$ and $\lambda = 1 + \max\{\lambda_b, \lambda_\sigma, \lambda_f\}$, we have:

$$\forall g \in \{b, \nabla b, \nabla^2 b, \sigma, \nabla\sigma, \nabla^2\sigma, \nabla f, \nabla^2 f, \nabla^3 f, \nabla^4 f\} : \|g(x)\| \le \lambda[1+\|x\|_2^s].$$

Finally, as done throughout this work, we assume that moments of sufficient order for both the discretization and diffusion are mutually bounded by a dimension-dependent constant M. In the preliminary section, we provided a recipe for computing M explicitly, assuming dissipativity and assuming the discretization is the Euler discretization.

Therefore, assuming these conditions, the terms above, each with a product of length 3, are bounded:

$$\begin{aligned}|E^{x_0}[f(Z_\eta)] - E[f(X_1)]| \le& \frac{\eta^2}{2}(4d^4 + 5d^3 + 3d^2)\lambda^3 M,\\ \le& \eta^2 6 d^4 \lambda^3 M.\end{aligned}$$

3.5 Extending to Finite Time Intervals

We now state a theorem due to Milstein [Mil94] which allows one to extend the local error of a discretization to any finite time. We make modifications to the result in order to apply the bounds just obtained.

Proposition 1. *One-step approximation and approximation on a finite interval [Mil94].*

Suppose the following conditions hold:

- *Lipschitz continuity and linearity of coefficients (Condition 1).*
- *Derivatives of the functions $f(x)$, $b(x)$ and $a(x)$ exist up to sufficient order, depending on the weak local order. Moreover, they are bounded by a polynomial of the form $\lambda[1+\|x\|_2^s]$.*
- *Bounded moments: for a sufficiently large positive integer l, moments up to l^{th} order are bounded by M, for both the discretization and the continuous diffusion.*

Let $T = N\eta$ be a finite time interval, elapsed over N steps where η is the step-size of the discretization. Then, if we have an integrator X_m of local weak order $k+1$, the following inequality holds:

$$|E^{x_0}[f(Z_T) - f(X_T)]| \leq CTM\eta^k,$$

where C is a constant.

For the Euler discretization, the result is that:

$$|E^{x_0}[f(Z_T) - f(X_T)]| \leq T6d^4\lambda^3 M\eta.$$

Hence the Euler discretization has a global weak error of order of one for a finite interval.

Proof We provide a sketch of the proof for the Euler case, accounting for modifications we have made to the conditions and methods. Similar arguments will hold for other discretizations.

First, we note that the conditions above can be stated specifically for the Euler discretization. As for bounded derivatives, Condition 3, which bounds derivatives of $a(x), b(x), f(x)$ suffices in our case. However, we will want to extend the number of derivatives for the drift and diffusion coefficients up to order four in order to proceed with the next step.

Given this condition, the partial derivatives of $u(x,t)$ with respect to x exist and are also bounded by a polynomial, up to the order four [Mil94]. Therefore the function $u(x,t)$ itself benefits from the local discretization bound we established. That is, it has a time-uniform local Euler discretization error of:

$$\forall t \in [0,T] : |E^x[u(Z_\eta, t)] - E[u(X_1,t)]| \leq 6d^4\lambda^3 M\eta^2.$$

Then, due to the arguments of Milstein we can obtain the bound as desired. Milstein constructs a telescoping series with $N = \frac{T}{\eta}$ terms. Each term in the summation is of the form of the weak local error: $E^x[u(Z_\eta, t)] - E[u(X_1,t)]$, for some t.

Thus we have the following bound, which then establishes the desired result:

$$|E^{x_0}[f(Z_T) - f(X_T)]| \leq \frac{T}{\eta} 6d^4\lambda^3 M\eta^2.$$

□

3.6 Symmetric Discretization of Second-Order Langevin Dynamics

We will show an example here of applying Taylor methods to a different discretization than the Euler discretization. Before we do this, we will need to introduce second-order Langevin dynamics. The order of the Langevin method is not to be confused with the order of the discretization. The Euler discretization, for example, could be applied to second-order Langevin dynamics.

3.6.1 Second-order Langevin Dynamics

In the literature [CDC15], second-order Langevin Dynamics is defined as:

$$\begin{cases} d\theta = & pdt, \\ dp = & -\nabla_\theta f(\theta)dt - Dpdt + \sqrt{2D}dW_t. \end{cases}$$

When physically interpreted, D is the friction coefficient, θ is thought of as the position vector, and p is thought of as the momentum vector.

Importantly, the generator for this diffusion can be split into three parts $\mathcal{A} = \mathcal{A}_A + \mathcal{A}_B + \mathcal{A}_O$ which correspond to three differential equations that make up the original differential equation:

$$A : \begin{cases} d\theta = & pdt \\ dp = & 0 \end{cases}, \qquad B : \begin{cases} d\theta = & 0 \\ dp = & -Dpdt \end{cases}, \qquad O : \begin{cases} d\theta = & 0 \\ dp = & -\nabla_\theta f(\theta)dt + \sqrt{2D}dW_t \end{cases},$$

where one can calculate that [CDC15]:

$$\begin{aligned} \mathcal{A}_A g =& \langle p, \nabla_\theta g\rangle, \\ \mathcal{A}_B g =& -D\langle p, \nabla_p g\rangle, \\ \mathcal{A}_O g =& -\langle \nabla_\theta f(\theta), \nabla_p g\rangle + 2D\,\mathrm{tr}(\nabla_p^2 g). \end{aligned}$$

Together, we have the expectation after time t has elapsed:

$$E^{(\theta_0, p_0)}[g(\theta_t)] = e^{t\mathcal{A}}g(\theta_0) = e^{t(\mathcal{A}_A + \mathcal{A}_B + \mathcal{A}_O)}g(\theta_0).$$

3.6.2 The Symmetric Splitting Discretization

We now discuss the ABOBA Symmetric Splitting discretization introduced in [CDC15], which is a higher-order integrator as we will now prove using more elementary techniques. The discretization is defined in five steps as follows (where $\bar{W}$ is an isotropic Gaussian vector independent of the current and previous time-steps):

1. $\theta^*_{m+1} = \theta_m + \frac{\eta}{2}p_m$,
2. $p^*_{m+1} = e^{-D\frac{\eta}{2}}p_m$,
3. $p^{**}_{m+1} = p^*_{m+1} - \nabla_\theta f(\theta^*_{m+1})\eta + \sqrt{2Dh}\bar{W}_{m+1}$,
4. $p_{m+1} = e^{-D\frac{\eta}{2}}p^{**}_{m+1}$,

5. $\theta_{m+1} = \theta^*_{m+1} + \frac{\eta}{2} p_{m+1}$.

We note that the order of these steps does not matter, as long as the algorithm is symmetric. For example, another valid symmetric discretization would be BAOAB.

We see that the expectation corresponding to this discretization is:

$$E^{(\theta_m, p_m)}[f(\theta_{m+1})] = e^{\frac{\eta}{2}\mathcal{A}_A} e^{\frac{\eta}{2}\mathcal{A}_B} e^{\eta\mathcal{A}_O} e^{\frac{\eta}{2}\mathcal{A}_B} e^{\frac{\eta}{2}\mathcal{A}_A} f(\theta_m).$$

Through Taylor expansions of each exponential map, and distributing the product by multiplying out terms, one can collect all first and second order terms to verify that this matches the exact expectation up to order two:

$$e^{\frac{\eta}{2}\mathcal{A}_A} e^{\frac{\eta}{2}\mathcal{A}_B} e^{\eta\mathcal{A}_O} e^{\frac{\eta}{2}\mathcal{A}_B} e^{\frac{\eta}{2}\mathcal{A}_A} f(\theta_m) = I + \eta(\mathcal{A}_A + \mathcal{A}_B + \mathcal{A}_O) + \frac{\eta^2}{2}(\mathcal{A}_A + \mathcal{A}_B + \mathcal{A}_O)^2 + \eta^3 R + O(\eta^4).$$

The order of integration is therefore two for this discretization. This result is obtained in [CDC15] using the Baker–Campbell–Hausdorff (BCH) formula:

$$\begin{aligned} e^X e^Y =& e^Z, \\ Z =& X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}[X,[X,Y]] - \frac{1}{12}[Y,[X,Y]] + \dots \quad , \end{aligned}$$

where X, Y, Z are operators and $[X, Y]$ is the commutator of the two operators, which in general is non-zero.

In the continuous case, the third order term is simply $\frac{1}{6}(\mathcal{A}_A + \mathcal{A}_B + \mathcal{A}_O)^3$. As in the previous section, we can bound each of these remainders, based on polynomial bounds of derivatives and moment bounds. This would yield an upper bound on the difference between expectations, now at a local error of order three.

To inspire this choice of a symmetric discretization, we can analyze how two non-commutative operators X and Y interact when exponentiated, which is the natural operation that occurs to generators when we study expectations. Suppose we have two operators X and Y, which can be thought of as generators corresponding to steps in a discretization. Due to the BCH formula:

$$e^{\eta X} e^{\eta Y} = e^{\eta Z_1},$$

where $Z_1 = X + Y + \frac{\eta}{2}[X,Y] + \frac{\eta^2}{12}([X,[X,Y]] - [Y,[X,Y]]) + O(\eta^3)$. Hence a discretization that can be split into two, non-symmetric parts will in general contain discretization errors of order one.

On the other hand, due to the BCH formula, we have the following identity for a symmetric scheme [Zyg11]:

$$e^{\frac{\eta}{2}X} e^{\eta Y} e^{\frac{\eta}{2}X} = e^{\eta Z_2},$$

where $Z_2 = X + Y + \frac{\eta^2}{12}([Y,[Y,X]] - \frac{1}{2}[X,[X,Y]]) + O(\eta^4)$.

Chapter 4

Strong Error

4.1 Introduction to Strong Error

In the following analysis, we assume the diffusion converges to an invariant measure $Z^* \sim p$, so that its long-time behaviour can be meaningfully examined. This will be discussed further in the next sections. For now, we wish to motivate some new concepts in this chapter.

First, we define strong local error, which contrasts the weak local error studied earlier.

Definition 2. *Strong Local Error of Order k*

A discretization has k^{th}-order strong local error if, for any starting position X_m, the following is satisfied:

$$E^{X_m}[\|Z_\eta - X_{m+1}\|_2] \leq C(1 + \|X_m\|_2)\eta^k,$$

where in general C and k depend on the underlying diffusion and the discretization. For this chapter, we will simply absorb $(1 + \|X_m\|_2)$ into C since moment bounds can be handled separately.

4.2 Convergence of the Continuous-Time Process

To aid in analyzing the limiting behaviour of a discretization, we need to somehow control the convergence of the diffusion that it is approximating. We will work with a time decay of the following form, which ensures that any instance of the diffusion when run long enough will eventually converge to the invariant measure p. This is stated mathematically as:

$$E^{Z^*}E^{X_m}[\|Z_\eta - Z^*\|_2] \leq e^{-s\eta}E^{Z^*}E^{X_m}[\|X_m - Z^*\|_2],$$

for some constant $s > 0$, where the expectation on the right-hand-side can be thought of as the initial distance between the limiting distribution $Z^* \sim p$ and the distribution of X_m.

This type of decay is known as a Wasserstein decay, and will be motivated and explored in the following sections in more detail. For now, simply interpret this statement as saying that the distance between the distribution Z_t and Z^* is exponentially decaying as time passes.

4.3 A Result Due to Continuous-Time Convergence and Strong Local Error

Proposition 2. *Assume we have selected an appropriate integrator with strong local order $k+1$.*

Moreover, assume we have a Wasserstein decay.

Let m be the number of steps needed to achieve an error of ϵ with an appropriate step size η. Then $\delta_m \triangleq E^{Z^*}E^{X_{m-1}}[\|X_m - Z^*\|_2] \le \epsilon$ if $m \ge \frac{(2C)^{\frac{1}{k}}}{s^{1+\frac{1}{k}}\epsilon^{\frac{1}{k}}}\log(\frac{2\delta_0}{\epsilon})$ and $\eta \le (\frac{s\epsilon}{2C})^{\frac{1}{k}}$.

Therefore, there is a convergence rate of $m \in O(\frac{1}{\epsilon^{\frac{1}{k}}}\log(\frac{1}{\epsilon}))$. Since the discretization proposed in [CCBJ17] has a strong local order of 2, we are able to match their convergence results. However, this new result applies to a broader class of diffusions, not just second-order Langevin dynamics, assuming an appropriate discretization can be designed. Moreover, this result generalizes to any valid discretization, making it possible to achieve a faster convergence rate with higher-order methods, which can be explored in future work.

Proof

$$E^{Z^*}E^{X_m}[\|X_{m+1} - Z^*\|_2] \le E^{Z^*}E^{X_m}[\|Z_\eta - Z^*\|_2] + E^{X_m}[\|Z_\eta - X_{m+1}\|_2].$$

The first term on the right-hand side shrinks due to the decay:

$$\begin{aligned} E^{Z^*}E^{X_m}[\|Z_\eta - Z^*\|_2] \le& e^{-s\eta}E^{Z^*}E^{X_m}[\|X_m - Z^*\|_2], \\ =& e^{-s\eta}\delta_m, \end{aligned}$$

where we emphasize that X_m is not a single point but instead follows some distribution which is ultimately dependent on x_0. For convenience, we simply notate the initial conditions as $E^{X_m}[\cdot]$, although the full notation would be $E^{X_m \sim q}[\cdot]$ for some distribution.

The second term is controlled by the strong order of the integrator. We have:

$$\begin{aligned} \delta_{m+1} \le& e^{-s\eta}\delta_m + C\eta^{k+1}, \\ \le& e^{-(m+1)s\eta}\delta_0 + C\eta^{k+1}(1 + e^{-s\eta} + e^{-2s\eta} + ... + e^{-ms\eta}), \\ \le& e^{-(m+1)s\eta}\delta_0 + C\eta^{k+1}\frac{1}{1-e^{-s\eta}}, \\ \le& e^{-(m+1)s\eta}\delta_0 + CD\frac{1}{s}\eta^k, \end{aligned}$$

where in the last line we use the fact that η is very small so that $\frac{1}{\eta}$ dominates in the series representation of $\frac{1}{1-e^{-\eta}}$. We therefore may select some D such that $\frac{1}{1-e^{-s\eta}} \le D\frac{1}{s\eta}$. For simplicity, we absorb D into C.

Therefore, if we wish to have $\delta_m \leq \epsilon$, we can bound each term at a level of $\frac{\epsilon}{2}$, meaning the step size η must satisfy:

$$\begin{aligned} \frac{C}{s}\eta^k \leq & \frac{\epsilon}{2}, \\ \eta \leq & (\frac{s\epsilon}{2C})^{\frac{1}{k}}. \end{aligned}$$

Furthermore the number of steps m must satisfy:

$$\begin{aligned} e^{-ms\eta}\delta_0 \leq & \frac{\epsilon}{2}, \\ m \geq & \frac{1}{s\eta}\log(\frac{2\delta_0}{\epsilon}). \end{aligned}$$

Injecting the dependence of η on ϵ, we have:

$$m \geq \frac{(2C)^{\frac{1}{k}}}{s^{1+\frac{1}{k}}\epsilon^{\frac{1}{k}}}\log(\frac{2\delta_0}{\epsilon}).$$

□

Chapter 5

Wasserstein Rate

5.1 Motivation

In the previous section, we utilized a specific type of decay. In this section, we will discuss a more general concept called the Wasserstein rate of a diffusion and prove various important properties relating to it.

To begin, we define a coupling J between two distributions P and Q as a joint distribution whose marginals equal each distribution. There are several couplings that can satisfy this property.

A helpful physical interpretation of a coupling between $P(x)$ and $Q(y)$ is the amount of density from the point x that is moved from P to the point y in Q. For example, marginalization can be thought of as picking up all the density from one distribution and placing it in the other: $p(x) = \int j(x, y)dy$.

This inspires a symmetric distance between distributions known as the Earth-Movers distance:

$$d_{EM}(P, Q) \triangleq \inf_{J} \int \|x - y\|_2 dJ(x, y),$$

where J must be a valid coupling.

The intuition behind this name is that we can interpret the above integral as the total work someone must do to transport density from one distribution P to another Q. The work is a weighted integral which adds up how much density is being transported ($dJ(x, y)$) from x in P and y in Q, multiplied by the distance that mass must go ($\|x - y\|_2$). The minimizing joint distribution is called the optimal transport plan or optimal coupling.

This notion of distance has many advantages over other ones and has been applied in several machine learning algorithms.

5.2 Wasserstein Rate

In the following discussion, let $Z_t^{x_0}$ denote the diffusion Z_t with initial condition x_0. We will study how far apart the diffusion can get from itself when starting at different initial conditions.

For the set of all possible couplings ζ between $Z_t^{x_0}$ and $Z_t^{y_0}$, we say the diffusion Z_t has a L^p-Wasserstein rate of $r_p(t)$ if:

$$d_{W_p}(Z_t^x, Z_t^y) \triangleq \inf_{J \in \zeta} E_J[\|Z_t^{x_0} - Z_t^{y_0}\|_2^p]]^{\frac{1}{p}} \leq r_p(t)\|x_0 - y_0\|_2,$$

where the distance on the left is known as the Wasserstein distance between the two distributions. When $p = 1$, it is equal to the Earth-Movers distance.

5.2.1 Sufficient Conditions for a Decaying Wasserstein Rate

There are several ways to ensure that a diffusion obeys a Wasserstein rate that is exponentially decreasing. Below we introduce uniform dissipativity, a strong condition that is sufficient for a L^2-Wasserstein rate with exponential decay, which also implies the L^1 case. In Langevin Dynamics, this condition amounts to the strong convexity of the function of interest f.

Condition 4. *Uniform dissipativity.*

$$\forall x, y \quad 2\langle b(x) - b(y), x - y\rangle + \|\sigma(x) - \sigma(y)\|_F^2 \leq -k\|x - y\|_2^2.$$

Lemma 4. *[GDVM16] Suppose Condition 4 holds for Z_t. Then the diffusion Z_t has L^2-Wasserstein decay rate: $r(t) = e^{-\frac{kt}{2}}$.*

Proof We begin by applying Dynkin's formula to $g(x,t) = e^{kt}\|x\|_2^2$ for the difference diffusion $Z_t^x - Z_t^y$:

$$\begin{aligned} E[g(t, Z_t^x - Z_t^y)] =& \|x - y\|_2^2 + E\left[\int_0^t ke^{ks}\|Z_s^x - Z_s^y\|_2^2 ds\right], \\ &+ E\left[\int_0^t e^{ks}(2\langle b(Z_s^x) - b(Z_s^y), Z_s^x - Z_s^y\rangle + \|\sigma(Z_s^x) - \sigma(Z_s^y)\|_F^2)ds\right], \\ \leq& \|x - y\|_2^2. \end{aligned}$$

We also know that:

$$E[g(t, Z_t^x - Z_t^y)] = e^{kt}E[\|Z_t^x - Z_t^y\|_2^2] \geq e^{kt} \inf_{\zeta} E[\|Z_t^{x_0} - Z_t^{y_0}\|_2^2],$$

which combined with the previous statement gives the result. □

Another important condition which implies an exponentially decaying L^1-Wasserstein rate is called distant dissipativity [EMS18]. This condition is more general than uniform dissipativity and thus will be our condition of choice.

Condition 5. *Distant dissipativity. Define $\tilde{\sigma}(x) \triangleq (\sigma(x)\sigma(x)^T - s^2 I)^{\frac{1}{2}}$.*

Suppose the coefficients $b(x)$ and $\sigma(x)$ satisfy the following:

$$\frac{1}{s^2}\left[\frac{2\langle b(x) - b(y), x - y\rangle}{\|x - y\|_2^2} + \frac{\|\tilde{\sigma}(x) - \tilde{\sigma}(y)\|_F^2}{\|x - y\|_2^2} - \frac{\left\|(\tilde{\sigma}(x) - \tilde{\sigma}(y))^T(x - y)\right\|_2^2}{\|x - y\|_2^4}\right] \leq \begin{cases} -K, & \textit{if } \|x - y\|_2 > R \\ L, & \textit{if } \|x - y\|_2 \leq R \end{cases}$$

for $R, L \geq 0$, $K > 0$, and $s \in \left(0, \frac{1}{\mu_0(\sigma^{-1})}\right)$ where, for some matrix-valued function A, we define $\mu_0(A) \triangleq \sup_x \|A(x)\|_{op}$.

Lemma 5. *[EMS18] Suppose Condition 5 holds for Z_t. Then the diffusion Z_t has L^1-Wasserstein decay rate: $r(t) = 2e^{\frac{LR^2}{8}}e^{-\frac{kt}{2}}$ for k such that:*

$$s^2k^{-1} \leq \begin{cases} \frac{e-1}{2}R^2 + e\sqrt{\frac{8}{K}}R + \frac{4}{K}, & \text{if } LR^2 \leq 8 \\ \frac{8\sqrt{2\pi}}{RL^{\frac{1}{2}}}(\frac{1}{L} + \frac{1}{K})e^{\frac{LR^2}{8}} + \frac{32}{R^2K^2}, & \text{if } LR^2 > 8 \end{cases}$$

The term $e^{\frac{LR^2}{8}}$ has an effect on the rate of convergence. We see that when R is relatively small, convergence to the invariant measure is rapid. When R is relatively large, this term grows exponentially in R^2. Moreover, we see an influence on k based on s^2. If s^2 is large, then the exponential decay $e^{-\frac{kt}{2}}$ is faster.

5.2.2 Applications of Wasserstein Decay

The convergence rates guaranteed in the previous section (Proposition 2) can be achieved by assuming a Wasserstein decay. Hence, we can select a diffusion that satisfies either Condition 4 or 5 in order to achieve an efficient convergence rate in practice.

Lastly, we present one useful result due to a L^1-Wasserstein rate which provides us with Lipschitz continuity of the semi-group operator. This will be applied in the following chapters.

Lemma 6. *Suppose f is Lipschitz continuous with constant $M_1(f)$ and that the diffusion has a L^1-Wasserstein rate $r(t)$. Then we have:*

$$\begin{aligned} |E[f(Z_t^x) - f(Z_t^y)]| \leq & M_1(f)E[\|Z_t^x - Z_t^y\|_2], \\ \leq & M_1(f)d_{W_1}(Z_t^x, Z_t^y), \\ \leq & M_1(f)r(t)\|x-y\|_2, \end{aligned}$$

which implies that the semi-group is also Lipschitz continuous with a time-decaying Lipschitz constant.

Chapter 6

A Global Weak Error Result

We present a theorem on the relation between one-step weak approximation and the approximation on any time interval, including infinite time.

Let us start by defining the function $v(x,s)$ which is closely related to $u(x,s)$ introduced in the previous sections. It is the expectation of f at a final time T when starting the diffusion at the point x and time s:

$$\begin{aligned} v(x,s) \triangleq & E^{s,x}[f(Z_T)], \\ = & E^{x}[f(Z_{T-s})], \\ = & u(x, T-s). \end{aligned}$$

Through similar arguments to the ones for $u(x,t)$, $v(x,t)$ satisfies the Kolmogorov backward equation, this time with time derivatives reversed:

$$\begin{aligned} v(x,T) = & f(x), \\ \partial_t v(x,s) + \mathcal{A} v(x,s) = & 0. \end{aligned}$$

To aid with the below proof, we will use the following notation for denoting a diffusion evaluated at time t, starting at time t_0 and position x_0: $Z^t_{t_0,x_0}$, and likewise for the discrete setting. With this notation, we have that $v(x,s) = E[f(Z^T_{s,x})]$.

Theorem 1. *Suppose that f is Lipschitz continuous with coefficient $M_1(f)$.*

Suppose the strong local error is of order $k+1$ for any starting point x_0:

$$E^{x_0}[\|Z_\eta - X_1\|_2] \leq C(1 + \|x_0\|_2)\eta^{k+1},$$

Suppose the diffusion has a L^1-Wasserstein rate of $r(t) = Ke^{-mt}$ (which is implied by Condition 4 or Condition 5).

Lastly, assume the moments of the discretization are bounded by M.

Then we have the following result for any $T > 0$, $N\eta = T$, including the case as $N \to \infty$:

$$|E^{x_0}[f(Z_T) - f(X_N)]| \leq CKMM_1(f)\frac{1}{m}\eta^k.$$

Proof We will modify the proof from Milstein [Mil94] where the global error rate is inferred from the local one in the weak sense. The key modification to the proof in order to extend time to infinity will be using Wasserstein decay of the diffusion and Lipschitz continuity of the generator applied to $f(x)$. This in turn will allow us to utilize the local strong error of the discretization, instead of the local weak error as in Milstein.

First, let us clarify notation and begin with a simple identity. The following is true since we can split an expectation at time T into an expectation at time T after time η has elapsed.

$$E^{x_0}[f(Z_T)] = E[f(Z^T_{t_0,x_0})] = E[f(Z^T_{t_1,Z^{t_1}_{t_0,x_0}})].$$

Now let us generate a sum, just as in Milstein. We start with the quantity of interest and add and subtract terms:

$$E[f(Z^T_{t_0,x_0})] = E[f(Z^T_{t_1,Z^{t_1}_{t_0,x_0}})] - E[f(Z^T_{t_1,X_1})] + E[f(Z^T_{t_1,X_1})].$$

Now, replace the last term above with a new difference:

$$E[f(Z^T_{t_1,X_1})] = E[f(Z^T_{t_2,Z^{t_2}_{t_1,X_1}})] - E[f(Z^T_{t_2,X_2})] + E[f(Z^T_{t_2,X_2})].$$

Continuing this way we have:

$$E[f(Z^T_{t_0,x_0})] - f(X_N) = \sum_{i=0}^{N-1} E[f(Z^T_{t_{i+1},Z^{t_{i+1}}_{t_i,X_i}})] - E[f(Z^T_{t_{i+1},X_{i+1}})],$$

noting that the last index has redundancy since $Z^T_{t_N,Z^{t_N}_{t_{N-1},X_{N-1}}} = Z^{t_N}_{t_{N-1},X_{N-1}}$ and $Z^T_{t_N,X_N} = X_N$.

We notice that each term is of the following form, due to the definition of $v(x,s)$ and via the law of total expectation:

$$E[f(Z^T_{t_{i+1},Z^{t_{i+1}}_{t_i,X_i}})] - E[f(Z^T_{t_{i+1},X_{i+1}})] = E[v(Z^{t_{i+1}}_{t_i,X_i}, t_{i+1}) - v(X^{t_{i+1}}_{t_i,X_i}, t_{i+1})],$$

where we have applied the same notation $X^t_{t_0,x_0}$ to denote the discretization starting at time t_0 and position x_0, evaluated at some time t corresponding to a certain step. The conditional expectation is with respect to $Z^{t_{i+1}}_{t_i,X_i}$ and X_{i+1}. For simplicity we do not notate this condition explicitly.

Now, by Lemma 6, we apply the Wasserstein decay with rate $r(t) = Ke^{-mt}$ to $u(x,t)$, using a change of variables: $v(x,t) = u(x, T-t)$. We first apply the law of total expectation, where conditional expectations are with respect to X_i. We do this since the strong error of the diffusion depends on the moments of the diffusion.

$$\begin{aligned}
|E[u(Z^{t_{i+1}}_{t_i,X_i}, T-t_{i+1}) - u(X^{t_{i+1}}_{t_i,X_i}, T-t_{i+1})]|, &\leq E\big[|E[u(Z^{t_{i+1}}_{t_i,X_i}, T-t_{i+1}) - u(X^{t_{i+1}}_{t_i,X_i}, T-t_{i+1})]| \ |X_i\big], \\
&\leq M_1(f) r(T-t_{i+1}) E\big[\big\|Z^{t_{i+1}}_{t_i,X_i} - X^{t_{i+1}}_{t_i,X_i}\big\|_2\big], \\
&\leq CKE[1 + \|X_i\|_2] M_1(f) e^{-m(T-t_{i+1})}\eta^{k+1}, \\
&\leq CKMM_1(f) e^{-m(T-t_{i+1})}\eta^{k+1},
\end{aligned}$$

where in the last step we take advantage of the local strong error of the discretization. Returning to the sum of interest, we bound the quality as desired:

$$\begin{aligned}|E[f(Z^T_{t_0,x_0}) - f(X_N)]| \leq & CKMM_1(f)\eta^{k+1}\sum_{i=0}^{N-1} e^{-m(T-t_{i+1})},\\ \leq & CKMM_1(f)\eta^{k+1}\frac{1}{1-e^{-m\eta}},\\ \leq & CKMM_1(f)\frac{1}{m}\eta^k.\end{aligned}$$

□

This is a novel result that will enable the extension of local weak error to any number of time-steps, without the error diverging to infinity as the number of steps increase to infinity. Moreover, this result will enable for a new decomposition of the optimization error that is expected from simulating a diffusion. Not only will the discretization errors be modified in this optimization error estimate, but also it will be possible to directly utilize the long-term behaviour of a diffusion without the need for averaging.

Chapter 7

Optimization via Ito Diffusions

7.1 Invariant Measure

In this section, we motivate how diffusions can be used for optimization. We will center most of our motivation around Langevin Dynamics, but mention an important generalization recently made in the literature that we will apply in the following section.

To perform optimization, the first goal is to construct a diffusion that has an invariant measure which has desirable properties. As we will see soon, Langevin Dynamics is a good choice due to its invariant measure. Recall that Langevin Dynamics is the special case of an Ito Diffusion when $b(x) = -\nabla f(x)$ and $\sigma(x) = \sqrt{\frac{2}{\gamma}}I$ so $a(x) = \frac{2}{\gamma}I$.

It is well-known that Langevin Dynamics has an invariant measure that is the Gibbs measure of the function $f(x)$. We prove this for completeness.

Proposition 3. *Langevin Dynamics has an invariant measure $p_\gamma(x) \propto e^{-\gamma f(x)}$.*

Proof Consider the Fokker–Planck equation applied to Langevin diffusion, and assume there is an invariant measure such that $\frac{\partial}{\partial t}p(x,t) = 0$:

$$\begin{aligned}\frac{\partial}{\partial x}[\nabla f(x)p(x)] + \frac{\partial^2}{\partial x^2}[\frac{1}{\gamma}p(x)] &=0,\\ \leftrightarrow \frac{\partial}{\partial x}[\nabla f(x)p(x) + \frac{1}{\gamma}\frac{\partial}{\partial x}p(x)] &=0.\end{aligned}$$

We see that $p(x) \propto e^{-\gamma f(x)}$ satisfies this equality, as desired:

$$\nabla f(x)e^{-\gamma f(x)} - \frac{\gamma}{\gamma}\nabla f(x)e^{-\gamma f(x)} = 0.$$

□

When γ is large, the distribution concentrates most of its density around the minimas of $f(x)$, which is favourable in the case of optimization.

In general, one can construct a diffusion which has a desired invariant measure $p(x)$ [EMS18]. Again this follows by application of the Fokker–Planck equation.

To do this, one selects a drift coefficient equal to:

$$b(x) = \frac{1}{2p(x)} \langle \nabla, p(x)[a(x) + c(x)] \rangle.$$

where $c(x)$ is some skew-symmetric matrix. Here the gradient represents the divergence operator applied row-wise to a matrix: $\langle \nabla, a(x) \rangle \triangleq \sum_i \boldsymbol{e}_i \sum_j \partial_j a_{ij}(x)$.

Suppose we wish to target the Gibb's measure with a diffusion. That is $p_\gamma(x) \propto e^{-\gamma f(x)}$. Moreover, assume we use a diffusion coefficient of the following form: $a_\gamma(x) = \frac{1}{\gamma} g(x) I$, where $g(x)$ is a scalar function that scales the identity matrix. For simplicity, also assume $c(x) = 0$. Then we have the following simple form of the drift coefficient:

$$b_\gamma(x) = \frac{1}{2\gamma} \nabla g(x) - \frac{1}{2} g(x) \nabla f(x).$$

7.2 Exponential Convergence of Expectations

We now define a notion of exponential convergence that will be applicable to optimization. We wish to determine how fast expectations under a diffusion converge to the desired expectation under the invariant measure.

Denote $p(f)$ as the expectation of f under the invariant measure p: $p(f) = E^{Z \sim p}[f(Z)]$.

Condition 6. *Exponential Convergence of Expectations*

1. *There exists a unique limiting distribution p of Z_∞.*

2. *Expectations exponentially converge to the expectation under the invariant measure:* $|E^{x_0}[f(Z_t)] - p(f)| \leq G e^{-gt}$.

A diffusion can be proved to exponentially converge in expectation under the appropriate conditions. In fact, for our purposes, the L^1-Wasserstein decay of the diffusion is a stronger condition which implies this type of convergence. All that must be assumed in addition is the Lipschitz continuity of f.

Lemma 7. *Z_t inherits exponentially decaying convergence of expectations from an exponentially decaying L^1-Wasserstein rate of $r(t)$ and from the Lipschitz continuity of $f(x)$.*

Proof

$$\begin{aligned}
|E[f(Z_t) - f(Z_\infty)]| \leq & M_1(f) E[\|Z_t - Z_\infty\|_2], \\
\leq & M_1(f) d_{W^1}(Z_t, Z_\infty), \\
\leq & M_1(f) r(t) d_{W^1}(x_0, Z_\infty), \\
= & M_1(f) r(t) d_0,
\end{aligned}$$

where we denote d_0 as the initial distance (in W_1 metric) between the initial conditions x_0 and the invariant measure. □

7.3 Expected Sub-Optimality

Lastly, we discuss a useful fact about the limiting distribution of Langevin Dynamics. For the Gibbs distribution, let us consider the expected sub-optimality: $p(f) - \min_x f(x)$, which tells us how far apart expectations from the invariant measure differ from the global optimizer of $f(x)$. We borrow a proposition:

Proposition 4. *[EMS18] Expected Sub-optimality*
Let x^* be the global optimizer of $f(x)$ and suppose Z_t is an $\alpha - \beta$ dissipative diffusion with invariant measure p.

Fix $C > 0$ and $\theta \in (0,1]$. If $\forall x \log(p(x^*)) - \log(p(x)) \leq C\|x - x^*\|_2^{2\theta}$, then:

$$-p(\log(p)) + \log(p((x^*)) \leq \frac{d}{2}\left[\frac{1}{\theta}\log(\frac{2C}{D}) + \log(\frac{e\beta}{\alpha})\right].$$

When the invariant measure takes the form of a Gibbs measure $e^{-\gamma f(x)}$:

$$p(f) - f(x^*) \leq \frac{d}{2\gamma}\left[\log(\frac{2\gamma}{d}) + \log(\frac{e\beta M_2(f)}{2\alpha})\right],$$

We note that $f(x)$ cannot be both Lipschitz and $\alpha - \beta$ dissipative, as we will require, when the diffusion is Langevin. If Langevin Dynamics is to be used, instead of the generalized diffusions used in this work, one should relax the Lipschitz assumption to pseudo-Lipschitz. See for example the work of [EMS18] to make this extension.

Chapter 8

Optimization Error

8.1 Langevin Dynamics Optimization Error

In this section we combine our results from previous chapters to obtain a new optimization error bound for Langevin Dynamics.

Let x^* be the global minimzer of $f(x)$. Given the objective:

$$\min_x f(x) \triangleq f(x^*),$$

we wish to design a diffusion that will rapidly reach an invariant measure with expectations that closely approximate the global minima. Moreover, we need to account for the error due to discretizing the diffusion.

Inspired by the work of [EMS18], we can split the optimization error after N steps of our discretization as follows:

$$E[f(X_N)] - f(x^*) \leq |E[f(X_N) - f(Z_T)]| + |E[f(Z_T)] - p(f)| + |p(f) - f(x^*)|.$$

The first term on the right-hand side is called the integration error and is due to the choice of discretization. The middle term is determined by how fast the diffusion converges to its invariant measure. The last term is called the expected sub-optimality. Let us combine results from Proposition 4 for expected sub-optimality and Theorem 1 for the global weak error of a discretization.

Theorem 2. *Suppose the function we wish to optimize is $f(x)$. Suppose that f is Lipschitz continuous, with constant $M_1(f)$.*

Let Z_t be a diffusion with invariant measure equal to the Gibb's measure of $f(x)$: $p(x) \propto e^{-\gamma f(x)}$. This can be achieved through the appropriate choice of drift and diffusion coefficient.

Assume we have selected an appropriate discretization such that the strong local error is of order $k+1$ and such that its moments of sufficient order are bounded by M. Moreover, assume M is a bound for moments of sufficient order for the continuous diffusion.

Furthermore, assume Condition 5 (Distant Dissipativity) or Condition 4 (Uniform Dissipativity). This ensures an exponentially decaying Wasserstein rate is achieved with $r(t) = Ke^{-mt}$. These dissipativity conditions also ensure the diffusion is regularly dissipative, and hence Proposition 4 holds for the expected sub-optimality of the Gibb's measure.

A Wasserstein rate allows us access to Theorem 1 for the global weak discretization error. Furthermore, Condition 6 (exponential convergence of expectations) is implied by a Wasserstein rate and the Lipschitz continuity of f, due to Lemma 7.

Then the following holds for the discretization optimization error:

$$\begin{aligned} E[f(X_N)] - f(x^*) \leq & CKMM_1(f)\frac{1}{m}\eta^k \\ & + KM_1(f)e^{-mT}d_0 \\ & + \frac{d}{2\gamma}\left[\log(\frac{2\gamma}{d}) + \log(\frac{e\beta M_2(f)}{2\alpha})\right]. \end{aligned}$$

Proof Under the conditions of Theorem 1, we have that:

$$|E[f(X_N) - f(Z_T)]| \leq CKMM_1(f)\frac{1}{m}\eta^k.$$

Due to a Wasserstein rate, we have by Lemma 7:

$$|E[f(Z_T)] - p(f)| \leq KM_1(f)e^{-mT}d_0.$$

Moreover, under the conditions of Proposition 4, we have:

$$p(f) - f(x^*) \leq \frac{d}{2\gamma}\left[\log(\frac{2\gamma}{d}) + \log(\frac{e\beta M_2(f)}{2\alpha})\right].$$

Together these statements give the desired result. □

8.2 Example: Application to the Student-t Regression

The Student-t Regression is a popular modification to standard regression models that is more robust to outliers. This is since standard regression models assume errors are distributed via a Gaussian distribution which is very sensitive to outliers. The Gaussian distribution is a special case of the Student-t distribution in the limit where the degrees of freedom become infinite. Lower values of the degrees of freedom allow the Student-t distribution to have a heavier tail and hence become more robust to outliers. Moreover, instead of manually excluding or weighting data points, which becomes increasing difficult in higher dimensions, the Student-t distribution offers a natural and practical way to adapt for outliers.

Let us assume that the likelihood function we wish to optimize includes the Student-t regression model, for some fixed v. Then the log likelihood that arises is:

$$\log(p(x)) = -\frac{v+1}{2}\log(1+\frac{\|x\|_2^2}{v}).$$

We will consider the case where $v > 2d+2$, where d is the dimension. This this will ensure our proposed diffusion will be dissipative. Similar functions were considered in [LWME19, EMS18], using the same framework of designing diffusions to perform optimization. The difference here is that we will apply specific guarantees to the global discretization error, thus modifying the optimization error.

This is an abstraction of the regression problem. Typically, x contains some data terms as constants and the parameters to be optimized for. For example, we typically have $\|x\|_2^2 = (y-X\beta)^T\sigma^{-1}(y-X\beta)^T$ for regression. However, we will abstract these data terms away to focus on the main properties of optimization. Moreover, σ can be optimized with an optimization method of choice, alternating between the optimization method to be proposed here for estimating the optimal parameters β.

Therefore our objective function to be minimized is simply:

$$f(x) = \frac{v+1}{2}\log(1+\frac{\|x\|_2^2}{v}).$$

This objective function is simple with an obvious minima corresponding to $x = 0$. In practice, the minima would be offset by some unknown data-related constant. Therefore, we will assume the minima is not known. One should note that despite this objective's simplicity and popularity, it is non-convex and would be handled poorly by gradient descent. Due to f's sub-linear growth in $\|x\|_2$, both gradient descent and Langevin dynamics would fail to optimize this function. In the case of Langevin dynamics, the dissipativity assumption is not met due to a weak gradient. In particular, we note that the gradient of f does not scale linearly with x as x distances itself from the minima (in fact, it shrinks arbitrarily):

$$\nabla f(x) = \frac{(v+1)x}{v+\|x\|_2^2}.$$

To design a diffusion that does satisfy dissipativity, while having an invariant measure corresponding to the Gibb's measure, we make the following choices. First, we select $g(x) = v + \|x\|_2^2$, such that $\sigma(x) = \sqrt{g(x)}I$ and $a(x) = g(x)I$, where I is the identity matrix. We see that $\sigma(x)$ and $a(x) \triangleq \sigma(x)\sigma(x)^T$ satisfy the growth conditions of Condition 1. Define $\sigma_\gamma(x) \triangleq \frac{1}{\sqrt{\gamma}}\sigma(x)$.

Using the recipe for obtaining an invariant measure corresponding to the Gibb's distribution $p_\gamma(x) \propto$

$e^{-\gamma f(x)}$, we have that:

$$b_\gamma(x) = \frac{1}{2\gamma}\nabla g(x) - \frac{1}{2}g(x)\nabla f(x).$$

In this case, we have: $\nabla g(x) = 2x$. Therefore:

$$\begin{aligned} b_\gamma(x) =& \frac{1}{\gamma}x - \frac{1}{2}(v + \|x\|_2^2)\frac{(v+1)x}{v+\|x\|_2^2}, \\ =& \frac{1}{\gamma}x - \frac{1}{2}(v+1)x, \\ =& -\left(\frac{v+1}{2} - \frac{1}{\gamma}\right)x. \end{aligned}$$

This coefficient is clearly linear as needed by Condition 1. An appropriately large value of γ must be selected to ensure this quantity remains negative.

We now verify that this diffusion is dissipative for our choice of v and for $\gamma \geq 1$:

$$\begin{aligned} \mathcal{A}\|x\|_2^2 =& -2\left(\frac{v+1}{2} - \frac{1}{\gamma}\right)\|x\|_2^2 + \frac{d}{\gamma}(v + \|x\|_2^2), \\ =& -\left(v + 1 - \frac{2+d}{\gamma}\right)\|x\|_2^2 + \frac{dv}{\gamma}, \end{aligned}$$

where the dissipativity coefficients are $\alpha = (v + 1 - \frac{2+d}{\gamma}) > 0$ and $\beta = \frac{dv}{\gamma}$.

Moreover, this diffusion satisfies uniform dissipativity:

$$\begin{aligned} 2\langle b_\gamma(x) - b_\gamma(y), x - y\rangle + \|\sigma_\gamma(x) - \sigma_\gamma(y)\|_F^2, =& -\left(\frac{v+1}{2} - \frac{1}{\gamma}\right)\|x-y\|_2^2 \\ &+ \frac{d}{\gamma}\left(\sqrt{v + \|x\|_2^2} - \sqrt{v + \|y\|_2^2}\right)^2, \\ \leq& -\alpha_{uniform}\|x-y\|_2^2, \end{aligned}$$

where $\alpha_{uniform} = (\frac{v+1}{2} - \frac{1+d}{\gamma})$. By Lemma 4, this gives us a specific Wasserstein rate of $e^{-\frac{\alpha_{uniform}t}{2}}$.

We also note that $f(x)$ is Lipschitz, as its gradient is bounded. Moreover, the Hessian of $f(x)$ is also bounded. Hence both $M_1(f)$ and $M_2(f)$ exist.

Suppose we have selected the Euler Discretization, which has a strong local error of order 1 [Mil94]. Finally, for the continuous diffusion and for the Euler method, we have moment bound guarantees we can use, as outlined in the preliminary section. Hence M can be used as the mutual moment bound between these methods.

Therefore by Theorem 2, the global expected optimization error for this function would be of the following order:

$$\begin{aligned} E[f(X_N)] - f(x^*) \leq& CMM_1(f)\frac{2}{\alpha_{uniform}} \\ &+ M_1(f)e^{-\alpha_{uniform}T}d_0 \\ &+ \frac{d}{2\gamma}\left[\log(\frac{2\gamma}{d}) + \log(\frac{e\beta M_2(f)}{2\alpha})\right]. \end{aligned}$$

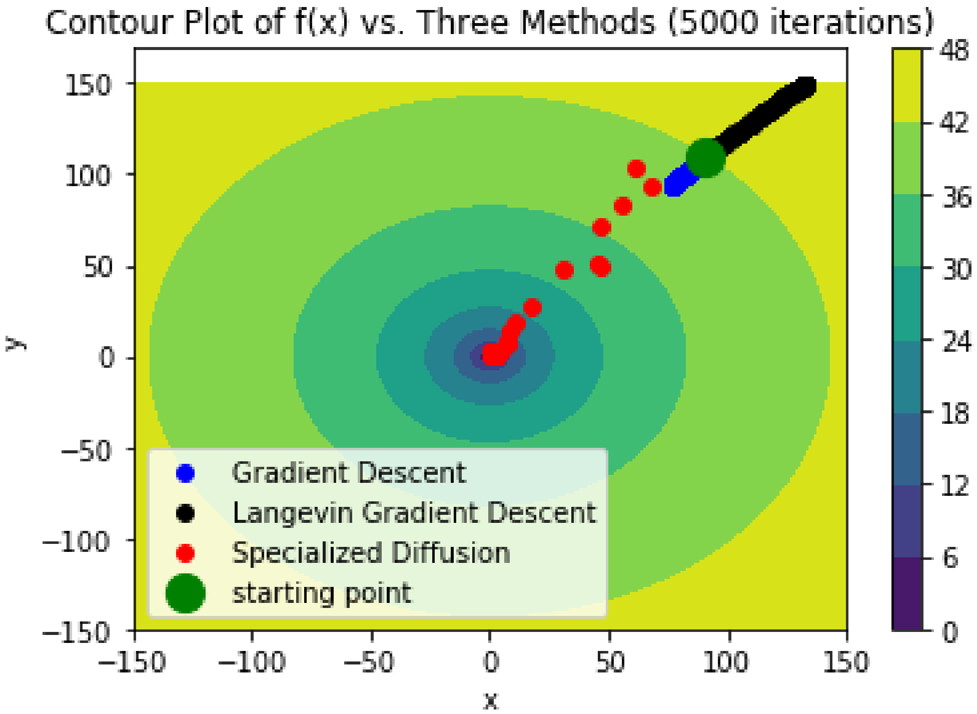

Figure 8.1: Comparison of Gradient Descent, the Langevin Gradient Descent (which diverges), and the diffusion designed in this example.

In Figure 8.1, we show how this diffusion performs compared to standard gradient descent and the Langevin algorithm for this choice of objective function $f(x)$. For illustration purposes, we choose a dimension of $d = 2$. The degrees of freedom is taken to be $v = 10$. The initial position is $x = (90, 110)$, the temperature is $\gamma = 1$, and the step size $\eta = 0.1$.

We see that only the specialized diffusion can optimize this non-convex objective, and that in fact the Langevin diffusion with noise diverges due to the weak gradient signal far from the minima. This is of practical interest since most gradient descent algorithms are stochastic, due to large data-set sizes. If the gradient is overpowered by noise, then the method would also fail to converge. Hence we have highlighted here the importance of a carefully crafted update equation.

Chapter 9

Conclusion

In this work, we studied how global optimization can be performed by using carefully crafted diffusions. We build on existing results for the expected optimization error by decomposing the error in a novel way. This decomposition relies on a newly developed infinite-time weak error result. Moreover, we review, validate and improve on prior work that analyzes the weak local error of the Euler discretization and strong local error of general discretizations. The former result will aid practitioners in calculating discretization error for their methods, and the latter result advances the theory for determining the convergence rates of various methods.

Areas for future work include computing upper bounds for higher-order methods such as the splitting scheme covered in this work. Specifically, the Taylor series expansion used here to analyze the weak error of the Euler discretization can be carried over to higher-order methods.

Lastly, we can extend the results of this work to the case where the full gradient is not accessible, as in stochastic gradient Langevin dynamics. Since the gradient can be estimated in an unbiased manner, we do not expect the analysis to change significantly and can follow the analysis in prior work [CDC15].

Interesting areas of future work would include designing diffusions with time-varying step-sizes [CCBJ17]. We also believe there is a way to infer the infinite-time weak error without use of strong error, but this remains an open area of research.

Bibliography

[CCAY$^+$18] 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 preprint arXiv:1805.01648 (2018).

[CCBJ17] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan, *Underdamped langevin mcmc: A non-asymptotic analysis*, arXiv preprint arXiv:1707.03663 (2017).

[CDC15] Changyou Chen, Nan Ding, and Lawrence Carin, *On the convergence of stochastic gradient mcmc algorithms with high-order integrators*, Advances in Neural Information Processing Systems, 2015, pp. 2278–2286.

[EMS18] Murat A Erdogdu, Lester Mackey, and Ohad Shamir, *Global non-convex optimization with discretized diffusions*, Advances in Neural Information Processing Systems, 2018, pp. 9671–9680.

[Erd17] Murat A Erdogdu, *Steins lemma and subsampling in large-scale optimization*, Ph.D. thesis, PhD thesis, Stanford University, 2017.

[GDVM16] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey, *Measuring sample quality with diffusions*, arXiv preprint arXiv:1611.06972 (2016).

[LWME19] Xuechen Li, Denny Wu, Lester Mackey, and Murat A Erdogdu, *Stochastic runge-kutta accelerates langevin monte carlo and beyond*, arXiv preprint arXiv:1906.07868 (2019).

[Mil94] Grigorii Noikhovich Milstein, *Numerical integration of stochastic differential equations*, vol. 313, Springer Science & Business Media, 1994.

[Øks03] Bernt Øksendal, *Stochastic differential equations*, Stochastic differential equations, Springer, 2003, pp. 65–84.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky, *Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis*, arXiv preprint arXiv:1702.03849 (2017).

[VZ$^+$15] Sebastian J Vollmer, Konstantinos C Zygalakis, et al., *(non-) asymptotic properties of stochastic gradient langevin dynamics*, arXiv preprint arXiv:1501.00438 (2015).

[Zyg11] KC Zygalakis, *On the existence and the applications of modified equations for stochastic differential equations*, SIAM Journal on Scientific Computing **33** (2011), no. 1, 102–130.