Large Scale Optimization and Sampling for Machine Learning and Statistics

Ayoub El Hanchi

February 2021

Department of Mathematics and Statistics
McGill University
Montreal, Quebec, Canada

A thesis submitted to McGill University in partial
fulfillment of the requirements of the degree of
Master of Science

Acknowledgements

I would like to thank my supervisor, Professor David Stephens, for his continuous support, and for giving me the chance to freely explore my interests. I am incredibly grateful for the opportunity to work on and learn about the fascinating topics I study. This would not have been possible without Professor Stephens.

To Cristina, thank you for always standing by me, encouraging me when things were not going as well as I wanted, and being my partner in this adventure. You are a constant source of inspiration and motivation for me, and I am so happy to have you by my side.

To my family, and particularly my parents and brother, thank you for supporting me through the best and the worst of this journey. I definitely would not have made it to this point without your sacrifices, and it would be right to say that you deserve as much credit as anybody for any successes I have had. I hope I continue to make you proud.

Abstract

Building systems capable of optimal decision-making under uncertainty is one of the great intellectual and engineering challenges of our time. Over the past century, two mathematical formulations of this problem have emerged as the main approaches to this problem: the Frequentist and Bayesian approaches. In many cases of interest, these two approaches naturally lead to two well-defined algorithmic problems: Optimization and Sampling. The growth of the size of datasets over the last few years put a strain on the previously developed methods for optimization and sampling, and a new set of algorithms was developed to adjust to the demands of modern machine learning and statistics. In this thesis, we review this newly developed set of algorithms and their convergence analyses, emphasizing the connection between the apparently separate algorithmic tasks of optimizing and sampling.

Résumé

Construire des systèmes capables de prendre des décisions optimales dans l'incertitude est l'un des grands défis intellectuels et techniques de notre temps. Au cours du siècle dernier, deux formulations mathématiques de ce problème ont émergé comme les principales approches à ce problème: l'approche Fréquentiste et l'approche Bayésienne. Dans de nombreux cas d'intérêt, ces deux approches conduisent naturellement à deux problèmes algorithmiques bien définis: l'optimisation et l'échantillonnage. La croissance de la taille des données ces dernières années a mis à l'épreuve les méthodes d'optimisation et d'échantillonnage développés précédemment, et un nouvel ensemble d'algorithmes a été développé pour s'adapter aux exigences de l'apprentissage automatique et des statistiques modernes. Dans cette thèse, nous révisons ces nouveaux algorithmes et leurs analyses de convergence, mettant l'accent sur la connexion entre les tâches algorithmiques, à priori distinctes, d'optimisation et d'échantillonnage.

Contents

Chapter 1

Introduction

One of the great modern intellectual and engineering challenges is the development of procedures and systems for optimal decision making under uncertainty. This is a deep problem, at the intersection of philosophy, mathematics, and computer science. Over the past century, two mathematical formulations of this problem, whose philosophical ramifications radically differ, have emerged as the principal contenders: the Frequentist approach and the Bayesian approach. For many problems of interest, the resulting decision making procedures from these two approaches naturally lead to two well defined algorithmic problems: optimization and sampling.

Perhaps the biggest promise of these systems is their ability to incorporate very large amounts of data into the decision making process, allowing the gathering and use of evidence on a scale unattainable before. A key ingredient in making such systems a reality is therefore the development of large scale optimization and sampling algorithms. Recent years have seen a flurry of research in this area, leading to the development of many new algorithms, with provable and explicit convergence guarantees. In particular, a few themes have stood out from this new literature compared to previous work.

First, the study of existing algorithms in continuous-time has led to many fruitful results. On the one hand, going to continuous-time has revealed structures that are hidden in discrete-time, leading to a better understanding of existing methods. On the other, it allowed the

development of new algorithms that are discretizations of known continuous-time processes. And perhaps just as importantly, it has allowed the use of well developed analytical tools in the study of convergence of these algorithms, connecting it to well developed topics such as optimal transport and dynamical systems.

Second, the use of controlled stochasticity has proven to be crucial in achieving state of the art results. From a purely computational point of view, stochasticity is a necessity when the size of the data is very large. If left uncontrolled however, it leads to a severe deterioration in performance. Luckily, the use of control variates and importance sampling strategies was provably shown to recover fast convergence rates using only cheap stochastic estimates.

Lastly, while at first glance very different, new connections between optimization and sampling were discovered and exploited, leading to a healthy flow of ideas between the two traditionally separate research communities. This led to significant advances in both areas, both on the algorithmic and analytical level.

The themes we have just discussed have led to a generic way of designing new optimization and sampling algorithms. One starts with a known continuous-time process converging to the desired solution. One then chooses a discretization method, giving rise to a deterministic algorithm. Finally, one replaces the quantities needed by the deterministic algorithm by stochastic estimates, and attempts to design control variates and importance sampling strategies to control the amount of stochasticity introduced.

In this thesis, I will attempt to carry out this construction starting from the two most basic processes. For optimization, I will consider gradient flow in Euclidean space, which, aside from being a very well studied process, has a very intuitive motivation behind its use for optimization: at each infinitesimal time-step, we move along the direction of steepest descent. For sampling, I will consider the Langevin diffusion process. Here, the initial motivation was based purely on the fact that this is a well studied stochastic process, known to converge to the desired solution. However, surprisingly, this process can be given the interpretation of a gradient flow in the space of probability measures equipped with the appropriate structures. We will explore this point of view as well, although most of the

analyses will rely on coupling techniques more closely related to the probabilistic point of view, since they yield the currently best known convergence rates in our setting.

I should note that attempting to cover all advances in this area is both out of my reach and almost impossible to cover in a single thesis. Instead, I will focus on the case of unconstrained optimization and sampling in Euclidean space, assuming strong convexity and smoothness of the function to be minimized or the potential to be sampled from. This is the scenario where the theory is most complete, and the results are the strongest. Furthermore, I will not cover the accelerated form of either process, which is admittedly the most interesting case since it achieves the oracle lower bound in optimization, and is known to converge faster in sampling. Nevertheless, my aim will be to give a complete treatment for the case I consider.

The rest of this chapter is a very short summary of statistical decision theory and one of its important applications: supervised learning. The goal of these summaries is to show how optimization and sampling problems naturally arise, and how the finite sum structure of the function to minimize or potential to sample from comes into existence for supervised learning problems.

1.1 Statistical Decision Theory

Statistical decision theory is a mathematical framework to analyze and construct decision rules under uncertainty. In the Frequentist approach, this only gives a framework for analysis: decision rules are constructed independently and then analyzed using the framework. In the Bayesian approach however, this framework automatically provides a method to construct an optimal decision rule.

The theory starts with the following components:

- $\mathcal{D}$: the set of possible observations.
- $\mathcal{P}$: A subset of the set of probability measures on $\mathcal{D}$.
- $\mathcal{A}$: the set of available actions.
- $\mathcal{L} : \mathcal{P} \times \mathcal{A} \to \mathbb{R}$: the loss function.

- $\delta : \mathcal{D} \to \mathcal{A}$: the decision rule.

The reasoning behind having these components goes as follows. We assume the state of the world can be summarized by a probability measure $P \in \mathcal{P}$. We observe $\mathcal{D} \ni D \sim P$, and we take action $\delta(D)$ with the aim of minimizing a loss function $\mathcal{L}(P, \delta(D))$ that measures how good the action $\delta(D)$ is in a particular state of the world P. If we knew what the state of the world P is, then we could ignore the observations, and simply pick the action that minimizes our loss. The problem, of course, is that we do not know what the state of world is. The way we deal with this uncertainty is what separates the Frequentist and the Bayesian approaches.

1.1.1 Frequentist approach

As previously mentioned, the Frequentist approach does not directly attempt to solve the problem of picking a decision rule. It simply states that *given* a decision rule δ, the appropriate measure of how good it is should be the frequentist risk:

$$R_F(P, \delta) \coloneqq \mathop{\mathbb{E}}_{D \sim P} [\mathcal{L}(P, \delta(D))] = \int_{\mathcal{D}} \mathcal{L}(P, \delta(d))\, dP(d)$$

In words, this means that to evaluate the effectiveness of a given decision rule, one should look at its performance when averaged over all possible observations. One is then free to come up with any decision rule, as long as one can show that it has small frequentist risk.

A consequence of using this criterion for evaluating decision rules is that in general there is no single decision rule that minimizes the frequentist risk across all possible states of the world. This approach can be succinctly summarized as: the state of the world P is fixed (but unknown), the observations X are random. Therefore, one should average over the observations to obtain a performance measure of a decision rule.

1.1.2 Bayesian approach

In contrast, the Bayesian approach asserts that the observations D are fixed (after we observe them), and that the state of the world P is uncertain. To express our uncertainty, we should therefore specify a probability measure on both the state of the world and the observations,

that is, on the set $\mathcal{D} \times \mathcal{P}$. This is usually specified as a distribution π on $\mathcal{P}$ referred to as the prior, and a conditional distribution $\rho(\cdot \mid P)$ on $\mathcal{D}$ referred to as the likelihood. Once the observations are made, we should update our beliefs about the state of the world by conditioning on the data to obtain the posterior distribution on $\mathcal{P}$:

$$\pi(\cdot \mid D) \propto \rho(D \mid \cdot)\pi(\cdot)$$

The appropriate measure of the quality of a decision rule is then given by the Bayesian posterior risk:

$$R_B(\delta \mid D) \coloneqq \mathop{\mathbb{E}}_{P \sim \pi(\cdot|D)} [\mathcal{L}(P, \delta(D))] = \int_{\mathcal{P}} \mathcal{L}(p, \delta(D))\, d\pi(p \mid D)$$

As oppose to the Frequentist approach where an optimal rule need not exist in general, an optimal rule for the Bayesian posterior risk can be easily characterized as:

$$\delta^*(D) \coloneqq \arg\min_{\delta} R_B(\delta \mid D)$$

1.2 Supervised Learning

One very important example of a decision that we might care about is that of predicting a real random variable $Y \in \mathbb{R}$ given a random variable $X \in \mathbb{R}^d$. This is usually known as regression in statistics and supervised learning in machine learning. We will use the latter for convenience.

Here we will frame this problem as a decision problem and embed it into the above framework. We will see that the Frequentist and Bayesian approaches give quite different methods, one leading to an optimization problem, and the other to a sampling problem (followed by an optimization problem).

The supervised learning problem can be formulated as a decision problem as follows. The observations $(X_i, Y_i)_{i=1}^N$ are assumed to be in $(\mathbb{R}^d \times \mathbb{R})^N$ for some $N \in \mathbb{N}$. The subset of probability measures $\mathcal{P}$ is given by those that are the N-times product of a single probability measure P on $\mathbb{R}^d \times \mathbb{R}$. The set of available actions is given by $\mathcal{F}$, a subset of the set of

functions from $\mathbb{R}^d$ to $\mathbb{R}$. Finally, the loss function is given by the generalization error:

$$\mathcal{L}\left(P, \delta((X_i, Y_i)_{i=1}^N)\right) \coloneqq \mathop{\mathbb{E}}_{(X,Y)\sim P}[l(f(X), Y)]$$

where $f \coloneqq \delta((X_i, Y_i)_{i=1}^N)$ and $l : \mathbb{R}^2 \to \mathbb{R}$ is a given function that evaluates the quality of a prediction.

1.2.1 Frequentist approach

The most widely used Frequentist decision rule for this problem is empirical risk minimization, and is given by:

$$\delta_F((X_i, Y_i)_{i=1}^N) \coloneqq \arg\min_{f\in\mathcal{F}} \left\{ \frac{1}{N} \sum_{i=1}^{N} l\left(f(X_i), Y_i\right) \right\}$$

In words, the intractable expectation in the generalization error is replaced by an expectation over the empirical measure coming from the data, which is then minimized. The theory showing that this decision rule has good Frequentist properties is statistical learning theory, and in particular, probably approximately correct (PAC) learning. This theory does not directly show good Frequentist risk, but rather, gives a high-probability bound that when using empirical risk minimization, the loss $\mathcal{L}$ will be as small as it can be within the class of functions $\mathcal{F}$ as the number of observations increases. This is the most widely employed decision rule in machine learning.

The class of functions $\mathcal{F}$ is usually given by:

$$\mathcal{F} \coloneqq \{f(\cdot, \theta) \mid \theta \in \mathbb{R}^d\}$$

so that the empirical risk minimization method can be stated as a finite sum optimization problem over Euclidean space:

$$\delta_F((X_i, Y_i)_{i=1}^N) \coloneqq \arg\min_{\theta\in\mathbb{R}^d} \left\{ \frac{1}{N} \sum_{i=1}^{N} l\left(f(X_i, \theta), Y_i\right) \right\} \tag{1.1}$$

1.2.2 Bayesian approach

In the Bayesian case, we follow the general decision-theoretic construction. We first further constrain the set of probability measures $\mathcal{P}$ to be given by those that are the N-times

product of a single probability measure P that is absolutely continuous with respect to Lebesgue measure λ on $\mathbb{R}^d \times \mathbb{R}$. Further we assume that the density of any such probability measure can be expressed as:

$$\frac{dP}{d\lambda} = \rho_x(x \mid \phi)\rho_y(y \mid x, \theta)$$

for some given density functions ρ_x and ρ_y parametrized by real vectors $\phi \in \mathbb{R}^k$ and $\theta \in \mathbb{R}^n$.

With these assumptions, we can now specify a prior over $\mathcal{P}$ by specifying a prior over the parameters ϕ and θ through their density π with respect to Lebesgue measure. We usually assume the factorized form:

$$\pi(\phi, \theta) \coloneqq \pi(\phi)\pi(\theta)$$

It is not hard to show that under our assumptions, the factorized form of the density of the prior is preserved in the density of the posterior:

$$\pi(\phi, \theta \mid (X_i, Y_i)_{i=1}^N) = \pi(\phi \mid (X_i, Y_i)_{i=1}^N)\pi(\theta \mid (X_i, Y_i)_{i=1}^N)$$

where:

$$\pi(\phi \mid (X_i, Y_i)_{i=1}^N) \propto \left\{ \prod_{i=1}^N \rho_x(X_i \mid \phi) \right\} \pi(\phi)$$
$$\pi(\theta \mid (X_i, Y_i)_{i=1}^N) \propto \left\{ \prod_{i=1}^N \rho_y(Y_i \mid X_i, \theta) \right\} \pi(\theta)$$

The class of functions $\mathcal{F}$ is usually assumed to be the set of all functions from $\mathbb{R}^d$ to $\mathbb{R}$, and the Bayesian decision rule can be shown to satisfy the pointwise equality:

$$\delta_B((X_i, Y_i)_{i=1}^N)(x) = \arg\min_{\hat{y} \in \mathbb{R}} \int_{\mathbb{R}^n} \mathop{\mathbb{E}}_{Y \sim \rho_y(\cdot \mid x, \theta)} [l(\hat{y}, Y)] \, \pi(\theta \mid (X_i, Y_i)_{i=1}^N) \, d\theta$$

The most difficult part of solving the above optimization problem is evaluating the expectation with respect to θ. It is usually estimated by sampling from the posterior and forming a Monte Carlo estimate. Writing the posterior density as:

$$\pi(\theta \mid (X_i, Y_i)_{i=1}^N) = e^{-U(\theta)}$$

we have:

$$U(\theta) \coloneqq \left\{ -\sum_{i=1}^{N} \log \rho_y(Y_i \mid X_i, \theta) \right\} - \log \pi(\theta) \tag{1.2}$$

so that the problem of sampling from the posterior is that of sampling from a distribution whose potential has a finite sum structure.

1.3 General Formulation and Organization

Motivated by the finite sum forms of (1.1) and (1.2), we are interested in the problem of optimizing a function or sampling from a potential with a finite sum structure. The reader is invited to ignore previously introduced notation as we will have no use for it.

Let $F : \mathbb{R}^d \to \mathbb{R}$ be a function. For the rest of this thesis, we will assume that F is twice differentiable. We define the optimization problem associated with F to be:

$$\text{Find} \quad x^* = \arg\min_{x \in \mathbb{R}^d} F(x) \tag{1.3}$$

Similarly, we define the sampling problem associated with F to be:

$$\text{Generate} \quad X \sim \rho^* \tag{1.4}$$

where ρ^* is defined by, for all Borel sets A:

$$\rho^*(A) \coloneqq \frac{\int_A \exp(-F(x))dx}{\int_{\mathbb{R}^d} \exp(-F(x))dx} \tag{1.5}$$

Of course, if we allow F to be any arbitrary twice differentiable function, neither (1.3) nor (1.4) are well defined in general. We therefore restrict our attention to strongly convex functions. We state this assumption formally next, which we also assume to hold for the rest of this thesis:

Assumption 1. *F is strongly convex, that is, there exists a $\mu > 0$ such that for all $x, y \in \mathbb{R}^d$:*

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2$$

We call μ the strong convexity constant of F.

Note that there are more general assumptions under which (1.3) and (1.4) are well defined. As we discussed in the introduction however, we will restrict ourselves to the strongly convex case as this is the scenario where the strongest results are available. Under strong convexity, we can show that both (1.3) and (1.4) are well defined. This follows from the following theorem and its corollary:

Theorem 1. *F has a unique minimizer $x^* \in \mathbb{R}^d$ and it is the unique solution to $\nabla F(x) = 0$.*

Proof.

Existence: Let $z \in \mathbb{R}^d$. Consider the set:

$$X = \{x \in \mathbb{R}^d \mid F(x) \leq F(z)\}$$

We claim that X is compact. Let $x \in X$. Then by strong convexity of F:

$$\begin{aligned} \|z - x\|_2^2 &\leq \frac{2}{\mu} \left[F(x) - F(z) - \langle \nabla F(z), x - z\rangle\right] \\ &\leq \frac{2}{\mu} \langle \nabla F(z), z - x\rangle \\ &\leq \frac{2}{\mu} \|\nabla F(z)\|_2 \|z - x\|_2 \end{aligned}$$

where the second line follows from the fact that $x \in X$, and the second from Cauchy-Schwarz. Therefore X is contained in the ball $B(z, \frac{2}{\mu} \|\nabla F(z)\|_2)$, and hence is bounded. To show that X is closed, let $(x_k)_{k\in\mathbb{N}}$ be a sequence in X converging to some $x \in \mathbb{R}^d$. We claim that $x \in X$. We have:

$$F(x_k) \leq F(z) \quad \forall k \in \mathbb{N} \Rightarrow \lim_{k\to\infty} F(x_k) \leq F(z) \Rightarrow F(x) \leq F(z)$$

where the second implication follows from the continuity of F. Therefore $x \in X$, and X is closed. By continuity of F and compactness of X, there exists a subset $X^* \subseteq X$ of minimizers of F over X. By definition of X, X^* is also the set of minimizers of F over all of $\mathbb{R}^d$. Let $x^* \in X^*$. We claim that $\nabla F(x^*) = 0$. Let $i \in [d]$ and e_i the i^{th} vector of the standard basis. By minimality of x^*, we have on the one hand:

$$[\nabla F(x^*)]_i = \lim_{h\to 0} \frac{F(x^* + he_i) - F(x^*)}{h} \geq 0$$

and on the other:

$$-\left[\nabla F(x^*)\right]_i = \lim_{h\to 0} \frac{F(x^* - he_i) - F(x^*)}{h} \geq 0$$

hence $\nabla F(x^*) = 0$.

Uniqueness: Let $x_1, x_2 \in X^*$. We have $F(x_1) = F(x_2)$ by their respective global minimalities. But by strong convexity of F and the fact that $\nabla F(x_2) = 0$ we have:

$$F(x_1) \geq F(x_2) + \frac{\mu}{2} \|x_1 - x_2\|_2^2$$

hence $x_1 = x_2$ and therefore $X^* = \{x^*\}$ contains only one element. Now suppose that there exists some $x \in \mathbb{R}^d \setminus \{x^*\}$ such that$\nabla F(x) = 0$. Then, again by strong convexity:

$$F(x^*) \geq F(x) + \frac{\mu}{2} \|x - x^*\|_2^2$$

but this contradicts the global minimality of x^*. □

Corollary 1. *The function* $\exp\left[-F(x)\right]$ *is integrable.*

Proof. Measurability of $\exp\left[-F(x)\right]$ follows from the continuity of F. By Theorem 1 and strong convexity of F we have for all $x \in \mathbb{R}^d$:

$$F(x) \geq F(x^*) + \frac{\mu}{2} \|x - x^*\|_2^2$$

so:

$$\begin{aligned}
\int_{\mathbb{R}^d} \exp\left[-F(x)\right] dx &\leq \exp\left[-F(x^*)\right] \int_{\mathbb{R}^d} \exp\left[-\frac{\mu}{2} \|x - x^*\|_2^2\right] dx \\
&\leq \exp\left[-F(x^*)\right] \left(\frac{2\pi}{\mu}\right)^{d/2} \\
&< \infty
\end{aligned}$$

□

Now that we have a precise formulation of both problems, and we are assured of their well-definedness, we give a short overview of the contents of the next chapters. We follow the generic recipe we mentioned in the introduction. In chapter 2, we introduce two continuous-time processes that solve the optimization and sampling problems respectively. We emphasis

the connection between these processes, and analyze their convergence in continuous-time. In chapter 3, we consider the Euler discretization of these processes, giving rise to our first algorithms. We study the effect of this discretization, and give explicit convergence rates. In chapter 4, we consider the optimization and sampling problems for functions F that can be expressed as expectations, and adapt the algorithms developed in chapters 2 and 3 to this case. Finally, in chapter 5, we focus on the special case where F is an expectation over a distribution of finite support, giving rise to a finite-sum structure. We develop specialized algorithms that take advantage of this additional structure, and show that they lead to significant performance improvements.

We will introduce additional assumptions as we need them. Similarly, we will make clear which model of complexity we are using whenever algorithms are discussed.

Chapter 2

Continuous Time Processes

In this chapter, we introduce two continuous-time processes that solve the optimization and sampling problems: gradient flow and Langevin diffusion. We start by showing that they indeed converge to the solutions of (1.3) and (1.4) respectively, and that this convergence is exponentially fast under our assumptions. We then establish two connections between optimization and sampling through these processes. The first connection stems from formulating the sampling problem as an optimization problem. In particular, we explore the interpretation of the Langevin dynamics as the gradient flow of relative entropy in the space of probability measures. The second connection comes from the opposite direction: we formulate the optimization problem as the problem of sampling from the Dirac measure concentrated on the minimizer.

Given the high level of technicality of these subjects, the overall tone of this chapter will be slightly informal. We will focus on the underlying ideas, and refer the reader to other sources where all the statements we mention are rigorously proved. Our goal in this chapter is simply to convince the reader that these processes are in some sense the "right" ones to consider.

As we will be discussing differential equations that make use on the gradient of F, we will find it useful to assume that F has Lipschitz gradients. In alignment with terminology used in the optimization literature, we will say that F is smooth if its gradient is Lipschitz

continuous. More explicitly we make the following assumption for the rest of this thesis:

Assumption 2. *F is smooth, that is, there exists an $L > 0$ such that for all $x, y \in \mathbb{R}^d$:*

$$\|\nabla F(y) - \nabla F(x)\|_2 \leq L \|y - x\|_2$$

We call L the smoothness constant of F.

For this section, the main use of this assumption will be to evoke existence and uniqueness theorems of ordinary and stochastic differential equations. In future chapters, this assumption will play different roles which we will discuss then.

Before introducing our first process, let us first state a simple lemma which will be useful throughout this chapter:

Lemma 1. *For all $x, y \in \mathbb{R}^d$:*

$$\langle \nabla F(y) - \nabla F(x), y - x \rangle \geq \mu \|y - x\|_2^2$$

Proof. By strong convexity of F we have:

$$F(y) - F(x) - \langle \nabla F(x), y - x \rangle \geq \frac{\mu}{2} \|y - x\|_2^2$$
$$F(x) - F(y) - \langle \nabla F(y), x - y \rangle \geq \frac{\mu}{2} \|x - y\|_2^2$$

Adding these two inequalities yields the result. □

2.1 Optimization through Gradient Flow

In the optimization problem (1.3), our goal is to find x^*, the unique minimizer of F. In light of Theorem 1, x^* is the unique solution to $\nabla F(x) = 0$. In most cases however, there is no analytic solution to this equation. One alternative is to start from some initial guess $x_0 \in \mathbb{R}^d$ and find a curve $(x_t)_{t \in \mathbb{R}^+}$ starting from x_0 and converging to x^*. How do we find such a curve ?

One reasonable possibility is the gradient flow of F starting at x_0. This curve is the solution to the initial value problem starting at x_0 and obeying:

$$dx_t = -\nabla F(x_t)\, dt \tag{2.1}$$

This is a natural curve to consider since heuristically, at each infinitesimal time step, we move in the direction of greatest decrease of the function F. The role of the magnitude of the gradient remains unclear at this point, but becomes clearer when this process is discretized. More precisely, we have the following convergence theorem:

Theorem 2. *The initial value problem starting at $x_0 \in \mathbb{R}^d$ and obeying (2.1) has a unique solution $(x_t)_{t\in\mathbb{R}^+}$ and it satisfies:*

$$\|x_t - x^*\|_2^2 \leq e^{-2\mu t} \|x_0 - x^*\|_2^2$$

for all $t \in \mathbb{R}$.

Proof. The existence and uniqueness of the solution $(x_t)_{t\in\mathbb{R}}$ follows from the smoothness of F and the standard theory of ordinary differential equations. For the convergence rate, we have:

$$\begin{aligned}
\frac{d}{dt} \|x_t - x^*\|_2^2 &= 2 \left\langle \frac{d}{dt}[x_t - x^*], x_t - x^* \right\rangle \\
&= -2 \langle \nabla F(x_t), x_t - x^* \rangle \\
&= -2 \langle \nabla F(x_t) - \nabla F(x^*), x_t - x^* \rangle \\
&\leq -2\mu \|x_t - x^*\|_2^2
\end{aligned}$$

where the second equality follows from the differential equation, the third from $\nabla F(x^*) = 0$, and the inequality follows from Lemma 1. Using Grönwall's inequality finishes the proof. □

2.2 Sampling through Langevin Diffusion

In the sampling problem (1.4), our goal is to generate a random variable whose distribution is ρ^* as defined in (1.5). Assuming that we have access to a source of uniform random variables, the computational equivalent of an analytical solution for a sampling problem would be to find an easily computable map $h : \mathbb{R} \to \mathbb{R}^d$ such that $h(u)$ is distributed according to ρ^* for a uniform random variable u. Just like in the optimization case however, we usually don't have an efficient way of finding such a map, particularly in high dimension. One alternative is to start with some random variable $x_0 \sim \rho_0$ for some initial probability measure ρ_0, and

find a stochastic process $(x_t)_{t\in\mathbb{R}^+}$ starting at x_0 such that the marginal distribution ρ_t of x_t converges to ρ^* in some appropriate sense.

One such stochastic process is the Langevin diffusion process associated with the potential F starting at $x_0 \sim \rho_0$. This stochastic process is the solution to the initial value problem starting at $x_0 \sim \rho_0$ and obeying:

$$dx_t = -\nabla F(x_t)\, dt + \sqrt{2}\, dW_t \tag{2.2}$$

where W_t is the standard Wiener process. We refer the reader to (Pavliotis (2014), chapters 2 and 3) for an introduction to diffusion processes and stochastic differential equations. We return to the issue of motivating the use of this process in the next section. We can nonetheless show that the marginals of the solution of this initial value problem converge to the target density ρ^*. First however, we need to equip the space of probability measures with a metric to formally have a notion of convergence. For reasons that will become more transparent in the next section, we select the 2-Wasserstein distance, which is defined on the set $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures on $\mathbb{R}^d$ with finite second moment. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance is defined by:

$$W_2(\mu,\nu) \coloneqq \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \mathop{\mathbb{E}}_{(y,x)\sim\gamma} \left[\|y-x\|_2^2 \right] \right)^{1/2} \tag{2.3}$$

where $\Gamma(\mu,\nu)$ is the collection of all probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. $\Gamma(\mu,\nu)$ is usually called the set of couplings of μ and ν. The fact that the quantity we defined is indeed a metric on $\mathcal{P}_2(\mathbb{R}^d)$ follows from optimal transport theory. See Villani (2003, 2009) for a detailed account. We are now ready to state and prove our second convergence theorem:

Theorem 3. *The initial value problem starting at $x_0 \sim \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and obeying (2.2) has a unique solution $(x_t)_{t\in\mathbb{R}^+}$ and it satisfies:*

$$W_2^2(\rho_t,\rho^*) \leq e^{-2\mu t} W_2^2(\rho_0,\rho^*)$$

for all $t \in \mathbb{R}^+$, where ρ_t is the distribution of x_t and ρ^* is as defined in (1.5).

Proof. The existence and uniqueness of the solution $(x_t)_{t\in\mathbb{R}^+}$ follows from the smoothness of F, the finiteness of the second moment of ρ_0, and the theory of stochastic differential

equations, see, e.g., Øksendal (2003); Evans (2012). Furthermore, from this same theory, we know that the marginals $(\rho_t)_{t\in\mathbb{R}^+}$ of $(x_t)_{t\in\mathbb{R}^+}$ all have finite second moments. This, combined with the finiteness of the second moment of ρ^* from Corollary 1, make the 2-Wasserstein distance $W_2(\rho_t, \rho^*)$ well defined for all $t \in \mathbb{R}^+$. In particular, the solution $(x_t)_{t\in\mathbb{R}^+}$ is an itô diffusion process, and as such the densities of its marginals ρ_t satisfy the forward Kolmogorov equation, also known as the Fokker-Planck equation, which reduces in our case to (see Pavliotis (2014), Section 4.5):

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla F) + \Delta \rho_t \tag{2.4}$$

where by abuse of notation we identified the densities with the corresponding measures.

Invariant measure: Notice that $\nabla F = -\nabla \log \rho^*$. If $\rho_0 = \rho^*$, then by the Fokker-Planck equation we have at $t = 0$:

$$\frac{\partial \rho_0}{\partial t} = -\nabla \cdot (\rho^* \nabla \log \rho^*) + \Delta \rho^* = -\Delta \rho^* + \Delta \rho^* = 0$$

so that if the process start at $x_0 \sim \rho^*$, then $x_t \sim \rho^*$ for all $t \in \mathbb{R}^+$. ρ^* is therefore an invariant measure of the process $(x_t)_{t\in\mathbb{R}^+}$. It is in fact unique, see (Pavliotis (2014), Proposition 4.2).

Convergence rate: We proceed using a coupling argument. Consider a second process $(y_t)_{t\in\mathbb{R}^+}$ starting at $y_0 \sim \rho^*$ and obeying the same stochastic differential equation:

$$dy_t = -\nabla F(y_t)\, dt + \sqrt{2}\, dW_t$$

Since ρ^* in the invariant measure of this process, we have $y_t \sim \rho^*$ for all $t \in \mathbb{R}^+$. Note that $(y_t)_{t\in\mathbb{R}^+}$ and $(x_t)_{t\in\mathbb{R}^+}$ are driven by the same Wiener process $(W_t)_{t\in\mathbb{R}^+}$. Therefore, the process $(y_t - x_t)_{t\in\mathbb{R}^+}$ is deterministic and differentiable, and we have:

$$\begin{aligned}
\frac{d}{dt} \|y_t - x_t\|_2^2 &= 2 \left\langle \frac{d}{dt}[y_t - x_t], y_t - x_t \right\rangle \\
&= -2 \langle \nabla F(y_t) - \nabla F(x_t), y_t - x_t \rangle \\
&\leq -2\mu \|y_t - x_t\|_2^2
\end{aligned}$$

where the second equality follows from the differential equation governing $(y_t - x_t)_{t\in\mathbb{R}^+}$, and the inequality from Lemma 1. Using Grönwall's inequality we get:

$$\|y_t - x_t\|_2 \leq e^{-2\mu t} \|y_0 - x_0\|_2^2$$
$$\mathbb{E}\left[\|y_t - x_t\|_2^2\right] \leq e^{-2\mu t}\mathbb{E}\left[\|y_0 - x_0\|_2^2\right]$$

By minimality of the coupling defining the 2-Wasserstein distance we have:

$$W_2^2(\rho_t, \rho^*) \leq \mathbb{E}\left[\|y_t - x_t\|_2^2\right]$$

Finally, we choose the initial coupling of x_0 and y_0 to be the optimal one to get:

$$\mathbb{E}\left[\|y_0 - x_0\|_2^2\right] = W_2^2(\rho_0, \rho^*)$$

and therefore:

$$W_2^2(\rho_t, \rho^*) \leq e^{-2\mu t} W_2^2(\rho_0, \rho^*)$$

□

2.3 Sampling as Minimization of Relative Entropy

The perspective we took in the previous section is probabilistic in nature, and is the one that we will use in subsequent chapters. However, there is a dual point of view one can take that better motivates the use of Langevin diffusion for sampling. This more analytic perspective comes from the following observation: the marginals of the stochastic process $(x_t)_{t\in\mathbb{R}^+}$ induce a curve $(\rho_t)_{t\in\mathbb{R}^+}$ in the space of probability measures.

In light of our discussion of gradient flow for optimization, we can try to look for a function on the space of probability measures that is strongly convex and is minimized at the target measure ρ^*. We could then consider its gradient flow, which heuristically should have the same linear convergence rate as the Euclidean one. Finally, we could try to look for a stochastic process that has the required marginals. This is very ambitious, for the space of probability measures is not even a vector space, so that many of the concepts we mentioned are not even defined. Still, amazingly, this can be done. This line of work was started by

Jordan et al. (1998) and culminated in the book Ambrosio et al. (2005). We follow the treatment of Ambrosio et al. (2005). Many of the statements we make in this subsection are not as precise as they can be. Our goal is only to give a glimpse into to how this construction is done.

Consider the complete metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ and a function $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$. Our goal will be to make sense of the gradient flow equation (2.1) of $\mathcal{F}$ in this metric space. That is, we would like to make sense of an equation of the form:

$$\text{“}\frac{d\rho_t}{dt} = -\nabla\mathcal{F}(\rho_t)\text{”} \tag{2.5}$$

where the left hand side should be viewed as the “derivative” of the function $\rho_t : \mathbb{R}^+ \to \mathcal{P}_2(\mathbb{R}^d)$ while the right hand side should be seen as the “gradient” of $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$. Clearly, neither of these are defined: the Fréchet derivative, which gives rises to the usual definitions of gradient and derivative we use in (2.1) only works for Banach spaces. The next two subsections will be concerned with making these notions more precise, while the last subsection treats the special case of the relative entropy functional and shows that the resulting gradient flow curve can be identified with the Langevin diffusion process.

2.3.1 Absolutely Continuous Curves in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

First, let us try to make sense of the left hand side of equation (2.5). We start by formalizing the notion of a curve.

Definition 1 (Curve)**.** *Let $I \subseteq \mathbb{R}$ be an interval. A continuous function $\gamma : I \to \mathcal{P}_2(\mathbb{R}^d)$ is called a curve.*

Our goal is to define the derivative of a curve. In Euclidean space, or more generally in a Banach space, this would normally be a vector. While the vector space structure is needed to define the direction, magnitudes can be defined using only the metric structure.

Definition 2 (Metric derivative)**.** *Let $\gamma : I \to \mathcal{P}_2(\mathbb{R}^d)$ be a curve. When it exists, we define its metric derivative at $t \in I$ to be:*

$$|\gamma'(t)| := \lim_{h\to 0} \frac{W_2(\gamma(t+h), \gamma(t))}{|h|}$$

The next step would be to define differentiable curves. For functions from $\mathbb{R}$ to $\mathbb{R}$, a slightly less constraining condition is that of absolute continuity. Inspired by the characterization of absolutely continuous functions on $\mathbb{R}$ given by the fundamental theorem of Lebesgue integration, one can define absolutely continuous curves on $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ as follows:

Definition 3 (Absolute continuity)**.** *A curve $\gamma : I \to \mathcal{P}_2(\mathbb{R}^d)$ is said to be absolutely continuous if there exists a $\beta \in L^1(I)$ such that:*

$$W_2(\gamma(s), \gamma(t)) \leq \int_s^t \beta(r)dr \quad \forall s < t \in I$$

The following proposition gives some further justification for this definition. See (Ambrosio et al. (2005), Theorem 1.1.2).

Proposition 1. *Let $\gamma : I \to \mathcal{P}_2(\mathbb{R}^d)$ be an absolutely continuous curve. Then for a.e. $t \in I$, γ has a metric derivative, $|\gamma'| \in L^1(I)$, and:*

$$W_2(\gamma(s), \gamma(t)) \leq \int_s^t |\gamma'|(r)dr \quad \forall s < t \in I$$

Furthermore, for any $\beta \in L^1(I)$ satisfying the condition of Definition 3:

$$|\gamma'|(t) \leq \beta(t) \quad \textit{for a.e.} \quad t \in I$$

To see why this justifies the above definition of absolute continuity, consider the following: if we replace $\mathcal{P}_2(\mathbb{R}^d)$ with $\mathbb{R}$ and assume γ is absolutely continuous (in the usual sense), then it would be almost everywhere differentiable and we would have equality in the condition of Definition 3. The above theorem says that a version of this holds on the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ with the above definition of absolute continuity.

Surprisingly, just like the derivative of an absolutely continuous function (in the usual sense) characterizes it, there exists a time dependent vector field that characterizes an absolutely continuous curve in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. The formal statement is the following. See (Ambrosio et al. (2005), Theorem 8.3.1).

Theorem 4. *Let I be an open interval in $\mathbb{R}$. Let $(\rho_t)_{t\in I}$ be a curve in $\mathcal{P}_2(\mathbb{R}^d)$. Then $(\rho_t)_{t\in I}$ is absolutely continuous if and only if there exists for each $t \in I$ a measurable vector field $v_t : \mathbb{R}^d \to \mathbb{R}^d$ such that:*

- $v_t \in L^2(\rho_t, \mathbb{R}^d)$ *for a.e.* $t \in I$.
- $\|v_t\|_{L^2(\rho_t, \mathbb{R}^d)} = |\rho'|(t)$ *for a.e.* $t \in I$.
- *The curve* $(\rho_t)_{t \in I}$ *satisfies the continuity equation:*

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t v_t) = 0 \tag{2.6}$$

In light of this result, it seems natural to associate the left hand side of equation (2.5) with the vector field v_t of Theorem 4.

2.3.2 Differentiation on $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

The goal of this section will be to make sense of the right-hand side of equation (2.5). In particular, consider a function $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$. Our goal will be to define the equivalent of a gradient of this function.

Let us first recall the definition of the gradient in Euclidean space. Note that this is a specialization of the definition of the Fréchet derivative on arbitrary Banach spaces. In particular, we make use of the Hilbert space structure of Euclidean space.

Definition 4. *The gradient of a differentiable function* $F : \mathbb{R}^d \to \mathbb{R}$ *at a point* $x \in \mathbb{R}^d$ *is the unique vector* $\nabla F(x) \in \mathbb{R}^d$ *satisfying:*

$$\lim_{y \to x} \frac{|F(y) - F(x) - \langle \nabla F(x), y - x \rangle|}{\|y - x\|_2} = 0$$

We can try to transpose this definition to the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the denominator can be replaced by $W_2(\mu, \nu)$ and the difference $F(y) - F(x)$ can be replaced by $\mathcal{F}(\mu) - \mathcal{F}(\nu)$. However, the inner product and the difference $y - x$, have no obvious candidates.

The following result from the theory of optimal transport points to a potential solution. For the remainder of this section, we will restrict ourselves to the metric space $(\mathcal{P}_2^{ab}(\mathbb{R}^d), W_2)$ where $\mathcal{P}_2^{ab}(\mathbb{R}^d)$ is the set of probability measures over $\mathbb{R}^d$ with finite second moments, and which are absolutely continuous with respect to Lebesgue measure. Before citing the result,

let us first introduce a piece of notation. For a measurable function $T : \mathbb{R}^d \to \mathbb{R}^d$, we define the pushforward measure $T_{\#}\mu$ to be:

$$T_{\#}\mu(B) := \mu(T^{-1}(B))$$

We are now ready to state the result. See (Ambrosio et al. (2005), section 6.2.3).

Theorem 5. *Let $\mu, \nu \in \mathcal{P}_2^{ab}(\mathbb{R}^d)$. Then there is a unique coupling $\gamma^* \in \Gamma(\mu, \nu)$ minimizing (2.3). Furthermore, there is an optimal transport map t_μ^ν such that $(t_\mu^\nu)_{\#}\mu = \nu$, and $\gamma^* = (Id, t_\mu^\nu)_{\#}\mu$ where Id is the identity map.*

To see why this result is useful to us, recall that we are trying to find a replacement to the term $y - x$ in the definition of the gradient of a function $\mathcal{F}$ at some reference probability measure μ. In light of Theorem 5, a natural candidate for this replacement is the map $t_\mu^\nu - Id$. We still need however some Hilbert space over which we can take an inner product. A natural candidate in this case is $L^2(\mu, \mathbb{R}^d)$. Based on these natural associations, we define the gradient of $\mathcal{F}$ as follows. See (Ambrosio et al. (2005), section 10.1).

Definition 5. *The gradient, if it exists, of a function $\mathcal{F} : \mathcal{P}_2^{ab}(\mathbb{R}^d) \to \mathbb{R}$ at a given probability measure μ is the unique function $\nabla\mathcal{F}(\mu) \in L^2(\mu, \mathbb{R}^d)$ satisfying:*

$$\lim_{\nu \to \mu} \frac{|\mathcal{F}(\nu) - \mathcal{F}(\mu) - \int_{\mathbb{R}^d} \langle \nabla\mathcal{F}(\mu)(x), t_\mu^\nu(x) - x\rangle \mu(x)\, dx|}{W_2(\nu, \mu)} = 0$$

If $\mathcal{F}$ has a gradient at all $\mu \in \mathcal{P}_2^{ab}(\mathbb{R}^d)$, then then we will say that $\mathcal{F}$ is differentiable.

With this definition, and assuming the function $\mathcal{F}$ is differentiable, we can define its gradient flow starting at some $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ by the absolutely continuous curve $(\rho_t)_{t \in \mathbb{R}^+}$ starting at ρ_0 and satisfying the continuity equation in Theorem 4 with $v_t = -\nabla\mathcal{F}(\rho_t)$:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla\mathcal{F}(\rho_t)) \tag{2.7}$$

2.3.3 Gradient Flow of relative entropy

Recall that our goal is to sample from the probability measure ρ^* defined in (1.5) which is absolutely continuous with respect to Lebesgue measure. One function that is known to be

minimized at the target measure is the relative entropy:

$$\mathcal{H}_{\rho^*}(\rho) \coloneqq \int_{\mathbb{R}^d} \rho(x) \log \frac{\rho(x)}{\rho^*(x)}\, dx$$

defined on $\mathcal{P}_2^{ab}(\mathbb{R}^d)$, and where we identify the elements of $\mathcal{P}_2^{ab}(\mathbb{R}^d)$ with their densities. It is known that $\mathcal{H}_{\rho^*}(\rho) \geq 0$ and $\mathcal{H}_{\rho^*}(\rho) = 0 \Leftrightarrow \rho = \rho^*$, so ρ^* is the unique minimizer of $\mathcal{H}_{\rho^*}(\rho)$. The gradient, in the sense of Definition 5, of relative entropy is given by (Ambrosio et al. (2005), Lemma 10.4.1):

$$\nabla \mathcal{H}_{\rho^*}(\rho) = \nabla \log \frac{\rho}{\rho^*}$$

where the gradient on the right-hand side is the usual Euclidean gradient. The gradient flow of $\mathcal{H}_{\rho^*}$ is therefore given by the absolutely continuous curve $(\rho_t)_{t \in \mathbb{R}^+}$ satisfying the equation:

$$\begin{aligned}
\frac{\partial \rho_t}{\partial t} &= \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\rho^*} \right) \\
&= \nabla \cdot (\rho_t \nabla [-\log \rho^*]) + \nabla \cdot (\rho_t \nabla \log \rho_t) \\
&= \nabla \cdot (\rho_t \nabla F) + \nabla \cdot (\nabla \rho_t) \\
&= \nabla \cdot (\rho_t \nabla F) + \Delta \rho_t
\end{aligned}$$

which is precisely the Fokker-Planck equation (2.4) governing the evolution of the marginals of the Langevin diffusion process !

2.4 Optimization as Sampling from Dirac Measure

The previous section shows how the sampling problem can be formulated as an optimization one over the space of probability measures, and how the Langevin diffusion process can be given the interpretation of the gradient flow of relative entropy. In this section, we go in the opposite direction.

Consider the Dirac measure defined by, for all Borel sets A:

$$\delta_{x^*}(A) = \begin{cases} 1 & \text{if} \quad x^* \in A \\ 0 & \text{otherwise} \end{cases}$$

Then it is clear that optimizing F is equivalent to sampling from δ_{x^*}. We have shown previously that the Langevin diffusion process can be viewed as a gradient flow. Here we ask: can we view the gradient flow of F as a Langevin diffusion process ?

We start by introducing a parameter $\beta > 0$ in the stochastic differential equation defining the Langevin diffusion process (2.2):

$$dx_t = -\nabla F(x_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t$$

For some initial condition $x_0 \sim \rho_0$, we refer to solutions of such stochastic differential equations by $(x_t^\beta)_{t\in\mathbb{R}^+}$ By an argument similar to the one we gave in the proof of Theorem 3, the invariant measures of $(x_t^\beta)_{t\in\mathbb{R}^+}$ are given by:

$$d\rho_\beta = \exp\left[-\beta F(x)\right] dx$$

Using the strong convexity of F, one can show that $\rho_\beta \to \delta_{x^*}$ weakly as $\beta \to \infty$. Similarly one can show that the solutions (x_t^β) converge to the gradient flow of F starting at x_0 as $\beta \to \infty$ in some appropriate sense. Therefore, one may view the gradient flow of F as a (limit of) Langevin diffusion process.

Chapter 3

Discrete Time Algorithms

In the previous chapter, we constructed two continuous time processes that solve the optimization and sampling problems, and showed that converge exponentially fast to their solutions in continuous time. Our task in this chapter will be to construct discretizations of these processes and study their discretization error and convergence rates.

3.1 Algorithms

3.1.1 Gradient Descent

Recall that our goal in optimization is to minimize F. We achieved this in continuous time by considering the gradient flow $(x_t)_{t\in\mathbb{R}^+}$ of F starting at some arbitrary point $x_0 \in \mathbb{R}^d$, and showing that x_t converges to x^* exponentially fast. To obtain an implementable algorithm, we need some way of evaluating x_t for a large enough time $t \in \mathbb{R}^+$. Unfortunately, in almost all cases, solving for the curve $(x_t)_{t\in\mathbb{R}^+}$ from its differential equation is not feasible. We instead rely on numerical methods to approximate it. In particular we use the simplest such method, namely Euler's.

We approximate the gradient flow of F starting at $x_0 \in \mathbb{R}^d$ as follows. Let $(\alpha_k)_{k=0}^{\infty}$ be a sequence of positive number with $\sum_{k=0}^{\infty} \alpha_k = \infty$. Define $t_k := \sum_{i=0}^{k-1} \alpha_i$ for $k \in \mathbb{N} \cup \{0\}$. Consider the partition $(t_k, t_{k+1})_{k=0}^{\infty}$ of $[0, \infty)$. We approximate the curve $(x_t)_{t\in\mathbb{R}^+}$ at the

points $(t_k)_{k=0}^{\infty}$ by the sequence $(x_k)_{k=0}^{\infty}$ defined as follows. We start by approximating x_{t_1} by:

$$\begin{aligned} x_{t_1} &= x_{t_0} - \int_{t_0}^{t_1} \nabla F(x_t)\, dt \\ &\approx x_0 - (t_1 - t_0)\nabla F(x_0) \\ &= x_0 - \alpha_0 \nabla F(x_0) \\ &=: x_1 \end{aligned}$$

Finally, we use this approximation to recursively approximate $x_{t_{k+1}}$ as:

$$\begin{aligned} x_{t_{k+1}} &= x_{t_k} - \int_{t_k}^{t_{k+1}} \nabla F(x_t)\, dt \\ &\approx x_{t_k} - (t_{k+1} - t_k)\nabla F(x_{t_k}) \\ &\approx x_k - \alpha_k \nabla F(x_k) \\ &=: x_{k+1} \end{aligned}$$

This sequence $(x_k)_{k=0}^{\infty}$ gives rise to Algorithm 1, known as gradient descent, whose convergence we study in at the end of this chapter.

Algorithm 1: Gradient Descent (GD)

Parameters: step sizes $(\alpha_k)_{k=1}^{\infty} > 0$

Initialization: $x_0 \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

  $x_{k+1} = x_k - \alpha_k \nabla F(x_k)$

end

3.1.2 Langevin Dynamics

We proceed in a similar fashion to derive the Langevin dynamics algorithm. Let $(\alpha_k)_{k=0}^{\infty}$ be a sequence of positive number with $\sum_{k=0}^{\infty} \alpha_k = \infty$. Define $t_k := \sum_{i=0}^{k-1} \alpha_i$ for $k \in \mathbb{N} \cup \{0\}$. Consider the partition $(t_k, t_{k+1})_{k=0}^{\infty}$ of $[0, \infty)$. We start by generating $x_0 \sim \rho_0$ (we will always

assume this is possible since we have the freedom to pick ρ_0), and approximate x_{t_1} by:

$$\begin{aligned}
x_{t_1} &= x_{t_0} - \int_{t_0}^{t_1} \nabla F(x_t)\, dt + \sqrt{2} \int_{t_0}^{t_1} dW_t \\
&= x_{t_0} - \int_{t_0}^{t_1} \nabla F(x_t)\, dt + \sqrt{2}\, [W(t_1) - W(t_0)] \\
&\approx x_0 - (t_1 - t_0)\nabla F(x_0) + \sqrt{2(t_1 - t_0)}\, \xi_0 \\
&= x_0 - \alpha_0 \nabla F(x_0) + \sqrt{2\alpha_0}\, \xi_0 \\
&=: x_1
\end{aligned}$$

Where $\xi_0 \sim \mathcal{N}(0, I_{d\times d})$. We then recursively approximate x_{t_k} by:

$$\begin{aligned}
x_{t_{k+1}} &= x_{t_k} - \int_{t_k}^{t_{k+1}} \nabla F(x_t)\, dt + \sqrt{2} \int_{t_k}^{t_{k+1}} dW_t \\
&= x_{t_k} - \int_{t_k}^{t_{k+1}} \nabla F(x_{t_k}) + \sqrt{2}\, [W(t_{k+1}) - W(t_k)] \\
&\approx x_{t_k} - (t_{k+1} - t_k)\nabla F(x_{t_k}) + \sqrt{2(t_{k+1} - t_k)}\, \xi_k \\
&\approx x_k - \alpha_k \nabla F(x_k) + \sqrt{2\alpha_k}\, \xi_k \\
&=: x_{k+1}
\end{aligned}$$

where $\xi_k \sim \mathcal{N}(0, I_{d\times d})$ and the collection $(\xi_k)_{k=0}^{\infty}$ is independent. The discrete time Markov chain $(x_k)_{k=0}^{\infty}$ gives rise to Algorithm 2, which we will call Langevin dynamics. We study the convergence of the marginals of this Markov chain in the section.

Algorithm 2: Langevin Dynamics (LD)

Parameters: step sizes $(\alpha_k)_{k=1}^{\infty} > 0$

Initialization: $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$

sample $x_0 \sim \rho_0$

for $k = 0, 1, 2, \ldots$ **do**

  sample $\xi_k \sim \mathcal{N}(0, I_{d\times d})$

  $x_{k+1} = x_k - \alpha_k \nabla F(x_k) + \sqrt{2\alpha_k}\, \xi_k$

end

3.2 Convergence Analysis

In this section we analyze the convergence of both gradient descent (GD) and Langevin dynamics (LD). We will evaluate the computational complexity of an algorithm by counting the number of iterations it requires to reach an ε accurate solution. In particular, for optimization problems, we use the criterion $\|x - x^*\|_2^2 \leq \varepsilon$, while for sampling we use $W_2^2(\rho, \rho^*) \leq \varepsilon$.

We do not introduce any additional assumptions here. We will however make use of the condition number of F given by $\kappa := L/\mu$. Compared to chapter 2, the role of the smoothness of F (Assumption 2) will be to allow us to bound the discretization error. In particular, we need the following refined version of Lemma 1. We refer the reader to (Nesterov (2004), Theorem 2.1.11) for a proof.

Lemma 2. *For all $x, y \in \mathbb{R}^d$:*

$$\langle \nabla F(y) - \nabla F(x), y - x \rangle \geq \frac{\mu L}{L + \mu} \|y - x\|_2^2 + \frac{1}{L + \mu} \|\nabla F(y) - \nabla F(x)\|_2^2$$

We will also need the "Peter-Paul" inequality, which we will use many times in subsequent chapters:

Lemma 3. *Let $x, y \in \mathbb{R}^d$. Then for all $\beta > 0$ we have:*

$$\|x \pm y\|_2^2 \leq (1 + \beta) \|x\|_2^2 + (1 + \beta^{-1}) \|y\|_2^2$$

Proof. We have for $x, y \in \mathbb{R}$ and $\beta > 0$:

$$\beta x^2 - 2xy + \beta^{-1} y^2 = \left(\sqrt{\beta} x - \sqrt{\beta^{-1}} y \right)^2 \geq 0 \Rightarrow 2xy \leq \beta x^2 + \beta^{-1} y^2$$
$$\beta x^2 + 2xy + \beta^{-1} y^2 = \left(\sqrt{\beta} x + \sqrt{\beta^{-1}} y \right)^2 \geq 0 \Rightarrow -2xy \leq \beta x^2 + \beta^{-1} y^2$$

Therefore:

$$2|xy| \leq \beta x^2 + \beta^{-1} y^2$$

Now by Cauchy-Schwarz inequality:

$$\begin{aligned}\|a \pm b\|_2^2 &= \|a\|_2^2 \pm 2\langle a, b\rangle + \|b\|_2^2 \\ &\leq \|a\|_2^2 + 2|\langle a, b\rangle| + \|b\|_2^2 \\ &\leq \|a\|_2^2 + 2\|a\|_2 \|b\|_2 + \|b\|_2^2 \\ &\leq (1+\beta)\|a\|_2^2 + (1+\beta^{-1})\|b\|_2^2\end{aligned}$$

□

3.2.1 Convergence of Gradient Descent

We now have all the tools to prove the following theorem.

Theorem 6. *Let $(x_k)_{k=0}^{\infty}$ be the gradient descent sequence generated by Algorithm 1 for a constant step size $\alpha_k = \alpha$ satisfying:*

$$\alpha \leq \frac{2}{L+\mu}$$

Then:

$$\|x_k - x^*\|_2^2 \leq \left(1 - 2\alpha\frac{\mu L}{L+\mu}\right)^k \|x_0 - x^*\|_2^2$$

Proof. Let $k \in \mathbb{N}$. We have:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|x_k - \alpha\nabla F(x_k) - x^*\|_2^2 \\ &= \|x_k - x^*\|_2^2 - 2\alpha\langle\nabla F(x_k), x_k - x^*\rangle + \alpha^2\|\nabla F(x_k)\|_2^2 \\ &= \|x_k - x^*\|_2^2 - 2\alpha\langle\nabla F(x_k) - \nabla F(x^*), x_k - x^*\rangle + \alpha^2\|\nabla F(x_k) - \nabla F(x^*)\|_2^2 \\ &\leq \left(1 - 2\alpha\frac{\mu L}{L+\mu}\right)\|x_k - x^*\|_2^2 + \alpha\left(\alpha - \frac{2}{L+\mu}\right)\|\nabla F(x_k) - \nabla F(x^*)\|_2^2 \\ &\leq \left(1 - 2\alpha\frac{\mu L}{L+\mu}\right)\|x_k - x^*\|_2^2\end{aligned}$$

where the third equality follows from $\nabla F(x^*) = 0$ from Theorem 1, the first inequality from Lemma 2, and the last inequality from the condition on the step size. The result then follows by induction. □

Based on this result, we can derive the complexity of gradient descent.

Corollary 2. *Let $(x_k)_{k=0}^{\infty}$ be the gradient descent sequence generated by Algorithm 1 with $\alpha_k = \alpha = 2/(L+\mu)$ and let $\varepsilon > 0$. If:*

$$k \geq \frac{\kappa+1}{4} \log\left(\frac{\|x_0 - x^*\|_2^2}{\varepsilon}\right)$$

Then:

$$\|x_k - x^*\|_2^2 \leq \varepsilon$$

Proof. With the chosen α it is easy to show:

$$1 - 2\alpha\frac{\mu L}{L+\mu} = \left(1 - \frac{2}{\kappa+1}\right)^2$$

Replacing in the bound of Theorem 6 we get:

$$\|x_k - x^*\|_2^2 \leq \left(1 - \frac{2}{\kappa+1}\right)^{2k} \|x_0 - x^*\|_2^2 \leq \exp\left(-\frac{4k}{\kappa+1}\right) \|x_0 - x^*\|_2^2$$

where the second inequality follows from $1 - x \leq \exp(-x)$. Bounding the right hand side by ε and solving for k by taking logarithms of both sides yields the result. □

Keeping only the dependence on the condition number κ and the precision ε, gradient descent therefore has complexity $O(\kappa \log(1/\varepsilon))$ for any $\varepsilon > 0$.

3.2.2 Convergence of Langevin Dynamics

To study the convergence of Langevin dynamics, we need a few more preliminary results. We start with a simple consequence of the Cauchy-Schwarz inequality.

Lemma 4. *Let $v : [a, b] \to \mathbb{R}^d$. Then:*

$$\left\| \int_a^b v(t)\, dt \right\|_2 \leq \int_a^b \|v(t)\|_2 \, dt$$

Proof. Let $u = \int_a^b v(t)\, dt$. We have:

$$\|u\|_2^2 = \sum_{i=1}^d u_i^2 = \sum_{i=1}^d u_i \int_a^b v_i(t)\, dt = \int_a^b \langle u, v(t) \rangle \, dt \leq \|u\|_2 \int_a^b \|v(t)\|_2 \, dt$$

where the inequality follows from Cauchy-Schwarz. □

Recall that for a real random variable X with finite second moment, its L_2 norm is given by:

$$\|X\|_{L_2} = \left(\mathbb{E}\left[\|X\|_2^2\right]\right)^{1/2}$$

We will make use of an application of Minkowski's inequality for integrals (see, e.g., Folland (2013), Theorem 6.19b).

Lemma 5. *Let $(X_t)_{t\in[a,b]}$ be a real valued integrable stochastic process with finite second moments. Then:*

$$\mathbb{E}\left[\left(\int_a^b X_t\,dt\right)^2\right]^{1/2} \leq \int_a^b \mathbb{E}\left[X_t^2\right]^{1/2}\,dt$$

Combining these lemmas gives us the following corollary which we will be useful to us.

Corollary 3. *Let $(X_t)_{t\in[a,b]}$ be a vector valued stochastic process with finite second moments. Then:*

$$\left\|\int_a^b X_t\,dt\right\|_{L_2} \leq \int_a^b \|X_t\|_{L_2}\,dt$$

Proof. We have:

$$\left\|\int_a^b X_t\,dt\right\|_{L_2} = \left(\mathbb{E}\left[\left\|\int_a^b X_t\,dt\right\|_2^2\right]\right)^{1/2} \leq \mathbb{E}\left[\left(\int_a^b \|X_t\|_2\,dt\right)^2\right]^{1/2} \leq \int_a^b \mathbb{E}\left[\|X_t\|_2^2\right]^{1/2}\,dt$$

where the first inequality follows from Lemma 4, and the second from Lemma 3. □

Finally, we will need the following bound on the second moment on the gradient:

Lemma 6. *Let $Y \sim \rho^*$. Then:*

$$\mathbb{E}\left[\|\nabla F(y)\|_2^2\right] \leq Ld$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\|\nabla F(Y)\|_2^2\right] &= \int_{\mathbb{R}^d} \|\nabla F(y)\|_2^2 \, \rho^*(y)\, dy \\
&= \int_{\mathbb{R}^d} \langle \nabla F(y), \nabla F(y)\rangle \rho^*(y)\, dy \\
&= \int_{\mathbb{R}^d} \langle \nabla \log \rho^*(y), \nabla \log \rho^*(y)\rangle \rho^*(y)\, dy \\
&= \int_{\mathbb{R}^d} \langle \nabla \rho^*(y), \nabla \log \rho^*(y)\rangle \, dy \\
&= \lim_{r\to\infty} \int_{B(x^*,r)} \langle \nabla \rho^*(y), \nabla \log \rho^*(y)\rangle \, dy
\end{aligned}$$

where the last line is justified by the positivity of the integrand and the monotone convergence theorem. Applying integration by parts we get in the first term, where we denote by $\hat{n}$ the unit normal vector to the surface of the ball $B(x^*, r)$:

$$\lim_{r\to\infty} \int_{\partial B(x^*,r)} \rho^*(y) \langle \nabla \log \rho^*(y), \hat{n}\rangle \, dS$$

We claim this term is zero since:

$$\begin{aligned}
\left| \int_{\partial B(x^*,r)} \rho^*(y) \langle \nabla \log \rho^*(y), \hat{n}\rangle \, dS \right| &\le \int_{\partial B(x^*,r)} |\rho^*(y) \langle \nabla \log \rho^*(y), \hat{n}\rangle| dS \\
&\le \int_{\partial B(x^*,r)} |\rho^*(y)| \, \|\nabla \log \rho^*(y)\|_2 \, dS \\
&\le \frac{1}{C_1} \exp\left[-F(x^*) - \frac{\mu}{2} r^2\right] \int_{\partial B(x^*,r)} \|\nabla F(y)\|_2 \, dS \\
&\le \frac{1}{C_1} \exp\left[-F(x^*) - \frac{\mu}{2} r^2\right] \int_{\partial B(x^*,r)} \|\nabla F(y) - \nabla F(x^*)\|_2 \, dS \\
&\le \frac{L}{C_1} \exp\left[-F(x^*) - \frac{\mu}{2} r^2\right] r \int_{\partial B(x^*,r)} dS \\
&= \frac{L C_2}{C_1} \exp\left[-F(x^*) - \frac{\mu}{2} r^2\right] r^d
\end{aligned}$$

where in the second line we used Cauchy-Schwarz and $\|\hat{n}\|_2 = 1$, in the third line we used the strong convexity of F, and C_1 is the normalization constant of ρ^*. In the fourth line we used $\nabla F(x^*) = 0$. In the fifth line we used the smoothness of F. In the last line we used that the surface area of the d-dimension sphere of radius r is $C_2 r^{d-1}$ for a constant C_2.

Taking the limit $r \to \infty$ give 0.Therefore, only the second term coming from integration by parts is non zero and we have:

$$\begin{aligned}\mathbb{E}\left[\|\nabla F(Y)\|_2^2\right] &= -\int_{\mathbb{R}^d} \Delta \log \rho^*(y)\rho^*(y)\, dy \\ &= \int_{\mathbb{R}^d} \Delta F(y)\rho^*(y)\, dy \\ &\leq Ld\end{aligned}$$

where in the last line we used:

$$\Delta F(y) = \sum_{i=1}^{d} \frac{\partial^2 F}{\partial y_i^2} = \mathrm{Tr}\left[\nabla^2 F(y)\right] \leq \sum_{i=1}^{d} \lambda_i(\nabla^2 F(y)) \leq Ld$$

where $\lambda_i(\nabla^2 F(y))$ is the i^{th} eigenvalue of the Hessian matrix and where the last inequality follows from the smoothness of F. □

We are now ready to state and prove our convergence theorem for the Langevin dynamics algorithm.

Theorem 7. *Let $(x_k)_{k=0}^{\infty}$ be the Markov chain simulated by Algorithm 2 with a constant step size $\alpha_k = \alpha$ satisfying:*

$$\alpha \leq \frac{2}{L+\mu}$$

Then:

$$W_2^2(\rho_k, \rho^*) \leq \left(1 - \alpha\frac{\mu L}{L+\mu}\right)^k W_2^2(\rho_0, \rho^*) + 12\alpha\kappa^2 d$$

where ρ_k is the distribution of x_k.

Proof. We proceed using a coupling argument. Consider a Langevin diffusion process $(y_t)_{t\in\mathbb{R}^+}$ satisfying:

$$dy_t = -\nabla F(y_t)\, dt + \sqrt{2}\, dW_t$$

and starting at $y_0 \sim \rho^*$. From the proof of Theorem 3, we know that this implies $y_t \sim \rho^*$ for all $t \in \mathbb{R}^+$. We also assume that the same Wiener process drives both $(y_t)_{t\in\mathbb{R}^+}$ and $(x_k)_{k=0}^{\infty}$. Furthermore we assume that x_0 and y_0 are optimally coupled so that:

$$W_2^2(\rho_0, \rho^*) = \mathbb{E}\left[\|y_0 - x_0\|_2^2\right]$$

Now we have:

$$
\begin{aligned}
&\left\|y_{(k+1)\alpha} - x_{k+1}\right\|_2^2 \\
&= \left\| y_{k\alpha} - x_k - \left(\int_{k\alpha}^{(k+1)\alpha} \nabla F(y_s)\, ds - \alpha \nabla F(x_k) \right) \right\|_2^2 \\
&= \left\| y_{k\alpha} - x_k - \alpha \left(\nabla F(y_{k\alpha}) - \nabla F(x_k) \right) - \int_{k\alpha}^{(k+1)\alpha} \left[\nabla F(y_s) - \nabla F(y_{k\alpha}) \right] ds \right\|_2^2 \\
&\leq (1+\beta) \left\| y_{k\alpha} - x_k - \alpha \left(\nabla F(y_{k\alpha}) - \nabla F(x_k) \right) \right\|_2^2 + (1+\beta^{-1}) \left\| \int_{k\alpha}^{(k+1)\alpha} \left[\nabla F(y_s) - \nabla F(y_{k\alpha}) \right] ds \right\|_2^2
\end{aligned}
$$

where in the first equality we used the fact that the same Wiener process drives both processes, and the last inequality follows from Lemma 3 for a free parameters $\beta > 0$. The first term can be bound as:

$$
\begin{aligned}
&\left\| y_{k\alpha} - x_k - \alpha \left(\nabla F(y_{k\alpha}) - \nabla F(x_k) \right) \right\|_2^2 \\
&= \left\| y_{k\alpha} - x_k \right\|_2^2 - 2\alpha \langle \nabla F(y_{k\alpha}) - \nabla F(x_k), y_{k\alpha} - x_k \rangle + \alpha^2 \left\| \nabla F(y_{k\alpha}) - \nabla F(x_k) \right\|_2^2 \\
&\leq \left(1 - 2\alpha \frac{\mu L}{L+\mu} \right) \left\| y_{k\alpha} - x_k \right\|_2^2 + \alpha \left(\alpha - \frac{2}{L+\mu} \right) \left\| \nabla F(y_{k\alpha}) - \nabla F(x_k) \right\|_2^2 \\
&\leq \left(1 - 2\alpha \frac{\mu L}{L+\mu} \right) \left\| y_{k\alpha} - x_k \right\|_2^2
\end{aligned}
$$

where the first inequality follows from Lemma 2, and the second from the condition on the

step size.We now bound the expectation of the second term:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\int_{k\alpha}^{(k+1)\alpha}\left[\nabla F(y_s)-\nabla F(y_{k\alpha})\right]\,ds\right\|_2^2\right]\\
&=\left\|\int_{k\alpha}^{(k+1)\alpha}\left[\nabla F(y_s)-\nabla F(y_{k\alpha})\right]\,ds\right\|_{L_2}^2\\
&\le\left(\int_{k\alpha}^{(k+1)\alpha}\left\|\nabla F(y_s)-\nabla F(y_{k\alpha})\right\|_{L_2}\,ds\right)^2\\
&\le L^2\left(\int_{k\alpha}^{(k+1)\alpha}\left\|y_s-y_{k\alpha}\right\|_{L_2}\,ds\right)^2\\
&\le L^2\left(\int_{k\alpha}^{(k+1)\alpha}\left\|\int_{k\alpha}^{s}\nabla F(y_t)\,dt\right\|_{L_2}ds+\int_{k\alpha}^{(k+1)\alpha}\left\|\int_{k\alpha}^{s}\sqrt{2}\,dW_t\right\|_{L_2}ds\right)^2\\
&\le L^2\left(\int_{k\alpha}^{(k+1)\alpha}\int_{k\alpha}^{s}\left\|\nabla F(y_t)\right\|_{L_2}\,dt\,ds+\sqrt{2d}\int_{k\alpha}^{(k+1)\alpha}\sqrt{s}\,ds\right)^2\\
&=L^2\left(\frac{1}{2}\alpha^2\left\|\nabla F(y_{k\alpha})\right\|_{L_2}+\frac{2\sqrt{2}}{3}\alpha^{3/2}\sqrt{d}\right)^2\\
&\le L^2\left(\frac{1}{2}\alpha^2\sqrt{Ld}+\frac{2\sqrt{2}}{3}\alpha^{3/2}\sqrt{d}\right)^2\\
&=L^2\alpha^3 d\left(\frac{1}{2}\sqrt{\alpha L}+\frac{2\sqrt{2}}{3}\right)^2\\
&\le L^2\alpha^3 d\left(\frac{\sqrt{2}}{2}+\frac{2\sqrt{2}}{3}\right)^2\\
&\le 3\alpha^3L^2d
\end{aligned}
$$

where the first line follows from the definition of the L_2 norm, the second from Lemma 3, the third from the smoothness of F, the fourth from the definition of the process $(y_t)_{t\in\mathbb{R}^+}$ and the triangle inequality, the fifth from Lemma 3, the sixth from the fact that the process $(y_t)_{t\in\mathbb{R}^+}$ is stationary, the seventh from Lemma 6, and finally the eighth from $\alpha\le 2/L$.Before putting the inequalities together, we choose $\beta=\alpha(\mu L)/(L+\mu)<1$ so that the coefficient of the first term is bounded by:

$$\left(1+\alpha\frac{\mu L}{L+\mu}\right)\left(1-2\alpha\frac{\mu L}{L+\mu}\right)\le 1-\alpha\frac{\mu L}{L+\mu}$$

while that of the second term is bounded by:

$$\left(1+\beta^{-1}\right) \leq 2\beta^{-1} = \frac{2(L+\mu)}{\alpha\mu L}$$

The overall bound is therefore:

$$\mathbb{E}\left[\left\|y_{(k+1)\alpha} - x_{k+1}\right\|_2^2\right] \leq \left(1-\alpha\frac{\mu L}{L+\mu}\right)\mathbb{E}\left[\left\|y_{k\alpha} - x_k\right\|_2^2\right] + 6\alpha^2\kappa(L+\mu)d$$

so that by induction we get:

$$\begin{aligned}
&\mathbb{E}\left[\left\|y_{(k)\alpha} - x_k\right\|_2^2\right] \\
&\leq \left(1-\alpha\frac{\mu L}{L+\mu}\right)^k \mathbb{E}\left[\left\|y_0 - x_0\right\|_2^2\right] + 6\alpha^2\kappa(L+\mu)d\sum_{i=0}^{k-1}\left(1-\alpha\frac{\mu L}{L+\mu}\right)^i \\
&\leq \left(1-\alpha\frac{\mu L}{L+\mu}\right)^k \mathbb{E}\left[\left\|y_0 - x_0\right\|_2^2\right] + 6\alpha^2\kappa(L+\mu)d\sum_{i=0}^{\infty}\left(1-\alpha\frac{\mu L}{L+\mu}\right)^i \\
&\leq \left(1-\alpha\frac{\mu L}{L+\mu}\right)^k \mathbb{E}\left[\left\|y_0 - x_0\right\|_2^2\right] + 12\alpha\kappa^2 d
\end{aligned}$$

By minimality of the coupling defining the Wasserstein distance we get:

$$W_2^2(\rho_k, \rho^*) \leq \mathbb{E}\left[\left\|y_{k\alpha} - x_k\right\|_2^2\right]$$

and by the the fact that x_0 and y_0 are optimally coupled we obtain the stated result. □

We can now derive the complexity of the Langevin dynamics algorithm.

Corollary 4. *Let $\varepsilon > 0$, and let $(x_k)_{k=0}^{\infty}$ be the Markov chain simulated by Algorithm 2 with step size:*

$$\alpha_k = \alpha = \min\left\{\frac{2}{L+\mu}, \frac{\varepsilon}{24\kappa^2 d}\right\}$$

If:

$$k \geq \max\left\{\frac{\kappa+1}{2}, \frac{24\kappa^2 d(L+\mu)}{\varepsilon\mu L}\right\}\log\left(\frac{2W_2^2(\rho_0,\rho^*)}{\varepsilon}\right)$$

Then:

$$W_2^2(\rho_k, \rho^*) \leq \varepsilon$$

where ρ_k is the distribution of x_k.

Proof. With the chosen α we have:

$$12\alpha\kappa^2 d \leq \frac{\varepsilon}{2}$$

combining this with Theorem 7 we have:

$$W_2^2(\rho_k, \rho^*) \leq \left(1 - \alpha\frac{\mu L}{L+\mu}\right)^k W_2^2(\rho_0, \rho^*) + \frac{\varepsilon}{2}$$

replacing α with its value, using the inequality $1 - x \leq \exp(-x)$, and solving for k we get the result. □

The convergence of Algorithm 2 has two regimes. In the low precision regime, $\varepsilon > O(\kappa^2 d)$, it has complexity $O(\kappa \log(1/\varepsilon))$, similar to gradient descent. In the high precision regime $\varepsilon < O(\kappa^2 d)$ however, this complexity deteriorates rapidly to $\tilde{O}(\kappa^2 d/\varepsilon)$. Notice also the linear dimension dependence, which is completely absent from the complexity of gradient descent. This makes the sampling problem significantly more difficult than the optimization one in the high precision high dimension regime. Note that by using decreasing step sizes, one can get rid of the logarithmic factor, but we won't pursue this further here.

Chapter 4

Stochastic Algorithms

In the last two chapters, we studied continuous time processes that solve the optimization and sampling problems, showed how to discretize them, and studied their convergence rates. We did all this in the general case where F has no particular structure besides strong convexity and smoothness. In this chapter, we will treat the case where F can be expressed as an expectation. Problems of this form occur often in machine learning and statistics.

The setup we will consider is the following. We assume that F is of the form:

$$F(x) := \mathbb{E}\left[f(x, \zeta)\right] \tag{4.1}$$

where ζ is a random variable taking values in some arbitrary space E over which the expectation is taken. If we could compute this expectation directly, we could then just use the algorithms from chapter 3 to solve the corresponding optimization or sampling problem. The underlying assumption here however is that ζ models the environment, in which case we do not know its distribution and are only able to obtain realizations of it.

Before introducing our model of computation, we make the following assumption on the component functions:

Assumption 3. *For any $\zeta \in E$, the function $f(\cdot, \zeta) : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex, that is, for all $x, y \in \mathbb{R}^d$:*

$$f(y, \zeta) \geq f(x, \zeta) + \langle \nabla f(x, \zeta), y - x \rangle$$

Note that our assumption on the differentiability of F from previous chapters combined with the convexity of the $f(\cdot, \zeta)$ for each $\zeta \in E$ already implies that the functions $f(\cdot, \zeta)$ are differentiable for almost all $\zeta \in E$, see (Bertsekas (1973), Proposition 2.3). We extend this to all $\zeta \in E$ in our assumption to avoid making almost sure statements. Perhaps surprisingly, our assumptions on F and the $f(\cdot, \zeta)$ are enough to allow the interchange of differentiation and integration. See (Bertsekas (1973), Proposition 2.2).

Lemma 7. *For all $x \in \mathbb{R}^d$:*

$$\nabla F(x) = \mathbb{E}\left[\nabla f(x, \zeta)\right]$$

In light of this, we will use the following model of computation. We assume that we have access to an oracle which takes as input a point $x \in \mathbb{R}^d$, samples ζ from its distribution (independently from other oracle calls), and returns $\nabla f(x, \zeta)$. This provides us with an unbiased estimate of $\nabla F(x)$ by Lemma 7. In the classical study of problems of this form, one usually assumes that the variance of this estimate is uniformly bounded over all of $\mathbb{R}^d$ by some constant. This assumption is however too strong, excluding many standard problems. Here instead we make an additional assumption on the functions $f(\cdot, \zeta)$.

Assumption 4. *For any $\zeta \in E$, the function $f(\cdot, \zeta)$ is smooth, that is, there exists an $L_\zeta > 0$ such that for all $x, y \in \mathbb{R}^d$:*

$$\left\|\nabla f(y, \zeta) - \nabla f(x, \zeta)\right\|_2 \leq L_\zeta \left\|y - x\right\|_2$$

Furthermore:

$$\sup_{\zeta \in E} L_\zeta = L_{sup} < \infty$$

This additional smoothness assumption is satisfied in many cases of interest, and allows much weaker assumptions on the variance of the gradient estimate generated by our oracle. We will state such assumptions when we need them. Finally, note that $\mathbb{E}\left[L_\zeta\right]$ and L_{sup} are valid smoothness constants of F.

Lemma 8. *For all $x, y \in \mathbb{R}^d$:*

$$\left\|\nabla F(y) - \nabla F(x)\right\|_2 \leq \mathbb{E}\left[L_\zeta\right] \left\|y - x\right\|_2 \leq L_{sup} \left\|y - x\right\|_2$$

Proof.

$$\begin{aligned}\|\nabla F(y) - \nabla F(x)\|_2 &= \|\mathbb{E}\left[\nabla f(y,\zeta) - \nabla f(x,\zeta)\right]\|_2 \\ &\leq \mathbb{E}\left[\|\nabla f(y,\zeta) - \nabla f(x,\zeta)\|_2\right] \\ &\leq \mathbb{E}\left[L_\zeta\right]\|y - x\|_2 \\ &\leq L_{sup}\|y - x\|_2\end{aligned}$$

where the second line follows from Jensen's inequality and the convexity of the Euclidean norm, and the third from the smoothness of $f(\cdot, \zeta)$. □

In light of this result, we define an alternative condition number $\kappa_{sup} := L_{sup}/\mu$, which we will use to characterize the complexity of SGD and SGLD.

4.1 Algorithms

Given that our oracle provides us with unbiased estimates of the gradient, it seems reasonable to simply replace the gradient by its unbiased estimate in gradient descent (Algorithm 1) and Langevin dynamics (Algorithm 2) for the purposes of solving the optimization or sampling problems associated with F. Doing this yields Algorithm 3 known as Stochastic Gradient Descent (SGD) for optimization, and Algorithm 4 known as Stochastic Gradient Langevin Dynamics (SGLD) for sampling. Note that by definition of our oracle, the random variables $(\zeta_k)_{k=0}^{\infty}$ are independent and identically distributed.

4.1.1 Stochastic Gradient Descent

Algorithm 3: Stochastic Gradient Descent (SGD)

Parameters: step sizes $(\alpha_k)_{k=1}^{\infty} > 0$

Initialization: $x_0 \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

  $x_{k+1} = x_k - \alpha_k \nabla f(x_k, \zeta_k)$

end

4.1.2 Stochastic Gradient Langevin Dynamics

Algorithm 4: Stochastic Gradient Langevin Dynamics (SGLD)

Parameters: step sizes $(\alpha_k)_{k=1}^{\infty} > 0$

Initialization: $\rho_0 \in \mathbb{R}^d$

sample $x_0 \sim \rho_0$

for $k = 0, 1, 2, \ldots$ **do**

  sample $\xi_k \sim \mathcal{N}(0, I_{d\times d})$

  $x_{k+1} = x_k - \alpha_k \nabla f(x_k, \zeta_k) + \sqrt{2\alpha_k}\,\xi_k$

end

4.2 Convergence Analysis

In this section we analyze the convergence of both Stochastic Gradient Descent and Stochastic Gradient Langevin Dynamics. To evaluate the complexity of each algorithm, we will count the number of oracle calls needed to reach an ε accurate solution. We will use the same criteria we used in chapter 3, namely $\|x - x^*\|_2^2 \leq \varepsilon$ and $W_2^2(\rho, \rho^*) \leq \varepsilon$.

In chapter 3, the smoothness assumption on F was used to control the discretization error. It will play the same role in this chapter. The smoothness of the functions $f(\cdot, \zeta)$ will play a different role here, namely that of bounding the variance of the gradient estimate. In particular, we will need the following result which is a consequence of the convexity and smoothness of the $f(\cdot, \zeta)$.

Lemma 9. *For all $x, y \in \mathbb{R}^d$, we have:*

$$\mathbb{E}\left[\|\nabla f(y,\zeta) - \nabla f(x,\zeta)\|_2^2\right] \leq 2L_{sup}\left[F(y) - F(x) - \langle \nabla F(x), y - x\rangle\right]$$

Proof. Let $\zeta \in E$. By convexity and smoothness of $f(\cdot, \zeta)$ and (Nesterov (2004), Theorem 2.1.5) we have:

$$\|\nabla f(y,\zeta) - \nabla f(x,\zeta)\|_2^2 \leq 2L_{\zeta}\left[f(y,\zeta) - f(x,\zeta) - \langle \nabla f(x,\zeta), y - x\rangle\right]$$

Using $L_{\zeta} \leq L_{sup}$ and taking expectation of both sides we get the result. □

As mentioned at the beginning of the chapter, the convexity and smoothness of the $f(\cdot, \zeta)$ allows much weaker assumptions on the variance of the gradient estimate generated by the oracle than uniform boundedness. To make the statement of these assumptions easier, we define the averaged variance function $\sigma^2 : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ given by, for $\rho \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\sigma^2(\rho) = \mathop{\mathbb{E}}_{x\sim\rho}\left[\mathbb{E}\left[\|\nabla f(x,\zeta)\|_2^2\right] - \|\nabla F(x)\|_2^2\right]$$

where the inner expectation is with respect to ζ. Recall that $\mathcal{P}_2(\mathbb{R}^d)$ is the space of probability measures in $\mathbb{R}^d$ with finite second moment. From this we can show the following sufficient condition for the finiteness of σ^2 over its domain.

Lemma 10. *If there exists a $z \in \mathbb{R}^d$ such that:*

$$\sigma^2(\delta_z) < \infty$$

then $\sigma^2(\rho) < \infty$ for all $\rho \in \mathcal{P}_2(\mathbb{R}^d)$

Proof.

$$\begin{aligned}
\sigma^2(\rho) &\le \mathop{\mathbb{E}}_{x\sim\rho}\left[\mathbb{E}\left[\|\nabla f(x,\zeta)\|_2^2\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim\rho}\left[\mathbb{E}\left[\|\nabla f(x,\zeta) - \nabla f(z,\zeta) + \nabla f(z,\zeta)\|_2^2\right]\right] \\
&\le 2\mathop{\mathbb{E}}_{x\sim\rho}\left[\mathbb{E}\left[\|\nabla f(x,\zeta) - \nabla f(z,\zeta)\|_2^2\right]\right] + 2\mathbb{E}\left[\|\nabla f(z,\zeta)\|_2^2\right] \\
&\le 2L_{sup}^2 \mathop{\mathbb{E}}_{x\sim\rho}\left[\|x - z\|_2^2\right] + 2\sigma^2(\delta_z) + 2\|\nabla F(z)\|_2^2 \\
&\le 4L_{sup}^2 \mathop{\mathbb{E}}_{x\sim\rho}\left[\|x\|_2^2\right] + 4L_{sup}^2\|z\|_2^2 + 2\sigma^2(\delta_z) + 2\|\nabla F(z)\|_2^2 \\
&< \infty
\end{aligned}$$

where in the third and fifth line we used the Peter-Paul inequality from Lemma 3 with $\beta = 1$, in the fourth line we used the smoothness of $f(\cdot, \zeta)$, and in the last line we used the hypothesis and the finiteness of the second moment of ρ. □

4.2.1 Convergence of Stochastic Gradient Descent

To prove our convergence result, we will need one last assumption. Unlike the other assumptions in this text, this one is exclusive to this subsection.

$$\sigma^2(\delta_{x^*}) < \infty$$

We are now finally ready to state our convergence theorem.

Theorem 8. *Let $(x_k)_{k=0}^{\infty}$ be the stochastic gradient descent sequence generated by Algorithm 3 with $\alpha_k = \alpha$ satisfying:*

$$\alpha \leq \frac{1}{2L_{sup}}$$

Then:

$$\mathbb{E}\left[\|x_k - x^*\|_2^2\right] \leq (1-\alpha\mu)^k \|x_0 - x^*\|_2^2 + \frac{2\alpha\sigma^2(\delta_{x^*})}{\mu}$$

where the expectation is taken with respect to the randomness of the oracle due to the sampling of $(\zeta_t)_{t=0}^{k-1}$.

Proof. Let $k \in \mathbb{N}$. Taking expectation over ζ_k, conditional on $(\zeta_t)_{t=0}^{k-1}$ we have:

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^*\|_2^2\right] &= \mathbb{E}\left[\|x_k - \alpha\nabla f(x_k, \zeta_k) - x^*\|_2^2\right] \\
&= \|x_k - x^*\|_2^2 - 2\alpha\langle \mathbb{E}\left[\nabla f(x_k, \zeta_k)\right], x_k - x^*\rangle + \alpha^2\mathbb{E}\left[\|\nabla f(x_k, \zeta_k)\|_2^2\right] \\
&= \|x_k - x^*\|_2^2 - 2\alpha\langle \nabla F(x_k), x_k - x^*\rangle + \alpha^2\mathbb{E}\left[\|\nabla f(x_k, \zeta_k)\|_2^2\right]
\end{aligned}$$

where the last line follows from Lemma 7. We bound the last term as follows:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(x_k, \zeta_k)\|_2^2\right] &= \mathbb{E}\left[\|\nabla f(x_k, \zeta_k) - \nabla f(x^*, \zeta_k) + \nabla f(x^*, \zeta_k)\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f(x_k, \zeta_k) - \nabla f(x^*, \zeta_k)\|_2^2\right] + 2\mathbb{E}\left[\|\nabla f(x^*, \zeta_k)\|_2^2\right] \\
&\leq 4L_{sup}\left[F(x_k) - F(x^*)\right] + 2\sigma^2(\delta_{x^*})
\end{aligned}$$

where the second line follows from Lemma 3, and the third from Lemma 9 and $\nabla F(x^*) = 0$.Replacing in the original bound, and using the strong convexity of F on the inner product term, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^*\|_2^2\right] &\leq (1-\alpha\mu)\|x_k - x^*\|_2^2 + 2\alpha\left(2\alpha L_{sup} - 1\right)\left[F(x_k) - F(x^*)\right] + 2\alpha^2\sigma^2(\delta_{x^*}) \\
&\leq (1-\alpha\mu)\|x_k - x^*\|_2^2 + 2\alpha^2\sigma^2(\delta_{x^*})
\end{aligned}$$

where the last line follows from the condition on the step size.Taking expectation over $(\zeta_t)_{t=0}^{k-1}$ on both sides, and applying the resulting inequality recursively we get:

$$\begin{aligned}\mathbb{E}\left[\|x_k - x^*\|_2^2\right] &\leq (1-\alpha\mu)^k \|x_0 - x^*\|_2^2 + 2\alpha^2\sigma^2(\delta_{x^*}) \sum_{i=0}^{k-1} (1-\alpha\mu)^i \\ &\leq (1-\alpha\mu)^k \|x_0 - x^*\|_2^2 + 2\alpha^2\sigma^2(\delta_{x^*}) \sum_{i=0}^{\infty} (1-\alpha\mu)^i \\ &= (1-\alpha\mu)^k \|x_0 - x^*\|_2^2 + \frac{2\alpha\sigma^2(\delta_{x^*})}{\mu}\end{aligned}$$

□

From this theorem, we can derive the complexity of Stochastic Gradient Descent. We leave the proof to the reader as it is a simple adaptation of the argument in the proof of Corollary 4.

Corollary 5. *Let $\varepsilon > 0$, and let $(x_k)_{k=0}^{\infty}$ be the sequence generated by Algorithm 3 with step size:*

$$\alpha_k = \alpha = \min\left\{\frac{1}{2L_{sup}}, \frac{\varepsilon\mu}{4\sigma^2(\delta_{x^*})}\right\}$$

If:

$$k \geq \max\left\{2\kappa_{sup}, \frac{4\sigma^2(\delta_{x^*})}{\varepsilon\mu^2}\right\} \log\left(\frac{2\|x_0 - x^*\|_2^2}{\varepsilon}\right)$$

Then:

$$\mathbb{E}\left[\|x_k - x^*\|_2^2\right] \leq \varepsilon$$

The complexity of SGD has therefore two regimes. In the low precision regime $\varepsilon < O(\sigma^2(\delta_{x^*})/L\mu)$, SGD convergence at the fast linear rate $O(\kappa_{sup}\log(1/\varepsilon))$ similar to gradient descent. In the high precision regime however, this complexity deteriorates to $\tilde{O}(\sigma^2(\delta_{x^*})/\varepsilon\mu^2)$. Note that by using decreasing step sizes, one can remove the logarithmic factor in the high precision regime, as well as use ε independent step sizes, but we do not pursue this further here.

4.2.2 Convergence of Stochastic Gradient Langevin Dynamics

We start by making the following assumption, which again is exclusive to this subsection.

$$\sigma^2(\rho^*) < \infty$$

The convergence theorem for SGLD follows.

Theorem 9. *Let $(x_k)_{k=0}^{\infty}$ be the Markov chain simulated by Algorithm 4 with a constant step size $\alpha_k = \alpha$ satisfying:*

$$\alpha \leq \frac{1}{2L_{sup}}$$

Then:

$$W_2^2(\rho_k, \rho^*) \leq \left(1 - \frac{\alpha\mu}{2}\right)^k W_2^2(\rho_0, \rho^*) + 24\alpha\kappa^2 d + \frac{4\alpha\sigma^2(\rho^*)}{\mu}$$

Proof. As usual, we proceed using a coupling argument. Consider a Langevin diffusion process $(y_t)_{t\in\mathbb{R}^+}$ satisfying:

$$dy_t = -\nabla F(y_t)\, dt + \sqrt{2}\, dW_t$$

and starting at $y_0 \sim \rho^*$. From the proof of Theorem 3, we know that this implies $y_t \sim \rho^*$ for all $t \in \mathbb{R}^+$. We also assume that the same Wiener process drives both $(y_t)_{t\in\mathbb{R}^+}$ and $(x_k)_{k=0}^{\infty}$. Furthermore we assume that x_0 and y_0 are optimally coupled so that:

$$W_2^2(\rho_0, \rho^*) = \mathbb{E}\left[\|x_0 - y_0\|_2^2\right]$$

Let $k \in \mathbb{N}$. We bound $\left\|x_{k+1} - y_{(k+1)\alpha}\right\|_2^2$ in the same way as in Theorem 7, replacing $\nabla F(x_k)$ by $\nabla f(x_k, \zeta_k)$ to get, for a free parameter $\beta > 0$:

$$\begin{aligned}
&\left\|x_{k+1} - y_{(k+1)\alpha}\right\|_2^2 \\
&\leq (1+\beta)\left\|x_k - y_{k\alpha} - \alpha\left[\nabla f(x_k, \zeta_k) - \nabla F(y_{k\alpha})\right]\right\|_2^2 + (1+\beta^{-1})\left\|\int_{k\alpha}^{(k+1)\alpha} \left[\nabla F(y_s) - \nabla F(y_{k\alpha})\right] ds\right\|_2^2
\end{aligned}$$

The expectation of the second term is bounded by $3\alpha^3 L^2 d$ as we showed in the proof of Theorem 7. For the first term, we take expectation over ζ_k, conditional on $(\zeta_t)_{t=0}^{k-1}$ to get:

$$\begin{aligned}
&\mathbb{E}\left[\left\|x_k - y_{k\alpha} - \alpha\left[\nabla f(x_k, \zeta_k) - \nabla F(y_{k\alpha})\right]\right\|_2^2\right] \\
&= \|x_k - y_{k\alpha}\|_2^2 - 2\alpha\langle \mathbb{E}\left[\nabla f(x_k, \zeta_k)\right] - \nabla F(y_{k\alpha}), x_k - y_{k\alpha}\rangle + \alpha^2 \mathbb{E}\left[\|\nabla f(x_k, \zeta_k) - \nabla F(y_{k\alpha})\|_2^2\right] \\
&= \|x_k - y_{k\alpha}\|_2^2 - 2\alpha\langle \nabla F(x_k) - \nabla F(y_{k\alpha}), x_k - y_{k\alpha}\rangle + \alpha^2 \mathbb{E}\left[\|\nabla f(x_k, \zeta_k) - \nabla F(y_{k\alpha})\|_2^2\right]
\end{aligned}$$

where the last line follows from Lemma 7. We now bound the last term as:

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(x_k,\zeta_k)-\nabla F(y_{k\alpha})\|_2^2\right]\\
&=\mathbb{E}\left[\|\nabla f(x_k,\zeta_k)-\nabla f(y_{k\alpha},\zeta_k)+\nabla f(y_{k\alpha},\zeta_k)-\nabla F(y_{k\alpha})\|_2^2\right]\\
&\leq 2\mathbb{E}\left[\|\nabla f(x_k,\zeta_k)-\nabla f(y_{k\alpha},\zeta_k)\|_2^2\right]+2\mathbb{E}\left[\|\nabla f(y_{k\alpha},\zeta_k)-\nabla F(y_{k\alpha})\|_2^2\right]\\
&\leq 4L_{sup}\left[F(x_k)-F(y_{k\alpha})-\langle\nabla F(y_{k\alpha}),x_k-y_{k\alpha}\rangle\right]+2\mathbb{E}\left[\|\nabla f(y_{k\alpha},\zeta_k)-\nabla F(y_{k\alpha})\|_2^2\right]
\end{aligned}$$

where we uses Lemma 3 for the first inequality, and Lemma 9 for the second. Replacing and using the strong convexity of F to bound the inner product term we obtain:

$$\begin{aligned}
&\mathbb{E}\left[\|x_k-y_{k\alpha}-\alpha\left[\nabla f(x_k,\zeta_k)-\nabla F(y_{k\alpha})\right]\|_2^2\right]\\
&\leq(1-\alpha\mu)\|x_k-y_{k\alpha}\|_2^2+\alpha\left(2\alpha L_{sup}-1\right)\left[F(x_k)-F(y_{k\alpha})-\langle\nabla F(y_{k\alpha}),x_k-y_{k\alpha}\rangle\right]+\\
&\quad 2\alpha^2\mathbb{E}\left[\|\nabla f(y_{k\alpha},\zeta_k)-\nabla F(y_{k\alpha})\|_2^2\right]\\
&\leq(1-\alpha\mu)\|x_k-y_{k\alpha}\|_2^2+2\alpha^2\mathbb{E}\left[\|\nabla f(y_{k\alpha},\zeta_k)-\nabla F(y_{k\alpha})\|_2^2\right]
\end{aligned}$$

where the last inequality follows from the positivity of the term in brackets due to the convexity of F and the condition on the step size. Replacing in the original bound and taking expectation with respect to all the randomness we obtain:

$$\mathbb{E}\left[\left\|x_{k+1}-y_{(k+1)\alpha}\right\|_2^2\right]\leq(1+\beta)(1-\alpha\mu)\mathbb{E}\left[\|x_k-y_{k\alpha}\|_2^2\right]+3(1+\beta^{-1})\alpha^3L^2d+2\alpha^2\sigma^2(\rho^*)$$

Taking $\beta=\alpha\mu/2<1$ the first coefficient is bounded by:

$$\left(1+\frac{\alpha\mu}{2}\right)(1-\alpha\mu)\leq\left(1-\frac{\alpha\mu}{2}\right)$$

while the second is bounded by:

$$(1+\beta^{-1})\leq 2\beta^{-1}=\frac{4}{\alpha\mu}$$

so that our bound is:

$$\mathbb{E}\left[\left\|x_{k+1}-y_{(k+1)\alpha}\right\|_2^2\right]\leq\left(1-\frac{\alpha\mu}{2}\right)\mathbb{E}\left[\|x_k-y_{k\alpha}\|_2^2\right]+12\alpha^2\kappa Ld+2\alpha^2\sigma^2(\rho^*)$$

Applying this inequality recursively, and bounding the resulting geometric sums we get:

$$\mathbb{E}\left[\|x_k-y_{k\alpha}\|_2^2\right]\leq\left(1-\frac{\alpha\mu}{2}\right)^k\mathbb{E}\left[\|x_0-y_0\|_2^2\right]+24\alpha\kappa^2d+\frac{4\alpha\sigma^2(\rho^*)}{\mu}$$

By minimality of the coupling defining the Wasserstein distance we get:

$$W_2^2(\rho_k, \rho^*) \leq \mathbb{E}\left[\|x_k - y_{k\alpha}\|_2^2\right]$$

and by the fact that x_0 and y_0 are optimally coupled we obtain the stated result. □

Corollary 6. *Let $\varepsilon > 0$, and let $(x_k)_{k=0}^{\infty}$ be the Markov chain simulated by Algorithm 4 with step size:*

$$\alpha_k = \alpha = \min\left\{\frac{1}{2L_{sup}}, \frac{\varepsilon\mu}{48\kappa^2 d\mu + 8\sigma^2(\rho^*)}\right\}$$

If:

$$k \geq \max\left\{2\kappa_{sup}, \frac{48\kappa^2 d\mu + 8\sigma^2(\rho^*)}{\varepsilon\mu^2}\right\}\log\left(\frac{2\|x_0 - x^*\|_2^2}{\varepsilon}\right)$$

Then:

$$W_2^2(\rho_k, \rho^*) \leq \varepsilon$$

where ρ_k is the distribution of x_k.

Chapter 5

Finite Sum Algorithms

In this chapter, we come back to the problem that motivated us from the start: the case where F has a finite sum structure. In particular, using the notation of the previous chapter, if ζ is uniformly distributed over $[n]$, then F can be written as, after defining $f_\zeta(x) := f(x, \zeta)$:

$$F(x) := \mathbb{E}\left[f_\zeta(x)\right] = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{5.1}$$

We will write L_i for the smoothness constant of f_i, and define $L_{max} := \max_{i\in[n]} L_i = L_{sup}$. The maximum condition number is $\kappa_{max} := L_{max}/\mu$.

As an aside, note that the particular form of F in (5.1) can be obtained for more general random variables ζ, all that is required is for ζ to have finite support $(\zeta_i)_{i=1}^n$, in which case we have:

$$F(x) = \mathbb{E}\left[f(x, \zeta)\right] = \sum_{i=1}^{n} p_i f(x, \zeta_i)$$

where $p_i = \mathbb{P}(\zeta = \zeta_i)$. We then recover the form (5.1) after defining $f_i(x) = np_i f(x, \zeta_i)$.

The underlying assumption we will be making in this chapter, which further separates our setup here from the one of chapter 4, is that we have the ability to pick any function from $(f_i)_{i=1}^n$ by its index. In particular, as opposed to the oracle of the previous chapter which took as input a point $x \in \mathbb{R}^d$, internally generated ζ, and returned $\nabla f(x, \zeta)$, here we will assume that our oracle takes as input a point $x \in \mathbb{R}^d$ and an index $i \in [n]$ and returns $\nabla f_i(x)$. This gives us extra freedom in how we generate the index i, and allows us to identify

a particular gradient estimate $\nabla f_i(x)$ with its index. The main point of this chapter will be to show that we can leverage the finite sum structure of F and this more powerful oracle to build better algorithms.

The main idea of SGD and SGLD from chapter 4 is that of replacing the gradient $\nabla F(x)$ by the unbiased estimate $\nabla f(x, \zeta)$ provided by the oracle. We can replicate this idea in our case using our more powerful oracle as follows. We first sample an index i uniformly from $[n]$, and then call our oracle to obtain the unbiased estimate $\nabla f_i(x)$ of $\nabla F(x)$. We then get the convergence guarantees of the previous chapter.

We can do better. In parallel to the sequence $(x_k)_{k=0}^{\infty}$, we maintain the sequence $((g_k^i)_{i=1}^n)_{k=0}^{\infty}$ defined recursively as follows:

$$g_{k+1}^i = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k \\ g_k^i & \text{otherwise} \end{cases}$$

for an arbitrary initialization $(g_0^i)_{i=1}^n$ and where i_k is the index sampled at iteration k. The goal of this sequence is to track the component gradients $\nabla f_i(x_k)$ as much as possible, while introducing no extra computational cost. Now consider the naive decomposition:

$$\nabla F(x_k) = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_k) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_k)$$

approximating $\nabla f_i(x_k) \approx g_k^i$, we arrive at the unbiased estimate:

$$\nabla F(x_k) \approx \nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i$$

Using this estimate we arrive at the following Algorithms.

5.1 Algorithms

5.1.1 Controlled Stochastic Gradient Descent

Algorithm 5: Controlled Stochastic Gradient Descent (CSGD)

Parameters: step sizes $(\alpha_k)_{k=1}^{\infty} > 0$

Initialization: $x_0 \in \mathbb{R}^d, (g_0^i)_{i=1}^n \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

sample i_k uniformly from $[n]$

$$x_{k+1} = x_k - \alpha_k \left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \tfrac{1}{n} \textstyle\sum_{i=1}^n g_k^i \right)$$

$$g_{k+1}^i = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k \\ g_k^i & \text{otherwise} \end{cases}$$

end

5.1.2 Controlled Stochastic Gradient Langevin Dynamics

Algorithm 6: Controlled Stochastic Langevin Dynamics (CSGLD)

Parameters: step sizes $(\alpha_k)_{k=1}^{\infty} > 0$

Initialization: $x_0 \in \mathbb{R}^d, (g_0^i)_{i=1}^n \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

sample $\xi_k \sim \mathcal{N}(0, I_{d\times d})$

sample i_k uniformly from $[n]$

$$x_{k+1} = x_k - \alpha_k \left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \tfrac{1}{n} \textstyle\sum_{i=1}^n g_k^i \right) + \sqrt{2\alpha_k}\, \xi_k$$

$$g_{k+1}^i = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k \\ g_k^i & \text{otherwise} \end{cases}$$

end

5.2 Convergence Analysis

5.2.1 Convergence of Controlled Stochastic Gradient Descent

Theorem 10. *Let $(x_k, (g_k^i)_{i=1}^n)_{k=0}^\infty$ be the sequence generated by Algorithm 5 with $\alpha_k = \alpha$ satisfying:*

$$\alpha \leq \frac{1}{5L_{max}}$$

Then:

$$\mathbb{E}\left[\|x_k - x^*\|_2^2\right] \leq (1-\lambda)^k T^0$$

where:

$$\lambda := \min\left\{\frac{1}{5n}, \alpha\mu\right\}$$

and:

$$T^k := \frac{\alpha}{2L_{max}} \sum_{i=1}^n \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + \|x_k - x^*\|_2^2$$

Proof. We bound $\mathbb{E}\left[T^{k+1}\right]$. Unless otherwise mentioned, all expectations in this proof are with respect to i_k conditional on $(i_t)_{t=0}^{k-1}$. The first term of $\mathbb{E}\left[T^{k+1}\right]$ is bounded by:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^n \left\|g_{k+1}^i - \nabla f_i(x^*)\right\|_2^2\right] \\
&= \sum_{j=1}^n \mathbb{P}(i_k = j) \left(\sum_{\substack{i=1 \\ i\neq j}}^n \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + \|\nabla f_j(x_k) - \nabla f_j(x^*)\|_2^2\right) \\
&= \sum_{j=1}^n \frac{1}{n} \left(\sum_{i=1}^n \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 - \left\|g_k^j - \nabla f_j(x^*)\right\|_2^2 + \|\nabla f_j(x_k) - \nabla f_j(x^*)\|_2^2\right) \\
&= \left(1 - \frac{1}{n}\right) \sum_{i=1}^n \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + \mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x^*)\|_2^2\right] \\
&\leq \left(1 - \frac{1}{n}\right) \sum_{i=1}^n \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + 2L_{max}\left[F(x_k) - F(x^*)\right]
\end{aligned}$$

where the first equality follow from the update of $(g_k^i)_{i=1}^n$ in Algorithm 5, and the last line

follows from Lemma 9. The second term of $\mathbb{E}\left[T^{k+1}\right]$ is bounded by:

$$
\begin{aligned}
&\mathbb{E}\left[\|x_{k+1}-x^*\|_2^2\right] \\
&= \mathbb{E}\left[\left\|x_k - \alpha\left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right) - x^*\right\|_2^2\right] \\
&= \|x_k - x^*\|_2^2 - 2\alpha\left\langle \mathbb{E}\left[\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right], x_k - x^*\right\rangle + \mathbb{E}\left[\left\|\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right\|_2^2\right] \\
&= \|x_k - x^*\|_2^2 - 2\alpha\langle \nabla F(x_k), x_k - x^*\rangle + \mathbb{E}\left[\left\|\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right\|_2^2\right]
\end{aligned}
$$

We bound the last term as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x^*) + \nabla f_{i_k}(x^*) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x^*)\|_2^2\right] + 2\mathbb{E}\left[\left\|\nabla f_{i_k}(x^*) - g_k^{i_k} - \left(\nabla F(x^*) - \frac{1}{n}\sum_{i=1}^n g_k^i\right)\right\|_2^2\right] \\
&\leq 4L_{max}\left[F(x_k) - F(x^*)\right] + 2\mathbb{E}\left[\left\|\nabla f_{i_k}(x^*) - g_k^{i_k}\right\|_2^2\right]
\end{aligned}
$$

where the first inequality follows from Lemma 3 with $\beta = 1$ and $\nabla F(x^*) = 0$, and the second from Lemma 9 and the fact that for a random vector X:

$$
\mathbb{E}\left[\|X - \mathbb{E}[X]\|_2^2\right] = \mathbb{E}\left[\|X\|_2^2\right] - \|\mathbb{E}[X]\|_2^2 \leq \mathbb{E}\left[\|X\|_2^2\right]
$$

Putting together all the inequalities we get:

$$
\begin{aligned}
\mathbb{E}\left[T^{k+1}\right] \leq& \left(1 - \frac{1}{n} + \frac{4\alpha L_{max}}{n}\right)\frac{\alpha}{2L_{max}}\sum_{i=1}^n \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + (1-\alpha\mu)\|x_k - x^*\|_2^2 + \\
&\alpha\left(4\alpha L_{max} - 1\right)\left[F(x_k) - F(x^*)\right] \\
&\leq \left(1 - \frac{1}{5n}\right)\frac{\alpha}{2L_{max}}\sum_{i=1}^n \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + (1-\alpha\mu)\|x_k - x^*\|_2^2 \\
&\leq (1-\lambda)\,T^k
\end{aligned}
$$

where the second line follows from the condition on the step size, and the last from the definition of λ. Taking expectation over all of the randomness of both sides, applying this inequality recursively, and noticing that:

$$\mathbb{E}\left[\|x_k - x^*\|_2^2\right] \leq \mathbb{E}\left[T^k\right]$$

we get the result. □

5.2.2 Convergence of Controlled Stochastic Gradient Langevin Dynamics

Theorem 11. *Let $(x_k, (g_k^i)_{i=1}^n)$ be the Markov chain simulated by Algorithm 6 with a constant step size $\alpha_k = \alpha$ satisfying:*

$$\alpha \leq \frac{1}{14 L_{max}}$$

Then:

$$W_2^2(\rho_k, \rho^*) \leq (1-\lambda)^k \left(W_2^2(\rho_0, \rho^*) + \frac{\alpha}{2L_{max}} \mathbb{E}\left[\sum_{i=1}^n \left\| g_k^i - \nabla f_i(y) \right\|_2^2 \right] \right) + 96\alpha\kappa_{max} n^2 d \left(2\kappa_{max} + 7\alpha n L_{max}\right)$$

where $y \sim \rho^*$ and:

$$\lambda := \min\left\{ \frac{1}{7n}, \frac{\alpha\mu}{2} \right\}$$

Proof. As usual, we proceed using a coupling argument. Consider a Langevin diffusion process $(y_t)_{t\in\mathbb{R}^+}$ satisfying:

$$dy_t = -\nabla F(y_t)\, dt + \sqrt{2}\, dW_t$$

and starting at $y_0 \sim \rho^*$. From the proof of Theorem 3, we know that this implies $y_t \sim \rho^*$ for all $t \in \mathbb{R}^+$. We also assume that the same Wiener process drives both $(y_t)_{t\in\mathbb{R}^+}$ and $(x_k)_{k=0}^\infty$. Furthermore we assume that x_0 and y_0 are optimally coupled so that:

$$W_2^2(\rho_0, \rho^*) = \mathbb{E}\left[\|x_0 - y_0\|_2^2\right]$$

Finally we define the sequence $((h_k^i)_{i=1}^n)_{k=0}^\infty$ with respect to $(y_{k\alpha})_{k=0}^\infty$ similarly to how $((g_k^i)_{i=1}^n)_{k=0}^\infty$ is defined with respect to $(x_k)_{k=0}^\infty$. In particular, we initialize:

$$h_0^i = \nabla f_i(y_0)$$

and perform the update:

$$h_{k+1}^i = \begin{cases} \nabla f_i(y_{k\alpha}) & \text{if } i = i_k \\ h_k^i & \text{otherwise} \end{cases}$$

Let $k \in \mathbb{N}$. We study the evolution of the Lyapunov function:

$$T^k := \frac{\alpha}{2L_{max}} \sum_{i=1}^n \left\|g_k^i - h_k^i\right\|_2^2 + \|x_k - y_{k\alpha}\|_2^2$$

Let us bound $\mathbb{E}\left[T^{k+1}\right]$. We start with the first term. Taking expectation with respect to i_k conditional on $(i_t)_{t=1}^{k-1}$ have:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^n \left\|g_{k+1}^i - h_{k+1}^i\right\|_2^2\right] \\
&= \sum_{j=1}^n \mathbb{P}(i_k = j) \left(\sum_{\substack{i=1 \\ i\neq j}}^n \left\|g_k^i - h_k^i\right\|_2^2 + \|\nabla f_j(x_k) - \nabla f_j(y_{k\alpha})\|_2^2\right) \\
&= \sum_{j=1}^n \frac{1}{n} \left(\sum_{i=1}^n \left\|g_k^i - h_k^i\right\|_2^2 - \left\|g_k^j - h_k^j\right\|_2^2 + \|\nabla f_j(x_k) - \nabla f_j(y_{k\alpha})\|_2^2\right) \\
&= \left(1 - \frac{1}{n}\right) \sum_{i=1}^n \left\|g_k^i - h_k^i\right\|_2^2 + \mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(y_{k\alpha})\|_2^2\right] \\
&\leq \left(1 - \frac{1}{n}\right) \sum_{i=1}^n \left\|g_k^i - h_k^i\right\|_2^2 + 2L_{max}\left[F(x_k) - F(y_{k\alpha}) - \langle \nabla F(y_{k\alpha}), x_k - y_{k\alpha}\rangle\right]
\end{aligned}$$

For the second term $\left\|x_{k+1} - y_{(k+1)\alpha}\right\|_2^2$ we bound it in the same way as in Theorem 7, replacing $\nabla F(x_k)$ by $\left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right)$ to get, for a free parameter $\beta > 0$:

$$\begin{aligned}
&\left\|x_{k+1} - y_{(k+1)\alpha}\right\|_2^2 \\
&\leq (1+\beta)\left\|x_k - y_{k\alpha} - \alpha\left[\left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right) - \nabla F(y_{k\alpha})\right]\right\|_2^2 + \\
&\quad (1+\beta^{-1})\left\|\int_{k\alpha}^{(k+1)\alpha} \left[\nabla F(y_s) - \nabla F(y_{k\alpha})\right] ds\right\|_2^2
\end{aligned}$$

The expectation of the second term is bounded by $3\alpha^3 L^2 d$ as we showed in the proof of

Theorem 7. For the first term, we take expectation over i_k, conditional on $(i_t)_{t=0}^{k-1}$ to get:

$$\begin{aligned}
&\mathbb{E}\left[\left\|x_k - y_{k\alpha} - \alpha\left[\left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right) - \nabla F(y_{k\alpha})\right]\right\|_2^2\right] \\
&= \|x_k - y_{k\alpha}\|_2^2 - 2\alpha\left\langle \mathbb{E}\left[\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right] - \nabla F(y_{k\alpha}), x_k - y_{k\alpha}\right\rangle \\
&\quad + \alpha^2\mathbb{E}\left[\left\|\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i - \nabla F(y_{k\alpha})\right\|_2^2\right] \\
&= \|x_k - y_{k\alpha}\|_2^2 - 2\alpha\langle \nabla F(x_k) - \nabla F(y_{k\alpha}), x_k - y_{k\alpha}\rangle \\
&\quad + \alpha^2\mathbb{E}\left[\left\|\left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right) - \nabla F(y_{k\alpha})\right\|_2^2\right]
\end{aligned}$$

We now bound the last term as follows. First let us rewrite the term inside the squared norm as:

$$\begin{aligned}
&\left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right) - \nabla F(y_{k\alpha}) \\
&= \left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right) - \left(\nabla f_{i_k}(y_{k\alpha}) - h_k^{i_k} + \frac{1}{n}\sum_{i=1}^n h_k^i\right) + \\
&\quad \left(\nabla f_{i_k}(y_{k\alpha}) - h_k^{i_k} + \frac{1}{n}\sum_{i=1}^n h_k^i\right) - \nabla F(y_{k\alpha}) \\
&= [\nabla f_{i_k}(x_k) - \nabla f_{i_k}(y_{k\alpha})] + \left[h_k^{i_k} - g_k^{i_k} - \left(\frac{1}{n}\sum_{i=1}^n h_k^i - \frac{1}{n}\sum_{i=1}^n g_k^i\right)\right] + \\
&\quad \left[\nabla f_{i_k}(y_{k\alpha}) - h_k^{i_k} - \left(\nabla F(y_{k\alpha}) - \frac{1}{n}\sum_{i=1}^n h_k^i\right)\right]
\end{aligned}$$

Using Lemma 3, and the fact that $\mathbb{E}\left[\|X - \mathbb{E}[X]\|_2^2\right] = \mathbb{E}\left[\|X\|_2^2\right] - \|\mathbb{E}[X]\|_2^2 \leq \mathbb{E}\left[\|X\|_2^2\right]$, we therefore have the bound:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\nabla f_{i_k}(x_k) - g_k^{i_k} + \frac{1}{n}\sum_{i=1}^n g_k^i\right) - \nabla F(y_{k\alpha})\right\|_2^2\right] \\
&\leq 3\mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(y_{k\alpha})\|_2^2\right] + 3\mathbb{E}\left[\left\|g_k^{i_k} - h_k^{i_k}\right\|_2^2\right] + 3\mathbb{E}\left[\left\|\nabla f_{i_k}(y_{k\alpha}) - h_k^{i_k}\right\|_2^2\right] \\
&\leq 6L_{max}\left[F(x_k) - F(y_{k\alpha}) - \langle \nabla F(y_{k\alpha}), x_k - y_{k\alpha}\rangle)\right] + 3\mathbb{E}\left[\left\|g_k^{i_k} - h_k^{i_k}\right\|_2^2\right] + 3\mathbb{E}\left[\left\|\nabla f_{i_k}(y_{k\alpha}) - h_k^{i_k}\right\|_2^2\right]
\end{aligned}$$

The last term is:

$$\mathbb{E}\left[\left\|\nabla f_{i_k}(y_{k\alpha}) - h_k^{i_k}\right\|_2^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(y_{k\alpha}) - h_k^i\right\|_2^2$$

We can bound the expectation of each term as follows. Taking expectation with respect to all sources of randomness we have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla f_i(y_{k\alpha}) - h_k^i\right\|_2^2\right] \\
&= \sum_{j=0}^{k-1}\mathbb{P}(h_k^i = \nabla f_i(y_{j\alpha}))\mathbb{E}\left[\|\nabla f_i(y_{k\alpha}) - \nabla f_i(y_{j\alpha})\|_2^2\right] \\
&= \sum_{j=0}^{k-1}\mathbb{P}(i_j = i, i_{j+1}\neq i, \ldots, i_{k-1}\neq i)\mathbb{E}\left[\|\nabla f_i(y_{k\alpha}) - \nabla f_i(y_{j\alpha})\|_2^2\right] \\
&= \sum_{j=0}^{k-1}\frac{1}{n}\left(1-\frac{1}{n}\right)^{k-j-1}\mathbb{E}\left[\|\nabla f_i(y_{k\alpha}) - \nabla f_i(y_{j\alpha})\|_2^2\right] \\
&\leq \frac{L_{max}^2}{n}\sum_{j=0}^{k-1}\left(1-\frac{1}{n}\right)^{k-j-1}\mathbb{E}\left[\|y_{k\alpha} - y_{j\alpha}\|_2^2\right]
\end{aligned}$$

Let us bound the inner expectation:

$$\begin{aligned}
\mathbb{E}\left[\|y_{k\alpha} - y_{j\alpha}\|_2^2\right] &= \mathbb{E}\left[\left\|\int_{j\alpha}^{k\alpha} -\nabla F(y_t)\,dt + \sqrt{2}\left(W(k\alpha) - W(j\alpha)\right)\right\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\int_{j\alpha}^{k\alpha}\nabla F(y_t)\,dt\right\|_2^2\right] + 4\mathbb{E}\left[\|W(k\alpha) - W(j\alpha)\|_2^2\right] \\
&\leq 2\left\|\int_{j\alpha}^{k\alpha}\nabla F(y_t)\,dt\right\|_{L_2}^2 + 4\mathbb{E}\left[\|W(k\alpha) - W(j\alpha)\|_2^2\right] \\
&\leq 2\left(\int_{j\alpha}^{k\alpha}\left\|\nabla F(y_t)\right\|_{L_2}\,dt\right)^2 + 4\alpha(k-j)d \\
&= 2\mathbb{E}\left[\|\nabla F(y_0)\|_2^2\right]\left(\int_{j\alpha}^{k\alpha}dt\right)^2 + 4\alpha(k-j)d \\
&\leq 2\alpha^2 Ld(k-j)^2 + 4\alpha(k-j)d
\end{aligned}$$

replacing we get:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\nabla f_i(y_{k\alpha}) - h_k^i\right\|_2^2\right] \\
&\leq \frac{L_{max}^2}{n}\left[2\alpha^2 Ld\sum_{j=0}^{k-1}\left(1-\frac{1}{n}\right)^{k-j-1}(k-j)^2 + 4\alpha d\sum_{j=0}^{k-1}\left(1-\frac{1}{n}\right)^{k-j-1}(k-j)\right] \\
&\leq \frac{L_{max}^2}{n}\left[2\alpha^2 Ld\sum_{j=1}^{k} j^2\left(1-\frac{1}{n}\right)^{j-1} + 4\alpha d\sum_{j=1}^{k} j\left(1-\frac{1}{n}\right)^{j-1}\right] \\
&\leq \frac{L_{max}^2}{n}\left[2\alpha^2 Ld\sum_{j=1}^{\infty} j^2\left(1-\frac{1}{n}\right)^{j-1} + 4\alpha d\sum_{j=1}^{\infty} j\left(1-\frac{1}{n}\right)^{j-1}\right] \\
&\leq 4\alpha^2 n^2 L_{max}^2 Ld + 4\alpha n L_{max}^2 d \\
&\leq 4\alpha n L_{max}^2 d\left(\alpha n L + 1\right)
\end{aligned}
$$

Collecting all the bounds and taking $\beta = (\alpha\mu)/2$ we obtain:

$$
\begin{aligned}
\mathbb{E}\left[T^{k+1}\right] \leq &\left(1 - \frac{1}{n} + \frac{12\alpha L_{max}}{n}\right)\frac{\alpha}{2L_{max}}\sum_{i=1}^{n}\left\|g_k^i - h_k^i\right\|_2^2 + \left(1-\frac{\alpha\mu}{2}\right)\left\|x_k - y_{k\alpha}\right\|_2^2 + \\
&\alpha\left(6\alpha L_{max} - 1\right)\left[F(x_k) - F(y_{k\alpha}) - \left\langle\nabla F(y_{k\alpha}), x_{k\alpha} - y_{k\alpha}\right\rangle\right] + \\
&12\alpha^2\kappa L d + 48\alpha^2 n\kappa_{max} L_{max} d\left(\alpha n L + 1\right) \\
\leq &\left(1-\lambda\right)\mathbb{E}\left[T^k\right] + 96\alpha^2 n^2 \kappa_{max} L_{max} d
\end{aligned}
$$

Applying this inequality recursively, and bounding the resulting geometric sum we get the result after noticing that:

$$
W_2^2(\rho_k, \rho^*) \leq \mathbb{E}\left[\left\|x_k - y_{k\alpha}\right\|_2^2\right] \leq \mathbb{E}\left[T^k\right]
$$

□

Bibliography

Agarwal, A. and Bottou, L. A Lower Bound for the Optimization of Finite Sums. In *International Conference on Machine Learning*, pp. 78–86. PMLR, June 2015.

Ahn, K., Yun, C., and Sra, S. SGD with shuffling: optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33, 2020.

Alain, G., Lamb, A., Sankar, C., Courville, A., and Bengio, Y. Variance Reduction in SGD by Distributed Importance Sampling. *arXiv:1511.06481 [cs, stat]*, April 2016.

Allen-Zhu, Z. Katyusha: The First Direct Acceleration of Stochastic Gradient Methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.

Ambrosio, L., Gigli, N., and Savare, G. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2005.

Arjevani, Y. Limitations on Variance-Reduction and Acceleration Schemes for Finite Sums Optimization. *Advances in Neural Information Processing Systems*, 30:3540–3549, 2017.

Bertsekas, D. P. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, August 1973.

Borsos, Z., Krause, A., and Levy, K. Y. Online Variance Reduction for Stochastic Optimization. In *Conference On Learning Theory*, pp. 324–357. PMLR, July 2018.

Borsos, Z., Curi, S., Levy, K. Y., and Krause, A. Online Variance Reduction with Mixtures. In *International Conference on Machine Learning*, pp. 705–714. PMLR, May 2019.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, January 2018.

Bouchard, G., Trouillon, T., Perez, J., and Gaidon, A. Online Learning to Sample. *arXiv:1506.09016 [cs, math, stat]*, March 2016.

Canevet, O., Jose, C., and Fleuret, F. Importance Sampling Tree for Large-scale Empirical Expectation. In *International Conference on Machine Learning*, pp. 1454–1462. PMLR, June 2016.

Cevher, V. and Vũ, B. C. On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, 13(5):1177–1187, July 2019.

Chatterji, N., Flammarion, N., Ma, Y., Bartlett, P., and Jordan, M. On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo. In *International Conference on Machine Learning*, pp. 764–773. PMLR, July 2018.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to Algorithms*. MIT Press, July 2009.

Csiba, D. and Richtárik, P. Importance Sampling for Minibatches. *Journal of Machine Learning Research*, 19(27):1–21, 2018.

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

Dalalyan, A. S. and Karagulyan, A. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, December 2019.

Dalalyan, A. S. and Riou-Durand, L. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, August 2020.

Defazio, A. A Simple Practical Accelerated Method for Finite Sums. *Advances in Neural Information Processing Systems*, 29:676–684, 2016.

Defazio, A. and Bottou, L. On the Ineffectiveness of Variance Reduced Optimization for Deep Learning. *Advances in Neural Information Processing Systems*, 32:1755–1765, 2019.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. *Advances in Neural Information Processing Systems*, 27:1646–1654, 2014a.

Defazio, A., Domke, J., and Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pp. 1125–1133. PMLR, June 2014b.

Dubey, K. A., J. Reddi, S., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. Variance Reduction in Stochastic Gradient Langevin Dynamics. *Advances in Neural Information Processing Systems*, 29, 2016.

Durmus, A. and Moulines, E. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, June 2017.

Durmus, A. and Moulines, E. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, November 2019.

El Hanchi, A. and Stephens, D. Adaptive Importance Sampling for Finite-Sum Optimization and Sampling with Decreasing Step-Sizes. *Advances in Neural Information Processing Systems*, 33, 2020.

Evans, L. C. *An Introduction to Stochastic Differential Equations*. American Mathematical Soc., December 2012.

Folland, G. B. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, June 2013.

Frostig, R., Ge, R., Kakade, S., and Sidford, A. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *International Conference on Machine Learning*, pp. 2540–2548. PMLR, June 2015.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD:

General Analysis and Improved Rates. In *International Conference on Machine Learning*, pp. 5200–5209. PMLR, May 2019.

Gower, R. M., Schmidt, M., Bach, F., and Richtárik, P. Variance-Reduced Methods for Machine Learning. *Proceedings of the IEEE*, 108(11):1968–1983, November 2020.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv:1706.02677 [cs]*, April 2018.

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. A. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, October 2019.

Haochen, J. and Sra, S. Random Shuffling Beats SGD after Finite Epochs. In *International Conference on Machine Learning*, pp. 2624–2633. PMLR, May 2019.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance Reduced Stochastic Gradient Descent with Neighbors. *Advances in Neural Information Processing Systems*, 28:2305–2313, 2015.

Johnson, R. and Zhang, T. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. *Advances in Neural Information Processing Systems*, 26:315–323, 2013.

Johnson, T. B. and Guestrin, C. Training Deep Models Faster with Robust, Approximate Importance Sampling. *Advances in Neural Information Processing Systems*, 31:7265–7275, 2018.

Jordan, R., Kinderlehrer, D., and Otto, F. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.

Katharopoulos, A. and Fleuret, F. Not All Samples Are Created Equal: Deep Learning with Importance Sampling. In *International Conference on Machine Learning*, pp. 2525–2534. PMLR, July 2018.

Kovalev, D., Horváth, S., and Richtárik, P. Don't Jump Through Hoops and Remove Those Loops: SVRG and Katyusha are Better Without the Outer Loop. In *Algorithmic Learning Theory*, pp. 451–467. PMLR, January 2020.

Lan, G. and Zhou, Y. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1):167–215, September 2018.

Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 32:10462–10472, 2019.

Lei, L. and Jordan, M. Less than a Single Pass: Stochastically Controlled Stochastic Gradient. In *Artificial Intelligence and Statistics*, pp. 148–156. PMLR, April 2017.

Li, M., Zhang, T., Chen, Y., and Smola, A. J. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '14, pp. 661–670, New York, NY, USA, August 2014. Association for Computing Machinery.

Lin, H., Mairal, J., and Harchaoui, Z. Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice. *Journal of Machine Learning Research*, 18(212):1–54, 2018.

Liu, R., Wu, T., and Mozafari, B. Adam with Bandit Sampling for Deep Learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Loshchilov, I. and Hutter, F. Online Batch Selection for Faster Training of Neural Networks. November 2015.

Ma, S., Bassily, R., and Belkin, M. The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, July 2018.

Ma, Y.-A., Chen, T., and Fox, E. A Complete Recipe for Stochastic Gradient MCMC. *Advances in Neural Information Processing Systems*, 28, 2015.

Mairal, J. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 25(2):829–855, January 2015.

Mishchenko, K., Khaled Ragab Bayoumi, A., and Richtarik, P. Random Reshuffling: Simple

Analysis with Vast Improvements. *Advances in Neural Information Processing Systems*, 33, 2020.

Moulines, E. and Bach, F. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. *Advances in Neural Information Processing Systems*, 24: 451–459, 2011.

Nagaraj, D., Jain, P., and Netrapalli, P. SGD without Replacement: Sharper Rates for General Smooth Convex Functions. In *International Conference on Machine Learning*, pp. 4703–4711. PMLR, May 2019.

Namkoong, H., Sinha, A., Yadlowsky, S., and Duchi, J. C. Adaptive Sampling Probabilities for Non-Smooth Optimization. In *International Conference on Machine Learning*, pp. 2574–2583. PMLR, July 2017.

Needell, D., Ward, R., and Srebro, N. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. *Advances in Neural Information Processing Systems*, 27:1017–1025, 2014.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, January 2009.

Nemirovski, A. S. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley series in discrete mathematics. Wiley, New York, 1983.

Nemirovsky, A. S., Yudin, D. B., and Dawson, E. R. *Problem Complexity and Method Efficiency in Optimization*. John Wiley, Chichester, 1983.

Nesterov, Y. *Introductory Lectures On Convex Programming*. 1998.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer US, 2004.

Nesterov, Y. *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer International Publishing, 2 edition, 2018.

Nguyen, L., Nguyen, P. H., Dijk, M., Richtarik, P., Scheinberg, K., and Takac, M. SGD and Hogwild! Convergence Without the Bounded Gradients Assumption. In *International Conference on Machine Learning*, pp. 3750–3758. PMLR, July 2018.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. In *International Conference on Machine Learning*, pp. 2613–2621. PMLR, July 2017.

Nguyen, L. M., Tran-Dinh, Q., Phan, D. T., Nguyen, P. H., and van Dijk, M. A Unified Convergence Analysis for Shuffling-Type Gradient Methods. *arXiv:2002.08246 [cs, math, stat]*, February 2020.

Nitanda, A. Stochastic Proximal Gradient Descent with Acceleration Techniques. *Advances in Neural Information Processing Systems*, 27:1574–1582, 2014.

Pavliotis, G. A. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Texts in Applied Mathematics. Springer-Verlag, New York, 2014.

Pesme, S., Dieuleveut, A., and Flammarion, N. On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent. In *International Conference on Machine Learning*, pp. 7641–7651. PMLR, November 2020.

Rajput, S., Gupta, A., and Papailiopoulos, D. Closing the convergence gap of SGD without replacement. In *International Conference on Machine Learning*, pp. 7964–7973. PMLR, November 2020.

Robbins, H. and Monro, S. A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

Roux, N., Schmidt, M., and Bach, F. A Stochastic Gradient Method with an Exponential Convergence _rate for Finite Training Sets. *Advances in Neural Information Processing Systems*, 25:2663–2671, 2012.

Safran, I. and Shamir, O. How Good is SGD with Random Shuffling? In *Conference on Learning Theory*, pp. 3250–3284. PMLR, July 2020.

Salehi, F., Celis, L. E., and Thiran, P. Stochastic Optimization with Bandit Sampling. August 2017.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized Experience Replay. November 2015.

Schmidt, M. and Roux, N. L. Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition. *arXiv:1308.6370 [math]*, August 2013.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, March 2017.

Sebbouh, O., Gazagnadou, N., Jelassi, S., Bach, F., and Gower, R. Towards closing the gap between the theory and practice of SVRG. *Advances in Neural Information Processing Systems*, 32:648–658, 2019.

Shalev-Shwartz, S. and Zhang, T. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Shalev-Shwartz, S. and Zhang, T. Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization. In *International Conference on Machine Learning*, pp. 64–72. PMLR, January 2014.

Shamir, O. Without-Replacement Sampling for Stochastic Gradient Methods. *Advances in Neural Information Processing Systems*, 29:46–54, 2016.

Shen, R. and Lee, Y. T. The Randomized Midpoint Method for Log-Concave Sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, C., Jiang, Y., and Ma, Y. Variance Reduction via Accelerated Dual Averaging for Finite-Sum Optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Stich, S. U., Raj, A., and Jaggi, M. Safe Adaptive Importance Sampling. *Advances in Neural Information Processing Systems*, 30:4381–4391, 2017.

Su, W., Boyd, S., and Candès, E. J. A Differential Equation for Modeling Nesterov's Ac-

celerated Gradient Method: Theory and Insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

Vaswani, S., Bach, F., and Schmidt, M. Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204. PMLR, April 2019a.

Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G., and Lacoste-Julien, S. Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates. *Advances in Neural Information Processing Systems*, 32:3732–3745, 2019b.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Soc., 2003.

Villani, C. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin Heidelberg, 2009.

Woodworth, B. E. and Srebro, N. Tight Complexity Bounds for Optimizing Composite Objectives. *Advances in Neural Information Processing Systems*, 29:3639–3647, 2016.

Zhang, Y. and Xiao, L. Stochastic Primal-Dual Coordinate Method for Regularized Empirical Risk Minimization. *Journal of Machine Learning Research*, 18(84):1–42, 2017.

Zhao, P. and Zhang, T. Stochastic Optimization with Importance Sampling for Regularized Loss Minimization. In *International Conference on Machine Learning*, pp. 1–9. PMLR, June 2015.

Zhou, K., Shang, F., and Cheng, J. A Simple Stochastic Variance Reduced Algorithm with Fast Convergence Rates. In *International Conference on Machine Learning*, pp. 5980–5989. PMLR, July 2018.

Øksendal, B. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer-Verlag, Berlin Heidelberg, 6 edition, 2003.