Large scale optimization and sampling for machine learning and statistics: algorithms and non-asymptotic rates

Ayoub El Hanchi

November 2020

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 had the chance to explore. 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 certainly 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 explosion 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. L'explosion 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, apparemment distinctes, d'optimisation et d'échantillonnage.

Contents

1 Introduction **6**
- 1.1 Statistical decision theory . . . 8
  - 1.1.1 Frequentist approach . . . 9
  - 1.1.2 Bayesian approach . . . 9
- 1.2 Supervised learning . . . 10
  - 1.2.1 Frequentist approach . . . 11
  - 1.2.2 Bayesian approach . . . 11
- 1.3 General formulation . . . 13

2 Continuous-time processes **14**
- 2.1 Assumptions . . . 14
- 2.2 Optimization through gradient flow . . . 15
- 2.3 Sampling through Langevin diffusion . . . 16
- 2.4 Langevin diffusion as gradient flow of relative entropy . . . 18
  - 2.4.1 Absolutely continuous curves in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ . . . 19
  - 2.4.2 Differential calculus in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ . . . 21
  - 2.4.3 Gradient flow of relative entropy . . . 22

3 Discrete-time algorithms **25**
- 3.1 Gradient descent . . . 25
  - 3.1.1 Derivation . . . 25
  - 3.1.2 Convergence analysis . . . 26
- 3.2 Unadjusted Langevin algorithm . . . 27
  - 3.2.1 Derivation . . . 27
  - 3.2.2 Convergence analysis . . . 28

4 Stochastic algorithms **34**

4.1 Oracle model of complexity 34
4.2 Further Assumptions 35
4.3 Stochastic Gradient Descent 36
4.4 Stochastic Gradient Langevin Dynamics 38

5 Variance reduced algorithms **44**
5.1 SAGA 44
5.2 SAGA-LD 47

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 most 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 the fast rates of deterministic methods using only cheap stochastic estimates. On another note, stochasticity was found to be advantageous in many settings where deterministic methods exhibit pathological behavior. One salient example of this is the superior ability of stochastic methods to avoid saddle points in high-dimensional optimization compared to deterministic methods.

Lastly, while at first glance very different, optimization and sampling were found to be very closely related. In fact, it is not too hard to formulate one problem in terms of the other. This has led to a healthy flow of ideas between the two traditionally separate research communities, leading to significant advances in both areas.

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 functional 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 functional 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 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 (see, for e.g., [19]). 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:

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

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

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

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, can now specify a prior over $\mathcal{P}$ by specifying a prior over the parameters ϕ and θ through their density π. We usually assume the factorized form:

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

It is not hard to show that the factorized form of the prior is preserved in 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

Motivated by the finite-sum forms of (1.1) and (1.2), we consider the general 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.

Consider a functional $F : \mathbb{R}^d \to \mathbb{R}$ of the form:

$$F(x) \coloneqq \sum_{i=1}^{N} f_i(x) \tag{1.3}$$

When discussing optimization problems, we will assume that this is the objective function. When discussing a sampling problem, we will assume that the target density is proportional to $e^{-F(\theta)}$. In the next chapters, we will introduce assumptions on the function F and the functions $\{f_i\}_{i=1}^N$ as we need them.

Chapter 2

Continuous-time processes

2.1 Assumptions

We work in $\mathbb{R}^d$ equipped with the usual inner product $\langle x, y\rangle := \sum_{i=1}^d x_i y_i$. We consider a differentiable functional $F : \mathbb{R}^d \to \mathbb{R}$ satisfying the following assumptions.

Assumption 2.1. *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.

Assumption 2.2. *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.

An important consequence of the strong-convexity of F is the existence and uniqueness of the minimizer x^* of F. See ([15], Theorem 2.3.2) for the proof.

Lemma 2.1. *Suppose F satisfies assumption 2.1. Then there exists a unique $x^* \in \mathbb{R}^d$ such that $F(x^*) < F(x)$ for all $x^* \neq x \in \mathbb{R}^d$. Furthermore, $\nabla F(x) = 0 \Leftrightarrow x = x^*$.*

The following inequality can be derived from assumptions 2.1 and 2.2. See ([15], Theorem 2.1.11) for the proof.

Lemma 2.2. *Suppose F satisfies assumptions 2.1 and 2.2. Then for all $x, y \in \mathbb{R}^d$:*

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

Combining Lemma 2.1 and 2.2 with $y = x$ and $x = x^*$ we obtain:

Corollary 2.1.

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

2.2 Optimization through gradient flow

In the optimization problem, our goal is to find x^*, the minimizer of F. In light of Lemma 2.1, x^* is the unique solution to $\nabla F(x) = 0$. In most cases however, there is no analytic solution to this equation. The alternative is to start from some initial guess x_0 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 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:

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

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

We are now ready to state and prove our first convergence theorem.

Theorem 2.1. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Then, 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:*

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

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

Proof. The existence and uniqueness of the solution $(x_t)_{t\in\mathbb{R}}$ follows from Assumption 2.2 and the standard theory of ordinary differential equations. For the convergence rate, we have:

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

where the first inequality follows from Corollary 2.1 and the second follows from $\mu \leq L$. Using Grönwall's inequality finishes the proof. □

2.3 Sampling through Langevin diffusion

In the sampling problem, our goal is to simulate a random variable X^* whose distribution has density ρ^* satisfying $\rho^*(x) \propto e^{-F(x)}$. Assuming that we have access to a source of uniform random variables, the 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 an map, particularly in high-dimension. The alternative is to start with some random variable $X_0 \sim \rho_0$, 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^*.

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 and obeying:

$$dX_t = -\nabla F(X_t)dt + \sqrt{2}dW_t \tag{2.2}$$

One motivation for the choice of this process is the following result. See, for e.g., ([16], Proposition 4.2).

Lemma 2.3. *ρ^* is the invariant probability measure of the process (2.2). Thats is, if $X_0 \sim \rho^*$, then $X_t \sim \rho^*$ for all $t \in \mathbb{R}^+$.*

We return to the issue of motivating the use of this process in the next section. We can nonetheless show that the marginals of this process converge to the target density ρ^* exponentially fast. 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 as:

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

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. Note that this metric is defined on $\mathcal{P}_2(\mathbb{R}^d)$, the set of probability measures on $\mathbb{R}^d$ with finite second moment. The fact that the quantity we defined is indeed a metric follows from optimal transport theory. See [21, 22] for a detailed account.

We are now ready to state and prove our second convergence theorem.

Theorem 2.2. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Then, 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:*

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

for all $t \in \mathbb{R}$, where ρ_t is the distribution of X_t, and $\rho^* \propto e^{-F}$.

Proof. The existence and uniqueness of the solution $(X_t)_{t\in\mathbb{R}}$ follows from Assumption 2.2 and the theory of stochastic differential equations, see, for e.g., [11]. It also follows from this theory that the marginals $(\rho_t)_{t\in\mathbb{R}^+}$ have moments of all order, and therefore the 2-Wasserstein distance is well defined.

For the 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 SDE:

$$dY_t = -\nabla F(Y_t)dt + \sqrt{2}dW_t$$

In light of Lemma 2.3, $Y_t \sim \rho^*$ for all $t \in \mathbb{R}^+$ We further assume that $(Y_t)_{t\in\mathbb{R}^+}$ and $(X_t)_{t\in\mathbb{R}^+}$ are driven by the same Wiener process $(W_t)_{t\in\mathbb{R}^+}$. With this

assumption, the process $Y_t - X_t$ is deterministic and differentiable, and we have:

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

where the first inequality follows from Lemma 2.2 and the second from $\mu \leq L$. Using Grönwall's inequality we get:

$$\begin{aligned}\left\|Y_t - X_t\right\|_2^2 &\leq e^{-\mu t}\left\|Y_0 - X_0\right\|_2^2 \\ \mathbb{E}\left[\left\|Y_t - X_t\right\|_2^2\right] &\leq e^{-\mu t}\mathbb{E}\left[\left\|Y_0 - X_0\right\|_2^2\right]\end{aligned}$$

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

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

Finally, we choose the initial coupling of X_0 and Y_0 to be the optimal one to get:

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

□

2.4 Langevin diffusion as gradient flow of relative entropy

The perspective we took in the previous section was that of stochastic processes, and is the one that we will use in subsequent proofs. However, there is a dual point of view one can take that shows that the Langevin diffusion is the "right" process to consider. This dual perspective comes from the following observation: the marginals of any 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 functional on the space of probability measures that is strongly-convex and is minimized at the target density 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. Amazingly, this can be done. This line of work was started by [13] and culminated in the book [1]. We follow the treatment of [1].

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

$$\frac{d\rho_t}{dt} = -\nabla \mathcal{F}(\rho_t) \tag{2.4}$$

for a given functional $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$.

2.4.1 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 (2.4). We start by formalizing the notion of a curve.

Definition 2.1 (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 this curve. In Euclidean 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 2.2 (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 2.3. *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:*

$$W_2(\gamma(s), \gamma(t)) \leq \int_s^t \beta(r)dr \quad \forall s < t \in I \tag{2.5}$$

The following theorem gives some further justification for this definition. See ([1], Theorem 1.1.2)

Theorem 2.3. *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 (2.5):

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

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 integral (2.5). 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 characterizes it, there exists a vector field that characterizes an absolutely continuous curve in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. The precise statement is the following, see ([1], Theorem 8.3.1):

Theorem 2.4. *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, it seems natural to associate the left-hand side of equation (2.4) with the vector field v_t of Theorem 2.4.

2.4.2 Differential calculus in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

The goal of this section will be to make sense of the right-hand side of equation (2.4). In particular, let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a functional. Our goal will be to define the equivalent of a gradient of this functional.

Let us first recall the definition of the gradient in Euclidean space:

Definition 2.4. *The gradient of a differentiable functional $F : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ is the unique vector $\nabla F(x)$ 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 $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. 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 term, and in particular, the difference $y - x$, has no obvious candidate.

The following result from the theory of optimal transport points to a potential solution. From now on, 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:

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

With this notation in place, we are now ready to state the result. See ([1], section 6.2.3)

Theorem 2.5. *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 a unique 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 functional $\mathcal{F}$ at some reference probability measure μ. In light of Theorem 2.5, a natural candidate for this replacement is the map $t_\mu^\nu - Id$. With this association, we define the gradient of $\mathcal{F}$ as follows.

Definition 2.5. *The gradient of a functional $\mathcal{F} : \mathcal{P}_2^{ab}(\mathbb{R}^d) \to \mathbb{R}$ at a given probability measure μ is the unique function $\nabla\mathcal{F} \in L^2(\mu, \mathbb{R}^d)$ satisfying:*

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

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, we can now define strongly convex functionals.

Definition 2.6. *A functional $\mathcal{F} : \mathcal{P}_2^{ab}(\mathbb{R}^d) \to \mathbb{R}$ is said to be μ-strongly geodesically convex if it satisfies for all $\rho, \nu \in \mathcal{P}_2^{ab}(\mathbb{R}^d)$:*

$$\mathcal{F}(\nu) \geq \mathcal{F}(\rho) + \int_{\mathbb{R}^d} \langle \nabla\mathcal{F}(x), t_\rho^\nu(x) - x \rangle \rho(x)\, dx + \frac{\mu}{2} W_2^2(\rho, \nu)$$

Finally, we have the following chain rule. See ([1], section 10.1.2).

Lemma 2.4. *Let $(\rho_t)_{t \in \mathbb{R}^+}$ be an absolutely continuous curve in $\mathcal{P}_2^{ab}(\mathbb{R}^d)$, and let v_t be the vector field with which it satisfies the continuity equation (2.4). Let $\mathcal{F} : \mathcal{P}_2^{ab}(\mathbb{R}^d) \to \mathbb{R}$ be a differentiable functional. Then we have:*

$$\frac{d}{dt}\mathcal{F}(\rho_t) = \int_{\mathbb{R}^d} \langle \nabla\mathcal{F}, v_t(x) \rangle \rho_t(x)\, dx$$

With these definitions, and assuming the functional $\mathcal{F}$ is differentiable, we can define its gradient flow by the absolutely continuous curve $(\rho_t)_{t \in \mathbb{R}^+}$ satisfying the continuity equation (2.4) with $v_t = -\nabla\mathcal{F}$.

2.4.3 Gradient flow of relative entropy

Recall that our goal is to sample from a probability measure with density $\rho^* \propto e^{-F}$. One functional that is known to be minimized at the target density

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. We have $\mathcal{H}_{\rho^*}(\rho) \geq 0$ and $\mathcal{H}_{\rho^*}(\rho) = 0 \Leftrightarrow \rho = \rho^*$, so ρ^* is the unique minimizer of $\mathcal{H}_{\rho^*}(\rho)$.

The gradient of relative entropy is given by ([1], 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. Furthermore, by strong convexity of F and ([1], Lemma 9.4.7), we have that $\mathcal{H}_{\rho^*}$ is μ-strongly geodesically convex. Finally, we have for $\rho, \rho^* \in \mathcal{P}_2^{ab}(\mathbb{R}^d)$ ([1], Corollary 10.2.7):

$$\frac{1}{2} \nabla W_2^2(\rho, \rho^*) = -(t_\rho^{\rho^*} - Id)$$

where the gradient is with respect to ρ, that is, ρ^* is fixed. $t_\rho^{\rho^*}$ is the optimal transport map from ρ to ρ^* of Theorem 2.5, and Id is the identity map.

Let us now consider the gradient flow of $\mathcal{H}_{\rho^*}$, that is, the absolutely continuous curve $(\rho_t)_{t \in \mathbb{R}^+}$ satisfying the continuity equation (2.4) with $v_t = -\nabla \log \frac{\rho_t}{\rho^*}$. We have the following convergence result.

Theorem 2.6. *Let $(\rho_t)_{t \in \mathbb{R}^+}$ be the gradient flow of $\mathcal{H}_{\rho^*}$. Then we have:*

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

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

Proof. We have:

$$\begin{aligned}
\frac{d}{dt} W_2^2(\rho_t, \rho^*) &= 2 \int_{\mathbb{R}^d} \langle \nabla \log \frac{\rho(x)}{\rho^*(x)}, t_\rho^{\rho^*}(x) - x \rangle \rho(x)\, dx \\
&\leq -2 \left[\mathcal{H}_{\rho^*}(\rho) - \mathcal{H}_{\rho^*}(\rho) + \frac{\mu}{2} W_2^2(\rho_t, \rho^*) \right] \\
&\leq -\mu W_2^2(\rho_t, \rho^*)
\end{aligned}$$

where the first line follows by the chain rule of Lemma 2.4, the second follows from strong geodesic convexity of $\mathcal{H}_{\rho^*}$, and the last from the fact that $\mathcal{H}_{\rho^*}(\rho^*) = 0$, and $\mathcal{H}_{\rho^*}(\rho) \geq 0$. Applying Grönwall's inequality finishes the proof. $\square$

The last thing that remains is to find a stochastic process whose marginals match with the gradient flow of the relative entropy. Recall that by construction the gradient flow $(\rho_t)_{t\in\mathbb{R}^+}$ of $\mathcal{H}_{\rho^*}$ satisfies the continuity equation:

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

which is exactly the Fokker-Planck equation (see, e.g., [16], section 2.5) for the stochastic differential equation:

$$\begin{aligned}dX_t &= \nabla \log \rho^*(X_t) + \sqrt{2} dW_t \\ &= -\nabla F(X_t) + \sqrt{2} dW_t\end{aligned}$$

which is precisely the Langevin diffusion process (2.2).

Chapter 3

Discrete-time algorithms

At this stage, we have two continuous-time processes that solve the optimization and sampling problems, and that converge exponentially fast to their solutions. Our next task will be to construct discretizations of these processes that preserve the fast convergence rate of the continuous-time processes as much as possible.

3.1 Gradient descent

3.1.1 Derivation

Recall that our goal in optimization is to minimize a functional $F : \mathbb{R}^d \to \mathbb{R}$. We achieved this in continuous-time by considering the gradient-flow $(x_t)_{t\in\mathbb{R}^+}$ of F, and showing that x_t converges to x^* exponentially fast. For the purpose of finding x^*, it is enough to evaluate the curve at some large time $t \in \mathbb{R}^+$ so that it is close enough to x^*, the curve itself is of little interest to us. This is in contrast to the standard study of numerical methods for ODEs which usually cares about how well an interpolating curve approximates the true curve.

Let $\alpha > 0$, and let $T \in \mathbb{N}$. Our goal is to approximate $x_{T\alpha}$. We start by partitioning the interval $[0, T\alpha)$ into $[t\alpha, (t+1)\alpha)_{t=0}^{T-1}$. We then start from

the given x_0 and approximate x_α as:

$$x_\alpha = x_0 + \int_0^\alpha -\nabla F(x_s)\, ds \approx x_0 - \alpha \nabla F(x_0) =: \tilde{x}_\alpha$$

Finally, we use this approximation to recursively approximate $x_{(t+1)\alpha}$ as:

$$\begin{aligned} x_{(t+1)\alpha} &= x_{t\alpha} + \int_{t\alpha}^{(t+1)\alpha} -\nabla F(x_s)\, ds \\ &\approx x_{t\alpha} - \alpha \nabla F(x_{t\alpha}) \\ &\approx \tilde{x}_{t\alpha} - \alpha \nabla F(\tilde{x}_{t\alpha}) \\ &:= \tilde{x}_{(t+1)\alpha} \end{aligned}$$

Upon defining $x_k \coloneqq \tilde{x}_{t\alpha}$ for all $k = t \in [T]$, we obtain the gradient descent sequence $(x_k)_{k=0}^T$, which can be restated as:

$$x_{k+1} = x_k - \alpha \nabla F(x_k) \tag{3.1}$$

3.1.2 Convergence analysis

We made numerous approximations in the derivation of the gradient descent algorithm for F from the gradient flow of F, and it is a priori unclear whether the sequence $(x_k)_{k=0}^T$ (i) converges to x^* as $T \to \infty$. (ii) preserves the linear convergence rate that the gradient flow of F enjoys in continuous-time. Fortunately, both of these properties hold as the following theorem shows.

Theorem 3.1. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Then, starting from x_0, the gradient descent sequence (3.1) with $\alpha \leq \frac{2}{L+\mu}$ satisfies:*

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

for all $k \in \mathbb{N}$.

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

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

where the first line follows from the definition of the gradient descent sequence (3.1), the second by expanding the square, the third from Corollary 2.1, and the last from the assumption $\alpha \leq \frac{2}{L+\mu}$. Applying this inequality recursively gives the stated result. □

3.2 Unadjusted Langevin algorithm

3.2.1 Derivation

We proceed similarly to the derivation of gradient descent. Let $\alpha > 0$, and let $T \in \mathbb{N}$. Our goal is to approximate $X_{T\alpha}$. We start by partitioning the interval $[0, T\alpha)$ into $[t\alpha, (t+1)\alpha)_{t=0}^{T-1}$. We then generate $X_0 \sim \rho_0$ and approximate X_α as:

$$\begin{aligned}
X_\alpha &= X_0 + \int_0^\alpha -\nabla F(X_s)\, ds + \sqrt{2} \int_0^\alpha dW_s \\
&\approx X_0 - \alpha \nabla F(X_0) + \sqrt{2\alpha}\, \xi_0 \\
&=: \tilde{X}_\alpha
\end{aligned}$$

Where $\xi_0 \sim \mathcal{N}(0, I_d)$. Finally, we use this approximation to recursively approximate $X_{(t+1)\alpha}$ as:

$$\begin{aligned}
X_{(t+1)\alpha} &= X_{t\alpha} + \int_{t\alpha}^{(t+1)\alpha} -\nabla F(X_s)\, ds + \sqrt{2} \int_{t\alpha}^{(t+1)\alpha} dW_s \\
&\approx X_{t\alpha} - \alpha \nabla F(X_{t\alpha}) + \sqrt{2\alpha}\, \xi_t \\
&\approx \tilde{X}_{t\alpha} - \alpha \nabla F(\tilde{X}_{t\alpha}) + \sqrt{2\alpha}\, \xi_t \\
&:= \tilde{X}_{(t+1)\alpha}
\end{aligned}$$

Where $\xi_t \sim \mathcal{N}(0, I_d)$ and where all the ξ_t are independent. Upon defining $X_k \coloneqq \tilde{X}_{t\alpha}$ for all $k = t \in [T]$, we obtain the discrete-time stochastic process $(X_k)_{k=0}^T$ of the unadjusted Langevin algorithm, which can be restated as:

$$X_{k+1} = X_k - \alpha \nabla F(X_k) + \sqrt{2\alpha}\, \xi_k \tag{3.2}$$

3.2.2 Convergence analysis

Just like in the derivation of gradient descent, we made numerous approximations in the derivation of the unadjusted Langevin algorithm. It is a priori unclear whether the marginals $(\rho_k)_{k=0}^T$ of the stochastic process $(X_k)_{k=0}^T$ (i) converge to ρ^* as $T \to \infty$. (ii) preserve the linear convergence rate that the marginals of the Langevin diffusion enjoy.

Perhaps surprisingly, the first property does not hold, while the second does. In particular, for a given $\alpha > 0$, the marginals of $(X_k)_{k=0}^T$ converge linearly to a probability measure ρ_α^*, which is in general different from the target measure ρ^*. Nevertheless, we can show that ρ_α^* and ρ^* are close for small α.

We start by first showing the existence of ρ_α^* and the convergence of the marginals $(\rho_k)_{k=0}^T$ to it.

Theorem 3.2. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Then, starting from $X_0 \sim \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$, the marginals $(\rho_k)_{k=0}^T$ of the Unadjusted Langevin algorithm (3.2) with $\alpha \leq \frac{2}{L+\mu}$ satisfy:*

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

for some $\rho_\alpha^* \in \mathcal{P}_2(\mathbb{R}^d)$. Furthermore, ρ_α^* is the stationary distribution of the Unadjusted Langevin algorithm process (3.2).

Proof. Let $X_0 \sim \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $Y_0 \sim \rho_0' \in \mathcal{P}_2(\mathbb{R}^d)$, and evolve each according to the Unadjusted Langevin algorithm:

$$\begin{aligned} X_{k+1} &= X_k - \alpha \nabla F(X_k) + \sqrt{2\alpha}\, \xi_k \\ Y_{k+1} &= Y_k - \alpha \nabla F(Y_k) + \sqrt{2\alpha}\, \xi_k \end{aligned}$$

where the ξ_k are independent $\mathcal{N}(0, I_d)$ random variables, and are the same for both processes. Denote by $(\rho_k)_{k=0}^T$ the marginals of $(X_k)_{k=0}^T$ and $(\rho'_k)_{k=0}^T$ the marginals of $(Y_k)_{k=0}^T$. Let $k \in \mathbb{N}$. We have:

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

Where the first line follows from the definition of the processes, the second by expanding the square, the third from Lemma 2.2, and the last by the condition $\alpha \leq \frac{2}{L+\mu}$. Applying this inequality recursively we get:

$$\|Y_{k+1} - X_{k+1}\|_2^2 \leq \left(1 - 2\alpha\frac{\mu L}{L+\mu}\right)^k \|Y_0 - X_0\|_2^2$$

Taking Y_0, X_0 so that they are optimally coupled, we get after taking expectation on both sides:

$$\mathbb{E}\left[\|Y_{k+1} - X_{k+1}\|_2^2\right] \leq \left(1 - 2\alpha\frac{\mu L}{L+\mu}\right)^k W_2^2(\rho'_0, \rho_0)$$

by minimality of the coupling defining the 2-Wasserstein distance, the left-hand side is an upper bound on $W_2^2(\rho'_{k+1}, \rho_{k+1})$, so we get:

$$W_2^2(\rho'_{k+1}, \rho_{k+1}) \leq \left(1 - 2\alpha\frac{\mu L}{L+\mu}\right)^k W_2^2(\rho'_0, \rho_0)$$

Taking $k = 0$ in the above inequality shows that the map $T : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$ induced by the Markov kernel defining the Unadjusted Langevin algorithm is a contraction mapping. Recalling that $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a complete metric space, we have by the Banach fixed point theorem the existence of a

unique probability measure ρ_α^* which is a fixed point of T. In other words, ρ_α^* is the unique stationary distribution of (3.2). Applying our derived inequality with the choice $\rho_0' = \rho_\alpha^*$, and noticing that this implies $\rho_k' = \rho_\alpha^*$ for all $k \in \mathbb{N}$, we get our stated result. □

The result we have shown would be useless for the purpose of sampling from ρ^* if we cannot show that ρ^* and ρ_α^* are close. Heuristically, however, as $\alpha \to 0$, the Unadjusted Langevin algorithm becomes the Langevin diffusion process, which we know to converge to the desired probability measure ρ^*. For small α, we therefore expect ρ^* and ρ_α^* to be close. The following is the precise statement.

Theorem 3.3. *Let ρ_α^* be the stationary distribution of the unadjusted Langevin algorithm when run with $\alpha \leq \frac{2}{L+\mu}$. Then:*

$$W_2(\rho_\alpha^*, \rho^*) \leq 4\kappa(\alpha d)^{1/2}$$

where $\kappa = \frac{L}{\mu}$.

Proof. **Useful inequalities:** Define the following norm on random vectors:

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

Recall the integral form of Minkowski's inequality:

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

for an integrable real valued stochastic process $(X_t)_{t\in[a,b]}$. Finally, we have the inequality:

$$\left\|\int_a^b v_t\, dt\right\|_2 \leq \int_a^b \|v_t\|_2\, dt$$

for a vector valued function $v_t : [a, b] \to \mathbb{R}^d$, which can be derived from the Cauchy-Schwarz inequality. Combining the previous two inequalities gives us:

$$\left\|\int_a^b V_t\, dt\right\|_{L_2} \leq \int_a^b \|V_t\|_{L_2}\, dt \tag{3.3}$$

for a vector valued stochastic process $(V_t)_{t\in[a,b]}$.

Proof of statement: Let $X_0 \sim \rho_\alpha^*$ and $Y_0 \sim \rho^*$, and assume that X_0 and Y_0 are optimally coupled. Define:

$$X_\alpha \coloneqq X_0 - \alpha \nabla F(X_0) + \sqrt{2\alpha}\, W_\alpha$$
$$dY_t \coloneqq -\nabla F(Y_t)dt + \sqrt{2}\, dW_t$$

By stationarity of the initial probability measures, we have $Y_\alpha \sim \rho^*$ and $X_\alpha \sim \rho_\alpha^*$, so by minimality of the coupling defining the Wasserstein distance, we have:

$$W_2(\rho^*, \rho_\alpha^*) \leq \|Y_\alpha - X_\alpha\|_{L_2}$$

On the other hand, we have by the triangle inequality:

$$\begin{aligned}\|Y_\alpha - X_\alpha\|_{L_2} &= \left\| Y_0 - X_0 - \int_0^\alpha \nabla F(Y_t)dt + \alpha \nabla F(X_0) \right\|_{L_2} \\ &\leq \|Y_0 - X_0 - \alpha(\nabla F(Y_0) - \nabla F(X_0))\|_{L_2} + \left\| \int_0^\alpha \nabla F(Y_t) - \nabla F(Y_0)\, dt \right\|_{L_2}\end{aligned}$$

The first term can be bound as usual using Lemma 2.2 and the condition $\alpha \leq \frac{2}{L+\mu}$ to yield:

$$\|Y_0 - X_0 - \alpha(\nabla F(Y_0) - \nabla F(X_0))\|_{L_2} \leq (1 - \alpha\mu)^{1/2} \|Y_0 - X_0\|_{L_2}$$

Combining the two bounds gives us:

$$\left(1 - (1 - \alpha\mu)^{1/2}\right) W_2(\rho_\alpha^*, \rho^*) \leq \left\| \int_0^\alpha \nabla F(Y_t) - \nabla F(Y_0)\, dt \right\|_{L_2}$$

Using the inequality $1 - \sqrt{1-x} \geq \frac{1}{2}x$ we get:

$$W_2(\rho_\alpha^*, \rho^*) \leq \frac{2}{\alpha\mu} \left\| \int_0^\alpha \nabla F(Y_t) - \nabla F(Y_0)\, dt \right\|_{L_2}$$

We bound the term on the right-hand side as follows:

$$\begin{aligned}
\left\| \int_0^\alpha \nabla F(Y_t) - \nabla F(Y_0)\, dt \right\|_{L_2} &\leq \int_0^\alpha \left\| \nabla F(Y_t) - \nabla F(Y_0) \right\|_{L_2} dt \\
&\leq L \int_0^\alpha \left\| Y_t - Y_0 \right\|_{L_2} dt \\
&\leq L \int_0^\alpha \left\| \int_0^t \nabla F(Y_s)\, ds \right\|_{L_2} dt + L \int_0^\alpha \left\| \sqrt{2} W_s \right\|_{L_2} dt \\
&\leq L \int_0^\alpha \int_0^t \left\| \nabla F(Y_s) \right\|_{L_2} ds\, dt + L\sqrt{2d} \int_0^\alpha \sqrt{t}\, dt \\
&= L \left\| \nabla F(Y_0) \right\|_{L_2} \int_0^\alpha \int_0^t ds\, dt + L\sqrt{2d} \int_0^\alpha \sqrt{t}\, dt \\
&= \frac{1}{2} \alpha^2 L \left\| \nabla F(Y_0) \right\|_{L_2} + \frac{2}{3} \alpha^{3/2} L \sqrt{2d}
\end{aligned}$$

where the first line follows from inequality (3.3), the second from Assumption 2.2, the third by definition of the process $(Y_t)_{t\in[0,\alpha]}$ and the triangle inequality, the fourth again by inequality (3.3), and the fifth from the stationarity of $(Y_t)_{t\in[0,\alpha]}$. It remains to bound the L_2 norm of the gradient. This can be done using integration by parts and the smoothness of F as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\| \nabla F(Y_0) \right\|_2^2 \right] &= \int_{\mathbb{R}^d} \left\| \nabla F(y) \right\|_2^2 \rho^*(y)\, dy \\
&= \int_{\mathbb{R}^d} \langle \nabla F(y), \nabla F(y) \rangle \rho^*(y)\, dy \\
&= \int_{\mathbb{R}^d} \langle \nabla \log \rho^*(y), \nabla \log \rho^*(y) \rangle \rho^*(y)\, dy \\
&= \int_{\mathbb{R}^d} \langle \nabla \rho^*(y), \nabla \log \rho^*(y) \rangle\, dy \\
&= -\int_{\mathbb{R}^d} \Delta \log \rho^*(y) \rho^*(y)\, dy \\
&= \int_{\mathbb{R}^d} \Delta F(y) \rho^*(y)\, dy \\
&\leq Ld
\end{aligned}$$

where in the fifth line we used integration by parts, which can be justified by the strong-convexity of F, and in the last line we used the L-smoothness of F.

Relacing we get:

$$\begin{aligned}\left\|\int_0^\alpha \nabla F(Y_t) - \nabla F(Y_0)\, dt\right\|_{L_2} &\leq \frac{1}{2}\alpha^2 L \left\|\nabla F(Y_0)\right\|_{L_2} + \frac{2}{3}\alpha^{3/2} L\sqrt{2d} \\ &\leq \alpha^{3/2} L\sqrt{d}\left(\frac{1}{2}\alpha^{1/2}L^{1/2} + \frac{2\sqrt{2}}{3}\right) \\ &\leq \alpha^{3/2} L\sqrt{d}\left(\frac{\sqrt{2}}{2} + \frac{2\sqrt{2}}{3}\right) \\ &\leq 2\alpha^{3/2} L\sqrt{d}\end{aligned}$$

Replacing in the original bound we get the result. □

Combining Theorem 3.2 and Theorem 3.3 we obtain our final convergence result.

Corollary 3.1. *Let $F : \mathbb{R}^d \to \mathbb{R}$ satisfy Assumptions 2.1 and 2.2, then the marginals $(\rho_k)_{k=0}^T$ of the iterates of the unadjusted Langevin algorithm (2.2) with $\alpha \leq \frac{2}{L+\mu}$ satisfy:*

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

where $\kappa = \frac{L}{\mu}$.

Proof. Apply the triangle inequality and use the bounds of Theorems 3.2 and 3.3. □

Chapter 4

Stochastic algorithms

Up to this point, we have brushed issues of computational complexity to the side, and only strived to construct algorithms that enjoy fast convergence to their solutions in terms of the number of iterations they take, without regard to the iteration cost. As we have discussed in chapter 1 however, we are interested in solving optimization and sampling problem that have a finite-sum structure, with the underlying assumption that the number of elements in the sum is very large. In this scenario, it becomes a necessity to only use stochastic estimates of the gradient of the functional F to keep the computational load manageable. In this chapter we explore the two basic algorithms that implement this idea: Stochastic Gradient Descent (SGD) for optimization, and Stochastic Gradient Langevin Dynamics (SGLD) for sampling.

4.1 Oracle model of complexity

Before delving into the algorithms themselves and their convergence proofs, let us first formalize a simple yet effective notion of computational complexity that will make it easier to compare the performance of algorithms. Recall that we assume that the functional $F : \mathbb{R}^d \to \mathbb{R}$ is of the form:

$$F(x) = \sum_{i=1}^{N} f_i(x) \tag{4.1}$$

so that its gradient is given by:

$$\nabla F(x) = \sum_{i=1}^{N} \nabla f_i(x)$$

To take into account the high cost of computing the sum of all N gradients, we will assume that we have access to an oracle which, given an index $i \in [N]$ and a point $x \in \mathbb{R}^d$, returns the gradient $\nabla f_i(x)$. We will assume that one call to the oracle has unit cost, so that for example evaluating the full gradient $\nabla F(x)$ has cost N. We will also ignore the cost of other operations such as, say, the cost of generating normal random variables. This is justified since in most cases the cost of gradient evaluations dominates that of all other operations.

With this oracle model of complexity, a common way to specify the cost of an algorithm is by counting the number of gradient evaluations it requires to reach a solution ε-away from the true solution. For example, gradient descent needs $O(N\kappa \log(\frac{1}{\varepsilon}))$ gradient evaluations to reach a point ε-away from the optimum. The unadjusted Langevin algorithm on the other hand has two regimes. In the low precision regime $\varepsilon > 4\kappa(\frac{2d}{L+\mu})^{1/2}$, it only needs $O(N\kappa \log(\frac{1}{\varepsilon'}))$ for $\varepsilon' = \varepsilon - 4\kappa(\frac{2d}{L+\mu})^{1/2}$ gradient evaluations to generate a sample ε-away from the target. In the high precision regime $\varepsilon < 4\kappa(\frac{2d}{L+\mu})^{1/2}$, this complexity deteriorates to $O(N\kappa^2 d\varepsilon^{-2})$. When N is large, it dominates the complexity, particularly when it interacts with other parameters in a multiplicative way. The goal of stochastic methods will be to reduce the total cost by only using cheap stochastic estimates at each iteration.

4.2 Further Assumptions

In order to be able to exploit the finite-sum structure of the functional F, we will have to make further assumptions about the individual functions f_i. In particular, we assume that they are differentiable and satisfy the following assumptions.

Assumption 4.1. *The functions $(f_i)_{i=1}^N$ are convex, that is, for all $x, y \in \mathbb{R}^d$:*

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

Assumption 4.2. *The functions $(f_i)_{i=1}^N$ are smooth, that is, for each $i \in [N]$ there exists an $L_i > 0$ such that for all $x, y \in \mathbb{R}^d$:*

$$\|\nabla f_i(y) - \nabla f_i(x)\|_2 \leq L_i \|y - x\|_2$$

we also define $L_{max} \coloneqq \max_{i\in[N]} L_i$.

It is not hard to show that the following inequalities hold:

$$L \leq \sum_{i=1}^{N} L_i \leq N L_{max}$$

where L is the smoothness constant of F.

With these assumptions we have the following inequality (see [15], Theorem 2.1.5):

Lemma 4.1. *Suppose the functions $(f_i)_{i=1}^N$ satisfy assumptions 4.1 and 4.2. Then for each $i \in [N]$, for all $x, y \in \mathbb{R}^d$ we have:*

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

4.3 Stochastic Gradient Descent

The idea of stochastic gradient descent (SGD) is simple: instead of evaluating the full gradient at each iteration, we compute only an unbiased estimate. The stochastic gradient descent sequence is given by:

$$x_{k+1} = x_k - \alpha N \nabla f_{I_k}(x_k) \tag{4.2}$$

where I_k is a uniformly distributed random variable on $[N]$, and the collection $(I_k)_{k=0}^{T-1}$ is independent. Note that:

$$\mathbb{E}\left[N \nabla f_{I_k}(x_k)\right] = \sum_{i=1}^{N} \frac{1}{N} N \nabla f_{I_k}(x_k) = \sum_{i=1}^{N} \nabla f_i(x_k) = \nabla F(x_k)$$

SGD therefore replaces the full gradient with an unbiased estimate. As oppose to gradient descent, each iteration has only unit cost. The downside is that the iterates of SGD do not converge to the solution, but only to a neighborhood of it, as shown by the following theorem.

Theorem 4.1. *Suppose that $F : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2, and has the finite-sum form given by equation (4.1). Further, assume the functions $(f_i)_{i=1}^N$ satisfy Assumptions 4.1 and 4.2. Then the elements of the SGD sequence (4.2) with $\alpha \leq \frac{1}{2NL_{max}}$ satisfy:*

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

where $\sigma^2 \coloneqq N \sum_{i=1}^N \|\nabla f_i(x^*)\|_2^2$.

Proof. Let $k \in \mathbb{N}$. Conditioning on $(I_t)_{t=1}^{k-1}$, we have:

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

Let us bound the last term first:

$$\begin{aligned}\mathbb{E}\left[\|N\nabla f_{I_k}(x_k)\|_2^2\right] &= \mathbb{E}\left[\|N\nabla f_{I_k}(x_k) - N\nabla f_{I_k}(x^*) + N\nabla f_{I_k}(x^*)\|_2^2\right] \\ &\leq 2\mathbb{E}\left[\|N\nabla f_{I_k}(x_k) - N\nabla f_{I_k}(x^*)\|_2^2\right] + 2\mathbb{E}\left[\|N\nabla f_{I_k}(x^*)\|_2^2\right]\end{aligned}$$

Expanding the first expectation we get:

$$\begin{aligned}&\mathbb{E}\left[\|N\nabla f_{I_k}(x_k) - N\nabla f_{I_k}(x^*)\|_2^2\right] = \sum_{i=1}^N N\|\nabla f_i(x_k) - \nabla f_i(x^*)\|_2^2 \\ &\leq \sum_{i=1}^N 2NL_i\big(f_i(x_k) - f_i(x^*) + \langle \nabla f_i(x^*), x - x^*\rangle\big) \\ &\leq 2NL_{max}\left(F(x_k) - F(x^*) + \langle \sum_{i=1}^N \nabla f_i(x^*), x - x^*\rangle\right) \\ &= 2NL_{max}\left(F(x_k) - F(x^*)\right)\end{aligned}$$

We can expand the second expectation to get:

$$\mathbb{E}\left[\|N\nabla f_{I_k}(x^*)\|_2^2\right] = N\sum_{i=1}^N \|\nabla f_i(x^*)\|_2^2 =: \sigma^2$$

Replacing in the original bound and using the strong convexity of F we obtain:

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

Taking expectation with respect to $(I_t)_{t=0}^{k-1}$ and recursively applying this inequality we get:

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

□

In the low precision regime, SGD is therefore more efficient than gradient descent, requiring only $O(\frac{NL_{max}}{\mu}\log\frac{1}{\varepsilon'})$ gradient evaluations for $\varepsilon' = \varepsilon - \frac{\sigma^2}{2NL_{max}\mu} > 0$. On the other hand, SGD performs worse in the high precision regime, where the complexity deteriorates to $O(\frac{\sigma^2}{\mu\varepsilon})$.

4.4 Stochastic Gradient Langevin Dynamics

The idea of stochastic gradient Langevin dynamics (SGLD) is similar to that of SGD: instead of using the full gradient at each iteration of the unadjusted Langevin algorithm, we use only an unbiased estimate. The resulting process is given by:

$$X_{k+1} = X_k - \alpha N\nabla f_{I_k}(X_k) + \sqrt{2\alpha}\,\xi_k$$

where $\xi_k \sim \mathcal{N}(0, I_d)$, I_k is a uniformly distributed random variable on $[N]$, and the collections $(\xi_k)_{k=0}^{T-1}$ and $(I_k)_{k=0}^{T-1}$ are independent, both within themselves and between them. Similar to the unadjusted Langevin algorithm,

SGLD converges to a unique stationary distribution. It is however, in general farther from the target ρ^* than the stationary distribution of the unadjusted Langevin algorithm. Following the analysis of the ULA, we start by showing the existence of a unique stationary measure of (4.4) and the linear convergence of the marginals of SGLD to it.

Theorem 4.2. *Suppose that $F : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2, and has the finite-sum form given by equation (4.1). Further, assume the functions $(f_i)_{i=1}^N$ satisfy Assumptions 4.1 and 4.2. Then, starting from $X_0 \sim \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$, the marginals $(\rho_k)_{k=0}^T$ of SGLD (4.4) with $\alpha \leq \frac{1}{NL_{max}}$ satisfy:*

$$W_2^2(\rho_k, \rho^*_{SGLD(\alpha)}) \leq (1 - \alpha\mu)^k W_2^2(\rho_0, \rho^*_{SGLD(\alpha)})$$

for some $\rho^*_{SGLD(\alpha)} \in \mathcal{P}_2(\mathbb{R}^d)$. Furthermore, $\rho^*_{SGLD(\alpha)}$ is the unique stationary distribution of SGLD (4.4).

Proof. Let $X_0 \sim \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $Y_0 \sim \rho'_0 \in \mathcal{P}_2(\mathbb{R}^d)$, and evolve each according to SGLD:

$$\begin{aligned} X_{k+1} &= X_k - \alpha N \nabla f_{I_k}(X_k) + \sqrt{2\alpha}\,\xi_k \\ Y_{k+1} &= Y_k - \alpha N \nabla f_{I_k}(Y_k) + \sqrt{2\alpha}\,\xi_k \end{aligned}$$

where the ξ_k are independent $\mathcal{N}(0, I_d)$ random variables, and the I_k are independent uniform random variables over $[N]$, and are the same for both processes. Denote by $(\rho_k)_{k=0}^T$ the marginals of $(X_k)_{k=0}^T$ and $(\rho'_k)_{k=0}^T$ the marginals of $(Y_k)_{k=0}^T$. Let $k \in \mathbb{N}$. We have:

$$\begin{aligned} \|Y_{k+1} - X_{k+1}\|_2^2 &= \|Y_k - X_k - \alpha N(\nabla f_{I_k}(Y_k) - \nabla f_{I_k}(X_k))\|_2^2 \\ &= \|Y_k - X_k\|_2^2 - 2\alpha\langle N\nabla f_{I_k}(Y_k) - N\nabla f_{I_k}(X_k), Y_k - X_k\rangle + \\ &\quad \alpha^2 \|N\nabla f_{I_k}(Y_k) - N\nabla f_{I_k}(X_k)\|_2^2 \end{aligned}$$

Taking expectation with respect to I_k conditioned on all other randomness, we get:

$$\begin{aligned} \mathbb{E}\left[\|Y_{k+1} - X_{k+1}\|_2^2\right] &= \|Y_k - X_k\|_2^2 - 2\alpha\langle \nabla F(Y_k) - \nabla F(X_k), Y_k - X_k\rangle \\ &\quad + \alpha^2 \mathbb{E}\left[\|N\nabla f_{I_k}(Y_k) - N\nabla f_{I_k}(X_k)\|_2^2\right] \end{aligned}$$

We bound the last term as follows:

$$\mathbb{E}\left[\|N\nabla f_{I_k}(Y_k) - N\nabla f_{I_k}(X_k)\|_2^2\right] = N\sum_{i=1}^{N}\|\nabla f_i(Y_k) - \nabla f_i(X_k)\|_2^2$$
$$\leq 2NL_{max}\sum_{i=1}^{N}(f_i(Y_k) - f_i(X_k) - \langle\nabla f_{I_k}(X_k), Y_k - X_k\rangle)$$
$$= 2NL_{max}\left(F(Y_k) - F(X_k) - \langle\nabla F(x_k), Y_k - X_k\rangle\right)$$

Where the second line uses Lemma 4.1. We bound the inner product term as follows:

$$-\langle\nabla F(Y_k) - \nabla F(X_k), Y_k - X_k\rangle$$
$$\leq \langle\nabla F(Y_k), X_k - Y_k\rangle + \langle\nabla F(X_k), Y_k - X_k\rangle$$
$$\leq -\left(F(Y_k) - F(X_k) + \frac{\mu}{2}\|Y_k - X_k\|_2^2\right) + \langle\nabla F(X_k), Y_k - X_k\rangle$$
$$= -\left(F(Y_k) - F(X_k) - \langle\nabla F(X_k), Y_k - X_k\rangle\right) - \frac{\mu}{2}\|Y_k - X_k\|_2^2$$

where the line before last uses the μ-strong convexity of F.

Replacing into the original bound we get:

$$\mathbb{E}\left[\|Y_{k+1} - X_{k+1}\|_2^2\right] \leq (1-\alpha\mu)\|Y_k - X_k\|_2^2 +$$
$$2\alpha\left(\alpha NL_{max} - 1\right)\left(F(Y_k) - F(X_k) - \langle\nabla F(X_k), Y_k - X_k\rangle\right)$$
$$\leq (1-\alpha\mu)\|Y_k - X_k\|_2^2$$

where the second inequality follows from the condition $\alpha \leq \frac{1}{NL_{max}}$ and the positivity of the second term which is due to the convexity of F.

Taking expectation and applying the above inequality recursively we get:

$$\mathbb{E}\left[\|Y_{k+1} - X_{k+1}\|_2^2\right] \leq (1-\alpha\mu)^k\,\mathbb{E}\left[\|Y_0 - X_0\|_2^2\right]$$

Taking X_0 and Y_0 to be optimally coupled, and noticing that the left-hand side is an upper bound on $W_2^2(\rho'_{k+1}, \rho_{k+1})$ we get:

$$W_2^2(\rho'_{k+1}, \rho_{k+1}) \leq (1-\alpha\mu)^k\, W_2^2(\rho_0, \rho'_0)$$

Taking $k = 0$ in the above inequality shows that the map $T : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$ induced by the Markov kernel defining SGLD is a contraction mapping. Recalling that $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a complete metric space, we have by the

Banach fixed point theorem the existence of a unique probability measure $\rho^*_{SGLD(\alpha)}$ which is a fixed point of T. In other words, $\rho^*_{SGLD(\alpha)}$ is the unique stationary distribution of (3.2). Applying our derived inequality with the choice $\rho'_0 = \rho^*_{SGLD(\alpha)}$, and noticing that this implies $\rho'_k = \rho^*_{SGLD(\alpha)}$ for all $k \in \mathbb{N}$, we get our stated result. □

While SGLD preserves the exponential convergence of SGLD, its stationary distribution can be quite far from the target density. We now give a precise quantitative bound on this deviation.

Theorem 4.3. *Let $\rho^*_{SGLD(\alpha)}$ be the stationary distribution of SGLD with step size $\alpha \leq \frac{1}{NL_{max}}$. Then:*

$$W_2(\rho^*_{SGLD(\alpha)}, \rho^*) \leq 4\kappa(\alpha d)^{1/2} + \sqrt{\frac{\alpha\sigma^2}{\mu}}$$

where $\kappa = \frac{L}{\mu}$ and :

$$\sigma^2 := 2\mathbb{E}\left[\|N\nabla f_I(Y) - \nabla F(Y)\|_2^2\right]$$

where the expectation is over $Y \sim \rho^*$ and the index I which is uniform over $[N]$.

Proof. Let $X_0 \sim \rho^*_{SGLD(\alpha)}$, $Y_0 \sim \rho^*$ and assume they are optimally coupled. Define:

$$X_\alpha = X_0 - \alpha N\nabla f_{I_0}(X_0) + \sqrt{2\alpha}W_\alpha$$
$$dY_t = -\nabla F(Y_t) + \sqrt{2}dW_t$$

By stationarity of the initial probability measures, we have $X_\alpha \sim \rho^*_{SGLD(\alpha)}$ and $Y_\alpha \sim \rho^*$, so by minimality of the coupling defining the Wasserstein distance we get:

$$W_2(\rho^*_{SGLD(\alpha)}, \rho^*) \leq \|Y_\alpha - X_\alpha\|_{L_2}$$

On the other hand we have:

$$\begin{aligned}
&\|X_\alpha - Y_\alpha\|_{L_2} = \left\| Y_0 - X_0 - \int_0^\alpha \nabla F(Y_t)\, dt + \alpha N \nabla f_{I_0}(X_0) \right\|_{L_2} \\
&= \left\| Y_0 - X_0 - \int_0^\alpha \left(\nabla F(Y_t) - \nabla F(Y_0)\right)\, dt - \alpha \nabla F(Y_0) + \alpha N \nabla f_{I_0}(X_0) \right\|_{L_2} \\
&\leq \|Y_0 - X_0 - \alpha \nabla F(Y_0) + \alpha N \nabla f_{I_0}(X_0)\|_{L_2} + \left\| \int_0^\alpha \nabla F(Y_t) - \nabla F(Y_0)\, dt \right\|_{L_2}
\end{aligned}$$

From the proof of Theorem 3.3, the second term is bounded by:

$$\left\| \int_0^\alpha \nabla F(Y_t) - \nabla F(Y_0)\, dt \right\|_{L_2} \leq 2\alpha^{3/2} L \sqrt{d}$$

For the first term we have, taking expectation over the index I_0:

$$\begin{aligned}
&\mathbb{E}\left[\|Y_0 - X_0 - \alpha(\nabla F(Y_0) - N \nabla f_{I_0}(X_0))\|_2^2\right] \\
&\leq \|Y_0 - X_0\|_2^2 - 2\alpha \langle \nabla F(Y_0) - \mathbb{E}\left[N \nabla f_{I_0}(X_0)\right], Y_0 - X_0 \rangle + \\
&\quad \alpha^2 \mathbb{E}\left[\|N \nabla f_{I_0}(X_0) - \nabla F(Y_0)\|\right] \\
&\leq \|Y_0 - X_0\|_2^2 - 2\alpha \langle \nabla F(Y_0) - \nabla F(X_0), Y_0 - X_0 \rangle + \\
&\quad 2\alpha^2 \mathbb{E}\left[\|N \nabla f_{I_0}(X_0) - N \nabla f_{I_0}(Y_0)\|_2^2\right] + \\
&\quad 2\alpha^2 \mathbb{E}\left[\|N \nabla f_{I_0}(Y_0) - \nabla F(Y_0)\|_2^2\right]
\end{aligned}$$

The first three terms can be bound as in Theorem 4.2 assuming $\alpha \leq \frac{1}{2NL_{max}}$ to give:

$$\begin{aligned}
&\mathbb{E}\left[\|Y_0 - X_0 - \alpha(\nabla F(Y_0) - N \nabla f_{I_0}(X_0))\|_2^2\right] \\
&\leq (1 - \alpha\mu) \|Y_0 - X_0\|_2^2 + 2\alpha^2 \mathbb{E}\left[\|N \nabla f_{I_0}(Y_0) - \nabla F(Y_0)\|_2^2\right]
\end{aligned}$$

Define:

$$\begin{aligned}
\Delta &\coloneqq \mathbb{E}\left[\|Y_0 - X_0\|_2^2\right] \\
c &\coloneqq 2\alpha^2 \mathbb{E}\left[\|N \nabla f_{I_0}(Y_0) - \nabla F(Y_0)\|_2^2\right] \\
b &\coloneqq 2\alpha^{3/2} L \sqrt{d}
\end{aligned}$$

The initial bound can then be rewritten as:

$$\Delta^{1/2} \leq \left((1 - \alpha\mu)\Delta + c\right)^{1/2} + b$$

By concavity of $\sqrt{x}$ we have:

$$\sqrt{x} - \sqrt{y} \geq \frac{1}{2\sqrt{x}}(x - y)$$

Applying this inequality to the one above we get:

$$\frac{1}{2\Delta^{1/2}} \left(\alpha\mu\Delta - c\right) - b \leq 0$$

which gives after multiplying by $\Delta^{1/2}$

$$\frac{\alpha\mu}{2}\Delta - b\Delta^{1/2} - \frac{c}{2} \leq 0$$

implying:

$$\Delta^{1/2} \leq \frac{b + \sqrt{b^2 + \alpha\mu c}}{\alpha\mu} \leq \frac{2b}{\alpha\mu} + \sqrt{\frac{c}{\alpha\mu}}$$

Recalling that $\Delta^{1/2} = W_2(\rho^*_{SGLD(\alpha)}, \rho^*)$ and replacing the variables with their values we get:

$$W_2(\rho^*_{SGLD(\alpha)}, \rho^*) \leq 4\kappa(\alpha d)^{1/2} + \sqrt{\frac{\alpha\sigma^2}{\mu}}$$

□

Chapter 5

Variance reduced algorithms

The results of the previous chapter show that both SGD and SGLD suffer from worse convergence guarantees than their deterministic counterpart. In a low precision regime, they enjoy a linear rate of convergence, but in the high precision regime, the variance of the stochastic gradients causes a significant slowdown. In this section we will explore an algorithm that provably overcome this problem, recovering a linear convergence rate in both cases.

5.1 SAGA

The main idea of variance-reduced methods is to use control-variates to reduce the variance of the naive SGD estimator. In SAGA, every time we evaluate a gradient $\nabla f_{I_k}(x_k)$ we store it, so that at each iteration k we have a table $(g_k^i)_{i=1}^N$ of the last seen gradients, where g_k^i is the last seen gradient of f_i at iteration k. The SAGA sequence is then defined by:

$$\begin{aligned} x_{k+1} &= x_k - \alpha \hat{g}_k \\ \hat{g}_k &= N\left(\nabla f_{I_k}(x_k) - g_k^{I_k}\right) + \sum_{i=1}^N g_k^i \end{aligned} \tag{5.1}$$

where as usual I_k is a uniform random variable on $[N]$, and we will assume that $g_0^i = 0$ for all $i \in [N]$. It is not hard to show that the SAGA estimator

is unbiased:

$$\begin{aligned}\mathbb{E}\left[\hat{g}_k\right] &= \mathbb{E}\left[N\nabla f_{I_k}(x_k)\right] - \mathbb{E}\left[Ng_k^{I_k}\right] + \sum_{i=1}^{N} g_k^i \\ &= \nabla F(x_k) - \sum_{i=1}^{N} g_k^i + \sum_{i=1}^{N} g_k^i \\ &= \nabla F(x_k)\end{aligned}$$

More interestingly, its variance is given by:

$$\begin{aligned}\mathbb{E}\left[\|\hat{g}_k - \nabla F(x_k)\|_2^2\right] &= \mathbb{E}\left[\left\|N\left(\nabla f_{I_k}(x_k) - g_k^{I_k}\right) - \left(\nabla F(x_k) - \sum_{i=1}^{N} g_k^i\right)\right\|_2^2\right] \\ &= N\sum_{i=1}^{N}\left\|\nabla f_i(x_k) - g_k^i\right\|_2^2 - \left\|\nabla F(x_k) - \sum_{i=1}^{N} g_k^i\right\|_2^2\end{aligned}$$

which we expect to be smaller than the variance of the SGD estimator if the g_k^i are close to $\nabla f_i(x_k)$.

With this intuition in hand, we now show that SAGA converges linearly to the optimum.

Theorem 5.1. *Suppose that $F : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2, and has the finite-sum form given by equation (4.1). Further, assume the functions $(f_i)_{i=1}^N$ satisfy Assumptions 4.1 and 4.2. Then the elements of the SAGA sequence (5.1) with $\alpha \leq \frac{1}{8NL_{max}}$ satisfy:*

$$\mathbb{E}\left[T(x_k, (g_k^i)_{i=1}^N)\right] \leq (1-\gamma)^k T(x_0, (g_0^i)_{i=1}^N)$$

where:

$$\begin{aligned}T(x_k, (g_k^i)_{i=1}^N) &= cN\sum_{i=1}^{N}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + \|x_k - x^*\|_2^2 \\ c &= \frac{\alpha}{2L_{max}} \\ \gamma &= \min\left\{\frac{1}{2N}, \alpha\mu\right\}\end{aligned}$$

.

Proof. Let $k \in \mathbb{N}$. We start by examining

$$\mathbb{E}\left[T(x_{k+1}, (g_k^i)_{i=1}^N)\right]$$

where the expectation is over I_k and is conditional on $I_0, \ldots, I_{k-1}$. The first term expands as:

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

The second term expands as:

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

We decompose the last term as:

$$\begin{aligned}
\mathbb{E}\left[\|\hat{g}\|_2^2\right] &= \mathbb{E}\left[\left\|N(\nabla f_{I_k}(x_k) - g_k^i) + \sum_{i=1}^N g_k^i\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|N(\nabla f_{I_k}(x_k) - \nabla f_i(x^*)) + N(\nabla f_i(x_k) - g_k^i) + \sum_{i=1}^N g_k^i\right\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\|N(\nabla f_{I_k}(x_k) - \nabla f_i(x^*))\|_2^2\right] + 2\mathbb{E}\left[\left\|N(g_k^i - \nabla f_i(x^*)) - \sum_{i=1}^N g_k^i\right\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\|N(\nabla f_{I_k}(x_k) - \nabla f_i(x^*))\|_2^2\right] + 2\mathbb{E}\left[\left\|N(g_k^i - \nabla f_i(x^*)\right\|_2^2\right]
\end{aligned}$$

where the last inequality follows from:

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

Replacing we get:

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1}-x^*\|_2^2\right] \leq \|x_k-x^*\|_2^2 - 2\alpha\langle\nabla F(x_k), x_k-x^*\rangle + \\ 2\alpha^2\mathbb{E}\left[\|N(\nabla f_{I_k}(x_k)-\nabla f_i(x^*))\|_2^2\right] + \\ 2\alpha^2\mathbb{E}\left[\left\|N(g_k^{I_k}-\nabla f_i(x^*))\right\|_2^2\right]\end{aligned}$$

Using strong-convexity to bound the inner product term, and using Lemma 4.1 to bound the term $\mathbb{E}\left[\|N(\nabla f_{I_k}(x_k)-\nabla f_i(x^*))\|_2^2\right]$, we get the global bound:

$$\begin{aligned}&\mathbb{E}\left[T(x_{k+1},(g_{k+1}^i)_{i=1}^N)\right] \leq (1-\alpha\mu)\|x_k-x^*\|_2^2 \\ &+ c\left(1-\frac{1}{N}+\frac{2\alpha^2}{c}\right)N\sum_{i=1}^N\left\|g_k^i-\nabla f_i(x^*)\right\|_2^2 \\ &+ 2\alpha NL_{max}\left(2\alpha-\frac{1}{NL_{max}}+\frac{c}{\alpha N}\right)(F(x_k)-F(x^*))\end{aligned}$$

Taking $c \coloneqq \frac{\alpha}{2L_{max}}$ and $\alpha \leq \frac{1}{8NL_{max}}$ we get:

$$\mathbb{E}\left[T(x_{k+1},(g_{k+1}^i)_{i=1}^N)\right] \leq \left(1-\min\{\frac{1}{2N},\alpha\mu\}\right)T(x_k,(g_k^i)_{i=1}^N)$$

Taking successive expectations and applying the inequality recursively we get the result. □

5.2 SAGA-LD

We now explore the use of the SAGA estimator in sampling. The SAGA Langevin dynamics (SAGA-LD) process is given by:

$$\begin{aligned}X_{k+1} &= X_k - \alpha\hat{g} + \sqrt{2\alpha}\xi_k \\ \hat{g}_k &= N\left(\nabla f_{I_k}(X_k) - g_k^{I_k}\right) + \sum_{i=1}^N g_k^i\end{aligned} \tag{5.2}$$

Where $\xi_k \sim \mathcal{N}(0, I_d)$ are independent normal random variables.

As usual, we start by showing the existence and linear convergence to a unique stationary probability measure $\rho^*_{SAGA-LD(\alpha)}$ of the process (5.2).

Theorem 5.2. *Suppose that $F : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2, and has the finite-sum form given by equation (4.1). Further, assume the functions $(f_i)_{i=1}^N$ satisfy Assumptions 4.1 and 4.2. Then, starting from $X_0 \sim \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$, the marginals $(\rho_k)_{k=0}^T$ of SAGA-LD (4.4) with $\alpha \leq \frac{1}{8NL_{max}}$ satisfy:*

$$W_2^2(\rho_k, \rho^*_{SAGA-LD(\alpha)}) \leq (1-\gamma)^k W_2^2(\rho_0, \rho^*_{SAGA-LD(\alpha)})$$

where:

$$\gamma \coloneqq \min\{\frac{1}{2N}, \frac{\alpha\mu}{2}\}$$

and for some $\rho^*_{SAGA-LD(\alpha)} \in \mathcal{P}_2(\mathbb{R}^d)$. Furthermore, $\rho^*_{SAGA-LD(\alpha)}$ is the unique stationary distribution of SAGA-LD (5.2).

Proof. Let $X_0 \sim \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $Y_0 \sim \rho_0' \in \mathcal{P}_2(\mathbb{R}^d)$, and evolve each according to SAGA-LD:

$$X_{k+1} = X_k - \alpha\hat{g}_k + \sqrt{2\alpha}\,\xi_k$$
$$Y_{k+1} = Y_k - \alpha\hat{h}_k + \sqrt{2\alpha}\,\xi_k$$

where $\hat{g}_k$ is as in (5.2), and:

$$\hat{h}_k = N(\nabla f_{I_k}(Y_k) - h_k^{I_k}) + \sum_{i=1}^N h_k^i$$

where $(h_k^i)_{i=1}^N$ is defined similarly to $(g_k^i)_{i=1}^N$. We consider the Lyapunov function:

$$T(X_k, (g_k^i)_{i=1}^N, Y_k, (h_k^i)_{i=1}^N) = cN\sum_{i=1}^N \left\|g_k^i - h_k^i\right\|_2^2 + \|Y_k - X_k\|_2^2$$

Following the same steps as those in the proof of Theorem 5.1 and replacing $\nabla f_i(x^*)$ with h_k^i we get:

$$\mathbb{E}\left[T(X_k, (g_k^i)_{i=1}^N, Y_k, (h_k^i)_{i=1}^N)\right] \leq (1-\gamma)^k\, T(x_0, (g_0^i)_{i=1}^N, y_0, (h_0^i)_{i=1}^N)$$

Replacing with the definition of T we get:

$$\mathbb{E}\left[\|Y_k - X_k\|_2^2\right] \leq (1-\gamma)^k \left[cN\sum_{i=1}^N \left\|g_0^i - h_0^i\right\|_2^2 + \mathbb{E}\left[\|Y_0 - X_0\|_2^2\right]\right]$$

Taking $g_0^i = h_0^i = 0$ and choosing X_0, Y_0 to be optimally coupled, we get:

$$W_2^2(\rho_k, \rho_k') \leq (1-\gamma)^k W_2^2(\rho_0, \rho_0')$$

By the Banach fixed point theorem we have the existence of $\rho^*_{SAGA-LD(\alpha)}$, and by its stationarity and the above inequality we get the linear rate of convergence. □

Finally, we bound the deviation of $\rho^*_{SAGA-LD(\alpha)}$ from ρ^*, showing that SAGA-LD is able to recover the precision of the unadjusted Langevin algorithm. As oppose to the proofs of Theorems 3.3 and 4.3, we are unable to prove an explicit bound on the Wasserstein distance between $\rho^*_{SGLD(\alpha)}$ and ρ^*. Instead, we directly prove a bound on the Wasserstein distance between ρ^* and the marginals ρ_k of the SAGA-LD process (5.2).

Theorem 5.3. *Let $(\rho_k)_{k=0}^{\infty}$ be the marginals of the process (5.2) with $\alpha \leq \frac{1}{8NL_{max}}$. Then:*

$$\begin{aligned} W_2^2(\rho_k^*, \rho^*) \leq (1-\gamma)^k \mathbb{E}\left[T(X_0, (g_0^i)_{i=1}^N, Y_0, (h_0^i)_{i=1}^N)\right] \\ + \left(32\alpha^3 L N^2 L_{max} d + 6\alpha^2 \kappa L d\right)/\gamma \end{aligned}$$

where $\kappa = \frac{L}{\mu}$ and:

$$\begin{aligned} &T(X_0, (g_0^i)_{i=1}^N, Y_0, (h_0^i)_{i=1}^N) = cN \sum_{i=1}^N \left\|g_0^i - h_0^i\right\|_2^2 + \|X_0 - Y_0\|_2^2 \\ &X_0 \sim \rho_0, Y_0 \sim \rho^*, h_0^i = \nabla f_i(Y_0) \\ &c = \frac{\alpha}{2L_{max}} \\ &\gamma = \min\left\{\frac{1}{2N}, \frac{\alpha\mu}{2}\right\} \end{aligned}$$

.

Proof. Let $X_0 \sim \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$, $Y_0 \sim \rho^*$ and assume they are optimally coupled. Define:

$$\begin{aligned} X_{k+1} &= X_k - \alpha \hat{g}_k + \sqrt{2\alpha}\,\xi_k \\ dY_t &= -\nabla F(Y_t) + \sqrt{2} dW_t \end{aligned}$$

Further, define the sequence:

$$Y_{k+1} = Y_k - \int_{k\alpha}^{(k+1)\alpha} \nabla F(Y_t)\, dt + \sqrt{2\alpha}\xi_k$$

and let $(h_k^i)_{i=1}^N$ be defined similarly to $(g_k^i)_{i=1}^N$ for the process $(Y_k)_{k=0}^\infty$, and assume that both $(h_k^i)_{i=1}^N$ and $(g_k^i)_{i=1}^N$ are updated at the same index I_k at iteration k. Furthermore assume the initialization $h_0^i = \nabla f_i(X_0)$.

We have:

$$\begin{aligned}
\|Y_{k+1} - X_{k+1}\|_2^2 &= \left\| Y_k - X_k - \int_{k\alpha}^{(k+1)\alpha} \nabla F(Y_t)\, dt + \alpha \hat{g}_k \right\|_2^2 \\
&\leq (1+\beta) \left\| Y_k - X_k - \alpha(\nabla F(Y_k) - \hat{g}_k) \right\|_2^2 \\
&\quad (1+\beta^{-1}) \left\| \int_{k\alpha}^{(k+1)\alpha} \nabla F(Y_t) - \nabla F(Y_0)\, dt \right\|_2^2
\end{aligned}$$

Taking expectation over I_k and conditioning on all other sources of randomness we get for the first term:

$$\begin{aligned}
&\mathbb{E}\left[\|Y_k - X_k - \alpha(\nabla F(Y_k) - \hat{g}_k)\|_2^2 \right] \\
&= \|Y_k - X_k\|_2^2 - \langle \nabla F(Y_k) - \nabla F(X_k), Y_k - X_k \rangle + \alpha^2 \mathbb{E}\left[\|\nabla F(Y_k) - \hat{g}_k\|_2^2 \right]
\end{aligned}$$

We decompose the last term as:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla F(Y_k) - \hat{g}_k\|_2^2 \right] &= \mathbb{E}\left[\left\| \hat{g}_k - \hat{h}_k + \hat{h}_k - \nabla F(Y_k) \right\|_2^2 \right] \\
&\leq 2\mathbb{E}\left[\left\| \hat{g}_k - \hat{h}_k \right\|_2^2 \right] + 2\mathbb{E}\left[\left\| \hat{h}_k - \nabla F(Y_k) \right\|_2^2 \right] \\
&= 2\mathbb{E}\left[\left\| \hat{g}_k - \hat{h}_k \right\|_2^2 \right] \\
&\quad + 2\mathbb{E}\left[\left\| N\nabla f_{I_k}(Y_k) - N h_k^{I_k} - (\nabla F(Y_k) - \sum_{i=1}^N h_k^{I_k}) \right\|_2^2 \right] \\
&\leq 2\mathbb{E}\left[\left\| \hat{g}_k - \hat{h}_k \right\|_2^2 \right] + 2\mathbb{E}\left[\left\| N\nabla f_{I_k}(Y_k) - N h_k^{I_k} \right\|_2^2 \right]
\end{aligned}$$

where $\hat{h}_k$ is defined as in the proof of Theorem 5.2, and the last line follows from:

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

replacing we get:

$$\begin{aligned}
&\mathbb{E}\left[\|Y_k - X_k - \alpha(\nabla F(Y_k) - \hat{g}_k)\|_2^2\right] \\
&= \|Y_k - X_k\|_2^2 - \langle \nabla F(Y_k) - \nabla F(X_k), Y_k - X_k \rangle + 2\alpha^2 \mathbb{E}\left[\left\|\hat{g}_k - \hat{h}_k\right\|_2^2\right] \\
&\quad + 2\alpha^2 \mathbb{E}\left[\left\|N\nabla f_{I_k}(Y_k) - N h_k^{I_k}\right\|_2^2\right]
\end{aligned}$$

Now consider the same Lyapunov function we studied in the proof of Theorem 5.2:

$$T(X_k, (g_k^i)_{i=1}^N, Y_k, (h_k^i)_{i=1}^N) = cN \sum_{i=1}^N \left\|g_k^i - h_k^i\right\|_2^2 + \|Y_k - X_k\|_2^2$$

The first term expands in the same way as it does in the proof of Theorem 5.2. From the inequalities we derived above, the second term expands as:

$$\begin{aligned}
&\mathbb{E}\left[\|Y_{k+1} - X_{k+1}\|_2^2\right] \\
&\leq (1+\beta)\left[\|Y_k - X_k\|_2^2 + \langle \nabla F(Y_k) - \nabla F(X_k), Y_k - X_k \rangle + 2\alpha^2 \mathbb{E}\left[\left\|\hat{g}_k - \hat{h}_k\right\|_2^2\right]\right] \\
&\quad + 2\alpha^2(1+\beta)\mathbb{E}\left[\left\|N\nabla f_{I_k}(Y_k) - N h_k^{I_k}\right\|_2^2\right] \\
&\quad + (1+\beta^{-1})\left\|\int_{k\alpha}^{(k+1)\alpha} \nabla F(Y_t) - \nabla F(Y_k)\, dt\right\|_2^2
\end{aligned}$$

The first term of this expansion is the same as the expansion in Theorem 5.2, and we can bound the third term as usual:

$$\mathbb{E}\left[\left\|\int_{k\alpha}^{(k+1)\alpha} \nabla F(Y_t) - \nabla F(Y_k)\, dt\right\|_2^2\right] \leq 2\alpha^3 L^2 d$$

Taking $\beta = \frac{\alpha\mu}{2}$ we get by using the contraction of the Lyapunov function:

$$\begin{aligned}
&\mathbb{E}\left[T(X_{k+1}, (g_{k+1}^i)_{i=1}^N, Y_{k+1}, (h_{k+1}^i)_{i=1}^N)\right] \\
&\leq (1 - \min\{\frac{\alpha\mu}{2}, \frac{1}{2N}\}) T(X_k, (g_k^i)_{i=1}^N, Y_k, (h_k^i)_{i=1}^N) + \Delta_k + \frac{6}{\alpha\mu}\alpha^3 L^2 d
\end{aligned}$$

where:

$$\Delta_k = 4\alpha^2 N \sum_{i=1}^{N} \left\|\nabla f_i(Y_k) - h_k^i\right\|_2^2$$

We now bound the terms of the sum in Δ_k. We have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla f_i(Y_k) - h_k^i\right\|_2^2\right] = \sum_{j=0}^{k-1} \mathbb{E}\left[\|\nabla f_i(Y_k) - \nabla f_i(Y_j)\|_2^2\right] \mathbb{P}(h_k^i = \nabla f_i(Y_j)) \\
&= \sum_{j=0}^{k-1} \mathbb{E}\left[\|\nabla f_i(Y_k) - \nabla f_i(Y_j)\|_2^2\right] \left(1 - \frac{1}{N}\right)^{k-j-1} \frac{1}{N} \\
&\leq \frac{L_{max}^2}{N} \sum_{j=0}^{k-1} \mathbb{E}\left[\|Y_k - Y_j\|_2^2\right] \left(1 - \frac{1}{N}\right)^{k-j-1} \\
&= \frac{L_{max}^2}{N} \sum_{j=0}^{k-1} \mathbb{E}\left[\left\|\int_{j\alpha}^{k\alpha} \nabla F(Y_t)\, dt + \sqrt{2}(W_{k\alpha} - W_{j\alpha})\right\|_2^2\right] \left(1 - \frac{1}{N}\right)^{k-j-1} \\
&\leq \frac{L_{max}^2}{N} \sum_{j=0}^{k-1} \left[2\alpha^2(k-j)^2 \mathbb{E}\left[\|\nabla F(Y)\|_2^2\right] + 4\alpha(k-j)d\right] \left(1 - \frac{1}{N}\right)^{k-j-1} \\
&= \frac{L_{max}^2}{N} \left(2\alpha^2 \mathbb{E}\left[\|\nabla F(Y)\|_2^2\right] \sum_{j=1}^{k} j^2 (1 - \frac{1}{N})^{j-1}\right) + \left(4\alpha d \sum_{j=1}^{k} j(1 - \frac{1}{N})^{j-1}\right) \\
&\leq 4\alpha^2 N^2 L_{max}^2 \mathbb{E}\left[\|\nabla F(Y)\|_2^2\right] + 4\alpha d N L_{max}^2 \\
&\leq 4\alpha^2 d N^2 L_{max}^2 L + 4\alpha d N L_{max}^2 \\
&\leq 4\alpha d N L_{max}\left(L + L_{max}\right) \\
&\leq 8\alpha d L N L_{max}
\end{aligned}$$

Replacing in the bound on the Lyapunov function we get:

$$
\begin{aligned}
&\mathbb{E}\left[T(X_{k+1},(g_{k+1}^i)_{i=1}^N,Y_{k+1},(h_{k+1}^i)_{i=1}^N)\right]\\
&\leq (1-\min\{\frac{\alpha\mu}{2},\frac{1}{2N}\})^k\mathbb{E}\left[T(X_0,(g_0^i)_{i=1}^N,Y_0,(h_0^i)_{i=1}^N)\right]\\
&\quad+32\alpha^3LN^2L_{max}d\sum_{j=1}^{\infty}(1-\min\{\frac{\alpha\mu}{2},\frac{1}{2N}\})^j\\
&\quad+\frac{6}{\alpha\mu}\alpha^3L^2d\sum_{j=1}^{\infty}(1-\min\{\frac{\alpha\mu}{2},\frac{1}{2N}\})^j\\
&\leq (1-\min\{\frac{\alpha\mu}{2},\frac{1}{2N}\})^k\mathbb{E}\left[T(X_0,(g_0^i)_{i=1}^N,Y_0,(h_0^i)_{i=1}^N)\right]\\
&\quad+\left(32\alpha^3LN^2L_{max}d+6\alpha^2\kappa Ld\right)\max\{\frac{2}{\alpha\mu},2N\}
\end{aligned}
$$

□

Bibliography

[1] L. Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows: In metric spaces and in the space of probability measures. 2005.

[2] N. Brosse, Alain Durmus, and E. Moulines. The promises and pitfalls of stochastic gradient langevin dynamics. *ArXiv*, abs/1811.10072, 2018.

[3] Niladri S. Chatterji, Nicolas Flammarion, Yian Ma, Y. Ma, P. Bartlett, and Michael I. Jordan. On the theory of variance reduction for stochastic gradient monte carlo. *ArXiv*, abs/1802.05431, 2018.

[4] A. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 79:651–676, 2014.

[5] A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *COLT*, 2017.

[6] A. Dalalyan and ssAvetik G. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *ArXiv*, abs/1710.00095, 2017.

[7] Aaron Defazio, Francis R. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *ArXiv*, abs/1407.0202, 2014.

[8] Kumar Avinava Dubey, S. Reddi, Sinead Williamson, B. Póczos, Alex Smola, and E. Xing. Variance reduction in stochastic gradient langevin dynamics. *Advances in neural information processing systems*, 29:1154–1162, 2016.

[9] Alain Durmus, S. Majewski, and B. Miasojedow. Analysis of langevin monte carlo via convex optimization. *J. Mach. Learn. Res.*, 20:73:1–73:46, 2019.

[10] Alain Durmus and É. Moulines. Sampling from a strongly log-concave distribution with the unadjusted langevin algorithm. *arXiv: Statistics Theory*, 2016.

[11] L. Evans. An introduction to stochastic differential equations. 2014.

[12] R. Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

[13] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker-planck equation. *Siam Journal on Applied Mathematics*, 1996.

[14] Wenlong Mou, Nicolas Flammarion, M. Wainwright, and P. Bartlett. Improved bounds for discretization of langevin diffusions: Near-optimal rates without convexity. *arXiv: Probability*, 2019.

[15] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.

[16] G. Pavliotis. Stochastic processes and applications: Diffusion processes, the fokker-planck and langevin equations. 2014.

[17] H. Robbins. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 2007.

[18] M. Schmidt, N. Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

[19] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[20] S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *NeurIPS*, 2019.

[21] C. Villani. Topics in optimal transportation. 2003.

[22] C. Villani. Optimal transport: Old and new. 2008.

[23] M. Welling and Y. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.