STOCHASTIC APPROXIMATION

on Riemannian Manifolds and the Space of Measures

Mohammad Reza Karimi Jaghargh

Diss. ETH No. 30588

DISS. ETH NO. 30588

Stochastic Approximation

on Riemannian Manifolds and the Space of Measures

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by

MOHAMMAD REZA KARIMI JAGHARGH

MSc ETH in Computer Science

born on 31.12.1992

accepted on the recommendation of

Prof. Dr. Andreas Krause	(supervisor)
Prof. Dr. Peter Bartlett	(co-examiner)
Prof. Dr. Hamed Hassani	(co-examiner)
Prof. Dr. Panayotis Mertikopoulos	(co-examiner)

2024

Cover image: Heptagonal tiling of the Poincaré half-plane model of the hyperbolic space, by Claudio Rocchini, licensed under CC-BY 3.0.

To Florina, Navid, and my parents.

ABSTRACT

Stochastic approximation methods are a class of iterative algorithms that play an essential role in applications involving noisy and incomplete observations. Rooted in the seminal works of Robbins and Monro (1951) and Kiefer and Wolfowitz (1952), this class of iterative processes drive a system towards a specified objective despite noise and bias. Stochastic approximation methods have become increasingly important in fields like statistics and machine learning due to their resilience to noise and low computational costs. They are especially useful in training neural networks and adaptive learning systems. The rise of big data has further heightened the significance of stochastic approximation, as it necessitates efficient and scalable algorithms for real-time decision-making.

This thesis embarks on a renewed exploration of stochastic approximation through a contemporary lens, focusing on its dynamics and long-term behavior in non-Euclidean spaces. Inspired by the dynamical systems approach introduced by Benaïm and Hirsch in the 1990s, this work tackles crucial questions surrounding the convergence of the iterates of a stochastic approximation algorithm, the characteristics of its limits, and the desirability of these limits. The thesis aims to provide profound theoretical insights into the stability and convergence of these algorithms amidst the challenges posed by their non-Euclidean structure and real-world conditions marked by noise and incomplete information.

This work examines stochastic approximation within three unconventional contexts where traditional methods are inadequate. Firstly, we delve into stochastic approximation on Riemannian manifolds, analyzing problems related to Riemannian optimization and strategic games. This non-Euclidean setting introduces complexities requiring novel approaches for proving convergence.

Secondly, the thesis investigates the discretization of stochastic differential equations by lifting the problem to the Wasserstein space. This is particularly relevant to sampling algorithms based on the Langevin diffusion, which are essential in Bayesian learning and generative modeling. We demonstrate how our findings can enhance the reliability and effectiveness of these sampling algorithms, ultimately supporting more robust Bayesian inference and generative processes.

Lastly, the thesis explores algorithms lacking a step-size parameter, focusing on the Sinkhorn algorithm for solving the entropic optimal transport and Schrödinger bridge problems. By proposing a variant with a step-size parameter and examining its continuous-time limit, we offer a stochastic approximation analysis that ensures convergence of this variant of the Sinkhorn algorithm even under noise and bias.

This thesis represents a harmonious blend of classical stochastic approximation and its extended applications in non-Euclidean setups. By integrating different areas of mathematics, we offer a thorough analysis that enriches the theoretical foundation of stochastic approximation and guarantees the robustness of these algorithms across a diverse array of scenarios.

ZUSAMMENFASSUNG

Stochastische Approximationsmethoden sind eine Klasse von iterativen Algorithmen, die bei Anwendungen mit verrauschten und unvollständigen Beobachtungen eine wesentliche Rolle spielen. Diese Klasse von iterativen Verfahren, die auf die bahnbrechenden Arbeiten von Robbins und Monro (1951) und Kiefer und Wolfowitz (1952) zurückgehen, steuern ein System trotz Rauschen und Verzerrungen auf ein bestimmtes Ziel zu. Stochastische Approximationsverfahren haben in Bereichen wie der Statistik und dem maschinellen Lernen zunehmend an Bedeutung gewonnen, da sie robust gegenüber Rauschen sind und geringe Rechenkosten verursachen. Sie sind besonders nützlich für das Training neuronaler Netze und adaptiver Lernsysteme. Mit dem Aufkommen von Big Data hat die Bedeutung der stochastischen Approximation weiter zugenommen, da effiziente und skalierbare Algorithmen erforderlich sind, um Entscheidungen in Echtzeit zu treffen.

In dieser Dissertation wird die stochastische Approximation aus einem modernen Blickwinkel erneut untersucht, wobei der Schwerpunkt auf ihrer Dynamik und ihrem langfristigen Verhalten in nichteuklidischen Räumen liegt. Inspiriert vom Ansatz der dynamischen Systeme, der von Benaïm und Hirsch in den 1990er Jahren eingeführt wurde, befasst sich diese Arbeit mit entscheidenden Fragen rund um die Konvergenz der Iterate eines stochastischen Approximationsalgorithmus, die Eigenschaften seiner Grenzen und die Erwünschtheit dieser Grenzen. Die Dissertation zielt darauf ab, tiefgreifende theoretische Einblicke in die Stabilität und Konvergenz dieser Algorithmen zu geben, inmitten der Herausforderungen, die sich aus ihrer nichteuklidischen Struktur und den realen Bedingungen ergeben, die durch Rauschen und unvollständige Informationen gekennzeichnet sind.

In dieser Arbeit wird die stochastische Approximation in drei unkonventionellen Kontexten untersucht, in denen konventionelle Methoden unzureichend sind. Zunächst befassen wir uns mit der stochastischen Approximation auf riemannschen Mannigfaltigkeiten und analysieren Probleme im Zusammenhang mit riemannscher Optimierung und strategischen Spielen. Diese nichteuklidische Umgebung führt zu komplexen Problemen, die neue Ansätze zum Nachweis der Konvergenz erfordern.

Zweitens untersucht die Dissertation die Diskretisierung stochastischer Differenzialgleichungen, indem das Problem auf den Wasserstein-Raum übertragen wird. Dies ist besonders relevant für Sampling-Algorithmen, die auf der Langevin-Diffusion basieren und die für das Bayes'sche Lernen und die generative Modellierung unerlässlich sind. Wir zeigen, wie unsere Erkenntnisse die Zuverlässigkeit und Effektivität dieser Sampling-Algorithmen verbessern können, was letztlich eine robustere Bayes'sche Inferenz und generative Prozesse unterstützt.

Schliesslich werden in der Dissertation Algorithmen untersucht, denen ein Schrittweitenparameter fehlt, wobei der Schwerpunkt auf dem Sinkhorn-Algorithmus zur Lösung des entropischen optimalen Transports und des Schrödinger-Brückenproblems liegt. Indem wir eine Variante mit einem Schrittweitenparameter vorschlagen und ihre zeitkontinuierliche Grenze untersuchen, bieten wir eine stochastische Approximationsanalyse an, die die Konvergenz dieser Variante des Sinkhorn-Algorithmus auch bei Rauschen und Verzerrungen gewährleistet.

Diese Dissertation stellt eine harmonische Mischung aus klassischer stochastischer Approximation und ihren erweiterten Anwendungen in nichteuklidischen Konstellationen dar. Durch die Integration verschiedener Bereiche der Mathematik bieten wir eine gründliche Analyse, die das theoretische Fundament der stochastischen Approximation bereichert und die Robustheit dieser Algorithmen in einer Vielzahl von Szenarien garantiert.

ACKNOWLEDGEMENTS

This thesis would not exist without the immense support and inspiration of many incredible people I have had the privilege to work with, learn from, and be supported by during my doctoral journey. First and foremost, a big thank-you goes to Andreas Krause, who has been a remarkable supervisor. Your approach, giving me the freedom to choose my research direction while providing invaluable insights, has been empowering. The way you helped me dive deeper into our field truly shaped my academic journey. Working under your guidance has been an honor and incredibly enriching.

Ya-Ping Hsieh has also been a tremendous influence, bringing a flood of new ideas and perspectives to my research. Our many discussions, especially those memorable sessions at Café Einstein in Bern, are moments I will cherish. They significantly contributed to the evolution of my research plan and made me learn a lot of interesting mathematics. Indeed, all the publications that form the backbone of this thesis are done in close collaboration with Ya-Ping. Thank you for all this and for giving me the best coffee ever.

In my early days in the LAS group as a master's student, I was fortunate to work with Hamed Hassani. Your support as I grew into a researcher, allowing me to learn by doing, has been invaluable. Our daily discussions shaped my vision for an academic career. I wish we could prove the "airy theorem" together someday. I also had the chance to get to know Kfir Levy. Our discussions on online learning were always enlightening. Your boundless energy, openness, and positivity have always inspired me.

Working with Panayotis Mertikopoulos has been an absolute delight. Our nice and fruitful collaboration, along with the time and guidance you provided when I needed it, was pivotal. I sincerely appreciate every bit of support and wisdom you shared. I would also like to thank Slava Borovitskiy, whom I learned so much about math. Thanks for bringing back the spark of math-olympiad-style problems and reigniting my passion. The Riemannian chapter of this thesis is better only because of your careful reading.

The journey would have never been the same without the support of my beloved LAS group friends. You all made the group a vibrant and supportive

place. I thank my officemates Anastasia, Mojmír, and Charlotte, whose presence was deeply appreciated. Thank you for making our office greener, funnier, and happier. I also thank Andisheh, Bhavy, Parnian, Lenart, and Vignesh for the many engaging chats and the positive vibes you brought into the group. A special thanks go to Rita, who has always been there for all sorts of problems and questions I had.

I wish to express my deepest gratitude to Florina. This thesis would not have been possible without you. Your unwavering support, energy, and love have made my life happier and fuller. Thank you for always being my cheerleader and constant anchor. You helped me grow into the person I am today, and I will always be grateful for everything you have done.

To my own family, your nurturing of my curiosity and passion for mathematics from the very beginning has been the foundation of my journey. Despite the distance, your support has always felt close. Thank you for all you have done for me since day zero.

Lastly, to Will Merry. Although I did not learn differential geometry in a conventional way, the lessons from your “Communication in Mathematics” course have profoundly improved the readability of this thesis. I only wish I could have learned more from you. May you rest in peace.

CONTENTS

CHAPTER ONE

INTRODUCTION

As with everything else, so with a mathematical theory:
beauty can be perceived, but not explained.
— ARTHUR CAYLEY

Stochastic approximation is an essential technique in the development and analysis of iterative algorithms. Established in the early 1950s by the pioneering studies of Robbins and Monro (1951) and Kiefer and Wolfowitz (1952), this methodology lays a fundamental framework for handling applications involving noisy data and incomplete observations. A core concept in stochastic approximation in its simplest form is the difference equation

$$\theta_{n+1} - \theta_n = a_n Y_n, \tag{1.1}$$

where θ_n represents the parameters of a system, Y_n is a function of the noise-corrupted observation at θ_n, and a_n is a diminishing step-size approaching zero as $n \to \infty$. The main idea is to adjust the parameters iteratively to achieve some desired goal asymptotically. This thesis investigates the qualitative and asymptotic properties of such recursive algorithms in the diverse forms they arise in applications.

The field of stochastic approximation started with the foundational work of Robbins and Monro [RM51] on the problem of finding the root of an unknown function, given noisy observations at arbitrary argument values. A classic example is to determine the correct dosage of a drug for some disease. Letting θ be the drug dosage level and $F(\theta)$ be the probability of success at dosage level θ, the experimenter's goal is to identify the correct dosage level at which the probability of success equals a specified value, say α. However, the experimenter can only observe F through experimental outcomes that are either successes or failures, rendering analytical solutions infeasible.

The first idea that comes to mind might be to perform multiple experiments at a constant level θ and estimate $F(\theta)$ via averaging. However, many of these observations occur at dosage levels far from the optimum, leading to inefficient use of resources. Robbins and Monro propose a more efficient approach. Here is an excerpt from their article, with a slight change of notation, describing their method:

> We give a method for making successive experiments at levels $\theta_1, \theta_2, \ldots$ in such a way that θ_n will tend to the optimum in probability ... Let $\{a_n\}$ be a fixed sequence of positive constants such that $\sum_{n=1}^{\infty} a_n^2 < \infty$. We define a (non-stationary) Markov chain $\{\theta_n\}$ by taking θ_1 to be an arbitrary constant and defining
>
> $$\theta_{n+1} - \theta_n = a_n(\alpha - Y_n), \tag{$*$}$$
>
> where Y_n is a random variable such that $\mathbb{E}[Y_n \mid \theta_n] = F(\theta_n)$.

This simple approach ensures that the parameters θ_n move—on average—in the correct direction after each observation. Robbins and Monro's key insight was recognizing that if the step-sizes a_n approach zero appropriately as $n \to \infty$, an *implicit averaging* emerges, which mitigates the effects of noise over the long term. Indeed, a significant portion of the content of [RM51] consists of probabilistic arguments showing this.

While the original work of Robbins and Monro concerns finding the zeros of a scalar function, one can extend it to vector-valued functions defined on more complicated spaces. In this broader sense, the root-finding problem shows up in various scientific and practical applications. In optimization, for example, finding the critical points of a function f corresponds to finding the roots of its gradient. By substituting F with the gradient of f and α with 0, the iteration ($*$) transforms into the renowned stochastic gradient descent algorithm, widely used in practice for training neural networks. In this scenario, Y_n is typically the gradient calculated based on a random batch of data that adheres to an underlying (but unknown) distribution. We will see later in this introduction that other important problems, such as approximate sampling from probability distributions and finding equilibria in games are other instances of the general root-finding problem.

Over the decades, the scope of stochastic approximation has expanded. In the 1970s and 1980s, the method was pivotal in signal processing and adaptive control, aiding in adaptive filtering and system identification [KC78; LS83; Hay86]. In robotics, these algorithms have enabled systems to learn and adapt in real time from uncertain or imprecise sensory inputs [Ber96; SB98]. Stochastic approximation has become increasingly important in recent years, especially in statistics and machine learning, due to the rise of big data: Stochastic optimization methods are crucial for training large-scale neural networks and developing real-time decision-making algorithms; online algorithms and reinforcement learning frameworks are now ubiquitous in adaptive learning systems, where fast and iterative updates are necessary. Many of today's advanced models, even though named differently, rely fundamentally on stochastic approximation techniques, highlighting its central position in the architecture of modern data-driven and intelligent systems.

Dynamics and Long-time Behavior

The simplicity of the update rule (1.1) might mask its potentially complex behavior. For instance, the Robbins–Monro algorithm ($*$) is guaranteed to converge to a root of the equation $F(\theta) = \alpha$ provided that F is monotone and the deviation of Y_n from $F(\theta_n)$ remains uniformly bounded. However, applying this algorithm outside these ideal conditions may result in the algorithm failing to find a root. The complexity increases further when one considers a more intricate algorithm, such as various forms of stochastic gradient descent, or when modeling adaptive behavior. In real-world applications, these algorithms and models often do not function under ideal conditions and are prone to noisy and incomplete information in their updates. Therefore, understanding how these algorithms behave in non-ideal scenarios is crucial.

This necessitates a deeper exploration of the *dynamics* of stochastic approximation algorithms. Specifically, we need to examine the long-term behavior of the sequence $\{\theta_n\}$ defined by (1.1). Key questions include: Does this sequence converge? If so, what are the limits of this convergence, and are these limits desirable? How can we characterize these limits, should they exist? By analyzing these factors, we can gain significant insights into the stability and effectiveness of these algorithms across various practical applications. This understanding can lead to the development of more reliable algorithms, improving their performance in diverse scenarios.

Traditionally, dynamics of different instances of stochastic approximation algorithms were studied in isolation. Significant contributions to this field have been made by Benaïm and Hirsch [BH96], whose series of works has unified several results into one general framework, and has provided profound insights into the long-term behavior and convergence properties of such algorithms. One of the key pieces of this framework is the *ODE method* of Ljung [Lju77], which relates the discrete-time update (1.1) to a continuous-time one. The main idea of Ljung is the following: In many practical situations, we can rewrite $Y_n = F(\theta_n) + U_n$, where $F(\theta_n)$ is the main signal and U_n is "all the noise factored away." With this, we have

$$\frac{1}{a_n}(\theta_{n+1} - \theta_n) = F(\theta_n) + U_n.$$

Averaging U_n away and tending the step-size to zero, one recovers a continuous-time "mean dynamics" corresponding to the update (1.1), described by the ordinary differential equation (ODE)

$$\frac{d}{dt}\theta(t) = F(\theta(t)).$$

By associating a *deterministic* differential equation with the algorithm, the challenging task of convergence analysis of a stochastic algorithm is effectively reduced to a “stability analysis problem” of an ODE.

The methodology of Benaïm and Hirsch is a beautiful tandem of ideas from dynamical systems and probability theory, and consists of the following steps. First, the continuous-time mean dynamics of the algorithm is constructed, which is typically evident in root-finding problems, while in some other cases, one has to do some work to find it. Next, the limit sets of the mean dynamics are identified using a dynamical systems approach. Finally, one demonstrates that the original non-ideal algorithm converges to these identified limit sets by showing that the stochastic noise diminishes over time, resulting in the algorithm’s trajectory “shadowing” the orbits of the corresponding ODE. In this case, the algorithm’s trajectory is called an *asymptotic pseudo-trajectory* of the mean dynamics. Proving this property is usually where the probabilistic ideas lie. It is crucial, however, to determine which terminal limiting objects are favorable, ascertain the probability of converging to these favorable points, assess the likelihood of avoiding undesirable points (such as strict saddle points in optimization problems), and evaluate the rate at which the algorithm trends toward equilibrium.

Beyond Euclidean Spaces

Traditionally, stochastic approximation algorithms are analyzed in the context of systems with parameters residing in some Euclidean space. This classical approach, while powerful, proves inadequate for a broader range of applications where system states are more naturally represented on a Riemannian manifold or within the space of measures. In such contexts, even the standard recursion (1.1) does not necessarily make sense.

While it is feasible to mimic a Euclidean analysis for manifolds (for example, by isometrically embedding the manifold in a Euclidean space), this approach often falls short of a robust intrinsic analysis. An intrinsic perspective not only offers deeper insights but also addresses the inherent topological complexities that arise in the manifold and infinite-dimensional settings, which are beyond the reach of conventional Euclidean techniques. Building on the foundational work of Benaïm and Hirsch, this thesis explores these novel settings, offering a more comprehensive understanding of stochastic approximation in non-Euclidean spaces. Below, we describe the instances we consider in this thesis.

Riemannian manifolds

A Riemannian manifold is a smooth, curved space where one can measure distances of points and angles between tangent vectors. It generalizes concepts like curves and surfaces to higher dimensions, and facilitates studying geometric properties in an abstract way. Riemannian manifolds are used to model complex, non-linear structures in various scientific and engineering fields. Two such examples are the Stiefel manifold and the Hyperbolic space. The *Stiefel manifold* consists of all orthonormal k-frames in $\mathbb{R}^d$, and is used for optimization problems with orthonormality constraints such as Principal Component Analysis (PCA) and eigenvalue problems, as well as modeling orientation in robotics and aircraft control. The *Hyperbolic space* is a space with constant negative curvature, and is used to represent high-dimensional data in lower dimensions while preserving hierarchical relationships, making it suitable for modeling tree-structured data (e.g., evolutionary trees in Phylogenetics), robot localization, and navigation.

The generic root-finding problem in a Riemannian manifold $\mathcal{M}$ is to

$$\text{find } \theta \in \mathcal{M} \text{ so that } V(\theta) = 0. \tag{1.2}$$

Here, V is a smooth vector field, which assigns a tangent vector to every point on the manifold. An example of a vector field is the (Riemannian) gradient of a smooth, real-valued function on the manifold. Solving the root-finding problem (1.2) in this context turns into a Riemannian optimization problem. Additionally, the general form of this root-finding problem includes other complex scenarios like bi-level problems, saddle-point problems, dynamic programming, and equilibrium problems found in games and other practical applications.

Most methods for solving (1.2) are iterative and construct a sequence of successive approximations $\theta_1, \theta_2, \ldots$ of the root of V. Similar to the Robbins–Monro algorithm, θ_{n+1} is constructed based on a noise-corrupted evaluation Y_n of V at θ_n. However, we have to change the update rule and use the Riemannian exponential map (or more practically, a retraction)

$$\theta_{n+1} = \exp_{\theta_n}(a_n Y_n)$$

so that the iterates stay on the manifold.

To get a feeling of the challenges of the Riemannian setting, let us consider the simplest case where $V \equiv 0$, focusing solely on the effect of noise. First, let us recall what happens in the Euclidean case: Suppose $U_1, U_2, \ldots$ is a martingale difference sequence bounded in L^2, meaning $\mathbb{E}[U_n \,|\, U_1, \ldots, U_{n-1}] = 0$ for every n and $\sup_n \mathbb{E}\,\|U_n\|^2 < \infty$. If the step-sizes a_n are such that $\sum a_n = \infty$ and $\sum a_n^2 < \infty$, then the sum $\sum_{n=1}^{\infty} a_n U_n$ converges almost surely to some limit. In other words, defining the partial sums $M_n := \sum_{k=1}^{n} a_k U_k$, the martingale M

almost surely converges to a limit M_∞.

Now, consider the same setup in the hyperbolic space $\mathcal{H}^2$: Starting from an arbitrary point $p_0 \in \mathcal{H}^2$, at each step n, we take a random tangent vector U_n from the tangent plane T_{p_n} at p_n and move in that direction (along a geodesic) to reach p_{n+1}. Mathematically, $p_{n+1} = \exp_{p_n}(a_n U_n)$. The left part of Fig. 1.1 illustrates this procedure.

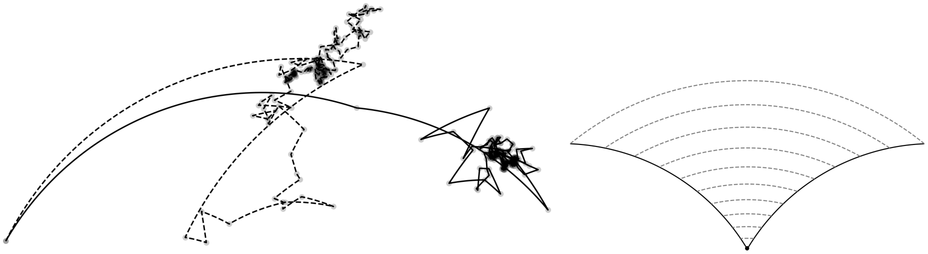


Figure 1.1. *Left:* Two sample trajectories of pure noise are depicted in the Poincaré half-plane model of the hyperbolic space. Consecutive points are connected via geodesics, which in this geometry, are arcs of circles centered on the x-axis, or vertical lines. As seen in the figure, the trajectories eventually converge to some (random) limit. *Right:* Exponential growth of the distance between two geodesics in the Hyperbolic space, emanating from the same point with the same speed. The dashed lines are the minimizing geodesics between corresponding points on the initial geodesics.

The question is whether the same result of Doob holds in this case—whether there exists a (random) point p_∞ such that $p_n \to p_\infty$ almost surely. As the distances grow exponentially in the Hyperbolic space due to its negative curvature, it is not at all obvious why such convergence should hold; see the right side of Fig. 1.1. The theory we develop later in Chapter 3 gives an affirmative answer to this question for a general class of Riemannian manifolds, including Hyperbolic spaces. Notice that our answer *cannot* use the linear structure of a Euclidean space, a property that is the defining property of a martingale, and essential for proving convergence results.

The Wasserstein space

Another important setting for studying stochastic approximation algorithms is discretization schemes for stochastic differential equations (SDEs). A prominent example in this category is Markov Chain Monte Carlo methods that are based on discretizing SDEs such as the Langevin diffusion

$$dX_t = V(X_t)\,dt + \sqrt{2}\,dW_t.$$

Here, V is usually the gradient of the log-density of some target probability distribution. These methods are widely used in practice to generate samples from complex, high-dimensional probability distributions, which is a critical task in Bayesian learning and statistics. Sampling is achieved by constructing a Markov chain whose stationary distribution matches (or approximates) the desired target distribution. Similar methods are used for generative modeling tasks via diffusion models.

Ordinary differential equations and stochastic differential equations exhibit distinct long-term behaviors. For instance, consider the function $f(x) = \frac{1}{2}x^2$, the differential equation $\dot{x} = -f'(x)$, and the stochastic differential equation $dX_t = -f'(X_t)\,dt + \sqrt{2}\,dW_t$. It is evident that from any initial point, the solution to the ODE converges to $x = 0$, the only zero of f'. In contrast, the solution to the SDE, influenced by Brownian motion, never settles and continues to fluctuate indefinitely.

At the macroscopic level, the behavior of an SDE tends to be deterministic. To illustrate this, consider a simple Brownian motion W_t representing the motion of a small particle in a suspension. By observing a large number of independent Brownian particles and focusing on their empirical distribution rather than individual positions, we can deduce that as the number of particles approaches infinity, the empirical distribution at each time t converges to a probability density $\varrho(t, x)$. This density function $\varrho(t, x)$ essentially satisfies the partial differential equation (PDE) known as the Fokker–Planck equation:

$$\frac{\partial \varrho(t,x)}{\partial t} = \frac{1}{2}\Delta_x \varrho(t,x),$$

where Δ_x represents the Laplacian operator with respect to the x variable. This PDE, which in this context is equivalent to the heat equation in physics, provides a deterministic framework to describe the macroscopic dynamics of Brownian motion or, more broadly, an SDE. Consequently, analyzing the macroscopic behavior of an SDE can be reduced to studying the corresponding Fokker–Planck equation. For example, the Fokker–Planck equation corresponding to the SDE $dX_t = -X_t\,dt + \sqrt{2}\,dW_t$ is:

$$\frac{\partial \varrho(t,x)}{\partial t} = \frac{\partial}{\partial x}(\varrho(t,x), x) + \frac{\partial^2 \varrho(t,x)}{\partial x^2}. \tag{1.3}$$

Further analysis reveals that, given mild regularity conditions on the initial

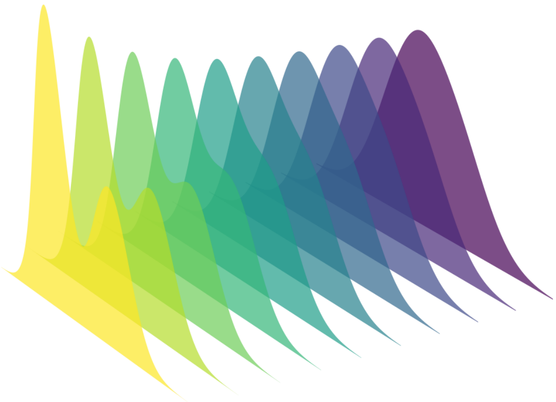


Figure 1.2. The solution of the Fokker–Planck equation (1.3) with the initial density $\varrho(0,\cdot)$ being the mixture of two Gaussians (yellow). As time progresses, the solution converges to the density of a standard Gaussian (purple).

density $\varrho(0,\cdot)$, the density $\varrho(t,\cdot)$ converges to the density of a standard Gaussian distribution as $t \to \infty$; see Fig. 1.2.

To interpret the Fokker–Planck equation for an SDE geometrically, one can view it as defining a curve $(\varrho_t)_{t\geq 0}$ within the space of probability measures. To rigorously discuss this curve, it is essential to define a corresponding metric space. In this context, it turns out that the Wasserstein metric, derived from optimal transport theory, is particularly suitable, as it turns the space of probability measures into a complete metric space, thus enabling a geometric analysis. The importance of the Wasserstein metric is underscored by the groundbreaking result of Jordan, Kinderlehrer, and Otto [JKO98]: The Fokker–Planck equation for the Langevin diffusion becomes the *gradient flow* of the relative entropy functional in the space of probability measures endowed with the Wasserstein metric. Otto [Ott01] further demonstrated that when the space of measures is equipped with the quadratic Wasserstein distance, it exhibits a Riemannian-like structure. This structure supports the definition of tangent spaces, geodesics, gradients, and curvature, and has become a powerful tool for formal computations within the Wasserstein space. In this manifold notation, the Fokker–Planck PDE becomes the following ODE in the Wasserstein space:

$$\dot{\varrho}_t = -\nabla_{W_2} H(\varrho_t \mid \mu),$$

where ∇_{W_2} is the gradient in the sense of Otto, H is the relative entropy, and μ is the stationary distribution of the Langevin diffusion.

Our primary take-away from this perspective is that discretization algorithms for SDEs can be conceptualized as stochastic approximation algorithms. In this framework, the algorithm's "iterates" correspond to their distributions within the Wasserstein space. The "mean dynamics" is also represented by the evolution curve derived from the Fokker–Planck equation associated with the SDE. Noise and biases in the discretization algorithm result in deviations from the mean dynamics, manifesting as a form of bias within the Wasserstein space. Consequently, examining the convergence properties in the Wasserstein space is equivalent to analyzing the macroscopic properties of the algorithm's iterates.

One can study other evolutionary equations using the same framework. A prominent example is *McKean–Vlasov processes*, which are often used for modeling the behavior of large systems of interacting particles (or agents), such as neurons in a wide neural network, or molecules following a kinetic equation. Investigating the dynamics of such evolutionary equations, their mean-field approximation, and their discretizations in the Wasserstein space is similar to that of a single SDE. We do not discuss this class of algorithms in this thesis and refer the reader to [KHK23b] and references therein for further explanation.

ODE Method, Revisited

In all examples above, the mean dynamics of a stochastic approximation algorithm is relatively clear from the algorithm itself. For instance, in the Riemannian setting, the vector field V induces a flow on the manifold, constituting the mean dynamics. Similarly, for SDE discretization, the corresponding Fokker–Planck equation identifies the mean dynamics in the Wasserstein space. More broadly, for any stochastic approximation algorithm characterized by a step-size parameter, it is feasible to recover the mean dynamics by performing averaging on the noise and considering the limiting behavior as the step-size approaches zero.

Certain algorithms, however, lack a defined step-size parameter. A notable example is the Sinkhorn algorithm—also known as the Iterative Proportional Fitting procedure—for solving entropic optimal transport and Schrödinger bridge problems. The goal of the entropic optimal transport problem is to find the most efficient way to transform one probability distribution into another while minimizing a given cost function, incorporating entropy to regularize the solution. This problem has gained traction for its applications in machine learning, economics, and physics. The Schrödinger bridge problem is the dynamic cousin of the entropic optimal transport problem, and involves finding the most likely evolution of a probability distribution subject to observed marginals at two different times. This problem has numerous applications in biology and molecular dynamics, to name a few. In the implementation of these iterative algorithms in practice, noise-induced updates are pervasive, necessitating a stochastic approximation analysis.

We propose a new approach to understanding these algorithms by framing them within the stochastic approximation framework, leveraging recent advancements that categorize these algorithms as optimization methods. Research has shown that the Sinkhorn algorithm can be viewed as a type of mirror descent algorithm, a well-known method typically used in constrained optimization and optimization in Banach spaces (as detailed in [Lég20; AKL22]).

Building on this understanding, we introduce a version of the Sinkhorn algorithm with adjustable step-sizes, along with a continuous-time version of the algorithm. Through a dynamic analysis using the ODE method, we develop a new stochastic approximation analysis for these algorithms. This analysis ensures that they converge to optimal solutions even when affected by noise and bias.

THESIS ROADMAP

This thesis is structured as follows. Chapter 2 is an introduction to the framework of Benaïm and Hirsch. This chapter establishes a foundation for a dynamical system viewpoint to examine the behavior of stochastic approximation algorithms. The subsequent three chapters (Chapters 3–5) address the core objectives of the thesis. Specifically, Chapter 3 offers a theoretical exploration of stochastic approximation algorithms within the context of Riemannian manifolds. In Chapter 4, the focus shifts to analyzing algorithms that can be interpreted as discretizations of stochastic differential equations, particularly within the Wasserstein space. Lastly, Chapter 5 presents findings related to the development of continuous-time counterparts of the Sinkhorn and Iterative Proportional Fitting algorithms, alongside a discussion of stochastic approximation methods for solving entropic optimal transport and Schrödinger bridge problems. Below, we detail this outline:

Chapter 2 starts with essentials of dynamical systems. Central notions such as flows, asymptotic pseudo-trajectories, and chain-recurrence are introduced, along with some of their basic properties. The main theorem of this chapter is the limit set theorem (Theorem 2.6), which identifies the limit sets of the iterates of a stochastic approximation algorithm. We outline the process of proving that a stochastic approximation algorithm meets the requirements of the limit set theorem by giving a detailed proof for the classic Euclidean case.

Chapter 3 studies our first non-Euclidean setting: Riemannian manifolds. This chapter lays the groundwork for a proper formulation and analysis of stochastic approximation algorithms in the Riemannian context. We begin by reminding foundational concepts and notations pertinent to differential and Riemannian geometry, followed by a discussion on adapting iterative methods to manifold settings. By examining specific examples related to machine learning and games, we aim to illustrate the practical relevance and challenges of nonlinear root-finding on manifolds. Next, we show that a wide class of algorithms—namely Riemannian Robbins–Monro algorithms—satisfy the asymptotic pseudo-trajectory property, and we show stability of these algorithms in Hadamard manifolds for weakly coercive vector fields. Our proof technique combines many ideas from differential and Riemannian geometry, parts of which might be of independent interest.

Chapter 4 studies our second non-Euclidean setting: the Wasserstein space. First, we review essential background knowledge, including aspects of stochastic calculus, the Fokker–Planck equation, and Wasserstein spaces. Next, we discuss the setup for analyzing SDE discretization algorithms, outlining the discretization template, interpolation methods, relevant metric spaces, and mean dynamics. The main theorem in this chapter states that the iterates of an SDE discretization converge to the same limits as the corresponding SDE. We then identify potential limit sets for two SDEs: Langevin and mirror Langevin diffusions. Following this, we demonstrate that a dissipativity condition is sufficient for stability. Finally, we present a set of practical sampling algorithms, showing that they adhere to the described template and proving their asymptotic convergence to the target distribution under noise and bias.

Chapter 5 shifts to exploring the linear structure and convexity in the space of signed measures, emphasizing the importance of the relative entropy functional. We start with an extensive review of the mirror descent algorithm, followed by essentials of analysis on topological vector spaces. We then introduce the Entropic Optimal Transport problem, the properties of its optimal solution, and the Sinkhorn algorithm. A variant of the Sinkhorn algorithm is derived based on the discrete-time mirror descent scheme. This variant transitions to a continuous-time flow in infinitesimal step-sizes, giving rise to the Sinkhorn flow. We analyze the convergence of both discrete- and continuous-time schemes using stochastic mirror descent analysis. Finally, we discuss the Schrödinger Bridge problem and interpret the Iterative Proportional Fitting procedure through mirror descent, linking these iterations to SDEs.

Chapter 6 concludes this thesis and brings future research directions and a few open problems.

Throughout the text, we mainly use *italics* for defining new mathematical objects and reserve **bold face** for headings and titles. Important theorems, corollaries, etc. are marked with a ▶ in their title. In several places in the thesis, the most technical parts of proofs are deferred to the appendix at the end, making it easier to follow the main arguments while reading the main text.

LIST OF PUBLICATIONS

Most of the results presented in this thesis have been published in the following conference proceedings:

> Mohammad Reza Karimi, Ya-Ping Hsieh, Panayotis Mertikopoulos, and Andreas Krause. "The Dynamics of Riemannian Robbins-Monro Algorithms". In: *Proceedings of 35th Conference on Learning Theory (COLT)*. 2022.
>
> Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. "A Dynamical System View of Langevin-Based Non-Convex Sampling". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.
>
> Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. "Sinkhorn Flow as Mirror Flow: a Continuous-Time Framework for Generalizing the Sinkhorn Algorithm". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2024.

The following conference proceedings are directly related to the material presented in this thesis but are not explicitly included:

> Ya-Ping Hsieh, Mohammad Reza Karimi, Andreas Krause, and Panayotis Mertikopoulos. "Riemannian Stochastic Optimization Methods Avoid Strict Saddle Points". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.
>
> Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. "Stochastic Approximation Algorithms for Systems of Interacting Particles". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

COLLABORATORS

This dissertation owes much to the invaluable contributions of three collaborators. The main content was developed through close collaboration with Ya-Ping Hsieh and Andreas Krause. Additionally, the material presented in Chapter 3 was created in collaboration with Panayotis Mertikopoulos.

GENERAL NOTATION

$\mathbb{R}_+$: the set of non-negative real numbers $[0, \infty)$.

$C^k(\Omega)$: the set of k-times continuously differentiable functions defined on the set Ω.

$C(\Omega; E)$: the set of E-valued continuous functions defined on the set Ω.

$C_c(\Omega; E)$: the set of E-valued continuous functions defined on the set Ω with compact support.

$C_b(\Omega; E)$: the set of E-valued continuous and bounded functions defined on the set Ω.

$L^\infty_+(\Omega)$: the set of non-negative functions in $L^\infty(\Omega)$.

$L^\infty_{++}(\Omega)$: the set of non-negative functions in $L^\infty(\Omega)$ that are bounded away from zero.

$A \coloneqq B$: A is defined by B.

$A(x) \equiv B(x)$: $A(x)$ and $B(x)$ are identical.

$A \doteq B$: A is equal to B, up to some additive constant.

$A \lesssim B$: A is less than B up to some multiplicative constant.

$\mathcal{P}(X)$: the set of all probability measures on the measurable set X.

a.s. : almost surely, or with probability 1.

$T_\# \mu$: the pushforward of the measure μ with the map T, i.e., $(T_\# \mu)(A) = \mu(T^{-1}(A))$.

$\mathbf{1}_A$: the indicator function of the set A, i.e., $\mathbf{1}_A(x) = 1$ if $x \in A$ and is otherwise 0.

$\sigma(C)$: the σ-algebra generated by the random variables in the set C.

$\mathcal{F} \vee \mathcal{G}$: the smallest σ-algebra containing both $\mathcal{F}$ and $\mathcal{G}$.

$\mu \ll \nu$: the measure μ is absolutely continuous with respect to ν.

$X \in \mathcal{F}$: the random variable Y is $\mathcal{F}$-measurable.

$\overline{A}$ or $\operatorname{cl} A$: the closure of the set A.

$B_\delta(p)$: (metric) open ball of radius δ centered at p.

$\lvert v\rvert$	:	Riemannian norm of a tangent vector v.
$\|v\|$ or $\|v\|_2$	:	Euclidean norm of a vector v.
$\nabla \cdot (v)$	:	divergence of a vector field v, i.e., $\nabla \cdot (v) = \sum \partial_i v_i$.
$\nabla^2 h$	:	Hessian of a function h.
Δh	:	Laplacian of a function h, i.e., $\Delta h = \sum \partial^2_{ii} h$.
$\mathrm{tr}(A)$	:	the trace of a matrix A.

CHAPTER TWO

DYNAMICAL SYSTEMS AND STOCHASTIC APPROXIMATION

In this chapter, we present those fundamental concepts from dynamical systems on metric spaces that are essential for stochastic approximation. The key notions discussed are asymptotic pseudo-trajectories and internally chain-transitive sets. An asymptotic pseudo-trajectory is a continuous curve that increasingly approximates some orbit of a flow over arbitrarily long time intervals. An internally chain-transitive set is a compact, connected, attractor-free invariant set, and is a potential limit of a perturbed orbit of a flow. The limit set theorem, Theorem 2.6, connects these two concepts, showing that the limits of a precompact asymptotic pseudo-trajectory is always an internally chain-transitive set.

We focus exclusively on aspects of dynamical systems that are directly relevant to this thesis. For comprehensive details, readers are encouraged to consult the outstanding work of Benaïm [Ben99] and the paper of Benaïm and Hirsch [BH96]. Nearly all the material, including theorem statements and proofs, is drawn from these sources. While some technical proofs are omitted, additional proofs are provided where they are not covered in the mentioned references.

2.1. FLOWS ON METRIC SPACES

Let $\mathcal{M}$ be a metric space equipped with the distance function (or metric) d. Metric spaces are one of the most general settings for studying many of the concepts of mathematical analysis and geometry. Many types of mathematical objects that have a natural notion of distance admit the structure of a metric space, including Euclidean spaces, Riemannian manifolds, normed vector spaces, and graphs. We always consider the topology on $\mathcal{M}$ induced by the metric d.

A *flow* on $\mathcal{M}$ is a continuous function $\Phi : \mathbb{R} \times \mathcal{M} \to \mathcal{M}$, mapping (t, x) to $\Phi(t, x) = \Phi_t(x)$, satisfying $\Phi_0 = \mathrm{Id}_{\mathcal{M}}$ and the semigroup property $\Phi_s \circ \Phi_t = \Phi_{t+s}$ for all $t, s \in \mathbb{R}$. If Φ is defined over $\mathbb{R}_+ \times \mathcal{M}$, it is called a *semi-flow*. An example to keep in mind is the flow of a vector field on some Euclidean space: Let $V : E \to E$ be a continuous vector field on the Euclidean space E, and for any initial point $x_0 \in E$, consider the solution of the initial value problem

$$\dot{x}(t) = V(x(t)), \quad x(0) = x_0.$$

If the solution is unique and exists for all $x_0 \in E$ and all $t \in \mathbb{R}$ (resp. $t \geq 0$), one can define the flow (resp. semi-flow) Φ corresponding to V as $\Phi_t(x_0) = x(t)$. See Fig. 2.1 for an illustration.

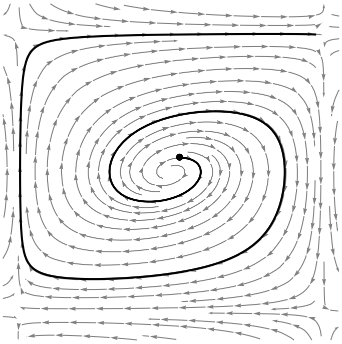


Figure 2.1. An illustration of the flow of a vector field in $\mathbb{R}^2$. The gray arrows depict the streamlines of the flow. The black curve is the forward orbit of the black point in the center. This example is taken from [CG24].

2.1.1. Invariant Sets and Attractors

Understanding the dynamics of a flow, namely if it converges towards some limit or avoids other sets, usually boils down to studying its invariant sets and attractors. In this section, we introduce such objects and bring a few relevant properties. To make the exposition easier, we let the time index set $\mathbb{T}$ to be $\mathbb{R}_+$ in case of semi-flows, and $\mathbb{R}$ in case of flows.

- A subset $A \subset \mathcal{M}$ is *positively invariant* for the flow (or semi-flow) Φ if $\Phi_t(A) \subset A$ for all $t \geq 0$, and is *invariant* if $\Phi_t(A) = A$ for all $t \in \mathbb{T}$.
- A point $p \in \mathcal{M}$ is at *equilibrium* if $\Phi_t(p) = p$ for all $t \in \mathbb{T}$.
- For $x \in \mathcal{M}$, the *forward orbit of* x is the set $\gamma^+(x) = \{\Phi_t(x) : t \geq 0\}$ and the orbit of x is $\gamma(x) = \{\Phi_t(x) : t \in \mathbb{T}\}$.
- A point $p \in \mathcal{M}$ is an *omega limit point of* x if $p = \lim_{t_k \to \infty} \Phi_{t_k}(x)$ for some sequence $t_k \to \infty$. We denote the set of all omega limit points of x by $\omega(x)$. See Fig. 2.2 for an illustration. An *alpha limit point* is defined similarly for the reverse flow.
- A nonempty subset $A \subset \mathcal{M}$ is an *attractor* if it is compact and invariant and has a neighborhood W, called a *fundamental neighborhood*, such that $d(\Phi_t(x), A) \to 0$ as $t \to \infty$ uniformly in $x \in W$. The *basin* of A is a positively invariant open set of all points x such that $d(\Phi_t(x), A) \to 0$ as $t \to \infty$.

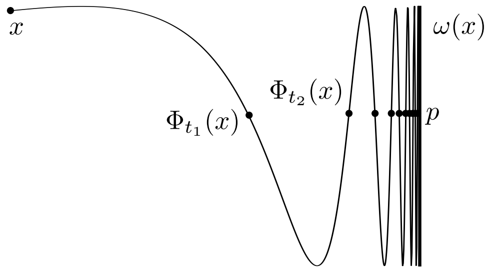


Figure 2.2. The forward orbit of the flow Φ starting at x is shown as the thin line, and the set $\omega(x)$ of all omega limit points of x is denoted by the thick line on the right. For each point $p \in \omega(x)$, there corresponds a sequence of times $t_1, t_2, \ldots \to \infty$ such that $\Phi_{t_i}(x)$ converges to p.

Let us now make use of the definitions above and show a few properties of invariant sets and attractors. We begin by showing that the omega limit set of a point is an invariant set.

Lemma 2.1. *The omega limit set of a point $x \in \mathcal{M}$ is an invariant set. Moreover, if the forward orbit $\gamma^+(x)$ of x has compact closure, then the omega limit set of x is also compact and connected.*

Proof. For any fixed $t \in \mathbb{T}$, $\Phi_t : \mathcal{M} \to \mathcal{M}$ is a continuous map on $\mathcal{M}$. Let $p \in \omega(x)$ and let $\{t_i\}_{i \geq 0}$ be a sequence such that $\Phi_{t_i}(x) \to p$. Applying Φ_t to both sides of the limit gives $\Phi_t(\Phi_{t_i}(x)) \to \Phi_t(p)$. This means that $\Phi_{t+t_i}(x) \to \Phi_t(p)$, implying $\Phi_t(p) \in \omega(x)$. Thus, $\omega(x)$ is an invariant set.

Let us now assume that $\gamma^+(x)$ is precompact. First, we show that $\omega(x)$ is compact. Clearly, $\omega(x)$ is precompact, as it is a subset of $\operatorname{cl} \gamma^+(x)$. To show that it is closed, consider a sequence of points $\{p_n\}$ in $\omega(x)$ converging to some point $p \in \mathcal{M}$. As $p_n \in \omega(x)$, there exists a convergent sequence in $\gamma^+(x)$ that converges to p_n. A diagonal argument on this sequence of sequences shows that the limit p is also in $\omega(x)$.

To see why $\omega(x)$ is connected, suppose on the contrary that it is included in the union of two disjoint open sets $A \cup B$. Take $p \in A$ and $q \in B$ and let $\{t_i\}$ and $\{s_j\}$ be a sequence of times for which $\Phi_{t_i}(x) \to p$ and $\Phi_{s_j}(x) \to q$. By taking subsequences, we can assume that $t_1 < s_1 < t_2 < s_2 < \cdots$ and $\Phi_{t_i}(x) \in A$ and $\Phi_{s_i}(x) \in B$ for all i. Since Φ is continuous, for each i, there exists a time $r_i \in (t_i, s_i)$ such that $\Phi_{r_i}(x) \notin A \cup B$. As $\gamma^+(x)$ is precompact, after taking a subsequence, we have that $\Phi_{r_i}(x) \to y$. Surely, $y \in \omega(x)$, but $y \notin A \cup B$ as $\mathcal{M} \setminus (A \cup B)$ is closed; a contradiction. □

Attractors are special invariant sets that absorb nearby points. One way to assess if there is some attractor inside an open set is provided below. Intuitively, if the flow, after a while, maps an open set into itself, there should be some attractor inside to make this happen. The nature of this result is similar to fixed-point theorems.

Lemma 2.2 (Ben99, Lem. 5.2)**.** *Let $U \subset \mathcal{M}$ be an open set with compact closure. If $\Phi_T(\overline{U}) \subset U$ for some $T > 0$, then there exists an attractor $A \subset U$ whose basin includes $\overline{U}$.*

Proof. The proof follows Benaïm's. The guiding intuition is the following: for large enough t, Φ_t maps $\overline{U}$ to the interior of U; taking the limit as $t \to \infty$ of the sets $\Phi_t(\overline{U})$ should give us the attractor, as one cannot escape from this limiting set, if started in $\overline{U}$.

Since U has compact closure, $\Phi_T(\overline{U})$ is compact and we can find an open set V such that $\Phi_T(\overline{U}) \subset V \subset \overline{V} \subset U$. Moreover, as the flow is continuous, there is some $\varepsilon > 0$ such that $\Phi_t(\overline{U}) \subset V$ for all $t \in [T - \varepsilon, T + \varepsilon]$. Thus, for all $t \geq T^2/\varepsilon$, writing $t = kT + r$ with $k \in \mathbb{N}$ and $0 \leq r/k < \varepsilon$, we see that $\Phi_t(\overline{U}) = (\Phi_{T+r/k} \circ \cdots \circ \Phi_{T+r/k})(\overline{U}) \in V$.

The "limit" of the sets $\Phi_t(\overline{U})$ is our candidate for the attractor. Concretely, let $A_t = \mathrm{cl}(\bigcup_{s\geq t}\Phi_s(\overline{U})) \subset \overline{V}$ and define $A = \bigcap_{t\geq 0} A_t$. By construction, A is compact and invariant, and it is the omega limit of a neighborhood of itself; as this neighborhood is contained in a compact set $\overline{U}$, uniform convergence to A started from $\overline{U}$ is also guaranteed. □

2.2. CHAIN RECURRENCE

Having stochastic approximation in mind, we rarely have access to exact evaluation of a flow and resort to constructing approximations of the orbit that are susceptible to noise and bias. An essential tool for analyzing such perturbed orbits of a flow is via the notion of chain-recurrence.

Let $\delta > 0$ and $T > 0$. A (δ, T)*-pseudo-orbit* from a to b is a chain of orbits

$$\{\Phi_t(y_i) : t \in [0, t_i]\}, \quad i = 0, 1, \ldots, k-1, \quad t_i \geq T$$

with

$$d(y_0, a) < \delta, \quad d(\Phi_{t_i}(y_i), y_{i+1}) < \delta, \quad \text{and} \quad y_k = b.$$

See Fig. 2.3 for an illustration. We write $\Phi : a \hookrightarrow_{\delta,T} b$ if there is a (δ, T)-pseudo-orbit from a to b, and drop Φ if it is clear from the context. We also write $a \hookrightarrow b$ if there is a (δ, T)-pseudo-orbit from a to b for all $\delta, T > 0$.

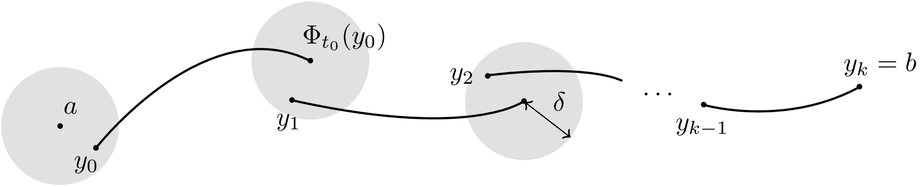


Figure 2.3. A (δ, T)-pseudo-orbit from a to b consists of a chain of trajectories of the flow Φ, each having duration at least T and the endpoint of each segment being δ-close to the start of the next one. In the figure, the gray circles are metric balls of radius δ and the black lines are orbits of the flow.

The flow Φ is called *chain-transitive* if $a \hookrightarrow b$ for all $a, b \in \mathcal{M}$. A point $a \in \mathcal{M}$ is called a *chain-recurrent point* if $a \hookrightarrow a$. We denote by $R(\Phi)$ the set of all chain-recurrent points of Φ. If every point in $\mathcal{M}$ is chain-recurrent, Φ is called a chain-recurrent flow.

Let $\Lambda \subset \mathcal{M}$ be an invariant set. We say Φ is chain-recurrent on Λ if $\Lambda = R(\Phi|_\Lambda)$. Here, $\Phi|_\Lambda$ is the restriction of the flow Φ to the set Λ. A compact invariant set on which Φ is chain-recurrent (resp. chain-transitive) is called an *internally chain-recurrent* (resp. *internally chain-transitive*) set. The following example gives more intuition on these concepts.

▷ **Example 2.1** (Ben99, Ex. 5.1)**.** Consider the unit circle $S^1 = \mathbb{R}/2\pi\mathbb{Z}$ and the flow Φ induced by the differential equation $\dot{\theta} = f(\theta)$ with $f(\theta) = \sin^2(\theta) \geq 0$, see Fig. 2.4.

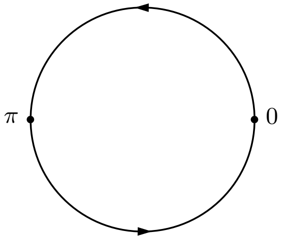


Figure 2.4. Flow on the unit circle.

Notice that the set of equilibria of this flow is $f^{-1}(0) = \{0, \pi\}$. Moreover, every point in S^1 is chain-recurrent, i.e., $R(\Phi) = S^1$. This is because we are allowed to "jump" over the points 0 and π and go from one side to the other.

Now, consider the upper half of the circle $\Lambda := [0, \pi]$, which is invariant for the flow. Note that Λ is a compact invariant set consisting of chain-recurrent points. However, it is not *internally* chain-recurrent: Choose T large and δ small, so that starting close to 0 always sends us to the left side, closer to π. ◁

The reason that in the previous example, $[0, \pi]$ failed to be internally chain-recurrent is that the flow restricted to $[0, \pi]$ has a proper attractor $\{\pi\}$. Proposition 2.3 below shows that this is the only reason that prevents a connected compact invariant set from being internally chain-recurrent.

Proposition 2.3 (Ben99, Prop. 5.3)**.** *Let $\Lambda \subset \mathcal{M}$. The following assertions are equivalent:*

(i) Λ *is an internally chain-transitive set.*

(ii) Λ *is connected and internally chain-recurrent.*

(iii) Λ *is a compact invariant set and $\Phi|_\Lambda$ has no proper attractors.*

Proof. (i) $\Rightarrow$ (ii): We only have to show that Λ is connected. Suppose on the contrary that $\Lambda \subset A \cup B$, where A, B are two disjoint open sets. Since Λ is compact, there exists some $\delta > 0$ such that $d(x, y) \geq \delta$ for all $x \in \Lambda \cap A$ and $y \in \Lambda \cap B$. This implies that there cannot be a $(\delta/2, T)$-pseudo-orbit starting from $\Lambda \cap A$ and ending in $\Lambda \cap B$.

The rest of the proof follows Benaïm's.

(ii) $\Rightarrow$ (iii): Suppose on the contrary that Λ admits a proper attractor A. We show that $A = \Lambda$ by showing that A is both closed and open relative to Λ. Let W be a relatively open fundamental neighborhood of A, and suppose that there is some $p \in W \setminus A$. What we show is that it is impossible to have $p \hookrightarrow p$. Since A is an attractor, it is compact, and there exists a $\delta > 0$ such that $U_\delta := \{x \in \Lambda : d(x, A) \leq \delta\}$ satisfies $U_\delta \subset U_{2\delta} \subset W$ and $B_\delta(p) \subset W \setminus U_{2\delta}$. Now, since W is a fundamental neighborhood of A, we can choose $T > 0$ in such a way that for all $x \in W$, $\Phi_T(x) \in U_\delta$. Therefore, any (δ, T)-pseudo-orbit starting from $B_\delta(p)$, ends up in U_δ in the first step. Then, the next starting point shall be in $U_{2\delta}$, and similarly for the rest of the pseudo-orbit. Thus, there is no chance to get δ-close to p; a contradiction.

(iii) $\Rightarrow$ (i): Take $x \in \Lambda$ and $\delta, T > 0$, and consider the set V of all points $y \in \Lambda$ such that $(\Phi|_\Lambda : x \hookrightarrow_{\delta,T} y)$. It is clear that the set V is open and satisfies $\Phi_T(\overline{V}) \subset V$. Lemma 2.2 then implies that V contains an attractor. However, since Λ has no proper attractors, it follows that $V = \Lambda$. Since this is true for all $\delta, T > 0$, the set Λ is internally chain-transitive. □

An important example of an internally chain-transitive set is given by the following proposition:

Proposition 2.4 (Ben99, Cor. 5.6)**.** *Let $x \in \mathcal{M}$. If $\gamma^+(x)$ has compact closure, then $\omega(x)$ is internally chain-transitive.*

2.3. ASYMPTOTIC PSEUDO-TRAJECTORIES AND THE LIMIT SET THEOREM

One of the key notions of this chapter is that of an asymptotic pseudo-trajectory.

Definition 2.5. We say that a continuous curve $\boldsymbol{x} : \mathbb{R}_+ \to \mathcal{M}$ is an *asymptotic pseudo-trajectory* of the flow Φ if, for all $T > 0$, it holds

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t))) = 0, \quad \text{almost surely.} \tag{2.1}$$

In other words, for each fixed $T > 0$, the curve $[0, T] \to \mathcal{M} : s \mapsto \boldsymbol{x}(t + s)$ *shadows* the forward orbit of the flow Φ started at $\boldsymbol{x}(t)$ over the interval $[0, T]$

with arbitrary accuracy for sufficiently large t. See Fig. 2.5 for an illustration. A large portion of this thesis is devoted to proving such property for different classes of algorithms.

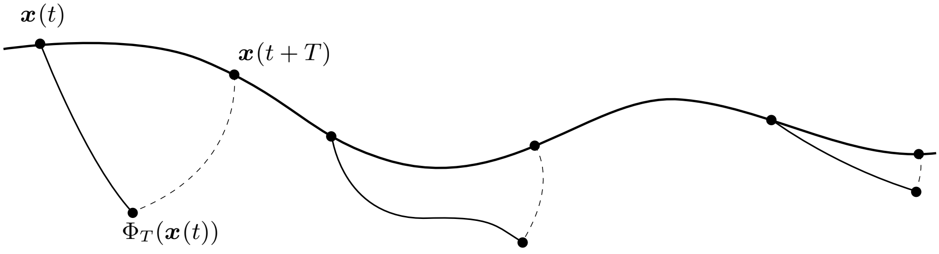


Figure 2.5. An illustration of an asymptotic pseudo-trajectory. The thick line depicts an asymptotic pseudo-trajectory $\boldsymbol{x}$. The thin lines are the orbits of the flow Φ for a time duration of T started at different $\boldsymbol{x}(t)$'s. As seen in the figure, as time t passes, the curve $\boldsymbol{x}$ gets closer to being an orbit of Φ.

The main reason to consider asymptotic pseudo-trajectories is the following theorem, known as the *limit set theorem* of Benaïm and Hirsch:

▶ **Theorem 2.6** (Ben99, Thm. 5.7)**.** *Let X be a precompact asymptotic pseudo-trajectory of the flow Φ. Then the limit set $\mathcal{L}(X) := \bigcap_{t\geq 0} \operatorname{cl}\{X(s) : s \geq t\}$ of X is an internally chain-transitive set.*

Denote by $C(\mathbb{R};\mathcal{M})$ the set of all $\mathcal{M}$-valued continuous functions (i.e., the space of all continuous curves on $\mathcal{M}$), endowed with the topology of uniform convergence on compact intervals. This topology is metrizable with the metric

$$d(f,g) := \sum_{k=1}^{\infty} \frac{1}{2^k} \min(1, d_k(f,g)), \tag{2.2}$$

where $d_k(f,g) := \sup_{t\in[-k,k]} d(f(t),g(t))$.

The main idea of the proof is to "lift" the problem the larger space $C(\mathbb{R};\mathcal{M})$ and to consider the *translation flow* Θ in that space: For any curve X and any $t \in \mathbb{R}$, Θ^t maps the curve X to $\Theta^t(X)$, defined as

$$\Theta^t(X)(s) = X(t+s).$$

As it turns out, the translation flow and Φ will be topologically conjugate when restricted to the set of orbits of Φ. This in turn implies that all topological properties of our interest, such as internally chain-transitive sets, are preserved by this conjugacy; we only have to prove things for the translation flow and automatically get results for the flow Φ. Below, we make this approach precise.

First, let us recall the notion of topological conjugacy for flows. Suppose Φ and Ψ are two flows on the metric spaces $\mathcal{M}$ and $\mathcal{N}$, respectively. We say Φ and Ψ are *topologically conjugate* if there is a homeomorphism (i.e., a continuous bijection with continuous inverse) $h : \mathcal{M} \to \mathcal{N}$ such that for all $t \in \mathbb{R}$,

$$\Psi(t, h(x)) = h(\Phi(t, x)), \quad \forall x \in \mathcal{M}.$$

Notice that in this case, the orbits of Φ are homeomorphically mapped to orbits of Ψ. Informally, topological conjugacy is a "change of variables" in the topological sense.

Let $S_\Phi \subset C(\mathbb{R}; \mathcal{M})$ be the set of all orbits $\phi^p : t \mapsto \Phi_t(p)$. Define the homeomorphism $H : \mathcal{M} \to S_\Phi$ by $H : p \mapsto \phi^p$, i.e.,

$$H(p)(t) = \phi^p(t) = \Phi_t(p).$$

It is then evident that H creates a topological conjugacy between the translation flow $\Theta|_{S_\Phi}$ restricted to S_Φ and the flow Φ:

$$\Theta^t|_{S_\Phi}(H(p)) = H(\Phi_t(p)), \quad \forall p \in \mathcal{M}.$$

Having this conjugacy at our disposal, our first task is to identify an equivalent notion for the asymptotic pseudo-trajectory property in the space $C(\mathbb{R}; \mathcal{M})$.

Lemma 2.7 (Ben99, Lem. 3.1)**.** *A continuous function $X : \mathbb{R}_+ \to \mathcal{M}$ is an asymptotic pseudo-trajectory for Φ if and only if*

$$\lim_{t\to\infty} d(\Theta^t(X), \hat{\Phi} \circ \Theta^t(X)) = 0,$$

where $\hat{\Phi}(X) = H(X(0)) = \phi^{X(0)}$.

Note that $\hat{\Phi} : C(\mathbb{R}; \mathcal{M}) \to S_\Phi$ is a retraction, that is, a continuous mapping from $C(\mathbb{R}; \mathcal{M})$ into the subspace S_Φ that preserves the position of all points in that subspace. Thus, we can interpret the lemma above in the following way: An asymptotic pseudo-trajectory of Φ is a point of $C(\mathbb{R}; \mathcal{M})$ whose forward trajectory under Θ is absorbed by S_Φ. If, moreover, the asymptotic pseudo-trajectory X is precompact, then the set of its translations will be precompact in $C(\mathbb{R}; \mathcal{M})$.

Theorem 2.8 (Ben99, Thm. 3.2)**.** *Let $X : \mathbb{R}_+ \to \mathcal{M}$ be a continuous curve with precompact image. If X is an asymptotic pseudo-trajectory, then the set $\{\Theta^t(X) : t \geq 0\}$ is precompact in $C(\mathbb{R}; \mathcal{M})$.*

We are now ready to prove Theorem 2.6.

Proof of Theorem 2.6. Since X is a precompact asymptotic pseudo-trajectory, Theorem 2.8 implies that $\{\Theta^t(X) : t \geq 0\}$ is precompact in $C(\mathbb{R}; \mathcal{M})$. Therefore, Proposition 2.4 implies that the omega limit set $\omega_\Theta(X)$ of X under the translation flow Θ is internally chain-transitive.

We now use the conjugacy (induced by H) between $\Theta|_{S_\Phi}$ and Φ. Note that chain-transitivity is preserved under conjugacy, and therefore, if we show that

$$H(\mathcal{L}(X)) = \omega_\Theta(X),$$

then we get that $\mathcal{L}(X)$ is internally chain-transitive, and we are done. We prove this by showing that each side is a subset of the other.

Take $p \in \mathcal{L}(X)$. This means that $p = \lim_{t_k \to \infty} X(t_k)$ for some sequence $\{t_k\}$. As $\{\Theta^t(X) : t \geq 0\}$ is precompact, after taking a subsequence, we can assume that $\Theta^{t_k}(X) \to Y \in C(\mathbb{R}; \mathcal{M})$. Since by Lemma 2.7, $\lim_{t_k \to \infty} d(\Theta^{t_k}(X), \hat{\Phi} \circ \Theta^{t_k}(X)) = 0$, we deduce that $Y = \hat{\Phi}(Y) = H(Y(0)) = H(p)$. Thus, $H(\mathcal{L}(X)) \subseteq \omega_\Theta(X)$.

To show the other direction, take $Y \in \omega_\Theta(X)$. That is, $Y = \lim_{t_k \to \infty} \Theta^{t_k}(X)$. Using Lemma 2.7 again shows that $\hat{\Phi}(Y) = Y$, which means that $Y = H(Y(0)) = \phi^q$ for some $q \in \mathcal{M}$. Moreover, $q = Y(0) = \lim_{t_k \to \infty} X(t_k)$. That is, $q \in \mathcal{L}(X)$. Thus, $\omega_\Theta(X) \subseteq H(\mathcal{L}(X))$, and we are done. □

2.4. GRADIENT-LIKE SYSTEMS

We now turn our attention to an important category of flows: those originating from gradient-like systems. Fundamentally, any gradient-like system is associated with a concept of "energy" that dissipates along the flow trajectories. When a flow is derived from the gradient of a potential function, it evidently adheres to this dissipation property. However, it is important to note that this is not the only way such systems can exhibit energy dissipation.

A broader framework to understand this dissipation property is given by Lyapunov functions. Lyapunov functions provide a generalized method to demonstrate that the energy of the system decreases along the flow, ensuring the system's stability. In this section, we will explore the concept of Lyapunov functions in detail and discuss their application in characterizing the internally chain-transitive sets of gradient-like systems.

Definition 2.9. Let $\Lambda \subset \mathcal{M}$ be a compact invariant set of the semi-flow Φ. A continuous function $f : \mathcal{M} \to \mathbb{R}$ is called a *Lyapunov function for* Λ if the function $t \in \mathbb{R}_+ \mapsto f(\Phi_t(x))$ is strictly descreasing for $x \in \mathcal{M} \setminus \Lambda$ and is constant for $x \in \Lambda$.

In the case where Λ is the set of equilibria of Φ, f is called a *strict Lyapunov function* and the flow Φ is called a *gradient-like system*.

Theorem 2.10. *Let $\Lambda \subset \mathcal{M}$ be a compact invariant set and let f be a Lyapunov function for Λ. If $f(\Lambda) \subset \mathbb{R}$ has empty interior, then any internally chain-transitive set L is contained in Λ. Moreover, $f|_L$ is constant.*

Proof. First, we show that $L \cap \Lambda \neq \emptyset$ and

$$\inf_{x \in L} f(x) = \inf_{x \in L \cap \Lambda} f(x).$$

For this, take $x \in L$. Since by assumption, $f(\Phi_t(x))$ is nonincreasing as a function of t and is bounded from below by the infimum of f over the compact set L, it has a limit $f_\infty(x) = \lim_{t\to\infty} f(\Phi_t(x)) \in \mathbb{R}$. Moreover, for any point $p \in \omega(x)$, continuity of f implies that $f(p) = f_\infty(x)$. Thus,

$$f(p) = f_\infty(x) \leq f(x), \quad \forall p \in \omega(x).$$

Since $\omega(x)$ is an invariant set (Lemma 2.1) and f is constant on trajectories in $\omega(x)$, we conclude that $\omega(x) \subset \Lambda$. Having in mind that L is compact and invariant itself, we know that $\omega(x) \subset L$. Therefore, we conclude that $\omega(x) \subset L \cap \Lambda$ and thus, $L \cap \Lambda \neq \emptyset$. Moreover, as $f(x) \geq f(p)$ for all $p \in \omega(x)$, we deduce that $\inf_{x \in L} f(x) = \inf_{x \in L \cap \Lambda} f(x) =: v^*$.

Since by assumption, $f(\Lambda)$ has empty interior, there exists a decreasing sequence $\{v_n\}$, so that $v_n \to v^*$ and $v_n \notin f(\Lambda)$. Let $L_n = \{x \in L : f(x) < v_n\}$. As f is a Lyapunov function, we find that $\Phi_t(\overline{L_n}) \subset L_n$ for all $t > 0$. Lemma 2.2 then implies that there exists some attractor $A \subseteq L_n$ whose basin contains $\overline{L_n}$. However, since L is internally chain-transitive, it cannot have any proper attractors (Proposition 2.3). The only possibility that remains is that $L_n = L$ for all n and $f(x) = v^*$ for all $x \in L$. As L is invariant, this also implies that $L \subset \Lambda$. □

2.5. EUCLIDEAN STOCHASTIC APPROXIMATION

In this section, we explain how to prove that a curve is an asymptotic pseudo-trajectory in the Euclidean setting. While this result is well-known and classical [see Ben99, Sec. 4], we present this proof to illustrate the core intuitions and arguments involved, along with the general structure of such proofs. Understanding this structure in its simplest form is beneficial for proofs that follow in the subsequent chapters.

Let us first describe the setup. Suppose V is a *complete* vector field in $\mathbb{R}^d$, that is, the ordinary differential equation $\dot{x}(t) = V(x(t))$ has a unique solution for all initial values and the solution exists for all $t \geq 0$. Consider the iterates $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots$ of a Robbins–Monro stochastic approximation algorithm

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \alpha_{n+1}(V(\boldsymbol{x}_n) + Z_{n+1}) \tag{2.3}$$

for finding the zeros of V, where Z_{n+1} is a deterministic or random perturbation, and α_{n+1} is a deterministic step-size. Let the filtration $\mathcal{F}_n$ encode all the information available at iteration n, that is, $\mathcal{F}_n = \sigma(\{\boldsymbol{x}_k, Z_k\}, k = 0, 1, \dots, n)$. With this, it is convenient to decompose Z_{n+1} as

$$Z_{n+1} = U_{n+1} + B_{n+1},$$

where $U_{n+1} = Z_{n+1} - \mathbb{E}[Z_{n+1} \,|\, \mathcal{F}_n]$ is the zero-mean stochastic *noise*, and $B_{n+1} = \mathbb{E}[Z_{n+1} \,|\, \mathcal{F}_n]$ is the (systematic) *bias*. The role of the noise and bias is going to be asymmetric and the set of assumptions for each is different.

Our main objective is to show—under some assumptions that will follow—that the asymptotic behavior of the sequence $\{\boldsymbol{x}_n\}_{n\geq 0}$ is determined by the flow of the vector field V. We already know one such result: the limit-set theorem (Theorem 2.6). To use this theorem, though, we have to first construct a continuous-time curve that interpolates the iterations $\{\boldsymbol{x}_n\}_{n\geq 0}$, and its convergence to some set implies the convergence of the iterates to that set. For the Euclidean setting, the interpolation is rather straightforward. Specifically, by defining the "effective time" variables

$$\tau_n = \sum_{k=1}^{n-1} \alpha_k,$$

we can construct the piecewise-linear interpolation

$$\boldsymbol{x}(t) = \boldsymbol{x}_n + \frac{t - \tau_n}{\tau_{n+1} - \tau_n}(\boldsymbol{x}_{n+1} - \boldsymbol{x}_n) \quad \text{for all } t \in [\tau_n, \tau_{n+1}],\ n \in \mathbb{N}. \tag{2.4}$$

We use the same symbol for this interpolation and the iterates and differentiate between them by using subscripts for the discrete-time iterates and parenthesis for the continuous-time counterpart. It is also convenient to define the "inverse" of the map $n \mapsto \tau_n$ as

$$m(t) = \sup\{n \geq 0 : \tau_n \leq t\}. \tag{2.5}$$

Let us now state the main assumptions and the result we are about to prove:

Theorem 2.11 (Ben99, Prop. 4.1)**.** *Let V be an L-Lipschitz vector field. Assume that*

(A1) $\sup_n \|\boldsymbol{x}_n\| < \infty$ *or* V *is bounded on the iterates* $\{\boldsymbol{x}_n\}$*,*

(A2) *the bias terms* B_n *vanish almost surely as* $n \to \infty$*,*

(A3) *the noise terms* U_n *have the* cancellation property*; that is, for any* $T > 0$*,*

$$\lim_{n\to\infty} \max\left\{\left\|\sum_{i=n}^{k-1} \alpha_{i+1} U_{i+1}\right\| : k = n+1, \ldots, m(\tau_n + T)\right\} = 0. \tag{2.6}$$

Then, the piecewise-linear interpolated curve $\boldsymbol{x}$ *defined in* (2.4) *is an asymptotic pseudo-trajectory of the flow* Φ *induced by* V*. That is, for any* $T > 0$*,*

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \|\boldsymbol{x}(t+h) - \Phi_h(\boldsymbol{x}(t))\| = 0. \tag{2.7}$$

Remark 2.1. A few remarks on the assumptions are in order:

(1) We take the vector field to be Lipschitz only for convenience and simplicity of the argument. Indeed, Benaïm proves the theorem above for continuous and complete vector fields. A more refined analysis is also provided in [Ben99, Prop. 4.1] for vector fields that are Lipschitz and bounded on a neighborhood of the iterates $\{\boldsymbol{x}_n : n \geq 0\}$.

(2) Boundedness of the iterates in (A1) is a rather practical assumption: in practice, the iterates of a stochastic approximation algorithm are expected not to blow up to infinity. In the literature, this property is also called *stability* of the algorithm, and there are specific criteria to assess such property. For example, having a Lyapunov function is most of the time sufficient to ensure stability.

(3) The assumption on the bias (A2) is required as there might be no cancellation of the bias terms in the long run, making convergence analysis rather challenging. The upside is that this assumption is satisfied by a large range of algorithms used in practice. Throughout this thesis, we will see many instances of practical algorithms, all of which satisfy (A2).

(4) We will discuss the assumption (A3) on the noise and ways to verify it in detail in Section 2.5.1 below. ◊

Let us now go ahead and see how these simple assumptions imply the strong asymptotic pseudo-trajectory property. The proof we present here is modified from Benaïm's proof.

Let us start by defining the continuous-time piecewise-constant processes

$$\overline{\boldsymbol{x}}(t) = \boldsymbol{x}_{m(t)}, \quad \overline{\alpha}(t) = \alpha_{m(t)+1}, \quad \overline{Z}(t) = Z_{m(t)+1}.$$

We also define $\overline{B}$ and $\overline{U}$ similar to $\overline{Z}$. Using this notation, we can rewrite the interpolation (2.4) in an integral form:

$$\boldsymbol{x}(t) = \boldsymbol{x}(0) + \int_0^t (V(\overline{\boldsymbol{x}}(s)) + \overline{Z}(s))\, ds. \tag{2.8}$$

As (2.7) suggests, for fixed $t \geq 0$ and $T > 0$, we need to compare—in supremum norm—the $[t, t+T]$-segment of the interpolation $\boldsymbol{x}$ with the orbit of the flow Φ starting at $\boldsymbol{x}(t)$ up to time T. We must subsequently show that this distance almost surely vanishes as $t \to \infty$.

A crucial idea for facilitating this comparison is to construct another curve based on the interpolation and the flow, which lies between the interpolation and the orbit of the flow. A suitable candidate for this curve is the *Picard iteration* applied to the interpolation. Given a continuous curve $X : [a, b] \to \mathbb{R}^d$, a Picard iteration is the result of applying the operator $\mathscr{L}_V : C([a, b]; \mathbb{R}^d) \to C([a, b]; \mathbb{R}^d)$ on the curve X. This operator keeps the initial point of the curve fixed, and satisfies

$$\mathscr{L}_V(X)(t) = X(a) + \int_a^t V(X(s))\, ds, \quad \forall t \in [a, b]. \tag{2.9}$$

From standard analysis concerning the existence and uniqueness of solutions to ODEs, we know that the Picard iteration introduces a contraction in the space of continuous functions. For our purposes, even a single Picard iteration is sufficient to provide information about the proximity of the interpolation to the flow orbit. A very important remark here, however, is that this notion only makes sense in the Euclidean setting: on other spaces, there is no *a priori* way to integrate vector fields arbitrarily.

Proof of Theorem 2.11. Let us fix $t, T > 0$ and define the curve $\lambda = \mathscr{L}_V(\boldsymbol{x}|_{[t,t+T]})$, which is the result of a Picard iteration applied to the $[t, t+T]$-segment of the interpolation $\boldsymbol{x}$. The proof follows by showing that the Picard curve is close to both interpolation and the orbit of flow; making the orbit and interpolation close by the triangle inequality.

For any $h \in [0, T]$, define $\phi(h) = \Phi_h(\boldsymbol{x}(t))$. We then have

$$\|\phi(h) - \boldsymbol{x}(t+h)\| \leq \|\phi(h) - \lambda(h)\| + \|\lambda(h) - \boldsymbol{x}(t+h)\|. \tag{2.10}$$

We bound each term from above individually. For the first term, we can use the integral representation of both the flow orbit and the Picard curve to get

$$\|\lambda(h) - \phi(h)\| = \left\| \int_0^h (V(\boldsymbol{x}(t+s)) - V(\phi(s)))\, ds \right\| \le L \int_0^h \|\boldsymbol{x}(t+s) - \phi(s)\|\, ds,$$

where we used the Lipschitzness of the vector field V. The similarity of the right-hand side of this bound with the left-hand side of (2.10) makes it convenient later to use Grönwall's lemma.

The second term of (2.10) is where the noise and bias come in. By the integral formulation of the interpolation (2.8) and Lipschitzness of V, we have

$$\begin{aligned} \|\boldsymbol{x}(t+h) - \lambda(h)\| &= \left\| \int_t^{t+h} V(\overline{\boldsymbol{x}}(s)) + \overline{Z}(s) - V(\boldsymbol{x}(s))\, ds \right\| \\ &= L \int_t^{t+h} \|\overline{\boldsymbol{x}}(s) - \boldsymbol{x}(s)\|\, ds + \left\| \int_t^{t+h} \overline{Z}(s)\, ds \right\|. \end{aligned} \tag{2.11}$$

Let us define

$$\Delta_Z(t, T) = \sup_{h \in [0,T]} \left\| \int_t^{t+h} \overline{Z}(s)\, ds \right\| \tag{2.12}$$

to be the worst-case effect of the accumulated noise and bias during the time interval $[t, t+T]$. We will treat this term later in Section 2.5.1, and show that it vanishes almost surely as $t \to \infty$. For now, let us assume that $\Delta_Z(t,T) \to 0$. Continuing with the first integral in (2.11), we have to bound the distance of the interpolation from the iterations inside each interpolating interval. For any $t \le s \le t + T$ we have

$$\|\overline{\boldsymbol{x}}(s) - \boldsymbol{x}(s)\| = \left\| \int_{\tau_{m(s)}}^s V(\overline{\boldsymbol{x}}(u)) + \overline{Z}(u)\, du \right\|.$$

Now we use the boundedness of the iterates or the vector field. Suppose that $\sup_n \|V(\boldsymbol{x}_n)\| \le K$, which holds in either case. Then, we can bound the last term further as

$$\left\| \int_{\tau_{m(s)}}^s V(\overline{\boldsymbol{x}}(u)) + \overline{Z}(u)\, du \right\| \le K\overline{\alpha}(s) + \left\| \int_{\tau_{m(s)}}^s \overline{Z}(u)\, du \right\|.$$

The first term vanishes as $t \to \infty$. The second term, however, needs some treatment. While it is merely the effect of a single noise and bias term (and vanishes as $t \to \infty$ because of our assumptions), as we later take the supremum over h, it becomes unclear if the supremum vanishes as well. Therefore, it is plausible to bound this term by something that only depends on t and T; taking the supremum in this case has no effect. For large enough t, we can assume that $\overline{\alpha}(s) < 1$, and we have

$$\left\| \int_{\tau_{m(s)}}^{s} \overline{Z}(u)\, du \right\| \leq \left\| \int_{t-1}^{\tau_{m(s)}} \overline{Z}(u)\, du \right\| + \left\| \int_{t-1}^{s} \overline{Z}(u)\, du \right\| \leq 2\Delta_Z(t-1, T+1).$$

Therefore,

$$\sup_{s \in [t, t+T]} \|\overline{\boldsymbol{x}}(s) - \boldsymbol{x}(s)\| \leq K \sup_{s \in [t, t+T]} \overline{\alpha}(s) + 2\Delta_Z(t-1, T+1).$$

Estimating the integral in (2.11) with the supremum of its argument gives

$$\|\boldsymbol{x}(t+h) - \lambda(h)\| \leq KLT \sup_{h \in [0,T]} \overline{\alpha}(t+h) + 2LT\Delta_Z(t-1, T+1) + \Delta_Z(t, T).$$

Let A_t be the right-hand side of the bound above. Putting it all together, we obtain using Grönwall's lemma that

$$\|\boldsymbol{x}(t+h) - \phi(h)\| \leq L \int_0^h \|\boldsymbol{x}(t+s) - \phi(s)\|\, ds + A_t \leq e^{hL} A_t.$$

Therefore,

$$\sup_{h \in [0,T]} \|\boldsymbol{x}(t+h) - \phi(h)\| \leq e^{TL} A_t.$$

Recalling that $A_t \to 0$ almost surely as $t \to \infty$, finishes the proof. □

2.5.1. Cancellation Property of Noise

The probabilistic part of the proof is controlling the worst-case (i.e., supremum) accumulation of noise and bias from time t to $t + T$ and showing that for every fixed $T > 0$, it almost surely vanishes as $t \to \infty$. We already have seen this term in the proof above, where we introduced Δ_Z in (2.12).

We begin by decomposing this error into separate noise and bias terms and deal with the bias first, which is easier. For any fixed $t \geq 0$ and $T > 0$, observe that

$$\Delta_Z(t, T) \leq \sup_{h \in [0,T]} \left\| \int_t^{t+h} \overline{U}(s)\, ds \right\| + \sup_{h \in [0,T]} \left\| \int_t^{t+h} \overline{B}(s)\, ds \right\|.$$

As the bias is assumed to vanish almost surely (see (A2) in Theorem 2.11), we have that $\sup_{s \geq t} \|\overline{B}(s)\| =: B^*(t) \to 0$ almost surely as $t \to \infty$. Therefore, as the second term in the bound above is bounded by $TB^*(t)$, it goes to zero with probability 1. Thus, we are left with the accumulation of the noise:

$$\Delta(t,T) \coloneqq \sup_{h\in[0,T]} \left\| \int_t^{t+h} \overline{U}(s)\,ds \right\|.$$

Note that controlling this term is essentially the same as (2.6) in (A3) by setting $n = m(t)$:

$$\max\left\{ \left\| \sum_{i=n}^{k-1} \alpha_{i+1} U_{i+1} \right\| : k = n+1, \dots, m(\tau_n + T) \right\} \tag{2.13}$$

The reason is that the curve $h \mapsto \int_t^{t+h} \overline{U}(s)\,ds$ is piecewise-linear and the supremum of its norm is only attained at either of its endpoints or one of the times $\{\tau_k : t \le \tau_k \le t+T\}$.

Let us outline the strategy for controlling this quantity. First, we realize that the sequence $\sum_{i=n}^{k} \alpha_n U_n$ is a martingale (for $k = n, n+1, \dots$). Then, we control one of its moments via the Burkholder inequality—a generalization of Doob's maximal inequality for moments other than 2. The choice of moment will be dictated by the rate of decrease of the step-size sequence $\{\alpha_n\}$. We then use Markov's inequality to control the probability that the maximum norm of this martingale goes above some threshold. Converting this to an almost sure convergence of the martingale is done via a simple Borel–Cantelli argument. Below, we explain these steps in more detail.

Recall that a sequence of integrable random variables $\{M_n\}_{n\in\mathbb{N}}$ is called a *martingale* adapted to the filtration $\{\mathcal{F}_n\}$, if

$$\mathbb{E}[M_{n+1} \,|\, \mathcal{F}_n] = M_n, \quad \text{a.s. for all } n.$$

A filtration is an increasing sequence of σ-algebras, i.e., $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for all n. For simplicity, we assume all martingales start at 0, that is, $M_0 = 0$, a.s.

Several martingale inequalities can be used to bound the excursions of stochastic approximation processes. The first one is the Doob's L^2 inequality [see, e.g., Str10], which states for all $N \in \mathbb{N}$,

$$\mathbb{E}\left[\sup_{n\le N} \|M_n\|^2\right] \le 4\,\mathbb{E}\,\|M_N\|^2.$$

This can be very useful, as it allows controlling the whole trajectory of the martingale up until N via the law of M_N only. A similar result for moments other than 2 is due to Burkholder [see Str10, Thm. 6.3.6] that states

$$\mathbb{E}\left[\sup_{n\le N}\|M_n\|^p\right] \le C_p^p\ \mathbb{E}\left[\left(\sum_{n=1}^{N}\|M_n - M_{n-1}\|^2\right)^{p/2}\right], \tag{2.14}$$

where C_p is some universal constant depending only on p. Notice that for $p=2$, the right-hand side of the Burkholder inequality becomes exactly the right-hand side of Doob's L^2 inequality after setting $C_p = 2$; this is because

$$\begin{aligned}\mathbb{E}[\|M_{n+1} - M_n\|^2 \,|\, \mathcal{F}_n] &= \mathbb{E}[\|M_{n+1}\|^2 \,|\, \mathcal{F}_n] - 2\langle M_n, \mathbb{E}[M_{n+1} \,|\, \mathcal{F}_n]\rangle + \|M_n\|^2 \\ &= \mathbb{E}[\|M_{n+1}\|^2 \,|\, \mathcal{F}_n] - \|M_n\|^2,\end{aligned}$$

and the summation becomes a telescopic sum.

Let us also remind the reader of the Borel–Cantelli's lemma:

Lemma 2.12 (Borel–Cantelli)**.** *Let $\{A_1, A_2, \ldots\}$ be a set of events in some probability space. If*

$$\sum_{n=1}^{\infty}\mathbb{P}[A_n] < \infty,$$

then the probability that A_n happens infinitely often is zero. That is, there exists some random N_0, so that for all $n \ge N_0$, A_n does not happen.

Below, we show that the Burkholder inequality in tandem with Borel–Cantelli's lemma implies that $\Delta(t,T)$ vanishes almost surely as $t\to\infty$.

Proposition 2.13 (Ben99, Prop. 4.2)**.** *Let U_n be a martingale difference sequence adapted to the filtration $\mathcal{F}_n$. Suppose that for some $p \ge 2$,*

$$\sup_n \mathbb{E}\,\|U_n\|^p < \infty, \quad \textit{and} \quad \sum_{n=1}^{\infty}\alpha_n^{1+p/2} < \infty.$$

Then, for every fixed $T>0$, the quantity $\Delta(t,T)$ defined in (2.13) vanishes almost surely as $t\to\infty$.

Proof. The proof is from Benaïm [Ben99]. Let $n = m(t)$ and define the martingale sequence for $k \ge n$

$$M_k := \alpha_n U_n + \cdots + \alpha_k U_k.$$

Recall the definition of $\Delta(t,T)$:

$$\Delta(t,T) := \max_{n\le k\le N}\|M_k\|,$$

where we have defined $N = m(\tau_n + T)$. Using the Burkholder inequality (2.14), we get

$$\mathbb{E}[\Delta(t,T)^p] = \mathbb{E}\left[\max_{n\le k\le N}\|M_k\|^p\right] \le C_p^p\,\mathbb{E}\left[\left(\sum_{k=n}^{N}\alpha_k^2\|U_k\|^2\right)^{p/2}\right]. \tag{2.15}$$

To get a simpler formulation of the right-hand side for the case where $p > 2$, we use Hölder's inequality[1] in the following way: Let $q = p/2$ and $\delta \in (0,1)$. Then, rewrite $\alpha_k^2\|U_k\|^2 = x_k y_k$, with $x_k = \alpha_k^{2(1-\delta)}\|U_k\|^2$ and $y_k = \alpha_k^{2\delta}$. Thus,

$$\left(\sum_{k=n}^{N}\alpha_k^2\|U_k\|^2\right)^q \le \left(\sum_{k=n}^{N}\alpha_k^{2q(1-\delta)}\|U_k\|^{2q}\right)\left(\sum_{k=n}^{N}\alpha_k^{2\delta q/(q-1)}\right)^{q-1}.$$

Choosing $\delta = (p-2)/(2p)$ makes the exponent $2\delta q/(q-1) = 1$, and we get

$$\begin{aligned}\left(\sum_{k=n}^{N}\alpha_k^2\|U_k\|^2\right)^q &\le \left(\sum_{k=n}^{N}\alpha_k^{1+p/2}\|U_k\|^p\right)\left(\sum_{k=n}^{N}\alpha_k\right)^{p/2-1}\\ &\le T^{p/2-1}\sum_{k=n}^{N}\alpha_k^{1+p/2}\|U_k\|^p.\end{aligned}$$

Plugging this estimate back into (2.15) gives

$$\mathbb{E}[\Delta(t,T)^p] \le C_p^pT^{p/2-1}\,\mathbb{E}\left[\sum_{k=n}^{N}\alpha_k^{1+p/2}\|U_k\|^p\right] = C_p^pT^{p/2-1}\sum_{k=n}^{N}\alpha_k^{1+p/2}\,\mathbb{E}\,\|U_k\|^p.$$

Since $\sup_n \mathbb{E}\,\|U_n\|^p < \infty$ by assumption, we have

$$\mathbb{E}[\Delta(t,T)^p] \le C(p,T)\sum_{k=n}^{N}\alpha_k^{1+p/2} = C(p,T)\int_t^{t+T}\overline{\alpha}(s)^{p/2}\,ds, \tag{2.16}$$

for some constant $C(p,T)$ depending on p and T. Notice that for $p = 2$, applying Hölder's inequality is not needed and one directly gets (2.16).

[1] For all $q > 1$, let $\bar{q} = q/(q-1)$ to be the Hölder conjugate of q, that is, $1/q + 1/\bar{q} = 1$. Then for any two vectors $x, y \in \mathbb{R}^d$, $|\langle x, y\rangle| \le \|x\|_q\|y\|_{\bar{q}}$. In other words,

$$\left|\sum x_i y_i\right|^q \le \left(\sum |x_i|^q\right)\left(\sum |y_i|^{q/(q-1)}\right)^{q-1}$$

Guided by our assumption that the series $\sum_n \alpha_n^{1+p/2}$ is summable, we are in the position to apply the Borel–Cantelli argument. Summing over all time windows of length T gives

$$\sum_{k\geq 0} \mathbb{E}[\Delta(kT,T)^p] \leq C(p,T) \int_0^\infty \overline{\alpha}(s)^{p/2}\, ds = C(p,T) \sum_{n=1}^\infty \alpha_n^{1+p/2} < \infty.$$

Let $\varepsilon > 0$ be arbitrary. We have by Markov's inequality

$$\mathbb{P}[\Delta(kT,T) \geq \varepsilon] = \mathbb{P}[\Delta(kT,T)^p \geq \varepsilon^p] \leq \frac{\mathbb{E}[\Delta(kT,T)^p]}{\varepsilon^p}$$

Therefore,

$$\sum_{k\geq 0} \mathbb{P}[\Delta(kT,T) \geq \varepsilon] < \infty.$$

By the Borel–Cantelli lemma, there exists a k_0 such that for all $k \geq k_0$, it holds with probability 1 that $\Delta(kT,T) < \varepsilon$. This means that

$$\lim_{k\to\infty} \Delta(kT,T) = 0$$

almost surely. This also implies that $\Delta(t,T) \to 0$ as $t \to \infty$, since for $t \in [kT,(k+1)T)$, it holds

$$\Delta(t,T) \leq 2\Delta(kT,T) + 2\Delta((k+1)T,T). \qquad \square$$

CHAPTER THREE

STOCHASTIC APPROXIMATION ON RIEMANNIAN MANIFOLDS

In this chapter, we examine stochastic approximation algorithms that are defined on Riemannian manifolds. These algorithms are either designed to find the zeros of a vector field on a Riemannian manifold given noisy and incomplete evaluations thereof, or model adaptive behavior on a Riemannian state space. Our main objective in this chapter is to formally define stochastic approximation on Riemannian manifolds, and prove convergence to desirable limits under suitable conditions. We specifically prove that under mild conditions, the iterates of these algorithms form an asymptotic pseudo-trajectory of the underlying flow, and find sufficient conditions that imply stability.

Originality. Main results of this chapter are published in the conference proceedings [Kar+22] as an extended abstract. However, there are considerable differences between this chapter and the mentioned publication, in notation, proofs, and content.

LIST OF IMPORTANT RESULTS

- **Theorem 3.4.** The iterates of a stochastic approximation algorithm on a Riemannian manifold constitutes an asymptotic pseudo-trajectory for the corresponding flow.

- **Corollary 3.5.** If the iterates of a stochastic approximation algorithm on a Riemannian manifold are precompact, they almost surely converge to an internally chain-transitive set of the flow.

- **Theorem 3.6.** For a Hadamard manifold and weakly coercive and bounded vector fields, the iterates of a stochastic approximation algorithm stay bounded.

- **Lemma 3.12.** A comparison theorem bounding the difference between the coordinates of a tangent vector in a normal and a parallel frame.

3.1. INTRODUCTION

In this chapter, we delve into one of the fundamental problems of nonlinear analysis: root-finding on Riemannian manifolds. Formally, our goal is to address the problem:

$$\text{Find } p \in \mathcal{M} \text{ such that } V(p) = 0, \tag{$*$}$$

where $\mathcal{M}$ denotes a d-dimensional Riemannian manifold and V is a vector field defined on $\mathcal{M}$. Root-finding problems of this nature are not only theoretically significant but also have a wide range of applications in various scientific fields. For instance, local optimization falls into this category by setting $V = -\nabla f$ for some smooth function f on $\mathcal{M}$. For comprehensive insights and foundational knowledge about optimization algorithms on manifolds and their applications, the reader is encouraged to consult the books by Absil, Mahony, and Sepulchre [AMS08] and Boumal [Bou23].

The importance of the generalized root-finding problem ($*$) extends beyond traditional optimization tasks, encapsulating countless applications in machine learning, game theory, and beyond. Practical instances include bi-level and saddle point problems, dynamic programming, and equilibrium finding in games.

As we transition from Euclidean to Riemannian settings, it is crucial to note the inherent complexities and differences that arise. Firstly, the iterative algorithms for solving ($*$) are influenced by the geometric structure of the manifold. For instance, in the Euclidean context ($\mathcal{M} = \mathbb{R}^d$) the Robbins–Monro stochastic approximation algorithm

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \alpha_n(V(\boldsymbol{x}_n) + Z_n) \tag{3.1}$$

is highly effective for root-finding and optimization. However, this update relies heavily on the linear structure of the Euclidean space, making its direct application in Riemannian settings infeasible. The curved geometry of Riemannian manifolds necessitates modifications to the update rule—as well as the theoretical proofs—to accommodate the manifold's intrinsic properties.

Our primary aim in this chapter is to bridge the gap between Euclidean and Riemannian stochastic approximation schemes. We achieve this by replacing the $+$ operation in (3.1) with the Riemannian exponential map, or more generally—and often more feasibly—a retraction. In Riemannian optimization, this methodology was first explored by Bonnabel [Bon13] for cases where the vector field V is the Riemannian gradient of an objective function. Subsequent studies expanded on Bonnabel's results specifically for Riemannian stochastic gradient descent (SGD) and Riemannian proximal point methods; see the bibliographic notes at the end of this chapter for references and pointers to these studies.

This body of literature exclusively considers cases where V is a gradient field and the results do not apply to general, non-gradient instances. There are, however, partial extensions to the non-gradient case. For instance, there is an array of works studying Riemannian extra-gradient methods under the assumption of geodesic monotonicity [see, e.g., FPN05]. Geodesic monotonicity is a strong, convexity-type assumption asserting that V globally points towards its roots in a suitable sense.

In our work, we do not assume geodesic monotonicity. Instead, we directly analyze the dynamics of Riemannian Robbins–Monro methods for general vector fields. Our main contributions are as follows:

(1) We propose a generalized framework that encompasses all previously mentioned methods (Riemannian SGD, extra-gradient, proximal point method, etc.) as special cases, and opens the possibility to introduce new stochastic approximation schemes for the root-finding problem ($*$).

(2) Under mild technical conditions on the manifold, we demonstrate that the sequence of the iterates of the algorithm forms an asymptotic pseudo-trajectory of an associated deterministic flow.

This result extends the foundational theory by Benaïm and Hirsch [BH96] for Euclidean Robbins–Monro schemes to Riemannian settings, enabling us to establish the almost sure convergence of Riemannian Robbins–Monro schemes to the internally chain-transitive sets of the underlying Riemannian dynamics. If V is gradient-like or strictly monotone, these internally chain-transitive sets correspond to the roots of V, thus recovering many asymptotic convergence results from earlier works, often under less restrictive assumptions. Furthermore, our framework, as discussed in Section 3.3, applies to various settings beyond gradient or monotone systems—such as non-convex potential games—and encompasses a wider class of stochastic approximation algorithms.

Given the absence of linear structure on $\mathcal{M}$, our primary challenge is the lack of a coordinate system suitable for analyzing the trajectories of Riemannian stochastic approximation algorithms. Unlike in $\mathbb{R}^d$, points and vectors on manifolds follow fundamentally different rules and must be compared with extra care. To overcome this challenge, we introduce a *Fermi coordinate system*, inspired by Manasse and Misner [MM63]. This framework allows us to demonstrate that Riemannian stochastic approximation schemes achieve similar error bounds to those in Euclidean spaces, albeit with some higher-order terms that diminish over time. Controlling the aggregation and propagation of these errors involves applying martingale theory, which ultimately results in the convergence properties discussed.

Chapter Roadmap

This chapter establishes the framework for formulating and analyzing stochastic approximation algorithms within a Riemannian context.

We begin in Section 3.2 by reviewing foundational concepts and notations in differential and Riemannian geometry. In Section 3.3, we provide specific examples of stochastic approximation on Riemannian manifolds, illustrating the practical relevance and challenges of nonlinear root-finding in machine learning and games. In Section 3.4, we discuss the adaptation of iterative methods to manifold settings, state our assumptions, and present two main theorems: one regarding the asymptotic pseudo-trajectory property and one about the stability of a stochastic approximation algorithm. An extensive proof of these two theorems is presented in Sections 3.5 and 3.6. In Section 3.7, we analyze the convergence behavior of four widely used stochastic approximation algorithms that satisfy our assumptions. Section 3.8 discusses two algorithmic variations—retractions and alternation—used in practice. Retraction is an approximation of the exponential map, and alternation is relevant in multiplayer game settings where players update their states sequentially. We show that these variations introduce extra bias, which can be managed similarly to other types of bias. We explore practical implications of our main theorems for optimization and games in Section 3.9 and conclude the chapter in Section 3.10. References and further reading are provided in the bibliographic notes at the end of the chapter.

3.2. A CRASH COURSE ON RIEMANNIAN GEOMETRY

In this section, we introduce some key concepts from differential and Riemannian geometry. Our primary purpose is to establish notational conventions and briefly mention essential ideas and notions that are used in the sequel. To maintain clarity and conciseness, we opt for a presentation style that does not delve into rigorous definitions for certain concepts. The content of this section reflects a highly subjective selection and is by no means comprehensive. For readers seeking a more thorough grounding in differential and Riemannian geometry, we strongly recommend consulting the masterpieces of Lee [Lee12; Lee18], Jost [Jos17], and do Carmo [Car92].

3.2.1. Charts and Tangent Vectors

We start with the basics of differential geometry. This includes the notion of a smooth manifold, charts, and tangent vectors. The reader that is not acquainted with differential geometry can safely think about a smooth manifold as a “smooth” subset of some Euclidean space.

Smooth manifolds. A d-dimensional *topological manifold* $\mathcal{M}$ is a topological space that locally looks like some open set in $\mathbb{R}^d$. This means that for every point $p \in \mathcal{M}$, there exists a neighborhood $\mathcal{U} \subset \mathcal{M}$ and a *homeomorphism* φ between $\mathcal{U}$ and some open set in $\mathbb{R}^d$; that is, φ is a continuous bijection with continuous inverse. We call $(\mathcal{U}, \varphi)$ a *coordinate chart* (or simply a *chart*), as it assigns d coordinates to each point in $\mathcal{U}$. A collection of charts whose domains cover $\mathcal{M}$ is called an *atlas* for $\mathcal{M}$.

To be able to do differential calculus on the manifold, we need to enforce some differentiability conditions on the atlas. An atlas is called a *smooth atlas* if any two charts $(\mathcal{U}, \varphi)$ and $(\mathcal{V}, \psi)$ in that atlas are *smoothly compatible*, which means that either the domains of φ and ψ are disjoint, or the map $\psi \circ \varphi^{-1}$ is a *diffeomorphism* on its domain, i.e., $\psi \circ \varphi^{-1} : \varphi(\mathcal{U} \cap \mathcal{V}) \to \psi(\mathcal{U} \cap \mathcal{V})$ is a smooth bijection with smooth inverse. The word “smooth” in this thesis means C^∞. However, in many contexts, one can replace C^∞ with C^k for some $k \geq 1$, and state the same results for manifolds with a C^k atlas.

Tangent vectors. A function $f : \mathcal{M} \to \mathbb{R}$ is *smooth* if for any chart φ, the function $f \circ \varphi^{-1}$ is smooth in the Euclidean sense. We denote by $C^\infty(\mathcal{M})$ the set of all smooth functions defined on the manifold $\mathcal{M}$. A linear map $v : C^\infty(\mathcal{M}) \to \mathbb{R}$ is called a *derivation at* p, if it satisfies the Leibniz rule:

$$v(fg) = f(p)v(g) + g(p)v(f), \quad \text{for all } f, g \in C^\infty(\mathcal{M}).$$

The set of all derivations at p is called the *tangent space at* p and is denoted by $T_p\mathcal{M}$. We call a derivation $v \in T_p\mathcal{M}$ a *tangent vector at* p. It turns out that the tangent space at any point is a vector space of the same dimension as $\mathcal{M}$.

In the special case where $\mathcal{M} = \mathbb{R}^d$, there are d derivations that we know from calculus: for $a \in \mathbb{R}^d$, define the derivations $\partial/\partial x^1|_a, \ldots, \partial/\partial x^d|_a$ via

$$\left.\frac{\partial}{\partial x^i}\right|_a f = \frac{\partial f}{\partial x^i}(a) \qquad \text{for } i = 1, \ldots, d \text{ and all } f \in C^\infty(\mathbb{R}^d), \tag{3.2}$$

where the right-hand side is the usual partial derivative. In this case, $\{\partial/\partial x^i|_a\}$ forms a basis of $T_a\mathbb{R}^d$.

Let $F : \mathcal{M} \to \mathcal{N}$ be a smooth map between two manifolds $\mathcal{M}$ and $\mathcal{N}$. That is, for any point $p \in \mathcal{M}$ along with a chart φ around p, and a chart ψ around $F(p)$ in $\mathcal{N}$, the function $\psi \circ F \circ \varphi^{-1}$ is a smooth function between two open sets in the Euclidean sense. The *differential* of F at the point $p \in \mathcal{M}$ evaluated at the tangent vector $v \in T_p\mathcal{M}$ is the derivation $dF_p(v)$, defined as

$$dF_p(v)(f) = v(f \circ F), \quad \text{for all } f \in C^\infty(\mathcal{N}). \tag{3.3}$$

This means that $dF_p : T_p\mathcal{M} \to T_{F(p)}\mathcal{N}$. As a coordinate chart $(\mathcal{U}, \varphi)$ is a diffeomorphism between $\mathcal{U}$ and $\varphi(\mathcal{U})$, its differential $d\varphi_p : T_p\mathcal{M} \to T_{\varphi(p)}\mathbb{R}^d$ at any point p becomes a vector space isomorphism. Let $\partial/\partial x^1|_{\varphi(p)}, \ldots, \partial/\partial x^d|_{\varphi(p)}$ be the basis for $T_{\varphi(p)}\mathbb{R}^d$ defined as in (3.2). The preimage of these tangent vectors under $d\varphi_p$ forms a basis for $T_p\mathcal{M}$, denoted by the similar notation $\partial/\partial x^i|_p$:

$$\left.\frac{\partial}{\partial x^i}\right|_p = (d\varphi_p)^{-1}\left(\left.\frac{\partial}{\partial x^i}\right|_{\varphi(p)}\right) = d(\varphi^{-1})_{\varphi(p)}\left(\left.\frac{\partial}{\partial x^i}\right|_{\varphi(p)}\right).$$

We also write $\partial_i|_p$ for $\partial/\partial x^i|_p$. The basis $\{\partial/\partial x^1|_p, \ldots, \partial/\partial x^d|_p\}$ is sometimes called the *basis of* $T_p\mathcal{M}$ *induced by the coordinate chart* φ. The action of these basis vectors on smooth functions is simply

$$\left.\frac{\partial}{\partial x^i}\right|_p f = \left.\frac{\partial}{\partial x^i}\right|_{\varphi(p)} (f \circ \varphi^{-1}) = \frac{\partial \widehat{f}}{\partial x^i}(\widehat{p}), \quad \text{for all } f \in C^\infty(\mathcal{U}),$$

where $\widehat{f} = f \circ \varphi^{-1}$ and $\widehat{p} = \varphi(p)$ are the coordinate representations of f and p, respectively. We can also express any tangent vector in this basis: $v = \sum v^i \frac{\partial}{\partial x^i}|_p$, and call v^i the ith *component* of v in this basis. The components can be computed using the coordinate functions: writing $\varphi(p) = (x^1(p), \ldots, x^d(p))$ with $x^i : \mathcal{U} \to \mathbb{R}$, we have that

$$v^i = v(x^i). \tag{3.4}$$

Curves and their velocity. By a *smooth curve* on $\mathcal{M}$ we mean a smooth function γ from an open interval $I \subset \mathbb{R}$ to $\mathcal{M}$. It is customary to denote the basis tangent vector of I at $t_0 \in I$ by $\partial/\partial t|_{t_0}$. Being a smooth map, the differential $d\gamma_{t_0} : T_{t_0}I \to T_{\gamma(t_0)}\mathcal{M}$ satisfies

$$d\gamma_{t_0}\left(\left.\frac{\partial}{\partial t}\right|_{t_0}\right)(f) = \left.\frac{\partial}{\partial t}\right|_{t_0}(f \circ \gamma) = \frac{d(f \circ \gamma)}{dt}(t_0), \tag{3.5}$$

where the last quantity is the usual time derivative of the function $f \circ \gamma : I \to \mathbb{R}$. Inspired by this, we call the tangent vector $d\gamma_{t_0}(\partial/\partial t|_{t_0})$ the *velocity vector of* γ *at time* t_0 and denote it by $\dot{\gamma}(t_0)$.

Now consider a coordinate chart around $\gamma(t_0)$, and for t close to t_0 let us denote the coordinates of $\gamma(t)$ as $\widehat{\gamma}(t) = (\gamma^1(t), \ldots, \gamma^d(t))$. Using Eqs. (3.4) and (3.5), we have

$$\dot{\gamma}(t) = \sum \dot{\gamma}^i(t) \frac{\partial}{\partial x^i}\bigg|_{\gamma(t)}, \quad \text{with} \quad \dot{\gamma}^i(t) = \dot{\gamma}(t)(x^i) = \frac{d(x^i \circ \gamma)}{dt} = \frac{d\gamma^i}{dt}(t). \tag{3.6}$$

This shows that $\dot{\gamma}(t)$ is essentially given by the same formula as in Euclidean spaces: it is the tangent vector whose components in a coordinate basis are the derivatives of the component functions of $\gamma(t)$.

Vector fields. By "gluing" all tangent spaces at different points of a manifold $\mathcal{M}$, one obtains the *tangent bundle* $T\mathcal{M}$. Concretely, the tangent bundle is the disjoint union

$$T\mathcal{M} = \bigsqcup_{p\in\mathcal{M}} T_p\mathcal{M},$$

and is a smooth manifold of dimension $2d$. The elements of $T\mathcal{M}$ can be thought of as tuples (p, v), where $p \in \mathcal{M}$ is called the base point of (p, v) and $v \in T_p\mathcal{M}$. We let π to be the function that projects a point (p, v) in the tangent bundle to its base point p. When the base point is clear from the context, we only use v to denote an element of $T\mathcal{M}$.

A *smooth vector field* is a "smooth assignment" of a tangent vector at every point of the manifold. More precisely, it is a smooth function V from $\mathcal{M}$ to $T\mathcal{M}$, such that $\pi \circ V$ is the identity map on $\mathcal{M}$. We denote by $\mathfrak{X}(\mathcal{M})$ the space of all smooth vector fields on $\mathcal{M}$.

There are two algebraic operations that relate vector fields to $C^\infty(\mathcal{M})$. Firstly, one can multiply a vector field V by a smooth function f to obtain a new vector field:

$$(fV)(p) := f(p)V(p).$$

This operation simply scales the vector field at each point as dictated by the value of the function at that point. Secondly, one can apply a vector field V to a smooth function f to obtain a new smooth function:

$$(Vf)(p) := V(p)(f).$$

That is, the value of the Vf at the point p is the result of applying the derivation $V(p)$ to f. We usually denote by $V(p)$ the value of the vector field at the point p. In some places (for example, when describing the Fermi coordinates in Section 3.5), we use the notation V_p, and reserve $V(f)$ for the applying a vector field to a function f for better readability.

For a smooth function $f : \mathcal{M} \to \mathbb{R}$ and a pair of vector fields $X, Y \in \mathfrak{X}(\mathcal{M})$, consider the functions $X(Yf)$ and $Y(Xf)$. In general, these functions cannot be described as the application of one vector field to f, as they involve derivatives of second order. However, if we consider $X(Yf) - Y(Xf)$, it turns out that the result is again a vector field, called the *Lie bracket* of X and Y and is denoted by $[X, Y]$. Lie brackets show up in different places, including the definition of curvature, which we discuss below.

3.2.2. Riemannian Metrics

We now aim to establish a metric structure on a smooth manifold with the goal of measuring length of curves and angles between tangent vectors. A *Riemannian metric* on a smooth manifold $\mathcal{M}$ is an assignment of an inner product to each tangent space $T_p\mathcal{M}$ which depends smoothly on the base point p. A *Riemannian manifold* is a smooth manifold equipped with a Riemannian metric. We will denote the scalar product of $v, w \in T_p\mathcal{M}$ by $\langle v, w\rangle_p$, and drop the subscript p if it is clear that v and w belong to which tangent space.

Consider a coordinate chart $(\mathcal{U}, \varphi)$ around p and let $\widehat{p} = \varphi(p)$. In these coordinates, the metric at the point p is defined via a symmetric positive definite matrix $(g_{ij}(\widehat{p}))_{i,j=1,\ldots,d}$, and for tangent vectors $v = \sum v^i \partial_i|_p$ and $w = \sum w^j \partial_j|_p$ in $T_p\mathcal{M}$ it holds

$$\langle v, w\rangle_p = \sum_i \sum_j g_{ij}(\widehat{p}) v^i w^j.$$

In particular, we have $g_{ij}(\widehat{p}) = \langle \partial_i|_p, \partial_j|_p\rangle$.

The *norm of a tangent vector* v is defined as

$$|v| = \langle v, v\rangle^{1/2}$$

and the *length of a curve* $\gamma : I \to \mathcal{M}$ is therefore defined as

$$L(\gamma) = \int_I |\dot{\gamma}(t)|\, dt = \int_I \langle \dot{\gamma}(t), \dot{\gamma}(t)\rangle^{1/2}\, dt.$$

Notice that this definition is independent of the parameterization of the curve, and can be extended to piecewise-smooth curves by defining the length of such curves to be the sum of the lengths of their smooth parts. Having a notion of length for curves, we can define the *distance function* on a Riemannian manifold:

$$d(p, q) = \inf\{L(\gamma) : \gamma : [a, b] \to \mathcal{M}, \gamma(a) = p, \gamma(b) = q\},$$

where the infimum is over all piecewise-smooth curves with endpoints p and q. One can show that this distance function satisfies the metric axioms, and the topology induced by this metric coincides with the topology of the manifold.

3.2.3. Connections and Covariant Derivatives

Geodesics, as we will see later, are generalizations of straight lines in Euclidean spaces to Riemannian manifolds. A defining property of a straight line is that it has zero acceleration. While this notion is easy to define in a Euclidean sense, it turns out to be nontrivial for Riemannian manifolds. To make sense of acceleration on a manifold, we have to introduce two new objects: an affine connection will give us a set of rules for taking directional derivatives of vector fields, and a covariant derivative allows us to make sense of time-derivative of a vector field along a curve.

Before getting into the definitions, let us build some intuition by considering a simple example. Take a two-dimensional surface $\mathcal{M} \subset \mathbb{R}^3$, as well as a parameterized smooth curve γ on $\mathcal{M}$. Let V be a vector field along γ and tangent to $\mathcal{M}$; that is, $V(t) \in T_{\gamma(t)}\mathcal{M}$ for all t that $\gamma(t)$ is defined. Our goal is to understand how V changes over time by making sense of $\frac{dV}{dt}(t)$. As everything lives in $\mathbb{R}^3$, we can think of $V(t)$ at different times as vectors in $\mathbb{R}^3$, and this allows us to define $\frac{dV}{dt}(t)$ in the usual Euclidean way: $\frac{dV}{dt}(t) = \lim_{h \to 0} h^{-1}(V(t+h) - V(t))$. However, as the surface is curved, the resulting vector may not belong to the tangent plane at the point $\gamma(t)$. This means that differentiating a vector field in this way is not an inherent geometric attribute within $\mathcal{M}$. To address this issue, we can project $\frac{dV}{dt}(t)$ onto the tangent space $T_{\gamma(t)}\mathcal{M}$ orthogonally, giving rise to what is known as the *covariant derivative*, and is denoted by $D_t V$. In other words, the covariant derivative of V is the time-derivative of V from the manifold's perspective.

To define the covariant derivative without resorting to embeddings and orthogonal projections, we first need to define an affine connection. An *affine connection* ∇ is a mapping $\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})$ denoted as $(X, Y) \mapsto \nabla_X Y$, satisfying the following properties:

(1) $\nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z$,

(2) $\nabla_X(aY + bZ) = a\nabla_X Y + b\nabla_X Z$,

(3) $\nabla_X(fY) = f\nabla_X Y + X(f)\, Y$,

in which $a, b \in \mathbb{R}$, $X, Y, Z \in \mathfrak{X}(\mathcal{M})$ and $f, g \in C^\infty(\mathcal{M})$. It turns out that the value of $\nabla_X Y|_p$ only depends on $X(p)$ and values of Y in an arbitrary small neighborhood around p; this makes $\nabla_X Y$ a *local operator*. In a local coordinate chart, it is customary to define

$$\nabla_{\partial_i} \partial_j = \sum_k \Gamma_{ij}^k \partial_k,$$

where Γ_{ij}^k are called the *connection coefficients* of ∇. With this notation we have

$$\nabla_X Y = \sum_k \Big(X(Y^k) + \sum_{i,j} X^i Y^j \Gamma_{ij}^k \Big) \partial_k .$$

Let us now go back to our initial problem: differentiating vector fields along curves. Suppose ∇ is an affine connection of $\mathcal{M}$. Then, there exists a unique correspondence that associates with every vector field V along a curve $\gamma : I \to \mathcal{M}$ another vector field $D_t V$ along γ, called *covariant derivative of* V *along* γ, such that for any other vector field W along γ and $a, b \in \mathbb{R}$, it holds

(1) $D_t(aV + bW) = aD_tV + bD_tW$, and

(2) $D_t(fV) = \frac{df}{dt}V + fD_tV$ for all smooth functions $f : I \to \mathbb{R}$.

Moreover, if V is the restriction of a vector field $X \in \mathfrak{X}(\mathcal{M})$ to γ, that is, $V(t) = X(\gamma(t))$, then

$$D_tV(t) = \nabla_{d\gamma/dt} X|_{\gamma(t)} .$$

Let us compute the covariant derivative in local coordinates. By taking a chart around $\gamma(t)$, we can write $V(t) = \sum V^i(t)\, \partial_i|_{\gamma(t)}$ and $\dot\gamma(t) = \sum \dot\gamma^i(t)\, \partial_i|_{\gamma(t)}$. By the properties of the covariant derivative, we have

$$D_tV(t) = \sum_i \Big(\frac{dV^i}{dt}(t)\, \partial_i|_{\gamma(t)} + V^i(t) D_t(\partial_i)|_{\gamma(t)} \Big).$$

As ∂_i along $\gamma(t)$ is the restriction of ∂_i, $D_t(\partial_i) = \nabla_{\dot\gamma(t)}(\partial_i)$. Writing $\dot V^k$ for $\frac{dV^k}{dt}$, we get the expression of covariant derivative in local coordinates:

$$D_tV(t) = \sum_k \Big(\dot V^k(t) + \sum_{i,j} \dot\gamma^i(t) V^j(t) \Gamma_{ij}^k(\gamma(t)) \Big) \frac{\partial}{\partial x^k}\bigg|_{\gamma(t)} . \tag{3.7}$$

3.2.4. The Levi-Civita Connection

The affine connection defines a notion of directional differentiation of vector fields on Riemannian manifolds. However, this notion is not unique. By imposing additional constraints that have to do with the Riemannian metric, a unique connection, called the Levi-Civita connection, emerges.

We say a connection ∇ is *compatible with the metric* g (or it preserves the metric g), if for any pair of vector fields V, W along a curve γ, the following holds:

$$\frac{d}{dt}\langle V, W\rangle = \langle D_tV, W\rangle + \langle V, D_tW\rangle .$$

Equivalently, for any vector fields $X, V, W \in \mathfrak{X}(\mathcal{M})$, it should hold that

$$X\langle V, W\rangle = \langle \nabla_X V, W\rangle + \langle V, \nabla_X W\rangle.$$

This condition holds for the Euclidean space along with Euclidean differentiation. We will see later that this condition has deep connections with the notion of parallelism on Riemannian manifolds.

The other condition that facilitates the relation between the connection and the Riemannian metric is symmetry. In Euclidean spaces, we can exchange $\partial/\partial x^i$ and $\partial/\partial x^j$, in the sense that for any smooth function f, we have $\frac{\partial}{\partial x^i}\frac{\partial}{\partial x^j} f = \frac{\partial}{\partial x^j}\frac{\partial}{\partial x^i} f$. This comes from the symmetry between differentiation along different coordinates. The equivalent notion for Riemannian manifolds is to require

$$\nabla_{\partial_i}\partial_j = \nabla_{\partial_j}\partial_i.$$

More generally, we call a connection ∇ *symmetric* (or torsion-free) if for any pairs of vector fields $X, Y \in \mathfrak{X}(\mathcal{M})$, it holds

$$\nabla_X Y - \nabla_Y X = [X, Y] = XY - YX.$$

This property implies the symmetric relation $\Gamma^k_{ij} = \Gamma^k_{ji}$ on the connection coefficients of ∇.

The fundamental theorem of Riemannian geometry states that every Riemannian manifold $(\mathcal{M}, g)$ has a unique connection ∇, called the *Levi-Civita connection*, that is both compatible to the metric g and is symmetric. From this point onwards, ∇ will always denote the Levi-Civita connection for a Riemannian manifold, and the metric coefficients Γ^k_{ij} are called the *Christoffel symbols*.

3.2.5. Parallel Transport

The covariant derivative defines a notion of parallelism in Riemannian manifolds. We say a vector field V along the curve $\gamma : I \to \mathcal{M}$ is *parallel* if its covariant derivative vanishes everywhere, that is,

$$D_t V(t) = 0, \quad \forall t \in I.$$

Let us examine this property in a coordinate chart $(\mathcal{U}, \varphi)$. Suppose $V(t) = \sum V^i(t)\,\partial_i|_{\gamma(t)}$ and $\dot\gamma(t) = \sum \dot\gamma^i(t)\,\partial_i|_{\gamma(t)}$. From (3.7) we obtain

$$\dot V^k(t) + \sum_{i,j} \dot\gamma^i(t) V^j(t) \Gamma^k_{ij}(\gamma(t)) = 0, \quad \text{for all } k = 1, \dots, d \text{ and } t \in I. \tag{3.8}$$

This is a system of linear ODEs, and given some initial data, one can show that it always has a unique solution. That is, for any tangent vector $v \in T_{\gamma(t_0)}\mathcal{M}$, there exists a unique parallel vector field V along γ such that $V(t_0) = v$. This vector field is called the *parallel transport of v along γ*. Supposing that $\gamma(t_0) = p$ and $\gamma(t_1) = q$, we refer to the vector $V(t_1)$ as $\mathrm{P}^{\gamma}_{p\to q}[v]$. If the curve γ is clear from the context, we just write $\mathrm{P}_{p\to q}[v]$ without mentioning the curve.

A key property of parallel transport which we use frequently is that it creates an isometry between tangent spaces. To see this, first observe that due to symmetry of the Levi-Civita connection, for two parallel vector fields V, W along a curve γ it holds

$$\frac{d}{dt}\langle V, W\rangle = \langle D_t V, W\rangle + \langle V, D_t W\rangle = 0.$$

This means that $|V(t)|$ is constant along the curve, and if $V(t)$ and $W(t)$ are orthogonal for some t, they are orthogonal for all t along the curve. Now, let $\gamma : [0,1] \to \mathcal{M}$ be a curve between $p, q \in \mathcal{M}$, and consider an orthonormal basis $\{E_1, \ldots, E_d\}$ of $T_p\mathcal{M}$; this basis can be obtained by starting from an arbitrary basis of $T_p\mathcal{M}$ and running the Gram–Schmidt algorithm. Parallel transporting each of these basis vectors along γ results in a basis $\{E_1(t), \ldots, E_d(t)\}$ for $T_{\gamma(t)}\mathcal{M}$ for all $t \in [0,1]$; this is because the parallel transport keeps $E_i(t)$ unit and orthogonal to the rest. This implies that the map $v \mapsto \mathrm{P}^{\gamma}_{p\to q}[v]$ from $T_p\mathcal{M}$ to $T_q\mathcal{M}$ is a vector space isomorphism, as well as an isometry. In short, for any $v, w \in T_p\mathcal{M}$, it holds $\langle v, w\rangle_p = \langle \mathrm{P}^{\gamma}_{p\to q}[v], \mathrm{P}^{\gamma}_{p\to q}[w]\rangle_q$ for all $v, w \in T_p\mathcal{M}$.

3.2.6. Geodesics and the Exponential Map

A geodesic is a generalization of a straight line in Euclidean spaces to Riemannian manifolds. As discussed before, the key defining property of a geodesic is that it has zero acceleration. Now that we have a working notion of a covariant derivative, this leads to the following definition: a curve $\gamma : I \to \mathcal{M}$ is a *geodesic* if

$$D_t\dot\gamma = \nabla_{\dot\gamma}\dot\gamma\big|_{\gamma(t)} = 0, \quad \text{for all } t \in I.$$

Note that this is equivalent to saying that the velocity vector of a geodesic is parallel along the geodesic. This implies that a geodesic is always constant speed; that is, $|\dot\gamma(t)|$ stays constant for $t \in I$.

A geodesic $\gamma : [a,b] \to \mathcal{M}$ is not necessarily the shortest curve that connects $\gamma(a)$ and $\gamma(b)$. Think of moving around the equator of Earth with constant speed until you reach where you started. This movement has zero acceleration (in the Riemannian sense), but surely is not the shortest curve connecting a point to itself. It is the case, however, the shortest curve property holds locally. That is, for t close enough to a, the curve γ restricted to $[a,t]$ is the shortest curve connecting $\gamma(a)$

to $\gamma(t)$. We call a geodesic that is also the shortest curve between its endpoints a *minimizing geodesic*.

In terms of coordinates, suppose we write $\dot{\gamma}(t) = \sum \dot{\gamma}^i(t)\, \partial_i|_{\gamma(t)}$. From (3.8) we have

$$\ddot{\gamma}^k(t) + \sum_{i,j} \dot{\gamma}^i(t)\dot{\gamma}^j(t)\Gamma^k_{ij}(\gamma(t)) = 0, \quad \text{for all } k = 1, \dots, d \text{ and all } t \in I. \tag{3.9}$$

By the usual existence and uniqueness results for ODEs, one can show that for any pair of initial values for position $\gamma(t_0)$ and velocity $\dot{\gamma}(t_0)$, there exists a unique and maximal solution to the system of ODEs above. Therefore, each initial point $p \in \mathcal{M}$ and each initial velocity vector $v \in T_p\mathcal{M}$ determine a unique maximal geodesic, which we denote by γ_v. We define the *exponential map* $\exp : \mathcal{E} \subseteq T\mathcal{M} \to \mathcal{M}$ as

$$\exp(v) = \gamma_v(1),$$

where the domain $\mathcal{E}$ is defined as

$$\mathcal{E} = \{v \in T\mathcal{M} : \gamma_v \text{ is defined on an interval containing } [0, 1]\}.$$

It turns out that exp is a smooth function between $\mathcal{E}$ and $\mathcal{M}$. In the sequel, we usually work with the restriction of exp to a tangent space of a point. We denote by $\exp_p$ the map

$$\exp_p(v) = \exp(v) \quad \text{for} \quad v \in \mathcal{E}_p,$$

where $\mathcal{E}_p = \{v \in \mathcal{E} : \pi(v) = p\}$. Note that for each $v \in T_p\mathcal{M}$, the geodesic γ_v is given by

$$\gamma_v(t) = \exp(tv)$$

for all t such that either side is defined.

A Riemannian manifold is called *complete*, if geodesics exist globally, or equivalently, the exponential map is defined on the whole tangent bundle $T\mathcal{M}$. Throughout this chapter, we only work with complete manifolds, and therefore, we adapt further notions and omit the discussions where the exponential map might not be defined in some places.

Another important property of the exponential map is that its differential $d(\exp_p)_0 : T_0(T_p\mathcal{M}) \to T_p\mathcal{M}$ at the vector 0 is the identity map of $T_p\mathcal{M}$. That is, by identifying $T_0(T_p\mathcal{M})$ with $T_p\mathcal{M}$, it holds

$$d(\exp_p)_0(v) = v.$$

This property will be crucial for using the inverse function theorem and the construction of normal coordinates, which we describe next.

3.2.7. Product Manifolds

The product $\mathcal{M} := \mathcal{M}_1 \times \mathcal{M}_2$ of two smooth manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ is also a smooth manifold. Most importantly, the tangent space at the point $p = (p_1, p_2) \in \mathcal{M}$ is (isomorphic to) the direct sum of the tangent spaces $T_{p_1}\mathcal{M}_1$ and $T_{p_2}\mathcal{M}_2$:

$$T_p\mathcal{M} \cong T_{p_1}\mathcal{M}_1 \oplus T_{p_2}\mathcal{M}_2.$$

Therefore, we can think of $v \in T_p\mathcal{M}$ as the pair (v_1, v_2) where $v_i \in T_{p_i}\mathcal{M}_i$ for $i = 1, 2$. If $\mathcal{M}_1, \mathcal{M}_2$ are Riemannian manifolds, $\mathcal{M}$ can be turned into a Riemannian manifold by defining the metric at the point $p = (p_1, p_2) \in \mathcal{M}$ as

$$\langle v, w\rangle_p = \langle v_1, w_1\rangle_{p_1} + \langle v_2, w_2\rangle_{p_2}, \quad \forall v = (v_1, v_2), w = (w_1, w_2) \in T_p\mathcal{M}.$$

Moreover, the exponential map at the point p for the tangent vector $v = (v_1, v_2) \in T_p\mathcal{M}$ is equal to

$$\exp_p(v) = (\exp_{p_1}(v_1), \exp_{p_2}(v_2)).$$

All these constructions can be generalized to an arbitrary product of finitely many Riemannian manifolds.

3.2.8. Normal Coordinates

Normal coordinates are a specific coordinate chart around a point p that is “aligned” with geodesics; in these coordinates, geodesics passing through p are mapped to lines passing through the origin in the Euclidean space. This makes computing the distance of p to the points in its neighborhood as easy as computing norms in Euclidean spaces.

Before defining such coordinate system, we have to make sense of the *inverse* of the exponential map; the map that assigns to each point q in the neighborhood of p a tangent vector v, so that the geodesic starting from p in the direction v meets q at time 1. As the differential $d(\exp_p)_0$ is the identity of $T_p\mathcal{M}$, by the inverse function theorem one can find a star-shaped[1] neighborhood $\mathcal{V} \subseteq T_p\mathcal{M}$ around 0 such that $\exp_p$ is a *diffeomorphism* between $\mathcal{V}$ and $\exp_p(\mathcal{V})$. This implies that the exponential map is invertible locally.

Now, choose an arbitrary orthonormal frame $(E_1, \dots, E_d)$ for $T_p\mathcal{M}$, that is, $\langle E_i, E_j\rangle = \delta_{ij}$. Using the isomorphism $B : \mathbb{R}^d \to T_p\mathcal{M}$ given by

$$B : (a^1, \dots, a^d) \mapsto a^1E_1 + \cdots + a^dE_d,$$

[1] A subset S of a vector space is said to be *star-shaped with respect to a point x* if for every $y \in S$, the line segment from x to y is also contained in S.

define the chart $(\mathcal{U}, \varphi)$ as

$$\varphi(q) = B^{-1}(\exp_p^{-1}(q)),$$

where $\mathcal{U}$ is the image of a star-shaped $\mathcal{V} \subseteq T_p\mathcal{M}$ under $\exp_p$ for which $\exp_p$ is a diffeomorphism. We call this chart a *normal coordinate system centered at* p. In simpler words, if $q = \exp_p(v)$ with $v = \sum v^i E_i \in \mathcal{V}$, the normal coordinates of q will be $(v^1, \dots, v^d) \in \mathbb{R}^d$.

A nice property of this coordinate system, as promised in the beginning, is that the geodesic equation for the geodesic γ_v passing through p at $t = 0$ with velocity v is simply

$$\widehat{\gamma}_v(t) = (tv^1, \dots, tv^d).$$

This also implies that for any other point $q \in \mathcal{U}$, we have

$$d(p, q) = \|\widehat{q}\|_2 \coloneqq ((\widehat{q}^1)^2 + \cdots + (\widehat{q}^d)^2)^{1/2}.$$

3.2.9. Conjugate and Cut Points

To define the normal coordinates centered at $p \in \mathcal{M}$, it was essential to work with star-shaped subsets of $T_p\mathcal{M}$ on which the exponential map is a diffeomorphism onto its image. This motivates considering the largest open ball (in $T_p\mathcal{M}$) centered at 0 on which $\exp_p$ is a diffeomorphism. The *injectivity radius at* $p \in \mathcal{M}$, denoted by $r_{\text{inj}}(p)$, is the supremum of all radii $r > 0$ such that $\exp_p$ is a diffeomorphism from $B_r(0) \subset T_p\mathcal{M}$ onto its image. The *injectivity radius of* $\mathcal{M}$, denoted by $\text{inj}(\mathcal{M})$, is defined to be the infimum of $r_{\text{inj}}(p)$ over all points p of the manifold. Therefore, if $\text{inj}(\mathcal{M}) = r > 0$, we are able to construct a normal neighborhood around an arbitrary point $p \in \mathcal{M}$ that is guaranteed to contain all points that are r-close to p.

When $\exp_p$ fails to be a diffeomorphism, interesting (and very often, problematic) things happen. Let $v \in T_p\mathcal{M}$ and $\gamma = \gamma_v$ be the geodesic segment $\gamma(t) = \exp_p(tv)$, and set $q = \gamma(1)$. We say p and q are *conjugate along* γ if v is a critical point[2] of $\exp_p$, that is, if $d(\exp_p)_v$ is not surjective.[3] It turns out that if q is the first point along γ that is conjugate to p, then γ is *not minimizing* past q. One example to keep in mind is that any pair of antipodal points on the two-dimensional sphere (such as Earth) are conjugate points.

[2] A point x is a *critical point* of a smooth map F, if the differential dF_x is not surjective.

[3] Conjugate points are usually defined using Jacobi fields. For example, Lee [Lee18] defines conjugate points as follows: Let $\gamma : I \to \mathcal{M}$ be a geodesic and $p = \gamma(a), q = \gamma(b)$ for some $a, b \in I$. We say that p and q are *conjugate along* γ if there is a nonzero Jacobi field along γ vanishing at $t = a$ and $t = b$. By [Lee18, Prop. 10.20, Thm. 10.26], our definition and the one from Lee are equivalent.

Let $\gamma : [a, b] \to \mathcal{M}$ be a geodesic in $\mathcal{M}$, and $p = \gamma(a)$ and $q = \gamma(b)$ be its endpoints. Whether γ is a minimizing geodesic is a subtle question. If q is close to p, then γ is the shortest curve connecting p to q. Moreover, as we saw above, γ will never be minimizing after the first conjugate point of p along γ. Before the first conjugate point of p, γ is shortest among nearby curves connecting p to q. However, even if p has no conjugate point along γ, it is still possible that γ is not the shortest curve connecting p to q. This leads to the definition of a cut point.

Let $p \in \mathcal{M}$ and $\gamma : [0, \infty) \to \mathcal{M}$ be a geodesic with $\gamma(0) = p$. If

$$t_0 \coloneqq \sup\{t \geq 0 : \gamma([0, t]) \text{ is a minimizing geodesic}\} < \infty,$$

then we call $\gamma(t_0)$ the *cut point of p along γ*. We will denote by $\operatorname{Cut}(p)$ the set of all cut points of p along all geodesics that emanate from p, and call it the *cut locus of p*.

Suppose $q = \exp_p(v)$ is in the cut locus of p and γ_v be the geodesic p to q. Then, a standard result [Lee18, Prop. 10.32] states that γ is minimizing, and one or both of the following conditions hold:

(a) q is conjugate to p along γ, or

(b) there are two or more minimizing geodesics from p to q.

Going back to our example, two antipodal points on a two-dimensional sphere satisfy both the conditions above, and therefore, are in the cut loci of each other. Note also that any point on a compact Riemannian manifold has a nonempty cut locus.

There is an interesting relation between the cut locus and the injectivity radius at a point: The injectivity radius at p is the distance from p to its cut locus provided that the cut locus is nonempty, and is infinite otherwise [Lee18, Prop. 10.36].

In general, one should always be aware of cut loci when working with Riemannian manifolds. The following proposition depicts an important example that how things can go wrong when considering the distance function on a Riemannian manifold [Pet16, Cor. 5.7.11]:

Proposition 3.1. *Fix a point $p \in \mathcal{M}$ and consider the (radial) distance to p, that is, the function $r(q) = d(p, q)$. Then r is differentiable on $\mathcal{M} \setminus (\operatorname{Cut}(p) \cup \{p\})$.*

3.2.10. Convex Neighborhoods

In many applications, normal neighborhoods are sufficient for a clean analysis of the problem at hand. However, when comparing distances of points near a point p, we also want the guarantee that (a) the points themselves are in the normal neighborhood centered at p, and (b) the minimizing geodesic connecting these points lie in that neighborhood. This requirement is crucial when computing the distance between two curves lying in a neighborhood of a point. We want all the minimizing geodesics connecting the corresponding points on the two curves lie in the neighborhood, constituting a smooth family of geodesics.

This motivates the definition of a convex subset. A subset $\mathcal{U} \subseteq \mathcal{M}$ is called *convex* if for any $p, q \in \mathcal{U}$, there exists a (not necessarily unique) minimizing geodesic connecting p and q whose image lies entirely in $\mathcal{U}$. We call $\mathcal{U}$ *strongly convex* if it is convex and the minimizing geodesic connecting each two of its points is unique. We also require that any ε-ball lying entirely in $\mathcal{U}$ to be convex.

Similar to the injectivity radius, we define the *convexity radius at p* as the radius of the largest geodesic ball centered at p that is strongly convex:

$$r_{\text{conv}}(p) = \sup\{r > 0 : B_s(p) \text{ is strongly convex for all } 0 < s < r\},$$

The *convexity radius of $\mathcal{M}$*, denoted by $r_{\text{conv}}(\mathcal{M})$, is then defined as the infimum of the convexity radii of all points of $\mathcal{M}$. The following theorem shows that if the sectional curvatures of $\mathcal{M}$ (see Section 3.2.11 below for a definition) is bounded from above and $\mathcal{M}$ has a positive injectivity radius, it also has a positive convexity radius.

Theorem 3.2 (CE08, Thm. 5.14)**.** *Suppose $\mathcal{M}$ is a complete, connected Riemannian manifold with sectional curvature bounded above by K. Then,*

$$r_{\text{conv}} \geq \frac{1}{2} \min\{\pi/\sqrt{K}, \text{inj}(\mathcal{M})\},$$

where $\pi/\sqrt{K}$ is interpreted as $+\infty$ if $K \leq 0$.

The intuition behind this theorem is as follows: If the sectional curvature is unbounded, points with high curvature will have cut loci that are closer to them. This proximity of the cut loci makes it difficult for a geodesic ball to be strongly convex. Recall that strong convexity requires unique minimizing geodesics between any two points within the ball, a condition that may not hold for points on the cut locus.

3.2.11. Curvature

Curvature is at the heart of Riemannian geometry and originates from the fundamental works of Gauss and Riemann. There are various notions of curvature for a Riemannian manifold, all of which can be derived from the Riemann curvature tensor. Define the map $R : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})$ by

$$R(X,Y)Z \coloneqq \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z,$$

and the *Riemann curvature tensor* as the map given by

$$\mathrm{Riem}(X,Y,Z,W) \coloneqq \langle R(X,Y)Z, W\rangle.$$

Let Π be a two-dimensional subspace of $T_p\mathcal{M}$ spanned by the vectors $v, w \in T_p\mathcal{M}$. The *sectional curvature* of Π is defined to be

$$\sec(\Pi) = \sec(v,w) \coloneqq \frac{\mathrm{Riem}(v,w,v,w)}{|v \wedge w|^2},$$

where

$$|v \wedge w| = \sqrt{|v|^2|w|^2 - \langle v, w\rangle}$$

is the area of the two-dimensional parallelogram determined by the pair of vectors v and w. Alternatively, the sectional curvature can be characterized by the circumference of small circles. For sufficiently small $r > 0$, let $C_r(p)$ denote the image under $\exp_p$ of the circle of radius r in Π, and let $\ell_r(p)$ denote the length of $C_r(p)$. Then it holds that

$$\ell_r(p) = 2\pi r\left(1 - \frac{r^2}{6}\sec(\Pi) + O(r^3)\right) \quad \text{as } r \to 0.$$

The sectional curvature also coincides with the Gaussian curvature of the surface $\exp_p(\Pi) \subseteq \mathcal{M}$ at point p; we omit the details of this correspondence for the sake of brevity.

3.2.12. Gradients and Hessians

Let $(\mathcal{U}, \varphi)$ be a coordinate chart of $\mathcal{M}$ and let $x^1, \ldots, x^d$ be its components. We saw in Section 3.2.1 that for any point $p \in \mathcal{U}$, this chart induces a basis $\{\partial/\partial x^1|_p, \ldots, \partial/\partial x^d|_p\}$ for $T_p\mathcal{M}$. Now consider the dual space of $T_p\mathcal{M}$ (in the sense of vector spaces), denoted by $T_p^*\mathcal{M}$. An element in the dual space is a linear functional on $T_p\mathcal{M}$, and is called a *covector* in differential geometry. The basis of $T_p\mathcal{M}$ induces a basis for the dual space, denote by $dx^1|_p, \ldots, dx^d|_p$ via

$$dx^i\left(\frac{\partial}{\partial x^j}\right) = \delta^i_j, \quad \text{where } \delta^i_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Similar to vector fields, a *covector field* is a smooth map that attaches to each point of the manifold a covector from its dual tangent space. The differential df of a smooth function $f : \mathcal{M} \to \mathbb{R}$ is an example of a covector field; recall from Section 3.2.1 that $df_p : T_p\mathcal{M} \to T_{f(p)}\mathbb{R} \cong \mathbb{R}$ is the map

$$df_p(v)(h) = v(h \circ f), \quad \forall v \in T_p\mathcal{M} \text{ and } h : \mathbb{R} \to \mathbb{R}.$$

Identifying the tangent space of $\mathbb{R}$ at any point with $\mathbb{R}$ and taking h above to be the identity map, we see that $df_p(v) = v(f)$, which implies $df_p \in T_p^*\mathcal{M}$. In terms of coordinates, we get

$$df = \sum_{i=1}^{d} df\left(\frac{\partial}{\partial x^i}\right) dx^i = \sum_{i=1}^{d} \frac{\partial f}{\partial x^i}\, dx^i.$$

To be more precise, by $\frac{\partial f}{\partial x^i}(p)$ we mean $\frac{\partial \widehat{f}}{\partial x^i}(\widehat{p})$, where we recall the notation $\widehat{f} = f \circ \varphi^{-1}$ and $\widehat{p} = \varphi(p)$ for the coordinate representation of f and p, respectively.

In Euclidean spaces, there is a natural correspondence between the differential of a smooth function and its gradient. Let us reiterate this correspondence with the hope of defining a Riemannian analogue of gradients. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, its gradient at a point $x \in \mathbb{R}^d$ is the (tangent) vector $\overline{\nabla} f(x)$ defined as

$$\overline{\nabla} f(x) = \sum_{i=1}^{d} \frac{\partial f}{\partial x^i}(x) \frac{\partial}{\partial x^i}\bigg|_x .$$

We use the "manifold notation" here to keep similarity with our previous discussions. Notice that in this case, the Euclidean gradient and differential are dual to each other: For any smooth vector field V, it holds that

$$df(V) = \langle \overline{\nabla} f, V \rangle.$$

We now extend this construct to the Riemannian case. Take a coordinate chart $(\mathcal{U}, \varphi)$ and suppose $f : \mathcal{M} \to \mathbb{R}$ is differentiable. We want to define the Riemannian gradient of f in such a way that a similar duality relation hold. That is, we want to define the vector field $\operatorname{grad} f$ such that for any smooth vector field $V \in \mathfrak{X}(\mathcal{M})$, it holds

$$\langle \operatorname{grad} f, V \rangle = df(V) = V(f).$$

The left-hand side of the equality above in terms of coordinates is

$$\langle \operatorname{grad} f, V \rangle = \sum_{i=1}^{d} \sum_{j=1}^{d} g_{ij} (\operatorname{grad} f)^i V^j,$$

and the right-hand side in terms of coordinates is

$$df(V) = \sum_{j=1}^{d} \frac{\partial f}{\partial x^j} V^j.$$

Setting these two expressions to be equal, we get an expression for the gradient:

$$\operatorname{grad} f = \sum_{i=1}^{d} \sum_{j=1}^{d} g^{ij} \frac{\partial f}{\partial x^i} \frac{\partial}{\partial x^j},$$

where g^{ij} is the (i, j) entry of the inverse of the matrix (g_{ij}). While this construction relied on a local coordinate system, it can be shown that $\operatorname{grad} f$ is independent of the choice of the local coordinates, and therefore, is an intrinsic notion.

Similarly, we define the *Hessian* of a smooth function $f : \mathcal{M} \to \mathbb{R}$ to be the covariant 2-tensor field[4] $\operatorname{Hess} f$, which for any pair of vector fields $V, W \in \mathfrak{X}(\mathcal{M})$, satisfies

$$\operatorname{Hess} f(V, W) = \langle \nabla_V \operatorname{grad} f, W \rangle.$$

To get a better feeling of the Hessian, consider a geodesic $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ with $\gamma(0) = p$ and $\dot{\gamma}(0) = v$. Then it turns out that

$$\left.\frac{d^2}{dt^2}\right|_{t=0} f(\gamma(t)) = (\operatorname{Hess} f)_p(v, v).$$

3.3. INTRODUCTORY EXAMPLES

In this section, we bring a handful of examples that are specifically related to stochastic approximation on Riemannian manifolds. Some of these problems are inherently defined on a Riemannian manifold (see Example 3.1 and Section 3.3.2), while others are described initially for Euclidean spaces but shown to have more computational benefits when considered on a Riemannian manifold (see Examples 3.2 and 3.3).

[4] A *covariant 2-tensor* is a multilinear map that eats two vectors and gives a number. A covariant 2-tensor field is a smooth assignment of a covariant 2-tensor to each point of the manifold.

3.3.1. Optimization Problems on Manifolds

We start with the more familiar optimization problem. Let $\mathcal{M}$ be a Riemannian manifold and f be a smooth function on $\mathcal{M}$. Consider the following minimization problem, which encapsulates a wide range of practical and theoretical problems:

$$\text{Minimize } f(p) \quad \text{such that } p \in \mathcal{M}. \tag{3.10}$$

A natural generalization of the Robbins–Monro scheme for solving (3.10) is

$$\boldsymbol{x}_{n+1} = \exp_{\boldsymbol{x}_n}(\alpha_n(-\nabla f(\boldsymbol{x}_n) + Z_n)), \tag{3.11}$$

where α_n is a sequence of step-sizes with $\alpha_n \to 0$ as $n \to \infty$ and $\sum \alpha_n = \infty$, and Z_n captures the noise and bias in evaluation of $\nabla f(\boldsymbol{x}_n)$. Below, we list three examples of such problems.

▷ **Example 3.1** (Rayleigh Quotient)**.** Let $\mathcal{M} = S^{d-1}$ be the $(d-1)$-dimensional sphere embedded in $\mathbb{R}^d$; that is,

$$\mathcal{M} = \left\{x \in \mathbb{R}^d : (x^1)^2 + \cdots + (x^d)^2 = 1\right\}.$$

The tangent space $T_x\mathcal{M}$ at $x \in \mathcal{M}$ is the $(d-1)$-dimensional hyperplane consisting of vectors $v \in \mathbb{R}^d$ that $v^\top x = 0$. We equip $\mathcal{M}$ with the metric induced by the embedding in $\mathbb{R}^d$; that is, the inner product of two tangent vectors is computed in the Euclidean sense. One can show that the geodesics of $\mathcal{M}$ are arcs of great circles of $\mathcal{M}$, and that the exponential map can be computed as

$$\exp_x(v) = x \cos\|v\| + \frac{v}{\|v\|} \sin\|v\|.$$

Let A be a $d \times d$ symmetric matrix, and consider the function

$$f : \mathcal{M} \to \mathbb{R}, \quad f(x) = x^\top A x,$$

known as the Rayleigh quotient. The gradient of f (in the Riemannian sense) can be obtained by projecting the Euclidean gradient of f on the tangent bundle of $\mathcal{M}$. Concretely, for $x \in \mathcal{M}$,

$$\nabla f(x) = (I - xx^\top)(2Ax),$$

where $(I - xx^\top)$ is the projection onto $T_x\mathcal{M}$, and $2Ax$ is the gradient of f in the Euclidean sense.

For simplicity, suppose that A has d distinct eigenvalues $\lambda_1 < \cdots < \lambda_d$, and suppose $v_1, \ldots, v_d$ are the corresponding eigenvectors which are assumed to be

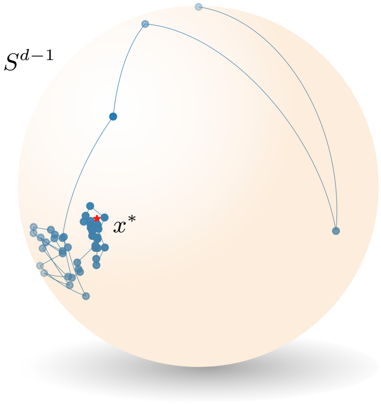


Figure 3.1. A few iterations of the stochastic approximation algorithm (3.11) for finding the smallest eigenvalue of a symmetric matrix A, given noisy access to A at each iteration. Due to the symmetry of eigenvectors, there are two optimal solutions (one is depicted by a star and the other is its antipodal point), and the algorithm converges to one of them.

orthonormal. It is well-known [see, e.g., AMS08, Prop. 4.6.2] that $\pm v_1$ are the global minimizers, and $\pm v_n$ are the global maximizers of f on $\mathcal{M}$. Moreover, $\pm v_i$ for $i = 2, \ldots, (d-1)$ are all saddle points of f. Therefore, to find the smallest eigenvalue and eigenvector of A, we can solve (3.10). Suppose we do so iteratively, and in each iteration we only observe a noisy version of A and compute the gradient based on the noisy A. Figure 3.1 shows a few iterations of this simple algorithm for $d = 3$. As observed, the iterates converge to an optimal solution, which is an eigenvector corresponding to the smallest eigenvalue. ◁

▷ **Example 3.2** (Natural Policy Gradient)**.** Consider a finite Markov Decision Process (MDP) and assume all policies π considered are ergodic, in the sense that they admit a unique stationary distribution ϱ^π over the states. Let the average (undiscounted) reward be defined as

$$\eta(\pi) = \sum_{s,a} \varrho^\pi(s)\pi(a \mid s)\, R(s,a),$$

where a and s denote an action and a state, respectively, and $R(s,a)$ is the reward obtained from choosing action a in the state s.

Consider the case where each policy is parameterized by some $\theta \in \mathbb{R}^d$, and write $\pi(a \mid s; \theta)$ for the probability assigned to the action a at the state s by the policy π_θ that is parameterized by θ. Our task is to find the parameters θ such that the corresponding policy π_θ maximizes the average reward. We solve this task using gradient ascent with respect to θ. The gradient of $\eta(\theta) := \eta(\pi_\theta)$ with respect to θ in the Euclidean sense is

$$\overline{\nabla}\eta(\theta) = \sum_{s,a} \varrho^\pi(s)\overline{\nabla}\pi(a \mid s, \theta)Q^\pi(s, a),$$

where $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^\infty R(s_t, a_t) - \eta(\pi) \mid s_0 = s, a_0 = a]$ is the state-action value and $\overline{\nabla}$ denotes the Euclidean gradient. Kakade [Kak01] realized that using such a gradient for gradient ascent is not "natural," as it depends on the specific parameterization of the policy space $\Pi := \{\pi_\theta : \theta \in \mathbb{R}^d\}$. In other words, different parameterizations of the set Π lead to different algorithms. Building up on ideas of Amari [Ama83], one can look at Π as a Riemannian manifold, whose metric is given by the Fisher information matrix. Specifically, for each state s, we can define a Riemannian metric (in the Euclidean chart)

$$G_s(\theta) = \mathbb{E}_{\pi(a|s,\theta)}\left[\overline{\nabla}\log\pi(a \mid s, \theta)\overline{\nabla}\log\pi(a \mid s, \theta)^\top\right],$$

which is the Fisher information matrix of the probability distribution over actions given by the policy π_θ at state s. This metric is defined for each state, and is independent of the transition probabilities of the underlying MDP. As shown by Amari [Ama83], the Fisher information matrix, up to a scale, is an invariant metric (in the sense that it defines the same distance between two distributions regardless of the choice of parameterization) on the space of the parameters of probability distributions—hence that name *natural* metric. Taking expectation with respect to the stationary distribution ϱ^{π_θ} over the states, we get the following metric on Π:

$$G(\theta) = \mathbb{E}_{\varrho^{\pi_\theta}}[G_s(\theta)].$$

Kakade uses this metric to compute Riemannian (or natural) gradients. When expressed in the Euclidean coordinates, the resulting update rule is

$$\theta_{n+1} = \theta_n + \alpha_n F(\theta_n)^{-1}\overline{\nabla}\eta(\theta_n). \qquad \triangleleft$$

In Section 3.7 we demonstrate various stochastic approximation algorithms that can be used to solve Riemannian optimization problems, and in Proposition 3.20 we prove a general result regarding convergence of these algorithms.

3.3.2. Games on Manifolds

We now consider games on Riemannian manifolds. By a *Riemannian game*, we mean a game between N players, the space of strategies of which is a Riemannian manifold. The space of all configurations of the game is then the product manifold $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_N$, where $\mathcal{M}_i$ is the strategy space of the ith player. Let $u_i : \mathcal{M} \to \mathbb{R}$ be the payoff function of the ith player, and the goal of the ith player is to maximize their payoff—for applications and a detailed discussion, see [RBS14] and references therein.

Let us note that a min-max optimization problem on a Riemannian manifold is a simple instance of the Riemannian game defined above: Let $\mathcal{M}$ and $\mathcal{N}$ be two manifolds, and $\ell : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ be a smooth function. Letting $u_1 = -\ell$ and $u_2 = \ell$ results in the following problem for the players:

$$\min_{p \in \mathcal{M}} \max_{q \in \mathcal{N}} f(p, q).$$

Besides min-max problems on Riemannian manifolds (such as those on matrix manifolds or hyperbolic spaces), we can also consider those problems that do not inherently possess a Riemannian structure. It turns out that endowing the strategy space of these problems with a Riemannian structure will be advantageous in many scenarios. These advantages include geodesic convexity of the strategy space and payoffs, as well as simplified determination of Nash equilibria. The Riemannian structure further facilitates the gradient-based methods by leveraging the manifold's geometry, leading to more efficient and accurate solutions.

Before delving deep into the examples, we have to introduce a generalization of Nash equilibria to the Riemannian setting: the Nash–Stampacchia equilibrium [Kri14]. In the following, define the vector field $V = (V_1, \ldots, V_N)$ with $V_i = \nabla_{p_i} u_i(p_1, \ldots, p_N)$ to be the direction of improvement for each player.

Definition 3.3. We say that the point $p^* := (p_1^*, \ldots, p_N^*) \in \mathcal{M}_1 \times \cdots \times \mathcal{M}_N$ is a *Nash–Stampacchia equilibrium* of V if it satisfies the system of variational inequalities

$$\langle V_i(p^*), \exp_{p_i^*}^{-1}(p_i) \rangle \geq 0, \quad \text{for } i = 1, \ldots, N \text{ and all } p_i \in \mathcal{M}_i. \tag{3.12}$$

Recall from Section 3.2.7 that $\mathcal{M}$ has a Riemannian structure and its exponential map is evaluated as

$$\exp_p(v) = (\exp_{p_1}(v_1), \ldots, \exp_{p_N}(v_N)).$$

Using this, Kristály [Kri14, Rem. 2.1 and Rem. 4.1] shows that the Nash–Stampacchia condition (3.12) is equivalent to the simpler condition

$$\langle V(p^*), \exp_{p^*}^{-1}(p) \rangle_{p^*} := \sum_{i=1}^{N} \langle V_i(p^*), \exp_{p_i^*}^{-1}(p_i) \rangle_{p_i^*} \geq 0, \quad \text{for all } p \in \mathcal{M}. \tag{3.13}$$

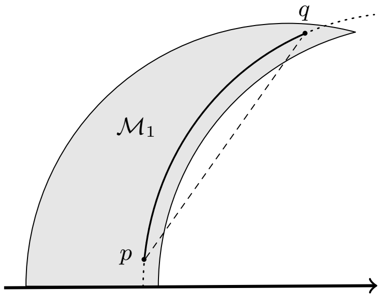


Figure 3.2. Example of a non-convex set in $\mathbb{R}^2$, which, if equipped with the metric of the hyperbolic space, becomes geodesically convex. The Euclidean line segment between points p and q leaves $\mathcal{M}_1$, while the geodesic connecting them—the arc of a circle centered on the x-axis—is contained in $\mathcal{M}_1$.

It also turns out that if the utility functions are Lipschitz, then the set of Nash–Stampacchia equilibria is a superset of the set of Nash equilibria, and if the utilities are convex, the two solution concepts coincide [see Kri14, Thm. 3.1].

Note that the choice of the Riemannian metric on $\mathcal{M}_i$ does not affect the concept of Nash equilibrium, while it is a defining piece of Nash–Stampacchia equilibrium. Below, we present an example inspired from Kristály [Kri14], which demonstrates this principle in action.

▷ **Example 3.3.** Let $\mathcal{M}_1 = \{(x, y) \in \mathbb{R}^2_+ : x^2 + y^2 \leq 4 \leq (x-1)^2 + y^2\}$ and $\mathcal{M}_2 = [-1, 1] \subset \mathbb{R}$ be the strategy space of two players. It is evident that $\mathcal{M}_1$ is not a convex set in the Euclidean sense. However, by looking at $\mathcal{M}_1$ as a subset of the upper-plane model of the hyperbolic space, $\mathcal{M}_1$ becomes geodesically convex; see Fig. 3.2. Guided by the variational inequalities (3.12) or (3.13), one can effectively find Nash–Stampacchia equilibria (and consequently, the set of possible Nash equilibria) in a game defined over $\mathcal{M}_1 \times \mathcal{M}_2$; something that is *a priori* a non-trivial task. ◁

This Riemannian viewpoint provides a profound understanding of the familiar *Replicator dynamics*, commonly used to model evolutionary systems. In their seminal paper, Mertikopoulos and Sandholm [MS18] illustrate that by defining a specific Riemannian metric, known as the *Shahshahani metric*, the replicator dynamics transforms into a Riemannian game dynamics. This transformation essentially reinterprets the replicator dynamics as the flow of the tangential component of the payoff vector field. This in turn broadens our comprehension of the underlying dynamics. For a detailed explanation and more implications, the reader is encouraged to consult the original article [MS18].

In Sections 3.7 and 3.8 we demonstrate various stochastic approximation algorithms, along with practical variations thereof, that can be used to solve Riemannian games, and in Propositions 3.21 and 3.22 we prove general results regarding convergence of such stochastic approximation schemes for a class of monotone games and non-convex potential games.

3.4. RIEMANNIAN ROBBINS–MONRO SCHEMES

In this section, define a class of stochastic approximation algorithms on Riemannian manifolds called Riemannian Robbins–Monro schemes, and translate the main objects of Chapter 2 into the language of Riemannian geometry. In Section 3.4.1, we bring our main assumptions and a discussion about their generality. Sections 3.4.2 and 3.4.3 state the main results of this chapter, namely the asymptotic pseudo-trajectory theorem and the stability theorem.

Let us start by defining the basic template of Riemannian Robbins–Monro algorithms. The main difference of a Riemannian stochastic approximation algorithm with its Euclidean counterpart is that addition along straight lines in a Euclidean space is replaced by the Riemannian exponential map, or in general, a retraction. We will discuss retractions later in Section 3.8.1, and focus on the exponential map for now.

Throughout this chapter, let $\mathcal{M}$ be a Riemannian manifold. We call the update rule of the form

$$\boldsymbol{x}_{n+1} = \exp_{\boldsymbol{x}_n}(\alpha_n(V(\boldsymbol{x}_n) + Z_n)) \tag{RRM}$$

a *Riemannian Robbins–Monro scheme*, where

(1) $\boldsymbol{x}_n \in \mathcal{M}$ denotes the state of the algorithm at each iteration $n = 1, 2, \ldots,$

(2) V is a vector field on the manifold,

(3) $Z_n \in T_{\boldsymbol{x}_n}\mathcal{M}$ is an error term, described in detail below,

(4) and $\alpha_n > 0$ is the method's step-size.

In the above, we assume that the error term Z_n is generated *after* $\boldsymbol{x}_n$. In addition, to differentiate between "random" (zero-mean) and "systematic" (non-zero-mean) errors, it will be convenient to further decompose Z_n as

$$Z_n = U_n + B_n, \tag{3.14}$$

where $U_n = Z_n - \mathbb{E}[Z_n \,|\, \mathcal{F}_n]$ captures the zero-mean part and $B_n = \mathbb{E}[Z_n \,|\, \mathcal{F}_n]$ represents the systematic component of Z_n. Here, the σ-algebra $\mathcal{F}_n = \sigma(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$

contains all the information until iteration n. We will see in Sections 3.7 and 3.8 that allowing for a nonzero systematic error B_n enables us to fit various algorithms into the (RRM) template. For brevity, we will also write

$$V_n = V(\boldsymbol{x}_n) + Z_n. \tag{3.15}$$

In this way, V_n can be seen as a noisy—and potentially biased—estimate of $V(\boldsymbol{x}_n)$.

Flows and orbits

A complete vector field V induces a flow on the manifold, which is the map $\Phi : \mathbb{R} \times \mathcal{M} \to \mathcal{M}$, with $\Phi_0 = \mathrm{Id}$ and

$$\frac{d}{dt}\Phi_t(p) = V(\Phi_t(p)), \quad \forall p \in \mathcal{M},\ t \in \mathbb{R}. \tag{3.16}$$

Recall that a vector field is *complete* if its corresponding flow is global, i.e., the solution of the ODE (3.16) exists for all $t \in \mathbb{R}$ and all $p \in \mathcal{M}$. The orbit of Φ starting at p is denoted by $\phi^p(t) \coloneqq \Phi_t(p)$, and we omit p and simply write $\phi(t)$ if the starting point p is clear from the context.

Geodesic interpolation

To compare the iterates $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ of (RRM) with the orbits of the flow Φ, we have to first construct a continuous-time interpolation of the iterates. In the Euclidean setting of Section 2.5, we connected the points in the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ by line segments. Concretely, by defining the effective time till iteration n as

$$\tau_1 = 0, \quad \text{and} \quad \tau_n = \sum_{k=1}^{n-1} \alpha_k,$$

we constructed the interpolation

$$\boldsymbol{x}(t) = \boldsymbol{x}_n + \frac{t-\tau_n}{\tau_{n+1}-\tau_n}(\boldsymbol{x}_{n+1} - \boldsymbol{x}_n) \quad \text{for all } t \in [\tau_n, \tau_{n+1}],\ n \in \mathbb{N}.$$

Of course, this definition is not meaningful in a Riemannian setting. Instead, guided by the formulation of (RRM), we construct the interpolation by following the *geodesic* emanating from $\boldsymbol{x}_n$ in the direction V_n until reaching $\boldsymbol{x}_{n+1}$. With this in mind, we define the *geodesic interpolation* $\boldsymbol{x}(t)$ of $\{\boldsymbol{x}_n\}$ as

$$\boldsymbol{x}(t) = \exp_{\boldsymbol{x}_n}((t-\tau_n)\, V_n) \quad \text{for all } t \in [\tau_n, \tau_{n+1}],\ n \in \mathbb{N}. \tag{GI}$$

By construction, $\boldsymbol{x}(\tau_n) = \boldsymbol{x}_n$ for all n, and each segment of $\boldsymbol{x}(t)$ is a geodesic.

3.4.1. Technical Assumptions

We now state our blanket assumptions that underlie the rest of this chapter. These are as follows:

▷ **Assumption 3.1** (on the manifold)**.** *The manifold $\mathcal{M}$ is a complete, connected, Riemannian manifold without boundary and has a positive injectivity radius* $\operatorname{inj}(\mathcal{M}) > 0$*. The sectional curvature of $\mathcal{M}$ at any point is bounded from below and above by κ_{low} and κ_{up}, respectively.*

Remark. A few remarks are in order:

(1) Recall that a manifold is complete if it is complete as a metric space. That is, every Cauchy sequence in $\mathcal{M}$ converges to a limit in $\mathcal{M}$. By the Hopf–Rinow theorem [Pet16, Thm. 16], a complete Riemannian manifold satisfies the Heine–Borel property—that is, any bounded closed set is compact—and is geodesically complete, i.e., the exponential map is defined on the entire tangent bundle. If the manifold is also connected, any two points can be connected via a minimizing geodesic. We use all these properties in our analysis: We connect points via minimizing geodesics to compute their distance, and require continuous functions to be bounded on bounded sets, needing precompactness of bounded sets.

(2) The injectivity radius at a point $p \in \mathcal{M}$ is the radius of the largest ball in $T_p\mathcal{M}$ centered at 0 such that the exponential map $\exp_p$ is a diffeomorphism onto its image. The injectivity radius of $\mathcal{M}$ is the infimum of these radii; see Section 3.2.9 for a reminder. Having positive injectivity radius ensures that the exponential map at every point is invertible on a ball of some fixed radius. This removes local topological complications and, together with bounded curvature, ensures a nonzero convexity radius (see Theorem 3.2 in Section 3.2.10); a property we use later to prove a local version of the asymptotic pseudo-trajectory property.

(3) Comparing vectors living in different tangent spaces usually involves studying certain Jacobi fields. Sectional curvature bounds allow us to use comparison theorems to bound these Jacobi fields by those on a manifold with constant curvature. Note, however, that curvature bounds are not sufficient to ensure the injectivity radius is bounded away from zero; see Berger spheres in [Pet16, Sec. 4.4.3] for a counterexample. ◊

▷ **Assumption 3.2** (on the vector field)**.** *The vector field V is differentiable, complete, and L-Lipschitz in the Riemannian sense.*

Remark. Let us make the following remarks:

(1) Recall that a vector field is complete if the solution of the corresponding ODE (3.16) exists for all times. This property is obviously required in a stochastic approximation analysis, as we need the existence of orbits of the flow for arbitrary long time intervals.

(2) We use the notion of Lipschitzness defined in [Bou23, Def. 10.44]: a vector field V is called *L-Lipschitz* in the Riemannian sense, if for all $p, q \in \mathcal{M}$ with q in the injectivity radius of p, it holds

$$|\mathrm{P}_{p\to q}[V(p)] - V(q)| \leq L \cdot d(p,q), \tag{3.17}$$

where the parallel transport is along the unique minimizing geodesic connecting p to q. If V is also C^1, we have for any $v \in T_p\mathcal{M}$,

$$|\nabla_v V(p)| \leq L|v|. \tag{3.18}$$

Consequently, for any smooth curve γ connecting p to q,

$$|\mathrm{P}^{\gamma}_{p\to q}[V(p)] - V(q)| \leq L \cdot L(\gamma),$$

where $L(\gamma)$ is the length of γ. See [Bou23, Prop. 10.46] for a proof.

(3) Smoothness of V allows us to bound the distance between the integral curves of its flow using (3.18); see Lemma A.4. ◊

▷ **Assumption 3.3** (on the step-sizes)**.** *The step-size sequence* $\{\alpha_n\}_{n\in\mathbb{N}}$ *of* (RRM) *satisfies the* Robbins–Monro summability conditions

$$\sum_{n=1}^{\infty} \alpha_n = \infty \quad \textit{and} \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty.$$

The step-size assumption is crucial for controlling the impact of noise and bias, as well as the effect of discretizing the flow.

Recall the error terms Z_n in (RRM) and their decomposition into the zero-mean part U_n and the systematic part B_n in (3.14). In the sequel, we refer to U_n and B_n as the *noise* and *bias* terms, respectively. In our analysis, we regularly work with upper bounds on the second moments of the noise and bias terms. Define U_n^* and B_n^* to be $\mathcal{F}_n$-adapted (real-valued) sequences, satisfying

$$\mathbb{E}[|B_n|^2 \,|\, \mathcal{F}_n] \leq (B_n^*)^2 \quad \text{and} \quad \mathbb{E}[|U_n|^2 \,|\, \mathcal{F}_n] \leq (U_n^*)^2 \quad \text{a.s.} \quad \forall n \in \mathbb{N}. \tag{3.19}$$

▷ **Assumption 3.4** (on the noise)**.** *When* $\operatorname{inj}(\mathcal{M}) = \infty$*, we assume*

$$\sum_{n=1}^{\infty} \alpha_n^2 \, \mathbb{E}[(U_n^*)^2] < \infty. \tag{3.20}$$

If $\operatorname{inj}(\mathcal{M}) < \infty$*, we assume that the noise terms* U_n *in* (3.14) *are a.s. bounded in norm: There exists some* $U^* \in \mathcal{F}_\infty$ *such that*

$$|U_n| \leq U^* \quad \text{a.s. for all } n \in \mathbb{N}. \tag{3.21}$$

▷ **Assumption 3.5** (on the bias)**.** *We assume that the upper bounds* B_n^* *for the bias terms satisfy*

$$\sum_{n=1}^{\infty} \alpha_n (\mathbb{E}[(B_n^*)^2])^{1/2} < \infty, \quad \text{and} \quad B_n^* \to 0 \quad \text{a.s.} \tag{3.22}$$

Moreover, if $\operatorname{inj}(\mathcal{M}) < \infty$*, we further assume that* $|B_n|$ *is a.s. uniformly bounded by some finite* $B^* \in \mathcal{F}_\infty$*.*

Remark. A few remarks are in order:

(1) Variants of the above assumptions are standard in the context of (Euclidean) Robbins–Monro methods; see, for example, [KC78; Ben99] and references therein. While Assumption 3.3 lies under the explicit control of the algorithm designer, Assumptions 3.4 and 3.5 are implicit and depend on the specific problem at hand, namely the mechanism providing access to V, the specific form of (RRM), and so on. We show in Section 3.7 that Assumptions 3.4 and 3.5 are indeed satisfied for a wide range of practical algorithms that adhere to the general template of (RRM).

(2) Assumptions 3.4 and 3.5 distinguish between cases where $\operatorname{inj}(\mathcal{M})$ is finite or infinite. The reason is that if $\operatorname{inj}(\mathcal{M}) < \infty$, unbounded noise or bias prevents ensuring that consecutive iterates of (RRM) remain close together within a short (but fixed) time window. This constraint is pivotal for subsequent analysis, raising whether the boundedness assumption may be relaxed. See Section 3.10 for further discussion. ◊

To exclude cases where (RRM) becomes unstable over time, a standard practice in the literature is to assume that the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ is contained in a compact subset of $\mathcal{M}$, a property known as *precompactness* or *stability* [see, e.g., KC78; BMP90; Ben99].

▷ **Assumption 3.6** (Precompactness)**.** *The set of iterates* $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ *has compact closure in* $\mathcal{M}$*, almost surely.*

Remark. Assumption 3.6 may be difficult to verify if $\mathcal{M}$ itself is not compact. To account for this, we introduce in Section 3.4.3 a set of sufficient conditions which guarantee that precompactness holds. ◊

3.4.2. Theorem on the Dynamics

The limit set theorem (Theorem 2.6) provides a fundamental link between asymptotic pseudo-trajectories and the long-run behavior of the flow (3.16) as captured by its internally chain-transitive sets. That being said, the asymptotic pseudo-trajectory property itself may be difficult to verify from first principles, so the application of Theorem 2.6 to Riemannian Robbins–Monro algorithms can be just as difficult. In the Euclidean case ($\mathcal{M} = \mathbb{R}^d$), we saw in Section 2.5 how Benaïm and Hirsch address this issue via a series of criteria under which standard (Euclidean) Robbins–Monro methods give rise to an asymptotic pseudo-trajectory of the associated mean dynamics. Unfortunately, however, in a Riemannian setting, these criteria cannot be used because they are deeply tied to the affine structure of $\mathbb{R}^d$; as a result, it is not clear how to leverage Theorem 2.6 to obtain a theory of stochastic approximation for Robbins–Monro methods on Riemannian manifolds. We tackle this question below:

▶ **Theorem 3.4.** *Suppose that Assumptions 3.1–3.6 hold. Then, with probability* 1, *the geodesic interpolation* $(\boldsymbol{x}(t))_{t\geq 0}$ *of the sequence of iterates* $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ *of* (RRM) *is an asymptotic pseudo-trajectory of the flow* Φ.

The proof of Theorem 3.4 is rather long and involved, and we dedicate the entire Section 3.5 to it. As we require the iterates to be precompact in Assumption 3.6, the following corollary is a straightforward application of the limit set theorem:

▶ **Corollary 3.5.** *Suppose that Assumptions 3.1–3.6 hold. Then,* $\boldsymbol{x}_n$ *almost surely converges to an internally chain-transitive set of the flow* Φ.

3.4.3. Stability Theorem

The main convergence result requires the algorithm's iterates to be contained in some compact set with probability 1. In the literature, this is known as the *stability* of the algorithm. There are several stability criteria for algorithms in Euclidean spaces, each applicable under specific restrictions and sometimes motivated by specific applications for which they are designed for. We refer to [Bor08, Ch. 3] for two such examples, and [Phe93; FP03] for more general criteria in the Euclidean setting. Below, we study stability in the context of Riemannian stochastic approximation algorithms.

Note that if the manifold itself is compact, there is nothing left to do. In this chapter, we establish a novel stability theorem that applies to *Hadamard manifolds*. These are a class of non-compact manifolds commonly used in applications; examples include Stiefel, Grassmannian, and other standard matrix manifolds, hyperbolic spaces, and so on.

▶ **Theorem 3.6.** *Suppose $\mathcal{M}$ is a complete Hadamard manifold and V is weakly coercive and bounded. Then, under Assumptions 3.4 and 3.5, the iterates of* (RRM) *are almost surely precompact.*

We postpone the related definitions and the proof of Theorem 3.6 to Section 3.6.

3.5. PROOF OF THE APT PROPERTY

In this section, we proof Theorem 3.4. In short, this theorem states that under some assumptions, the (geodesic) interpolation of the iterations of a Riemannian Robbins–Monro scheme is an asymptotic pseudo-trajectory of the underlying mean dynamics.

Because the proof of Theorem 3.4 and the geometric scaffolding required are long and delicate, we have divided the proof into seven steps outlined below. For each step, we give a high-level overview of the main difficulties and technical challenges involved.

(1) **Localization:** The definition of an asymptotic pseudo-trajectory requires comparison of the geodesic interpolation $\boldsymbol{x}$ and the flow orbit at *arbitrary large time-scales* T. As it is challenging to compare distances at large on an arbitrary Riemannian manifold, we prove that it is sufficient to show the asymptotic pseudo-trajectory property for an arbitrary small T.

(2) **The Picard curve:** The geodesic interpolation is a stochastic curve affected by noise and bias. Comparing this curve directly with the orbits of the flow turns out to be challenging. Inspired by the concept of Picard iteration and the Euclidean proof, we construct the Picard curve, which combines the favorable aspects of both flow orbits and the interpolation, and use it as a "bridge" between the orbit and the interpolation.

For any t, we consider the flow orbit ϕ and the Picard curve λ, both starting at $\boldsymbol{x}(t)$, and decompose the distance between the flow orbit and the interpolation into

$$d(\boldsymbol{x}([t,t+T]),\phi([0,T])) \\ \leq d(\boldsymbol{x}([t,t+T]),\lambda([0,T])) + d(\lambda([0,T]),\phi([0,T])).$$

Due to the desirable properties of the Picard curve λ, we show in the following steps that both terms of this decomposition vanish as $t \to \infty$.

(3) **Boundedness of the curves:** When $\operatorname{inj}(\mathcal{M}) < \infty$, we show that all these curves, namely the flow orbit, the Picard curve, and the $[t, t+T]$-segment of the interpolation, stay in a small geodesic ball around $\boldsymbol{x}(t)$ when T is small enough. By the localization argument in the first step, the proof reduces to a convex neighborhood of $\boldsymbol{x}(t)$, removing topological complexities arising from cut points.

(4) **The Fermi normal coordinates:** To compute the sup distance between two curves, it is natural to "move along" one curve and "look" at the other, while tracking the distance. This is precisely what the Fermi normal coordinate system allows us to do. We consider this coordinate system along the geodesic interpolation and "look at" the flow and Picard curves.

(5) **From Fermi to parallel:** Comparing tangent vectors (in our case, the velocity vector of curves) is challenging in normal coordinates while being an easy task in parallel coordinates. Since everything thus far is expressed in terms of Fermi normal coordinates, we show how to relate them with parallel coordinates. Using this idea, we control the distance of the flow orbit to the Picard curve. We develop along the way a comparison result (Lemma 3.12) based on sectional curvature bounds. This lemma might be of independent interest.

(6) **Picard curve vs. geodesic interpolation:** In this step, we deal with the noise and bias in the geodesic interpolation. We use a combination of Fermi-to-parallel and parallel transport arguments to give a bound on the distance of the Picard curve to the geodesic interpolation.

(7) **Finishing the proof:** We conclude the proof by putting all bounds together and applying Grönwall's lemma, showing the asymptotic pseudo-trajectory property.

Step 1. Localization

Recall that the curve $\boldsymbol{x}$ is an asymptotic pseudo-trajectory of the flow Φ if for all $T > 0$,

$$\lim_{t\to\infty} \sup_{0\le h\le T} d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t))) = 0. \tag{3.23}$$

To show this property, we have to compare the geodesic interpolation $\boldsymbol{x}$ and the flow orbit at *arbitrary large time-scales* T. Effectively, this comparison boils down to constructing a family of (minimizing) geodesics connecting $\boldsymbol{x}(t+h)$ and $\Phi_h(\boldsymbol{x}(t))$ for $h \in [0, T]$ and studying the evolution of the length of these geodesics.

This method works well in the Euclidean case but may cause significant topological challenges in a Riemannian manifold. Specifically, the points $\Phi_h(\boldsymbol{x}(t))$ may go beyond the cut locus of $\boldsymbol{x}(t+h)$, which can make the family of geodesics non-smooth. It is unclear if the set of non-differentiability times is even discrete, as the cut locus has a complex structure.[5]

To avoid these difficulties, we show in this step that it is sufficient to verify (3.23) for an arbitrary small T, chosen in a way that ensures no curve passes the cut locus of another. By an induction-like argument, we show that if (3.23) holds for some T, it also holds for $2T$. Our argument formalizes [Sha21, Claim 1].

Lemma 3.7. *Let V be a C^1 vector field and Φ be its corresponding flow. If a continuous piecewise-smooth curve $\boldsymbol{x}$ satisfies (3.23) for some $T > 0$, then it is an asymptotic pseudo-trajectory of the flow Φ.*

We prove this lemma in Appendix A.2. An essential piece of the proof is the fact that the integral curves of the flow of a Lipschitz, C^1 vector field cannot get too far from each other if started at different points. Lemma A.4 in the same appendix formalizes this intuition.

Step 2. The Picard Curve

In the Euclidean proof of Section 2.5, we used the Picard iteration to construct a better approximation of the flow orbit starting from the interpolation. Let us reiterate this construction here. Picard's iteration is a method of creating successive approximations to the solution of the initial value problem

$$\dot{x} = V(x), \quad x(0) = x_0. \tag{3.24}$$

One starts with an arbitrary initial curve $c : [0, 1] \to \mathbb{R}^d$ with initial value $c(0) = x_0$, and in each iteration, constructs a new curve $\lambda = \mathscr{L}_V(c)$ from c via

$$\lambda(t) = c(0) + \int_0^t V(c(s))\, ds. \tag{3.25}$$

Under some regularity conditions on V (such as Lipschitzness), the sequence of Picard iterations $c, \mathscr{L}_V(c), \mathscr{L}_V(\mathscr{L}_V(c)), \ldots$ converges (uniformly) to the solution

[5] The situation is more straightforward when one is interested in the distance of a curve to a *fixed point*. In this case, Figalli and Villani [FV08] show that one can arbitrarily approximate the curve in C^2 topology so that the resulting curve passes the cut locus of the fixed point in a discrete set of times. In our case, we are interested in two curves and the set of times that one curve passes the cut locus of the other.

of the initial value problem (3.24).

Inspired by this, our goal in this step of the proof is to perform one step of Picard iteration on the geodesic interpolation $\boldsymbol{x}$, hoping that the resulting curve is not far away from $\boldsymbol{x}$ and is closer (than $\boldsymbol{x}$) to the integral curve ϕ. This, however, needs some special treatment for Riemannian manifolds, as one cannot integrate vectors living in different tangent spaces as in (3.25).

In other words, our goal is to construct the "integral of V along c" in a Riemannian sense, namely

$$\text{``}\int_0^t V(c(s))\,ds\,\text{,''}$$

for a curve $c : [0,1] \to \mathcal{M}$ and a vector field $V \in \mathfrak{X}(\mathcal{M})$. Instead of defining this integral, we resort to a differential characterization. Notice that in (3.25), the curve λ can be described equivalently as the solution of the initial value problem

$$\dot{\lambda}(t) = V(c(t)), \quad \lambda(0) = c(0).$$

Unfortunately, this ODE also does not make sense in a Riemannian sense as $\dot{\lambda}(t)$ is a tangent vector at $\lambda(t)$, while $V(c(t))$ lives in $T_{c(t)}\mathcal{M}$. To resolve this issue, we parallel transport the vector $V(c(t))$ along the minimizing geodesic connecting $c(t)$ to $\lambda(t)$, that is, we formally consider the following ODE on the manifold:

$$\dot{\lambda}(t) = \mathrm{P}_{c(t)\to\lambda(t)}[V(c(t))], \quad \lambda(0) = c(0). \tag{3.26}$$

In Proposition 3.8 below, we show that this ODE is well-defined and has a unique solution. We call this solution the *Picard curve starting at* $c(0)$ and denote it by $\lambda^{c(0)}$. If $c(0)$ is clear from the context, we drop the superscript and only write λ.

To prove this proposition, we work in a strongly convex neighborhood of $c(0)$; see Section 3.2.10 for a definition. In short, strong-convexity of a neighborhood guarantees that the minimizing geodesics between any two points in the neighborhood is unique and is entirely contained in the neighborhood. In other words, for any two points p and q in the neighborhood, p is in a normal neighborhood centered at q and vice versa.

Proposition 3.8. *Let $c : [0,a] \to \mathcal{M}$ be a smooth curve and V be a smooth vector field along c; let $\mathcal{U}$ be a strongly convex neighborhood of $c(0)$ containing a ball $B_r(c(0))$. Define $t_{\text{exit}} = \inf\{s : c(s) \notin \mathcal{U}\}$ to be the first time of c from $\mathcal{U}$. Then there exists a unique solution to the ODE* (3.26) *defined on* $[0,t]$*, with*

$$t \geq \min\left\{\frac{r}{V^*}, t_{\text{exit}}\right\}, \quad \textit{where} \quad V^* = \max_{t\in[0,a]} |V(c(t))|.$$

Proof. Without loss of generality, let us assume that the image of c is contained entirely in $\mathcal{U}$ (otherwise, one can do the analysis below up to the first exit time of c from $\mathcal{U}$). For $p \in \mathcal{U}$ and $t \in [0, a]$ define $F(t, p) = \mathrm{P}_{c(t)\to p}[V(c(t))]$ to be the parallel transport of $V(c(t))$ from $c(t)$ to p along the minimizing geodesic. Note that the image of this geodesic is entirely contained in $\mathcal{U}$ due to strong-convexity. By Lemma A.1, F depends smoothly on both t and p. Therefore, by the existence and uniqueness theorem of smooth ODEs, there exists a unique curve λ defined on some maximal open interval I (containing 0), satisfying $\lambda(0) = c(0)$ and $\dot{\lambda}(t) = F(t, \lambda(t))$. We now show that the solution of this ODE cannot exit $B_r(c(0))$ for $t \le r/V^*$, that is, $r/V^* \in I$. For this, observe that at any $t \in I$, we have

$$|\dot{\lambda}(t)| = |\mathrm{P}_{c(t)\to\lambda(t)}[V(c(t))]| = |V(c(t))| \le V^*,$$

where in the second equality we used the fact that parallel transport is an isometry between tangent spaces. Therefore,

$$d(\lambda(t), c(0)) \le \int_0^t |\dot{\lambda}(s)|\, ds \le tV^*.$$

Thus, as long as $t < r/V^*$, we have that $\lambda(t) \in B_r(c(0)) \subset \mathcal{U}$. □

We use Proposition 3.8 to construct the Picard curve for the geodesic interpolation $\boldsymbol{x}$, with the property that for all $t > 0$, the Picard curve starting at $\boldsymbol{x}(t)$ is defined up to some arbitrary small but fixed $T > 0$. For this, we have to show that

(a) there is some $r > 0$ such that for all $p \in \mathcal{M}$, the ball $B_r(p)$ is contained in some strongly convex neighborhood of p,

(b) the vector field is uniformly bounded on the entire geodesic interpolation $\boldsymbol{x}$ by some V^*, and

(c) the Picard curve $\lambda^{\boldsymbol{x}(t)}$ is still well-defined for the piecewise-smooth curve $\boldsymbol{x}$.

For (a), it is enough to see that Assumption 3.1, together with Theorem 3.2 imply that the manifold $\mathcal{M}$ has a positive convexity radius $r_{\text{conv}}(\mathcal{M})$. Therefore, we can take r in the proposition above to be r_{conv}.

As the geodesic interpolation $\boldsymbol{x}$ is a continuous piecewise-smooth curve, we can construct the Picard curve on each smooth piece, and "glue" the pieces to obtain a continuous, piecewise-smooth curve λ, thus resolving (c).

To verify (b), we show that the image of the geodesic interpolation is precompact; we then take V^* to be the maximum of $|V|$ on the compact set containing $(\boldsymbol{x}(t))_{t\ge0}$. Recall that Assumption 3.6 requires the iterates $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ to be in a compact set. If we show that the geodesic segments in $\boldsymbol{x}$ are not too long, then including all the geodesic segments between the iterates would not violate precompactness. As the length of the nth segment of $\boldsymbol{x}$ is $\alpha_n|V_n|$ (recall the definition

(3.15) of V_n), we thus have to show that $\alpha_n |V_n|$ is uniformly bounded, almost surely. We actually show in the lemma below that $\alpha_n |V_n| \to 0$.

Lemma 3.9. *It holds that*

$$\lim_{n\to\infty} \alpha_n |V_n| = 0, \quad \textit{almost surely.} \tag{3.27}$$

As a result, the image of the geodesic interpolation $\boldsymbol{x}$ is a precompact set in $\mathcal{M}$.

Proof. Let us decompose the norm of V_n as

$$\alpha_n |V_n| \le \alpha_n |V(\boldsymbol{x}_n)| + \alpha_n |U_n| + \alpha_n |B_n|.$$

As the iterates $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ are precompact and V is continuous, $|V(\boldsymbol{x}_n)|$ is uniformly bounded. Therefore, $\alpha_n |V(\boldsymbol{x}_n)|$ vanishes as $n \to \infty$ because $\alpha_n \to 0$.

Let $\varepsilon > 0$ be arbitrary. By Markov's inequality,

$$\sum_{n=1}^{\infty} \mathbb{P}(\alpha_n |U_n| \ge \varepsilon) \le \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \alpha_n^2 \, \mathbb{E}[|U_n|^2] \le \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \alpha_n^2 \, \mathbb{E}[(U_n^*)^2] < \infty,$$

where we used Assumption 3.4 and the fact that $\mathbb{E}[|U_n|^2] = \mathbb{E}[\mathbb{E}[|U_n|^2 \,|\, \mathcal{F}_n]] \le \mathbb{E}[(U_n^*)^2]$. Thus, by Borel–Cantelli lemma, we can almost surely find an n_0 such that $\alpha_n |U_n| < \varepsilon$ for all $n \ge n_0$. Since ε was arbitrary, we see that $\alpha_n |U_n|$ converges to 0 as $n \to \infty$ with probability 1. The same argument works for bias terms B_n, after observing that Assumption 3.5 implies $\sum \alpha_n^2 \, \mathbb{E}[(B_n^*)^2] < \infty$.

Having (3.27), it is now not difficult to show that the geodesic interpolation $\boldsymbol{x}$ is a precompact subset of $\mathcal{M}$. To see this, fix some arbitrary point $p \in \mathcal{M}$ and choose a radius $R > 0$ such that $\boldsymbol{x}_n \in B_R(p)$ for all $n \in \mathbb{N}$ (one can do this as the manifold is assumed to have Heine–Borel property; see the remark after Assumption 3.1). Also choose n_0 such that $\alpha_n |V_n| < R$ for all $n \ge n_0$. Then, for any point $\boldsymbol{x}(t)$ on the geodesic interpolation with $t \ge \tau_{n_0}$ we have

$$d(\boldsymbol{x}(t), p) \le d(\boldsymbol{x}(t), \boldsymbol{x}_{m(t)}) + d(\boldsymbol{x}_{m(t)}, p) < R + R = 2R,$$

where $m(t)$ is the largest integer n such that $\tau_n \le t$. Thus,

$$\sup_{t\ge 0} d(\boldsymbol{x}(t), p) \le \max\left\{2R, \sup_{t\in[0,\tau_{n_0}]} d(\boldsymbol{x}(t), p)\right\} < \infty. \qquad \square$$

Step 3. Boundedness of the Constructs

Our next step is to show that we can choose $T > 0$ small enough, so that for every $t \geq 0$, the flow orbit $\phi^{\boldsymbol{x}(t)}$ and the Picard curve $\lambda^{\boldsymbol{x}(t)}$ started at $\boldsymbol{x}(t)$, as well as the $[t, t+T]$-segment of the geodesic interpolation $\boldsymbol{x}$ all stay in a strongly convex neighborhood of $\boldsymbol{x}(t)$. Note that this step is only required for manifolds $\mathcal{M}$ with $\operatorname{inj}(\mathcal{M}) < \infty$; the boundedness considerations of this step of the proof are not needed for manifolds that are diffeomorphic to $\mathbb{R}^d$.

Let $r(p) := d(p, \boldsymbol{x}(t))$ be the distance to $\boldsymbol{x}(t)$, also called the *radial distance* if we consider $\boldsymbol{x}(t)$ as the origin. A common technique to bound the distance of some curve to $\boldsymbol{x}(t)$ is to bound the derivative of the radial distance r along that curve, and then integrating that upper bound. Special care has to be taken, however, when considering the derivative of r along a curve, as the curve might pass through the cut locus of $\boldsymbol{x}(t)$; The radial distance might cease to be differentiable on the cut locus (see Proposition 3.1 in Section 3.2.9).

Let us fix t and choose $T > 0$ so that the Picard curve λ started at $\boldsymbol{x}(t)$ is defined up to time T. As the flow orbit and the Picard curve are absolutely continuous, we can use the metric derivative formula (see Lemma A.2) to obtain

$$r(\phi(h)) - r(\phi(0)) = r(\phi(h)) \leq \int_0^h |\dot{\phi}(s)|\, ds = \int_0^h |V(\phi(s))|\, ds.$$

Adding and removing $\mathrm{P}_{\phi(0)\to\phi(s)}[V(\phi(0))]$ from $V(\phi(s))$ gives

$$|V(\phi(s))| \leq |V(\phi(s)) - \mathrm{P}_{\phi(0)\to\phi(s)}[V(\phi(0))]| + |V(\phi(0))|$$

and since V is L-Lipschitz and bounded on the precompact curve $\boldsymbol{x}$,

$$\leq L\, r(\phi(s)) + V^*.$$

Using Grönwall's inequality, we obtain the bound

$$\sup_{0\leq h\leq T} r(\phi(h)) \leq TV^* + LV^* \int_0^T s\, e^{L(T-s)}\, ds = V^*(e^{LT} - 1)/L. \tag{3.28}$$

Similarly, for the Picard curve λ we have that

$$|\dot{\lambda}|(h) = \left|\mathrm{P}_{\boldsymbol{x}(t+h)\to\lambda(h)}[V(\boldsymbol{x}(t+h))]\right| = |V(\boldsymbol{x}(t+h))| \leq V^*$$

for almost every $h \in [0, T]$. Therefore, Lemma A.2 implies

$$\sup_{0\le h\le T} r(\lambda(h)) \le \sup_{0\le h\le T} \int_0^h V^* \, ds = TV^*. \tag{3.29}$$

Therefore, we can choose $T > 0$ small enough so that both of the bounds (3.29) and (3.28) are smaller than r_{conv}, implying that both the flow orbit and the Picard curve lie entirely a strongly convex neighborhood of $\boldsymbol{x}(t)$.

For the geodesic interpolation, we use Assumptions 3.4 and 3.5 for the case of $\operatorname{inj}(\mathcal{M}) < \infty$ and get

$$\begin{aligned}
&\sup_{0\le h\le T} d(\boldsymbol{x}(t), \boldsymbol{x}(t+h)) \\
&\qquad \le \{\alpha_n - (t-\tau_n)\}|V_n| + \alpha_{n+1}|V_{n+1}| + \cdots + \{\alpha_k + (t+T-\tau_k)\}|V_k| \\
&\qquad \le (\{\alpha_n - (t-\tau_n)\} + \alpha_{n+1} + \cdots + \{\alpha_k + (t+T-\tau_k)\}) \cdot (V^* + U^* + B^*) \\
&\qquad = T(V^* + U^* + B^*),
\end{aligned}$$

where $n = m(t)$ and $k = m(t+T)$.

To summarize, we claim that it is sufficient to take

$$T \le \min\left\{ \frac{1}{2L} \log\left(\frac{Lr_{\text{conv}}}{V^*} + 1 \right), \frac{r_{\text{conv}}}{2(V^* + U^* + B^* + 1)} \right\}$$

to ensures that the Picard curve is defined up to T, and that $\phi([0,T])$ and $\lambda([0,T])$, as well as $\boldsymbol{x}([t, t+T])$ are contained in $B_{r_{\text{conv}}}(\boldsymbol{x}(t))$, uniformly for all $t \ge 0$. This is because the exit time of the geodesic interpolation from $B_{r_{\text{conv}}}(\boldsymbol{x}(t))$ is at least $r_{\text{conv}}/(V^* + U^* + B^*)$, and by Proposition 3.8, the choice of T above ensures that the Picard curve is well-defined up to time T.

Remark. Since T is chosen in such a way that the flow orbits and Picard curves are at most r_{conv} far away from their initial points, we can enlarge the compact set containing the geodesic interpolation (resulting from Lemma 3.9) so that it includes, for all $t \ge 0$, the flow and Picard curves starting at $\boldsymbol{x}(t)$, while remaining compact. For brevity, we reuse the notation V^* for the maximum norm of V on this enlarged compact set. ♢

For the rest of the proof, we choose T in the following way: When $\operatorname{inj}(\mathcal{M}) < \infty$, we set

$$T \le \min\left\{ \frac{1}{2L} \log\left(\frac{Lr_{\text{conv}}}{V^*} + 1 \right), \frac{r_{\text{conv}}}{2(V^* + U^* + B^* + 1)}, \frac{\pi}{2\sqrt{\kappa_{\text{up}}}} \right\}, \tag{3.30a}$$

and when $\operatorname{inj}(\mathcal{M}) = \infty$, we only assume

$$T \leq \min\bigl\{\pi/2\sqrt{\kappa_{\text{up}}}, +\infty\bigr\}. \tag{3.30b}$$

In both cases, we set $1/\sqrt{\kappa_{\text{up}}} = +\infty$ if $\kappa_{\text{up}} \leq 0$.

Digression: The Fermi Normal Coordinates

Recall the notion of normal coordinates as discussed in Section 3.2.8. Normal coordinates are suitable for computing distance to a fixed point locally. However, in a situation like ours, we have to compute pointwise distance of two curves, namely the flow orbit and the geodesic interpolation. The *Fermi coordinate system* (FCS) [MM63], roughly speaking, is a system of normal coordinates along a curve, and turns out to be the right tool for our computations. As we encounter Fermi coordinates several times in the rest of the proof, it is no waste of time to make a short digression in order to give a short overview of their construction and their key attributes. It is imperative to direct the reader's attention to the foundational concepts and notations related to coordinate charts and tangent vectors, reviewed in Section 3.2.1. A thorough comprehension of these preliminary notions is essential for accurately interpreting and assimilating the following discussions.

Let $I = [0, T]$ and consider a smooth curve $c : I \to \mathcal{M}$ with an arbitrary orthonormal frame $(E_1(0), \ldots, E_d(0))$ for $T_{c(0)}\mathcal{M}$. We obtain a system of orthonormal frames $(E_i(t))$ for $t \in [0, T]$ by parallel transporting each $E_i(0)$ from $T_{c(0)}\mathcal{M}$ to $T_{c(t)}\mathcal{M}$ along c. For each $t \in I$, let $\mathcal{U}_t \subseteq \mathcal{M}$ be a normal neighborhood around $c(t)$. The *Fermi normal coordinates along the curve* c is then a diffeomorphism ψ between $\mathcal{U} = \bigcup_{t \in I}(\{t\} \times \mathcal{U}_t)$, which is a neighborhood of the curve $t \mapsto (t, c(t))$ in the product manifold $I \times \mathcal{M}$,[6] and some neighborhood $\mathcal{V}$ in $I \times \mathbb{R}^d$ of the curve $t \mapsto (t, 0)$, and is given by the relation[7]

$$\psi\bigl(t, \exp_{c(t)}\bigl(\textstyle\sum x^i E_i(t)\bigr)\bigr) = (t, (x^1, \ldots, x^d)). \tag{3.31}$$

In other words, for a fixed $t \in I$, $\psi(t, \cdot)$ is the normal coordinate chart centered at $c(t)$ with the basis $(E_i(t))$. We define $x^i = x^i(t, p)$ to be the ith coordinate of the point p in the normal coordinates $\psi(t, \cdot)$, and write $x = x(t, p) = (x^1(t, p), \ldots, x^d(t, p)) \in \mathbb{R}^d$. See Fig. 3.3 for an illustration.

The coordinate chart ψ induces the basis $\{\partial/\partial t, \partial/\partial x^1, \ldots, \partial/\partial x^d\}$ on the

[6] In the literature, $\mathcal{U}$ is sometimes called a *tubular neighborhood* of the (one-dimensional) submanifold $\{(t, c(t)) : t \in I\}$ of $I \times \mathcal{M}$ [Lee18, p. 139].

[7] There are two equivalent ways to define the Fermi coordinates. If one defines coordinate charts to be from open sets in $\mathbb{R}^d$ to neighborhoods on the manifold (which is the choice of do Carmo [Car92]), one has to replace ψ with ψ^{-1} in the definitions above. Our choice is merely for consistency with the rest of this chapter and that of Lee [Lee18].

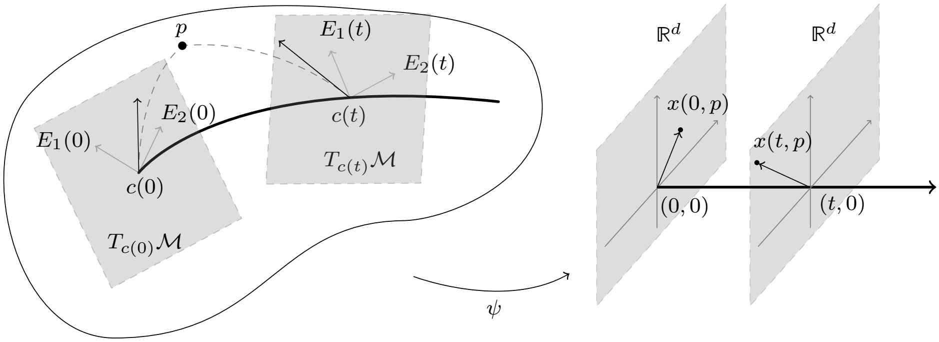


Figure 3.3. Fermi coordinates along the curve c. *Left:* An arbitrary orthonormal basis $\{E_i(0)\}$ is chosen for $T_{c(0)}\mathcal{M}$ (shown as gray rectangles), and then parallel transported along c to get $\{E_i(t)\}$. Any point $p \in \mathcal{M}$ that is in the normal neighborhood of $c(t)$ will get normal coordinates $(x^i(t,p))_{i=1}^d$. The dashed line represents the geodesic connecting points on c to p, and the black arrow represents the initial velocity of this geodesic. *Right:* The coordinates of the point p at two different times 0 and t.

tangent space of each point (t,p) in $\mathcal{U} \subset I \times \mathcal{M}$, in the sense that for any smooth function f on $\mathcal{U}$ we have

$$\left.\frac{\partial}{\partial t}\right|_{(t,p)} f = \left.\frac{\partial}{\partial t}\right|_{(t,x)} (f \circ \psi^{-1}) = \frac{\partial (f \circ \psi^{-1})}{\partial t}(t,x), \tag{3.32}$$

with the right-hand side interpreted in the Euclidean sense. Similar relation holds for $\partial/\partial x^i|_{(t,p)}$.

To do computations within the Fermi coordinate system, we have to be able to compute the components of tangent vectors in $T\mathcal{U}$. Let us start by computing the components of a vector field V on $\mathcal{M}$ in the Fermi coordinate chart ψ. As we work with the product manifold $I \times \mathcal{M}$, we first extend V to a vector field on $I \times \mathcal{M}$ by defining its action at (t_0, p_0) on smooth functions $f : I \times \mathcal{M} \to \mathbb{R}$ to be the same as the action of V_{p_0} on the function $p \mapsto f(t_0, p)$. We use the same symbol V to denote this extended vector field. Expanding $V_{(t_0,p_0)}$ in the basis of $T_{(t_0,p_0)}\mathcal{U}$ gives

$$V_{(t_0,p_0)} = a(t_0,p_0) \left.\frac{\partial}{\partial t}\right|_{(t_0,p_0)} + \sum_{i=1}^{d} V^i(t_0,p_0) \left.\frac{\partial}{\partial x^i}\right|_{(t_0,p_0)}.$$

Observe that the component corresponding to $\partial/\partial t$ is zero, as

$$a(t_0,p_0) = V_{(t_0,p_0)}((t,p) \mapsto t) = V_{p_0}(p \mapsto t_0) = 0,$$

and for the component corresponding to $\partial/\partial x^i$ we have

$$V^i(t_0,p_0) = V_{(t_0,p_0)}(x^i) = V_{p_0}(p \mapsto x^i(t_0,p)),$$

which is precisely the ith component of the vector field V in the normal coordinates $\psi(t_0,\cdot)$.

Now suppose $(t_0,p_0) \in \mathcal{U}$. Besides the tangent vector $\partial/\partial t|_{(t_0,p_0)}$, which corresponds to differentiation with respect to t in the coordinates defined by ψ (see (3.32)), we can also consider differentiation with respect to the time variable directly on the product manifold $I \times \mathcal{M}$. The corresponding tangent vector is the velocity vector of the curve $\beta : t \mapsto (t,p_0) \in I \times \mathcal{M}$ at $t = t_0$. We denote this tangent vector by $(\partial/\partial t)^{\mathbb{R}}$ to distinguish it from $\partial/\partial t$. Thus, for any smooth function f defined on $\mathcal{U}$ we have

$$\frac{\partial}{\partial t}^{\mathbb{R}}\bigg|_{(t_0,p_0)} f = \frac{d}{dt}\bigg|_{t=t_0} (f \circ \beta)(t) = \frac{d}{dt}\bigg|_{t=t_0} f(t,p_0).$$

The following lemma from Fujita and Kotani [FK82] gives the components of this tangent vector in the Fermi normal coordinates. We have adapted the notation of the original lemma to match ours.

Lemma 3.10 (FK82, Lem. 1.2)**.** *Let $c : I \to \mathcal{M}$ be a smooth curve and consider the Fermi normal coordinates $\psi : \mathcal{U} \to \mathcal{V}$ along c defined as above. Let $c^i(s;t)$ be the ith component of $c(s)$ in the normal coordinate $\psi(t,\cdot)$ and define*

$$(c')^i(t) = \frac{d}{ds}\bigg|_{s=t} c^i(s;t) = \lim_{s\to t} \frac{c^i(s;t) - c^i(t;t)}{s-t}. \tag{3.33}$$

Then it holds

$$\frac{\partial}{\partial t}^{\mathbb{R}}\bigg|_{(t,p)} = \frac{\partial}{\partial t}\bigg|_{(t,p)} - \sum_{i=1}^{d}\{(c')^i(t) + \varepsilon^i(t,x)\}\frac{\partial}{\partial x^i}\bigg|_{(t,p)}, \tag{3.34}$$

where $x = x(t,p)$ and $\varepsilon^i(t,x)$ are smooth functions satisfying

$$\max_{t\in I} |\varepsilon^i(t,x)| = O(\|x\|_2^2) \text{ as } \|x\| \to 0,$$

where $O(\cdot)$ hides constants that only depend on the Riemannian curvature tensor.

It is worth reiterating the key point from Lemma 3.10. When considering the time derivative in the Fermi coordinate chart, (3.34) tells us that we have to take the *movement of the center* $c(t)$ *of the coordinate system* into account. Intuitively, this means that the velocity of a moving point, as seen from a moving observer $c(t)$, is approximately its instantaneous velocity minus the observer's instantaneous velocity. Moreover, due to the manifold's nonlinearity, this approximation worsens when the moving point gets farther from the observer.

To make this intuition more precise, below we compute the expression of the velocity of one curve in the Fermi coordinates around the other. This computation turns out to be a simple corollary of Lemma 3.10, and is essential for the rest of the proof of Theorem 3.4.

Corollary 3.11. *Consider the Fermi coordinate system* $\psi : \mathcal{U} \to \mathcal{V}$ *around the curve* $c : [0, T] \to \mathcal{M}$. *Let* $\beta : [0, T] \to \mathcal{M}$ *be another smooth curve on* $\mathcal{M}$ *such that* $\beta(t) \in \mathcal{U}_t$ *for all* $t \in [0, T]$. *Let* $\hat{\beta}(t) = x(t, \beta(t))$ *be the normal coordinates of* $\beta(t)$ *in the chart* $\psi(t, \cdot)$. *Moreover, let* $\dot{\beta}^i(t)$ *be the* i*th component of the velocity vector of* $\beta(t)$ *in this normal coordinates. Then the velocity of the curve* $t \mapsto \hat{\beta}(t) \in \mathbb{R}^d$ *is given by the vector* $(\dot{\hat{\beta}}^1(t), \ldots, \dot{\hat{\beta}}^d(t)) \in \mathbb{R}^d$ *with*

$$\dot{\hat{\beta}}^i(t) = \dot{\beta}^i(t) - (c')^i(t) + O(\|\hat{\beta}(t)\|_2^2),$$

where $(c')^i(t)$ *is defined in* (3.33).

Proof. Let $I = [0, T]$; consider the curve $c : I \to \mathcal{U}$ defined as $c(t) = (t, \beta(t))$. The velocity vector of c at time $t_0 \in I$ is by definition $\dot{c}(t_0) = (dc)_{t_0}((\partial/\partial t)^{\mathbb{R}}|_{t_0})$. Since c is a smooth map to a product manifold, its differential has a product form $(dc)_{t_0} = \mathrm{Id} \oplus (d\beta)_{t_0}$.[8] Therefore, $\dot{c}(t_0) = (\partial/\partial t)^{\mathbb{R}}|_{(t_0, \beta(t_0))} + \dot{\beta}(t_0)$. The components of this tangent vector in the normal coordinates centered at $c(t_0)$ can be computed via Lemma 3.10 and the definition of $\dot{\beta}^i(t_0)$:

$$\begin{aligned}\dot{c}(t_0) &= \frac{\partial}{\partial t}^{\mathbb{R}}\bigg|_{(t_0,\beta(t_0))} + \sum_{i=1}^{d} \dot{\beta}^i(t_0) \frac{\partial}{\partial x^i}\bigg|_{(t_0,\beta(t_0))} \\ &= \frac{\partial}{\partial t}\bigg|_{(t_0,\beta(t_0))} + \sum_{i=1}^{d} \Big\{\dot{\beta}^i(t) - (c')^i(t_0) - \varepsilon^i(t_0, \hat{\beta}(t_0))\Big\} \frac{\partial}{\partial x^i}\bigg|_{(t_0,\beta(t_0))}.\end{aligned}$$

This implies that the velocity of the curve $(t, \hat{\beta}(t)) = \psi(t, \beta(t))$ in $I \times \mathbb{R}^d$ is the vector $(1, \dot{\hat{\beta}}^1(t), \ldots, \dot{\hat{\beta}}^d(t))$ with $\dot{\hat{\beta}}^i(t)$ given as in the statement of the corollary. □

[8] We are using the identification $T_t I \times T_p\mathcal{M} \cong T_{(t,p)}(I \times \mathcal{M})$ [Lee12, Prop. 3.14].

Remark 3.1. We can compute $(c')^i(t)$ defined in (3.33) explicitly when the curve c is a geodesic. Observe that in this case, for s close to t we have

$$c(s) = \exp_{c(t)}((s-t)\,\dot{c}(t)).$$

Therefore, in the normal coordinates $\psi(t,\cdot)$, the ith coordinate of $c(s)$ will be

$$c^i(s;t) = (s-t)\langle \dot{c}(t), E_i(t)\rangle.$$

Taking the derivative with respect to s at $s = t$ yields

$$(c')^i(t) = \langle \dot{c}(t), E_i(t)\rangle.$$

Moreover, as the frame $(E_i(t))$ is parallel along c, $\langle \dot{c}(t), E_i(t)\rangle$ is constant and independent of t. Therefore, for all $t \in [0,T]$ and $i = 1,\dots,d$,

$$(c')^i(t) = (c')^i(0) = \langle \dot{c}(0), E_i(0)\rangle. \tag{3.35}$$

We will use this formula in the proof that follows. ♢

Step 4. Controlling distances

After this brief digression, let us return to our original problem of computing pointwise distances between two curves. Recall that our objective is to bound

$$\sup_{h\in[0,T]} d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t)))$$

by a function of t that vanishes as $t \to \infty$. Following the argument in Section 3.5, we take T to be small enough (see (3.30)), so that all the curves of our interest (the geodesic interpolation $\boldsymbol{x}([t,t+T])$, the Picard curve $\lambda([0,T])$, and the flow orbit $\phi([0,T])$) fall in a strongly convex neighborhood of $\boldsymbol{x}(t)$. We bound the desired distance by decomposing it into the distance from $\boldsymbol{x}$ to λ and the distance from λ to ϕ, doing all the computations in an appropriate Fermi coordinate system.

Let us begin with an arbitrary orthonormal frame $(E_k(0))_{k=1}^d$ at $T_{\boldsymbol{x}(0)}\mathcal{M}$, and parallel transport this frame along $\boldsymbol{x}$ to obtain $(E_k(t))$ for all $t \geq 0$. This is possible since $\boldsymbol{x}$ is a piecewise-smooth curve [Lee18, Cor. 4.33]. We consider each time interval where $\boldsymbol{x}$ is smooth, and in each of these intervals, we construct the Fermi coordinates ψ centered around $\boldsymbol{x}$ using the frames $(E_k(\cdot))$. Similar to Corollary 3.11, we will express the coordinate representation of a curve c in the Fermi coordinate system with $\hat{c}$, so that $\hat{c}(t)$ is the normal coordinates of $c(t)$ in the chart $\psi(t,\cdot)$.

Suppose $\boldsymbol{x}$ is smooth in $I = [a, b] \subset [t, t+T]$. For any $h \in [a-t, b-t]$ we compute

$$\begin{aligned} d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t))) = d(\boldsymbol{x}(t+h), \phi(h)) &= \|\hat{\phi}(h)\|_2 \\ &\leq \|\hat{\phi}(h)\|_2 + \|\hat{\lambda}(h)\|_2 \\ &\leq \|\hat{\phi}(h) - \hat{\lambda}(h)\|_2 + 2\|\hat{\lambda}(h)\|_2. \end{aligned} \tag{3.36}$$

To avoid confusion with the (Riemannian) norm of tangent vectors, we use $\|\cdot\|_2$ for the usual Euclidean norm of a vector in $\mathbb{R}^d$.

As $\hat{\phi}$ and $\hat{\lambda}$ are smooth in I (recall from Proposition 3.8 that λ is smooth wherever $\boldsymbol{x}(t+\cdot)$ is smooth), we can bound their distance by integrating the difference of their velocities, that is,

$$\|\hat{\phi}(h) - \hat{\lambda}(h)\|_2 - \|\hat{\phi}(a) - \hat{\lambda}(a)\|_2 = \left\| \int_a^h (\dot{\hat{\phi}}(s) - \dot{\hat{\lambda}}(s))\, ds \right\|_2.$$

By Corollary 3.11 and the definition of the flow orbit (3.16) and the Picard curve (3.26), we can compute the velocities in Fermi coordinates

$$\begin{aligned} \dot{\hat{\phi}}^i(s) &= \dot{\phi}^i(s) - (\boldsymbol{x}')^i(t+s) + O(\|\hat{\phi}(s)\|_2^2), \quad \text{and} \\ \dot{\hat{\lambda}}^i(s) &= \dot{\lambda}^i(s) - (\boldsymbol{x}')^i(t+s) + O(\|\hat{\lambda}(s)\|_2^2), \end{aligned}$$

where $(\boldsymbol{x}')^i(t+s)$ is defined as in (3.33). We recall that $\dot{\phi}(s)$ (resp. $\dot{\lambda}(s)$) is the velocity vector of the flow orbit $\phi(\cdot)$ (resp. Picard curve $\lambda(\cdot)$) at time s, and its components in the normal coordinate system $\psi(t+s, \cdot)$ centered at $\boldsymbol{x}(t+s)$ is $\dot{\phi}^i(s)$ (resp. $\dot{\lambda}^i(s)$). In what follows, it is useful to pack the components of a tangent vector into a Euclidean vector and compare vectors in the Euclidean sense. For this, we use the notation $\operatorname{vec}(\cdot)$. For example, $\operatorname{vec}(\dot{\phi}^i(s))$ is a vector in $\mathbb{R}^d$ with components $(\dot{\phi}^1(s), \dots, \dot{\phi}^d(s))$. With this notation, we have

$$\begin{aligned} \|\hat{\phi}(h) - \hat{\lambda}(h)\|_2 - \|\hat{\phi}(a) - \hat{\lambda}(a)\|_2 &= \left\| \int_a^h (\dot{\hat{\phi}}(s) - \dot{\hat{\lambda}}(s))\, ds \right\|_2 \\ &\leq \left\| \int_a^h (\operatorname{vec}(\dot{\phi}^i(s)) - \operatorname{vec}(\dot{\lambda}^i(s)))\, ds \right\|_2 + \int_a^h R_1(s)\, ds, \end{aligned} \tag{3.37}$$

where the remainder term $R_1(s) = O(\|\hat{\phi}(s)\|_2^2 + \|\hat{\lambda}(s)\|_2^2)$. Similarly, for the other term in the decomposition (3.36) we have

$$\begin{aligned}&\|\hat{\lambda}(h)\|_2 - \|\hat{\lambda}(a)\|_2 \\ &\qquad \leq \left\| \int_a^h (\mathrm{vec}(\dot{\lambda}^i(s)) - \mathrm{vec}((\boldsymbol{x}')^i(t+s)))\, ds \right\|_2 + \int_a^h R_2(s)\, ds \qquad (3.38)\end{aligned}$$

with $R_2(s) = O(\|\hat{\lambda}(s)\|_2^2)$. Note that by our choice of T, we have

$$\|\hat{\lambda}(s)\|_2 = d(\boldsymbol{x}(t+s), \lambda(s)) \leq d(\boldsymbol{x}(t+s), \boldsymbol{x}(t)) + d(\boldsymbol{x}(t), \lambda(s)) \leq 2r_{\mathrm{conv}},$$

and similarly for $\|\hat{\phi}(s)\|_2$. Therefore,

$$R_1(s) = O(r_{\mathrm{conv}}(\|\hat{\phi}(s)\|_2 + \|\hat{\lambda}(s)\|_2)) = O(\|\hat{\phi}(s)\|_2 + \|\hat{\lambda}(s)\|_2),$$

and similarly, $R_2(s) = O(\|\hat{\lambda}(s)\|_2)$.

Remark 3.2. We remark that we can glue the curves $\hat{\phi}$ and $\hat{\lambda}$ defined on sub-intervals that $\boldsymbol{x}(t+\cdot)$ is smooth, to obtain the *continuous* piecewise-smooth curves $\hat{\phi}, \hat{\lambda} : [0,T] \to \mathbb{R}^d$. This is possible as we use the same normal coordinate system at the boundary of each sub-interval to define $\hat{\phi}$ and $\hat{\lambda}$. We can therefore combine the inequalities (3.37) and (3.38) by summing over all sub-intervals that $\boldsymbol{x}$ is smooth to obtain for all $h \in [0,T]$,

$$\|\hat{\phi}(h) - \hat{\lambda}(h)\|_2 \leq \left\| \int_0^h (\mathrm{vec}(\dot{\phi}^i(s)) - \mathrm{vec}(\dot{\lambda}^i(s)))\, ds \right\|_2 + \int_0^h R_1(s)\, ds, \qquad (3.39)$$

and

$$\|\hat{\lambda}(h)\|_2 \leq \left\| \int_0^h (\mathrm{vec}(\dot{\lambda}^i(s)) - \mathrm{vec}((\boldsymbol{x}')^i(t+s)))\, ds \right\|_2 + \int_0^h R_2(s)\, ds, \qquad (3.40)$$

where we used the fact that $\hat{\phi}(0) = \hat{\lambda}(0) = 0$ as $\phi(0) = \lambda(0) = \boldsymbol{x}(t)$. The only caveat to bear in mind is that we have to refer to the FCS at the corresponding time interval to make sense of the meaning of $\hat{\phi}$ and $\hat{\lambda}$. Moreover, $\mathrm{vec}(\dot{\phi}^i(s))$ and $\mathrm{vec}(\dot{\lambda}^i(s))$ are defined everywhere except on a discrete set of times. ◊

Step 5. From Fermi to Parallel Frames

So far, we have reduced the proof to bounding (3.39) and (3.40) from above. However, the components $\dot{\phi}^i$ and $\dot{\lambda}^i$ in these equations are not amenable to further computation as they are given by a normal coordinate system. This is the main downside of using normal coordinates, as normal coordinate systems do not necessarily induce an orthonormal basis for tangent spaces other than that of the center of the coordinate system.

To remedy this, we first consider frames at $\lambda(s)$ and $\phi(s)$ that are parallel to $(E_k(t+s))$. In these frames, we will see shortly that we can easily compare the vectors $\dot\lambda(s)$ and $\dot\phi(s)$. We then show that this comparison is close to the desired comparison in normal coordinates.

Concretely, let $(E'_k(s))$ be an orthonormal basis of $T_{\lambda(s)}\mathcal{M}$ obtained by parallel transporting $(E_k(t+s))$ from $\boldsymbol{x}(t+s)$ to $\lambda(s)$ along the minimizing geodesic. The ith component of $\dot\lambda(s)$ in this frame, denoted by $\dot\lambda^{i,\shortparallel}(s)$, is simply

$$\begin{aligned}
\dot\lambda^{i,\shortparallel}(s) &= \langle \dot\lambda(s), E'_i(s)\rangle \\
&= \langle \mathrm{P}_{\boldsymbol{x}(t+s)\to\lambda(s)}[V(\boldsymbol{x}(t+s))], E'_i(s)\rangle \\
&= \langle \mathrm{P}_{\boldsymbol{x}(t+s)\to\lambda(s)}[V(\boldsymbol{x}(t+s))], \mathrm{P}_{\boldsymbol{x}(t+s)\to\lambda(s)}[E_i(t+s)]\rangle \\
&= \langle V(\boldsymbol{x}(t+s)), E_i(t+s)\rangle.
\end{aligned}$$

Similarly define $(E''_k(s))$ to be an orthonormal basis of $T_{\phi(s)}\mathcal{M}$ obtained by parallel transporting $(E_k(t+s))$ from $\boldsymbol{x}(t+s)$ to $\phi(s)$ along the minimizing geodesic, and let $\dot\phi^{i,\shortparallel}(s)$ be the ith component of $\dot\phi(s)$ in this frame. Then,

$$\begin{aligned}
\dot\phi^{i,\shortparallel}(s) - \dot\lambda^{i,\shortparallel}(s) &= \langle V(\phi(s)), E''_i(s)\rangle - \langle V(\boldsymbol{x}(t+s)), E_i(t+s)\rangle \\
&= \langle \mathrm{P}_{\phi(s)\to\boldsymbol{x}(t+s)}[V(\phi(s))] - V(\boldsymbol{x}(t+s)), E_i(t+s)\rangle
\end{aligned}$$

Therefore, since $(E_i(t+s))$ is orthonormal,

$$\begin{aligned}
\|\mathrm{vec}(\dot\phi^{i,\shortparallel}(s)) - \mathrm{vec}(\dot\lambda^{i,\shortparallel}(s))\|_2 &= |\mathrm{P}_{\phi(s)\to\boldsymbol{x}(t+s)}[V(\phi(s))] - V(\boldsymbol{x}(t+s))| \\
&\le L\, d(\phi(s), \boldsymbol{x}(t+s)) \\
&= L\|\hat\phi(s)\|_2. \qquad (3.41)
\end{aligned}$$

We would like to replace $\mathrm{vec}(\dot\phi^{i}(s)) - \mathrm{vec}(\dot\lambda^{i}(s))$ in (3.39) with $\mathrm{vec}(\dot\phi^{i,\shortparallel}(s)) - \mathrm{vec}(\dot\lambda^{i,\shortparallel}(s))$. This is not straightforward, though, as these components are coming from two different frames, one being orthonormal, and the other coming from a normal coordinate system. In Lemma 3.12 below, we show a way to compare coordinates in these two frames.

▶ **Lemma 3.12.** *Suppose the sectional curvatures of $\mathcal{M}$ are bounded between κ_{low} and κ_{up}; let $p, q \in \mathcal{M}$ with q within the injectivity radius of p. If $\kappa_{\mathrm{up}} > 0$, assume further that $d(p,q) < \pi/2\sqrt{\kappa_{\mathrm{up}}}$. Consider an orthonormal frame (E_i) for $T_p\mathcal{M}$ and the parallel frame (E'_i) obtained by parallel transporting (E_i) from p to q along the minimizing geodesic. Let $v \in T_q\mathcal{M}$; let the components of v in the* *frames induced by the normal coordinates and the parallel frame* (E_i') *be* v^i *and* $v^{i,\shortparallel}$ *respectively. Then one has the estimate*

$$\|\operatorname{vec}(v^i) - \operatorname{vec}(v^{i,\shortparallel})\|_2 \leq \frac{\pi}{2} \cdot \kappa_{\max} \cdot f_{\kappa_{\text{low}}}(d(p,q)) \cdot |v|,$$

where $\kappa_{\max} = \max(|\kappa_{\text{up}}|, |\kappa_{\text{low}}|)$ *and* $f_{\kappa_{\text{low}}}$ *is defined as*

$$f_{\kappa_{\text{low}}}(a) = \begin{cases} \frac{a^2}{6} & \text{if } \kappa_{\text{low}} = 0, \\ r^2\left(1 - \frac{\sin(a/r)}{a/r}\right) & \text{if } \kappa_{\text{low}} = \frac{1}{r^2} > 0, \\ r^2\left(\frac{\sinh(a/r)}{a/r} - 1\right) & \text{if } \kappa_{\text{low}} = -\frac{1}{r^2} < 0. \end{cases}$$

Proof. Let $w \in T_p\mathcal{M}$ be such that $q = \exp_p(w)$ and $|w| = d(p,q)$. First, observe that $\partial/\partial x^i|_q = (d\exp_p)_w(E_i)$. Therefore,

$$v = (d\exp_p)_w\left(\sum v^i E_i\right).$$

Define the tangent vector $\tilde{v} \in T_p\mathcal{M}$ as $\tilde{v} = \sum v^i E_i$. From the linearity of parallel transport we obtain

$$\mathrm{P}_{p\to q}[\tilde{v}] = \sum v^i E_i'.$$

Thus, $v - \mathrm{P}_{p\to q}[\tilde{v}] = \sum(v^{i,\shortparallel} - v^i)E_i'$ and

$$\begin{aligned}\|\operatorname{vec}(v^i) - \operatorname{vec}(v^{i,\shortparallel})\|_2 &= |v - \mathrm{P}_{p\to q}[\tilde{v}]| \\ &= |(d\exp_p)_w(\tilde{v}) - \mathrm{P}_{p\to q}[\tilde{v}]|_q \\ &\leq \kappa_{\max} \cdot f_{\kappa_{\text{low}}}(|w|) \cdot |\tilde{v}|,\end{aligned}$$

where we used Lemma A.3. We are thus left with bounding $|\tilde{v}|_p$ in terms of $|v|_q$. For this, we decompose $\tilde{v}$ into $\tilde{v}_\perp + \tilde{v}_{\tan}$, where $\tilde{v}_\perp$ is perpendicular to w and $\tilde{v}_{\tan}$ is parallel to it. We have by Gauss's lemma that

$$\langle (d\exp_p)_w(\tilde{v}_{\tan}), (d\exp_p)_w(w)\rangle = \langle \tilde{v}_{\tan}, w\rangle$$

and

$$|(d\exp_p)_w(\tilde{v}_{\tan})| = |\tilde{v}_{\tan}|. \tag{3.42}$$

Moreover, as $|w| = d(p,q) \leq \pi/2\sqrt{\kappa_{\text{up}}}$ (if $\kappa_{\text{up}} > 0$), we can use Rauch's lower bound [BK81, Prop. 6.4] to obtain

$$|\tilde{v}_\perp| \leq |(d\exp_p)_w(\tilde{v}_\perp)| \cdot \frac{|w|}{s_{\kappa_{\text{up}}}(|w|)}, \tag{3.43}$$

where $s_\kappa : \mathbb{R} \to \mathbb{R}$ is defined as

$$s_\kappa(a) = \begin{cases} a & \text{if} \quad \kappa = 0, \\ \frac{1}{r}\sin(ra) & \text{if} \quad \kappa = \frac{1}{r^2} > 0, \\ \frac{1}{r}\sinh(ra) & \text{if} \quad \kappa = -\frac{1}{r^2} < 0. \end{cases}$$

Thus, when $\kappa_{\text{up}} > 0$, (3.43) reduces to

$$|\tilde{v}_\perp| \leq \frac{\pi}{2} \cdot |(d\exp_p)_w(\tilde{v}_\perp)|,$$

as $\frac{x}{\sin x} \leq \pi/2$ for $x \in [0, \pi/2]$, and when $\kappa_{\text{up}} < 0$, it becomes

$$|\tilde{v}_\perp| \leq |(d\exp_p)_w(\tilde{v}_\perp)|,$$

as $\frac{x}{\sinh x} \leq 1$ for $x \in \mathbb{R}$. By combining these two cases, we obtain

$$\begin{aligned} |\tilde{v}|^2 = |\tilde{v}_\perp|^2 + |\tilde{v}_{\tan}|^2 &\leq \frac{\pi^2}{4}\left(|(d\exp_p)_w(\tilde{v}_\perp)|^2 + |(d\exp_p)_w(\tilde{v}_{\tan})|^2\right) \\ &= \frac{\pi^2}{4}|(d\exp_p)_w(\tilde{v})|^2 = \frac{\pi^2}{4}|v|^2. \end{aligned} \qquad \square$$

Using Lemma 3.12 and the fact that the time horizon T is chosen in such a way that $d(\phi(s), \boldsymbol{x}(t+s)) < \pi/2\sqrt{\kappa_{\text{up}}}$ when $\kappa_{\text{up}} > 0$, we obtain the bound

$$\|\text{vec}(\dot{\phi}^i(s)) - \text{vec}(\dot{\phi}^{i,\shortparallel}(s))\|_2 \leq \frac{\pi}{2} \cdot \kappa_{\max} \cdot f_{\kappa_{\text{low}}}(\|\hat{\phi}(s)\|_2) \cdot |\dot{\phi}(s)|. \tag{3.44}$$

What is left is to bound $f_{\kappa_{\text{low}}}$. Lemma A.3 in the appendix shows that $f_{\kappa_{\text{low}}}$ is dominated by $f_{-\kappa_{\max}}$. Using the inequalities $(\sinh x)/x \leq e^x$ for all $x \geq 0$, and $e^{ax} - 1 \leq x \cdot (e^{aR} - 1)/R$ for $x \in [0, R]$, we get

$$\begin{aligned} f_{\kappa_{\text{low}}}(\|\hat{\phi}(s)\|_2) &\leq \frac{1}{\kappa_{\max}}\left(e^{\sqrt{\kappa_{\max}}\,\|\hat{\phi}(s)\|_2} - 1\right) \\ &\leq \frac{e^{2r_{\text{conv}}\sqrt{\kappa_{\max}}} - 1}{2r_{\text{conv}}\,\kappa_{\max}} \cdot \|\hat{\phi}(s)\|_2, \end{aligned} \tag{3.45}$$

where we used the fact that $\phi(s)$ is in the ball $B_{r_{\text{conv}}}(\boldsymbol{x}(t))$, and therefore, its distance to $\boldsymbol{x}(t+s)$ is at most $2r_{\text{conv}}$. Combining (3.45) with (3.44) and recalling that the vector field V (and hence $|\dot{\phi}(s)|$) is bounded from above by V^*, we get

$$\|\text{vec}(\dot{\phi}^i(s)) - \text{vec}(\dot{\phi}^{i,\shortparallel}(s))\|_2 =: R_3(s) = O(\|\hat{\phi}(s)\|_2), \tag{3.46}$$

where O hides constants depending on $V^*, \kappa_{\max}$, and r_{conv}. Similarly, for the Picard curve we obtain

$$\|\operatorname{vec}(\dot{\lambda}^i(s)) - \operatorname{vec}(\dot{\lambda}^{i,\parallel}(s))\|_2 =: R_4(s) = O(\|\hat{\lambda}(s)\|_2). \tag{3.47}$$

We are now in a position to use the results of this step and bound (3.39) further. Using Eqs. (3.41), (3.46) and (3.47), we obtain

$$\begin{aligned}
&\|\hat{\phi}(h) - \hat{\lambda}(h)\|_2 \\
&\qquad \leq \left\| \int_0^h (\operatorname{vec}(\dot{\phi}^i(s)) - \operatorname{vec}(\dot{\lambda}^i(s)))\, ds \right\|_2 + \int_0^h R_1(s)\, ds, \\
&\qquad \leq \int_0^h \|\operatorname{vec}(\dot{\phi}^i(s)) - \operatorname{vec}(\dot{\lambda}^i(s))\|_2\, ds + \int_0^h R_1(s)\, ds, \\
&\qquad \leq \int_0^h \|\operatorname{vec}(\dot{\phi}^{i,\parallel}(s)) - \operatorname{vec}(\dot{\lambda}^{i,\parallel}(s))\|_2\, ds + \int_0^h (R_1 + R_3 + R_4)(s)\, ds, \\
&\qquad \leq L \int_0^h \|\hat{\phi}(s)\|_2\, ds + \int_0^h (R_1 + R_3 + R_4)(s)\, ds.
\end{aligned} \tag{3.48}$$

The resulting bound is very promising, as its first term reminds us of the Grönwall's lemma. However, we have to keep the remainder terms under control.

Step 6. Distance of the Picard Curve to the Interpolation

As promised in the beginning of the proof, in this step, we deal with the noise and bias in the algorithm. Similar to the Euclidean proof in Section 2.5, it will be convenient to introduce objects that relate continuous-time constructs to their discrete-time counterpart. We have already seen the effective time τ_n and the continuous-to-discrete counter $m(t) = \sup\{n \geq 1 : \tau_n \leq t\}$. For an arbitrary process $A_1, A_2, \ldots$, let us define the continuous-time, piecewise-constant interpolation $\overline{A}(t)$ as

$$\overline{A}(t) = A_n, \quad t \in [\tau_n, \tau_{n+1}). \tag{3.49}$$

Using this notation, we can write the geodesic interpolation (GI) in differential form:

$$\dot{\boldsymbol{x}}(t) = \mathrm{P}^{\boldsymbol{x}}_{\boldsymbol{x}_n \to \boldsymbol{x}(t)} \big[V(\overline{\boldsymbol{x}}(t)) + \overline{U}(t) + \overline{B}(t) \big], \quad t \in [\tau_n, \tau_{n+1}).$$

In the spirit of the martingale convergence theorem in Euclidean spaces, we show that Assumptions 3.3 and 3.4 imply a similar property for the noise terms U_n. Unlike in a Euclidean setup, however, we cannot simply add up the noise terms and form a martingale. Instead, we consider a set of orthonormal frames at each iterate of (RRM) and verify the cancellation property for the *coordinates* of the noise terms in these frames. This turns out to be enough for our asymptotic pseudo-trajectory result to hold.

Concretely, For each $n \in \mathbb{N}$, consider an arbitrary $\mathcal{F}_n$-measurable orthonormal frame $(E_k(n))_{k=1}^d$ for $T_{\boldsymbol{x}_n}\mathcal{M}$ and consider the components U_n^i of U_n in this basis, that is,

$$U_n = \sum U_n^i E_i(n), \quad \text{with} \quad U_n^i = \langle U_n, E_i(n) \rangle,$$

and pack these components into a Euclidean vector $\operatorname{vec}(U_n^i) \in \mathbb{R}^d$. It is then evident that

$$\mathbb{E}[\operatorname{vec}(U_n^i) \mid \mathcal{F}_n] = 0$$

and

$$\mathbb{E}[\|\operatorname{vec}(U_n^i)\|_2^2 \mid \mathcal{F}_n] = \mathbb{E}[|U_n|^2 \mid \mathcal{F}_n].$$

In other words, $\operatorname{vec}(U_n^i)$ becomes a martingale difference sequence in $\mathbb{R}^d$. For $t, T \geq 0$, define the *cumulative noise* from time t up to $t+T$ as

$$\Delta(t,T) = \sup_{0 \leq h \leq T} \left\| \int_t^{t+h} \operatorname{vec}(\overline{U}(s)^i)\, ds \right\|_2 .$$

Given Assumption 3.4, a similar argument as in Section 2.5.1 implies that for all fixed $T > 0$,

$$\lim_{t \to \infty} \Delta(t,T) = 0, \quad \text{almost surely.} \tag{3.50}$$

Note that the property (3.50) is independent of the choice of the frames $(E_k(n))$.

Let us now consider the bias terms B_n. Recall from (3.19) that $\mathbb{E}[|B_n|^2 \mid \mathcal{F}_n] \leq (B_n^*)^2$. Define

$$B^*(t,T) := \sup_{0 \leq h \leq T} \overline{B^*}(t+h).$$

Assumption 3.5 readily implies that for any fixed $T > 0$,

$$\lim_{t \to \infty} B^*(t,T) = 0, \quad \text{almost surely.} \tag{3.51}$$

This control over the bias terms turns out to be enough for our purposes.

We now turn to the distance of the Picard curve to the interpolation, which we bounded by (3.40). Our first task is to obtain an expression for $(\boldsymbol{x}')^i(t+s)$. Let $n = m(t+s)$ and consider the interpolation between $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n+1}$. By construction of (GI), the curve $\boldsymbol{x}$ is a geodesic when restricted to $[\tau_n, \tau_{n+1}]$. It then follows from (3.35) in Remark 3.1 that

$$(\boldsymbol{x}')^i(t+s) = \langle \dot{\boldsymbol{x}}(\tau_n), E_i(\tau_n) \rangle.$$

Now observe that (GI) already specifies $\dot{\boldsymbol{x}}(\tau_n)$:

$$\dot{\boldsymbol{x}}(\tau_n) = V(\boldsymbol{x}(\tau_n)) + U_n + B_n.$$

Let $U_n^{i,\shortparallel}$ and $B_n^{i,\shortparallel}$ be the components[9] of the noise and bias vectors in the frame $(E_i(\tau_n))$, and define the Euclidean vectors $\operatorname{vec}(U_n^{i,\shortparallel})$ and $\operatorname{vec}(B_n^{i,\shortparallel})$. With the notation for piecewise-constant interpolations (3.49), we can write

$$\operatorname{vec}((\boldsymbol{x}')^i(t+s)) = \operatorname{vec}(V^i(\bar{\boldsymbol{x}}(t+s))) + \operatorname{vec}(\bar{U}^{i,\shortparallel}(t+s)) + \operatorname{vec}(\bar{B}^{i,\shortparallel}(t+s)). \tag{3.52}$$

Using (3.52), we can now bound (3.40) further:

$$\begin{aligned} \|\hat{\lambda}(h)\|_2 \leq & \int_0^h \|\operatorname{vec}(\dot{\lambda}^i(s)) - \operatorname{vec}(V^i(\bar{\boldsymbol{x}}(t+s)))\|_2 \, ds \\ & + \left\| \int_0^h \operatorname{vec}(\bar{U}^{i,\shortparallel}(t+s)) \, ds \right\|_2 + \left\| \int_0^h \operatorname{vec}(\bar{B}^{i,\shortparallel}(t+s)) \, ds \right\|_2 \\ & + \int_0^h R_2(s) \, ds. \end{aligned}$$

By the comparison (3.47) between Fermi and parallel coordinates and the definitions of cumulative noise $\Delta(t,T)$ and $B^*(t,T)$, we obtain

$$\begin{aligned} \|\hat{\lambda}(h)\|_2 \leq & \int_0^h \|\operatorname{vec}(\dot{\lambda}^{i,\shortparallel}(s)) - \operatorname{vec}(V^i(\bar{\boldsymbol{x}}(t+s))\|_2 \, ds \\ & + \Delta(t,T) + B^*(t,T) + \int_0^h (R_2 + R_4)(s) \, ds \end{aligned} \tag{3.53}$$

With an identical argument to (3.41), the first term in (3.53) can be bounded as

$$\begin{aligned} & \|\operatorname{vec}(\dot{\lambda}^{i,\shortparallel}(s)) - \operatorname{vec}(V^i(\bar{\boldsymbol{x}}(t+s))\|_2 \\ & \quad = |V(\boldsymbol{x}(t+s)) - \mathrm{P}_{\bar{\boldsymbol{x}}(t+s) \to \boldsymbol{x}(t+s)}[V(\bar{\boldsymbol{x}}(t+s))]| \\ & \quad \leq L \, d(\boldsymbol{x}(t+s), \bar{\boldsymbol{x}}(t+s)). \end{aligned} \tag{3.54}$$

The interpolations $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ agree at $m(t+s) = \tau_n$. Hence, as $\boldsymbol{x}$ is a constant-speed

[9] We kept the tradition of putting $\shortparallel$ next to the coordinate, suggesting that the coordinates are in the (parallel) frame $(E_i(\tau_n))$.

geodesic in the interval $[\tau_n, \tau_{n+1}]$, we have

$$\begin{aligned} d(\boldsymbol{x}(t+s), \boldsymbol{x}_n) &\leq \int_{\tau_n}^{t+s} |\dot{\boldsymbol{x}}(u)|\, du \\ &= (t+s-\tau_n)|V(\bar{\boldsymbol{x}}(t+s)) + \bar{U}(t+s) + \bar{B}(t+s)| \\ &\leq \alpha^*(t)(V^* + B^*(t,T)) + (t+s-\tau_n)|\bar{U}(t+s)|, \end{aligned}$$

where $\alpha^*(t) = \sup_{u \geq t} \bar{\alpha}(u)$. For t large enough, we can assume that $\alpha^*(t) < 1$, and from the definition of $\Delta(t,T)$, we have that

$$\begin{aligned} (t+s-\tau_n)|\bar{U}(t+s)| &= \left\| \int_{\tau_n}^{t+s} \operatorname{vec}(\bar{U}^{i,\shortparallel}(u))\, du \right\|_2 \\ &\leq \left\| \int_{t-1}^{\tau_n} \operatorname{vec}(\bar{U}^{i,\shortparallel}(u))\, du \right\|_2 + \left\| \int_{t-1}^{t+s} \operatorname{vec}(\bar{U}^{i,\shortparallel}(u))\, du \right\|_2 \\ &\leq 2\Delta(t-1, T+1). \end{aligned}$$

Combining these bounds with Eqs. (3.53) and (3.54) then gives

$$\begin{aligned} \|\hat{\lambda}(h)\|_2 \leq{}& Lh[\alpha^*(t)(V^* + B^*(t,T)) + 2\Delta(t-1, T+1)] \\ &+ \Delta(t,T) + B^*(t,T) + \int_0^h (R_2 + R_4)(s)\, ds. \end{aligned} \tag{3.55}$$

Finishing the Proof

We can now finish the proof. Recall the decomposition (3.36):

$$d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t))) \leq \|\hat{\phi}(h) - \hat{\lambda}(h)\|_2 + 2\|\hat{\lambda}(h)\|_2.$$

Using (3.48) and (3.55) we obtain

$$\begin{aligned} \|\hat{\phi}(h) - \hat{\lambda}(h)\|_2 + 2\|\hat{\lambda}(h)\|_2 \leq{}& L\int_0^h \|\hat{\phi}(s)\|_2\, ds + \int_0^h (R_1 + R_3 + R_4)(s)\, ds \\ &+ 2Lh[\alpha^*(t)(V^* + B^*(t,T)) + 2\Delta(t-1, T+1)] \\ &+ 2\Delta(t,T) + 2B^*(t,T) + 2\int_0^h (R_2 + R_4)(s)\, ds. \end{aligned}$$

As $(R_1+R_2+R_3+R_4)(s) = O(\|\hat{\phi}(s)\|_2 + \|\hat{\lambda}(s)\|_2)$ and $\Delta(t,T) \leq 2\Delta(t-1,T+1)$, we get for some $C_1, C_2 > 0$,

$$\|\hat{\phi}(h)\|_2 + \|\hat{\lambda}(h)\|_2 \leq C_1 \int_0^h (\|\hat{\phi}(s)\|_2 + \|\hat{\lambda}(s)\|_2)\, ds \\ + hC_2(\alpha^*(t) + B^*(t,T) + \Delta(t-1,T+1)). \quad (3.56)$$

Grönwall's inequality then implies

$$\|\hat{\phi}(h)\|_2 + \|\hat{\lambda}(h)\|_2 \leq hC_2(\alpha^*(t) + B^*(t,T) + \Delta(t-1,T+1))\, e^{hC_1}.$$

From this, we conclude that

$$\begin{aligned} &\lim_{t\to\infty} \sup_{h\in[0,T]} d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t))) \\ &\qquad \leq \lim_{t\to\infty} \sup_{h\in[0,T]} (\|\hat{\phi}(h)\|_2 + \|\hat{\lambda}(h)\|_2) \\ &\qquad \leq \lim_{t\to\infty} TC_2(\alpha^*(t) + B^*(t,T) + \Delta(t-1,T+1))\, e^{TC_1} \\ &\qquad = 0 \quad \text{a.s.}, \end{aligned}$$

since $\alpha^*(t) \to 0$, and by (3.51) and (3.50), $\Delta(t,T)$ and $B^*(t,T)$ vanish with probability 1 as $t \to \infty$. □

3.6. PROOF OF THE STABILITY THEOREM

In this section, we give a complete proof for the stability theorem, stated in Theorem 3.6, which gives conditions for the iterates of a Riemannian Robbins–Monro algorithm to be precompact. Recall that if the manifold is compact, stability holds without any further assumptions. In this section, we focus on a wide class of non-compact manifolds, called Hadamard manifolds.

Let us begin with some definitions. We say $\mathcal{M}$ is a *Hadamard manifold*, if it is a complete simply connected Riemannian manifold with non-positive sectional curvatures. Recall that a topological space $\mathcal{M}$ is *simply connected* if and only if it is path-connected, and whenever $c : [0,1] \to \mathcal{M}$ and $c' : [0,1] \to \mathcal{M}$ are two paths with the same start and endpoint, then c can be continuously deformed into c' while keeping both endpoints fixed. Fix an arbitrary base point $o \in \mathcal{M}$. Define the radial distance function $r(p) := d(p, o)$, and let

$$k(p) := \frac{1}{2} d^2(p, o) = \frac{1}{2} r^2(p). \tag{3.57}$$

We see below that on a Hadamard manifold, one has a notion of "outward-pointing radial direction" given by the gradient of squared radial distance ∇k, that is,

$$\nabla k(p) = -\exp_p^{-1}(o).$$

Using this notion, we can further formalize when a vector field tends to "keep iterates stable" by pushing them "inwards."

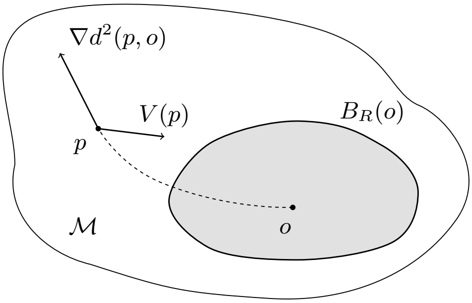


Figure 3.4. A weakly coercive vector field. For all points p outside the geodesic ball $B_R(o)$, the vector field V should have negative inner product with the radial velocity at p. The dashed line depicts the geodesic from o to p.

We say that the vector field V is *weakly coercive*, if for all $p \in \mathcal{M} \setminus B_R(o)$ outside a closed geodesic ball of radius $R > 0$ and centered at o it holds

$$\langle V(p), \nabla k(p) \rangle \leq 0. \tag{3.58}$$

See Fig. 3.4 for an illustration. Weak coercivity may be viewed as a Riemannian relaxation of the coercivity condition in Euclidean spaces:

$$\lim_{p \to \infty} \frac{\langle V(p), p \rangle}{\|p\|_2} = -\infty.$$

The condition above posits that the inward-pointing component of V grows unbounded at infinity, a property which is frequently used to ensure the stability of Euclidean iterative algorithms [Phe93; FP03]. In our Riemannian setting, the role of the radial field is played by the gradient of the squared distance function ∇k. In addition, it is important to bear in mind that weak coercivity does not impose any growth requirements on the radial component of V; it only requires that V does not have a consistent outward-pointing component that could lead the process to diverge. Therefore, it is significantly weaker in that respect than Euclidean coercivity (hence the adjective "weak").

Let us now turn our attention to the proof of Theorem 3.6. We heavily rely on the structure of the squared radial distance function k defined in (3.57) along with its gradient and Hessian. The following theorem, which is adapted from [Jos17], shows that k is smooth and gives a control on its Hessian.

Theorem 3.13 (Jos17, Thm. 6.6.1)**.** *Suppose $o \in \mathcal{M}$ is an arbitrary base point in a Hadamard manifold $\mathcal{M}$, and let r and k be defined as in* (3.57)*. Moreover, suppose that the sectional curvature of $\mathcal{M}$ is non-positive and bounded from below by $(-\kappa^2)$. Then k is smooth on $\mathcal{M}$ and*

$$\nabla k(p) = -\exp_p^{-1}(o).$$

Additionally, $|\nabla k(p)| = r(p)$ and

$$(\operatorname{Hess} k)_p(v, v) \leq \kappa \cdot r(p) \cdot \coth(\kappa \cdot r(p)) \cdot |v|^2$$

for all $p \in \mathcal{M}$ and $v \in T_p\mathcal{M}$.

Remark. Since $\mathcal{M}$ is simply connected and complete, it holds that $\operatorname{inj}(\mathcal{M}) = \infty$ [Jos17, Cor. 6.9.1], and we can deduce the result above from [Jos17, Thm. 6.6.1]. ♢

Proof of Theorem 3.6. Our proof relies on constructing a suitable energy function that serves as a proxy for the distance of the iterates of (RRM) from an arbitrarily chosen base point o. This function is of the form

$$E(p) = f(r(p)) \tag{3.59}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a C^∞ non-negative function with $f(x) = 0$ for all $x \leq R$ and satisfies

$$0 \leq f'(x) \leq C_1, \quad f''(x) \leq C_2 \tag{3.60}$$

for all $x \geq R$. Moreover, we require $f(x) = \Omega(x)$ as $x \to \infty$ so that an upper bound on f implies a bound on x; see Fig. 3.5. The gradient of E can be computed as

$$\nabla E(p) = \begin{cases} 0 & \text{if } r(p) \leq R, \\ \frac{f'(r(p))}{r(p)} \nabla k(p) & \text{if } r(p) > R. \end{cases} \tag{3.61}$$

Our first result is that $E = f \circ r$ has a bounded Hessian and is smooth. For a proof, see Appendix A.3.

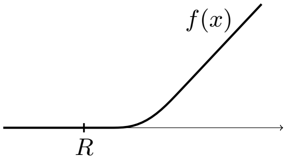


Figure 3.5. An example of a function f satisfying (3.60), with $f(x) = 0$ when $x \leq R$, and $f(x) \sim x$ when $x \to \infty$. Lemma A.5 gives a concrete formula for f.

Lemma 3.14. *Let E be defined as in (3.59). Then E is negatively correlated with V everywhere, in the sense that*

$$\langle \nabla E(p), V(p) \rangle \leq 0, \quad \forall p \in \mathcal{M}. \tag{3.62}$$

Moreover, there exists a constant $C > 0$ such that $(\operatorname{Hess} E)_p(v, v) \leq C|v|^2$ and

$$E(p') \leq E(p) + \langle \nabla E(p), \exp_p^{-1}(p') \rangle + \frac{C}{2} d^2(p, p'), \quad \forall p, p' \in \mathcal{M}. \tag{3.63}$$

We now proceed to the main argument, where we use E to control the distance of the iterates $\boldsymbol{x}_n$ to o. Letting $E_n = E(\boldsymbol{x}_n)$ and using Lemma 3.14, we have

$$\begin{aligned} E_{n+1} = E\left(\exp_{\boldsymbol{x}_n}(\alpha_n V_n)\right) &\leq E_n + \alpha_n \langle \nabla E(\boldsymbol{x}_n), V_n \rangle + \frac{C\alpha_n^2}{2} |V_n|^2 \\ &\leq E_n + \alpha_n \langle \nabla E(\boldsymbol{x}_n), U_n + B_n \rangle + \frac{3C\alpha_n^2}{2} \left(|V(\boldsymbol{x}_n)|^2 + |U_n|^2 + |B_n|^2\right), \end{aligned}$$

where the second line follows from the negative correlation of E and V, the definition (3.15) of V_n, and the Cauchy-Schwarz inequality. Conditioning on $\mathcal{F}_n = \sigma(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ and taking expectations, we obtain

$$\mathbb{E}[E_{n+1} \,|\, \mathcal{F}_n] \leq E_n + \alpha_n |\nabla E(\boldsymbol{x}_n)| \cdot \mathbb{E}[|B_n| \,|\, \mathcal{F}_n] + \tfrac{3}{2} C\alpha_n^2 \left[(V^*)^2 + (U_n^*)^2 + (B_n^*)^2\right],$$

where V^* is the bound on $|V|$, and U_n^* and B_n^* are defined in (3.19).

Using (3.61), we see that $|\nabla E(\boldsymbol{x}_n)| \leq \frac{C_1}{r(\boldsymbol{x}_n)} |\nabla k(\boldsymbol{x}_n)| = C_1$. Moreover, as $\mathbb{E}[|B_n| \,|\, \mathcal{F}_n] \leq B_n^*$, we obtain

$$\mathbb{E}[E_{n+1} \,|\, \mathcal{F}_n] \leq E_n + \alpha_n C_1 B_n^* + \tfrac{3}{2} C\alpha_n^2 \left[(V^*)^2 + (U_n^*)^2 + (B_n^*)^2\right]. \tag{3.64}$$

To proceed, let $\varepsilon_n = \alpha_n C_1 B_n^* + \frac{3}{2} C\alpha_n^2 \left[(V^*)^2 + (U_n^*)^2 + (B_n^*)^2\right]$ be the residual term in (3.64). Notice that

$$\sum_{n=1}^{\infty} \varepsilon_n \leq C_1 \sum_{n=1}^{\infty} \alpha_n B_n^* + \frac{3C}{2} \sum_{n=1}^{\infty} \alpha_n^2 ((V^*)^2 + (U_n^*)^2 + (B_n^*)^2), \tag{3.65}$$

and hence, by Assumptions 3.3–3.5 and the dominated convergence theorem, we infer that $\mathbb{E}[\sum \varepsilon_n] < \infty$.[10] Next, consider the auxiliary process

$$E_n' = E_n + \mathbb{E}[\textstyle\sum_{k=n}^{\infty} \varepsilon_k \,|\, \mathcal{F}_n]$$

adapted to the same filtration as E_n. By (3.64), we have

$$\begin{aligned} \mathbb{E}[E_{n+1}' \,|\, \mathcal{F}_n] &= \mathbb{E}\big[E_{n+1} + \mathbb{E}\big[\textstyle\sum_{k=n+1}^{\infty} \varepsilon_k \,\big|\, \mathcal{F}_{n+1}\big] \,\big|\, \mathcal{F}_n\big] \\ &\leq E_n + \varepsilon_n + \mathbb{E}\big[\mathbb{E}\big[\textstyle\sum_{k=n+1}^{\infty} \varepsilon_k \,\big|\, \mathcal{F}_{n+1}\big] \,\big|\, \mathcal{F}_n\big] \\ &= E_n + \mathbb{E}[\textstyle\sum_{k=n}^{\infty} \varepsilon_k \,|\, \mathcal{F}_n] \\ &= E_n'. \end{aligned}$$

This shows that E_n' is a supermartingale adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Therefore, $\mathbb{E}[E_n'] \leq \mathbb{E}[E_1'] < \infty$, which implies that E_n' is uniformly bounded in L^1, and by Doob's supermartingale convergence theorem, E_n' converges almost surely to some finite random limit E_∞'. Hence, $E_n = E_n' - \mathbb{E}[\sum_{k=n}^{\infty} \varepsilon_k \,|\, \mathcal{F}_n]$ converges almost surely to some random finite limit. From this and the fact that $E_n = \Omega(r(\boldsymbol{x}_n))$, we conclude that $\limsup_n r(\boldsymbol{x}_n) < \infty$, as claimed. □

3.7. EXAMPLES OF RIEMANNIAN STOCHASTIC APPROXIMATION ALGORITHMS

In this section, our main goal is to demonstrate how a variety of algorithms can be viewed as specific instances of (RRM), thus allowing us to apply Theorem 3.4 and Corollary 3.5 to analyze their convergence behaviors. For clarity, we will discuss algorithms that operate with indirect access to the vector field V through what is known as a *stochastic first-order oracle* (SFO). In essence, when an SFO is invoked at a point p on the manifold $\mathcal{M}$, given an independent random seed ω from a set of seeds Ω, it produces a random vector $\widetilde{V}(p;\omega)$ in the tangent space

[10] Note that Assumption 3.5 states that $\sum \alpha_n (\mathbb{E}[(B_n^*)^2])^{1/2} < \infty$. By Jensen's inequality, $\sum \alpha_n \, \mathbb{E}[B_n^*] < \infty$, and as the terms in the series are positive, we also get $\sum \alpha_n^2 \, \mathbb{E}[(B_n^*)^2] < \infty$.

$T_p\mathcal{M}$ which is of the form

$$\widetilde{V}(p;\omega) = V(p) + U(p;\omega), \tag{SFO}$$

where the error term $U(p;\omega) \in T_p\mathcal{M}$ is assumed to be zero-mean and have bounded second moments:

$$\mathbb{E}_\omega[U(p;\omega)] = 0 \quad \text{and} \quad \mathbb{E}_\omega[|U(p;\omega)|^2] \leq \sigma^2, \quad \forall p \in \mathcal{M}. \tag{3.66}$$

Below, we list four different algorithms that are used for Riemannian root-finding problems. We first cast each of the algorithms to the Riemannian Robbins–Monro (RRM) framework by stating what is the value of noise U_n and bias B_n in the error decomposition (3.14) of (RRM). Finally, we state a proposition, in which we show that the convergence result of Theorem 3.4 and Corollary 3.5 holds for these algorithms under the mentioned assumptions.

Remark. We note that certain algorithms presented below are intended for optimization purposes, hence the SFO yields a "stochastic gradient." It is important to emphasize that within our framework, there is no necessity for V to be a gradient. Nevertheless, we will employ the same terminology and occasionally refer to $\widetilde{V}$ as a "stochastic gradient." ◊

The first algorithm, which is the simplest of all four, is the Riemannian analogue of the famous stochastic gradient descent method:

▷ **Algorithm 3.1.** The *Riemannian stochastic gradient method* [Bon13] queries the SFO at each iteration and proceeds as

$$\boldsymbol{x}_{n+1} = \exp_{\boldsymbol{x}_n}(\alpha_n V(\boldsymbol{x}_n;\omega_n)). \tag{RSGM}$$

This is clearly an RRM scheme by setting $U_n = U(\boldsymbol{x}_n;\omega_n)$ and $B_n = 0$. ◁

The next algorithm is inspired by the proximal point method in Euclidean optimization [Mar70; Roc76].

▷ **Algorithm 3.2.** The (deterministic) *Riemannian proximal point method* [FO02] is an implicit update rule of the form

$$\exp^{-1}_{\boldsymbol{x}_{n+1}}(\boldsymbol{x}_n) = -\alpha_n V(\boldsymbol{x}_{n+1}). \tag{RPPM}$$

The RRM representation of (RPPM) may then be obtained by taking $U_n = 0$ and

$$B_n = \mathrm{P}_{\boldsymbol{x}_{n+1}\to\boldsymbol{x}_n}[V(\boldsymbol{x}_{n+1})] - V(\boldsymbol{x}_n). \qquad ◁$$

Remark 3.3. Our formulation of (RPPM) is inspired by the proximal point method for convex optimization. Concretely, let $f : \mathcal{M} \to \mathbb{R}$ be a differentiable, geodesically convex function. The proximal point method constructs its iterates by the so-called proximal step:

$$\boldsymbol{x}_{n+1} = \operatorname*{arg\,min}_{p \in \mathcal{M}} \left\{ f(p) + \frac{1}{2\alpha_n} d^2(p, \boldsymbol{x}_n) \right\}.$$

Ferreira and Oliveira [FO02, Thm. 5.1] show that when $\mathcal{M}$ is a Hadamard manifold, the optimal solution of the proximal step satisfies

$$\exp^{-1}_{\boldsymbol{x}_{n+1}}(\boldsymbol{x}_n) = \alpha_n \nabla f(\boldsymbol{x}_{n+1}).$$

Replacing $-\nabla f$ with V in this relation gives the update rule (RPPM). ◊

▷ **Algorithm 3.3.** Inspired by the original work of Korpelevich [Kor76], the *Riemannian stochastic extra-gradient* method [TH12; NSS16] proceeds as

$$\begin{aligned} \boldsymbol{x}_{n+1/2} &= \exp_{\boldsymbol{x}_n}\big(\alpha_n \widetilde{V}(\boldsymbol{x}_n; \omega_n)\big), \\ \boldsymbol{x}_{n+1} &= \exp_{\boldsymbol{x}_n}\big(\mathrm{P}_{\boldsymbol{x}_{n+1/2} \to \boldsymbol{x}_n}[\alpha_n \widetilde{V}(\boldsymbol{x}_{n+1/2}; \omega_{n+1/2})]\big) \end{aligned} \tag{RSEG}$$

where ω_n and $\omega_{n+1/2}$ are independent seeds for (SFO). To recast (RSEG) in the RRM framework, simply take

$$\begin{cases} U_n = \mathrm{P}_{\boldsymbol{x}_{n+1/2} \to \boldsymbol{x}_n}[U(\boldsymbol{x}_{n+1/2}; \omega_{n+1/2})], \text{ and} \\ B_n = \mathrm{P}_{\boldsymbol{x}_{n+1/2} \to \boldsymbol{x}_n}[V(\boldsymbol{x}_{n+1/2})] - V(\boldsymbol{x}_n). \end{cases}$$ ◁

Compared to (RSGM), the scheme (RSEG) involves two oracle queries per iteration. Building on an original idea by Popov [Pop80], the last oracle query can be "recycled," leading to a more efficient method, known in other contexts as *optimistic gradient* variant:

▷ **Algorithm 3.4.** The *Riemannian optimistic gradient* method proceeds as

$$\begin{aligned} \boldsymbol{x}_{n+1/2} &= \exp_{\boldsymbol{x}_n}\big(\alpha_n V(\boldsymbol{x}_{n-1/2}; \omega_{n-1})\big), \\ \boldsymbol{x}_{n+1} &= \exp_{\boldsymbol{x}_n}\big(\mathrm{P}_{\boldsymbol{x}_{n+1/2} \to \boldsymbol{x}_n}[\alpha_n V(\boldsymbol{x}_{n+1/2}; \omega_n)]\big). \end{aligned} \tag{ROG}$$

As such, (ROG) may be seen as a special case of (RRM) by taking

$$\begin{cases} U_n = \mathrm{P}_{\boldsymbol{x}_{n+1/2} \to \boldsymbol{x}_n}[U(\boldsymbol{x}_{n+1/2}; \omega_n)], \text{ and} \\ B_n = \mathrm{P}_{\boldsymbol{x}_{n+1/2} \to \boldsymbol{x}_n}[V(\boldsymbol{x}_{n+1/2})] - V(\boldsymbol{x}_n). \end{cases}$$ ◁

In view of Theorem 3.4 and Corollary 3.5, the convergence analysis of Algorithms 3.1–3.4 above boils down to verifying Assumptions 3.1–3.6. The following proposition summarizes this.

Proposition 3.15. *Suppose that*

(H1) $\mathcal{M}$ is either a compact or a Hadamard manifold satisfying Assumption 3.1,

(H2) the vector field V is bounded and satisfies Assumption 3.2,

(H3) V is weakly coercive (3.58) in case $\mathcal{M}$ is not compact,

(H4) and the errors U of the SFO for V are zero-mean and have bounded second moments (3.66). If $\mathcal{M}$ is compact, the errors are further assumed to be a.s. uniformly bounded in norm.

Then, with probability 1, the iterates of Algorithms 3.1–3.4 converge to an internally chain-transitive set of the flow (3.16).

Note that the precompactness of the iterates (Assumption 3.6) follows readily from (H1), (H2), (H3), and Theorem 3.6. Therefore, for each algorithm, we only have to verify the noise and bias conditions of Assumptions 3.4 and 3.5. The proof can be found in Appendix A.4.

3.8. ALGORITHMIC VARIATIONS

In the last section, we mentioned a few basic algorithms for solving Riemannian root-finding problems. However, these algorithms can be further modified to be more computationally efficient. In this section, we show two different ways to modify a Riemannian Robbins–Monro algorithm.

3.8.1. Retractions

One of the critical operations in (RRM) is computing the exponential map at each iteration. While the exponential map can be computed in closed form for standard manifolds (such as spheres, hyperbolic spaces, or some matrix manifolds), its computation for a general manifold is computationally prohibitive: one has to solve the geodesic equation (3.9), which is a second-order ODE. This can be a computation bottleneck, as the exponential map is needed for every iteration of (RRM).

A popular alternative to computing the exact exponential map is to use a so-called *retraction map* [AMS08; Bou23]. This method is especially effective for manifolds embedded in a Euclidean space. Let us explain the idea of a retraction map for this latter case. Suppose $\mathcal{M} \subset \mathbb{R}^N$ is a d-dimensional manifold embedded in $\mathbb{R}^N$. In this case, the tangent space $T_p\mathcal{M}$ at any point $p \in \mathcal{M}$ can be seen as a d-dimensional affine subspace of $\mathbb{R}^N$. The idea of a retraction map is that instead of computing the exponential map $\exp_p(v)$ for a tangent vector $v \in T_p\mathcal{M}$, we compute $p + v$ in $\mathbb{R}^N$ and project this point back onto the manifold. If v is small, the resulting point will be a good approximation of $\exp_p(v)$ [see AMS08, Sec. 4.1].

▷ **Example.** In our introductory examples, we saw the geodesic equation for a two-dimensional unit sphere in Example 3.1:

$$\exp_x(v) = x\cos\|v\| + \frac{v}{\|v\|}\sin\|v\|.$$

The following retraction, however, is much simpler to implement:

$$R_x(v) := \frac{x+v}{\|x+v\|}. \qquad \triangleleft$$

Concretely, a retraction map is a smooth map R from the tangent bundle to the manifold that agrees with the exponential map up to the first order, that is,

$$R_p(0) = p \quad \text{and} \quad \frac{d}{dt}\bigg|_{t=0} R_p(tv) = v \quad \text{for all } (p, v) \in T\mathcal{M}. \tag{3.67}$$

In our framework, replacing the exponential map with a retraction can be viewed as a source of bias in computing the vector field. To see this, suppose we are running Algorithm 3.1 for the vector field V using the stochastic first-order oracle $\widetilde{V}$, and at each iteration, instead of using the exponential map, we use a retraction R. The new update rule becomes

$$\boldsymbol{x}_{n+1} = R_{\boldsymbol{x}_n}(\alpha_n \widetilde{V}(\boldsymbol{x}_n; \omega_n)) = R_{\boldsymbol{x}_n}(\alpha_n\{V(\boldsymbol{x}_n) + U(\boldsymbol{x}_n; \omega_n)\}). \tag{3.68}$$

Defining the tangent vector $v_n \in T_{\boldsymbol{x}_n}\mathcal{M}$ via

$$\exp_{\boldsymbol{x}_n}(\alpha_n v_n) = \boldsymbol{x}_{n+1} \quad \text{or} \quad \alpha_n v_n = \exp^{-1}_{\boldsymbol{x}_n}(\boldsymbol{x}_{n+1}),$$

we see that the difference of v_n and $\widetilde{V}(\boldsymbol{x}_n; \omega_n)$ is a form of bias; that is, by defining $B_n := v_n - \widetilde{V}(\boldsymbol{x}_n; \omega_n)$, the iteration (3.68) can be written in (RRM) form as

$$\boldsymbol{x}_{n+1} = \exp_{\boldsymbol{x}_n}(\alpha_n\{V(\boldsymbol{x}_n) + U(\boldsymbol{x}_n; \omega_n) + B_n\}). \tag{3.69}$$

Our goal in the rest of this section is to find conditions so that the bias term B_n

satisfies Assumption 3.5. If that is the case, the same convergence guarantees hold for all the Riemannian Robbins–Monro schemes that use retractions, without any extra effort.

Before stating our results, we need to introduce the notion of a totally retractive neighborhood, as well as some local properties of retractions on these neighborhoods. Inspired by strongly convex neighborhoods, we call a neighborhood $\mathcal{U}$ *totally retractive* for a retraction R [HAG15], if there exists some $\delta > 0$ so that for any $p \in \mathcal{U}$, the retraction R is a diffeomorphism from $B_\delta(0) \subset T_p\mathcal{M}$ to its image under R, and $R_p(B_\delta(0)) \supseteq \mathcal{U}$. Existence of a totally retractive neighborhood can be shown along the lines of [Car92, Thm. 3.3.7].

In a compact neighborhood that is both strongly convex and totally retractive, one can show that the inverse of the exponential map and the retraction map are close to each other, in a sense made precise by the following lemma:

Lemma 3.16 (ZS20, Lem. 3)**.** *Let $\mathcal{U}$ be a compact, strongly convex, totally retractive neighborhood. Then there exists a constant $C > 0$ such that for all points $p, p' \in \mathcal{U}$,*

$$|R_p^{-1}(p') - \exp_p^{-1}(p')| \leq C\, d(p, p')^2 = C\, |\exp_p^{-1}(p')|^2.$$

As the retraction is a first-order approximation of the exponential map, it is rather intuitive to expect that the distance travelled by the retraction map is comparable to that of the exponential map. The following lemma makes this intuition precise:

Lemma 3.17 (HAG15, Lem. 3)**.** *Let R be a retraction on $\mathcal{M}$. For each $p \in \mathcal{M}$ there are constants $c_1, C_2 > 0$ and $\delta_{c_1,C_2} > 0$ such that for all q in a sufficiently small neighborhood of p and all $v \in T_q\mathcal{M}$ with $|v| \leq \delta_{c_1,C_2}$,*

$$c_1|v| \leq d(q, R_q(v)) \leq C_2|v|.$$

Note that $|v| = d(q, \exp_q(v))$.

Lemmas 3.16 and 3.17 describe the local behavior of a retraction; all the constants depend on a compact neighborhood of a point $p \in \mathcal{M}$. To streamline our discussion, let us assume that the local results of these lemmas hold globally over the manifold, with some global constants.

▷ **Assumption 3.7** (on the retraction)**.** *There exists some radius $r_{\text{retr}} > 0$ so that every point $q \in \mathcal{M}$ has a totally retractive neighborhood containing $B_{r_{\text{retr}}}(q)$. Moreover, there exists some global constant $C > 0$ such that for any two points $p, p' \in B_{r_{\text{retr}}}(q)$,*

$$|R_p^{-1}(p') - \exp_p^{-1}(p')| \leq C_1\, d(p, p')^2. \tag{3.70}$$

Additionally, there exists some global constant $C_2 > 0$ such that for any $p \in B_{r_{\mathrm{retr}}}(q)$ and $v \in T_p\mathcal{M}$ with $|v| \leq r_{\mathrm{retr}}$,

$$d(p, R_p(v)) \leq C_2|v|. \tag{3.71}$$

Note that if the manifold is compact, Assumption 3.7 is automatically satisfied. For non-compact manifolds, one has to impose sufficient regularity on the retraction so that these uniform bounds hold.

As it turns out, to replace the exponential map in Algorithms 3.1–3.4 with a retraction, we only need to slightly strengthen our assumptions on the noise of the SFO:

Proposition 3.18. *Suppose that the error term of* (SFO) *has bounded fourth moments, i.e., there is some $C > 0$ such that $\mathbb{E}_\omega[|U(p;\omega)|^4] \leq C^2$ for all $p \in \mathcal{M}$. Then Proposition 3.15 holds as stated if the exponential map in Algorithms 3.1–3.4 is replaced by a retraction satisfying Assumption 3.7.*

Proof. We proof this proposition only for Algorithm 3.1; the proof for the rest of the algorithms is similar. What we have to show is that the bias term in (3.69) satisfies Assumption 3.5.

Recall the definition of the tangent vector v_n in (3.69). From Assumption 3.7, we have

$$\begin{aligned}
|\alpha_n B_n| &= |\alpha_n v_n - \alpha_n \widetilde{V}(\boldsymbol{x}_n; \omega_n)| \\
&= |\exp^{-1}_{\boldsymbol{x}_n}(\boldsymbol{x}_{n+1}) - R^{-1}_{\boldsymbol{x}_n}(\boldsymbol{x}_{n+1})| \\
&\leq C_1 d(\boldsymbol{x}_n, \boldsymbol{x}_{n+1})^2 && \text{by (3.70)} \\
&\leq C_1 C_2^2 \, |\alpha_n \widetilde{V}(\boldsymbol{x}_n; \omega_n)|^2 && \text{by (3.71).}
\end{aligned}$$

Thus, $|B_n| \leq \mathcal{O}(\alpha_n|\widetilde{V}(\boldsymbol{x}_n;\omega_n)|^2) \leq \mathcal{O}(\alpha_n) + \mathcal{O}(\alpha_n|U(\boldsymbol{x}_n;\omega_n)|^2)$. A similar Borel–Cantelli argument as in the proof of Lemma 3.9 shows that the bounded fourth moments of $U(\boldsymbol{x}_n;\omega_n)$ imply $|B_n| \to 0$, almost surely. The summability condition in Assumption 3.5 is similarly satisfied. □

3.8.2. Alternating Algorithms

In the context of multiplayer games, the iterative process (RRM) plays the role of defining the evolution of players' strategies over time, where $\boldsymbol{x}_n$ symbolizes the array of strategies—or the *strategy profile*—adopted by all players at a given iteration n. The conventional interpretation of (RRM) suggests a simultaneous update mechanism, where the transition from one strategy profile to the next is instantaneous, affecting all players concurrently. Another common variant is where the strategy update process is sequential. In this setting, each player, in a predetermined order, revises their strategy based on the latest available information, which includes the most recent updates from preceding players. Consequently, what is theoretically modeled as a single, collective stride in (RRM), unfolds through a series of individual steps; one per player.

Let us explain the general idea for a two-player min-max game. The same reasoning works for N-player games with a more involved notation. Let $\mathcal{M}$ and $\mathcal{N}$ be Riemannian manifolds representing the space of strategies for each player. Consider the function $\ell : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ and the min-max game

$$\min_{p\in\mathcal{M}} \max_{q\in\mathcal{N}} \ell(p, q).$$

A strategy profile in this game is simply a pair $(p, q) \in \mathcal{M} \times \mathcal{N}$.

For a Riemannian Robbins–Monro algorithm for solving the min-max game above, instead of updating the strategy profiles $\boldsymbol{x}_n = (p_n, q_n) \in \mathcal{M} \times \mathcal{N}$ simultaneously, consider the following alternating update:

$$\begin{aligned} p_{n+1} &= \exp_{p_n}(\alpha_n\{V^{(1)}(p_n, q_n) + Z_n^{(1)}\}), \\ q_{n+1} &= \exp_{q_n}(\alpha_n\{V^{(2)}(p_{n+1}, q_n) + Z_n^{(2)}\}), \end{aligned} \tag{RRM-alt}$$

where $V^{(1)} = -\nabla_p \ell(p, q)$, $V^{(2)} = \nabla_q \ell(p, q)$, and $Z_n^{(i)}$ are the error terms for the update of each player's strategy.

The key idea to use our theory for this instance is to cast (RRM-alt) as a biased simultaneous update. Concretely, by defining the vector field V on the product manifold $\mathcal{M} \times \mathcal{N}$ as

$$V(p, q) = (V^{(1)}(p, q), V^{(2)}(p, q)) = (-\nabla_p \ell(p, q), \nabla_q \ell(p, q)),$$

it is easy to observe that (RRM-alt) is equivalent to the simultaneous update

$$\boldsymbol{x}_{n+1} = \exp_{\boldsymbol{x}_n}(V(\boldsymbol{x}_n) + Z_n), \quad \boldsymbol{x}_n = (p_n, q_n),$$

where the error term Z_n includes the error of each player, as well as the error induced by converting the sequential update to a simultaneous one:

$$Z_n = \begin{pmatrix} Z_n^{(1)} \\ Z_n^{(2)} + V^{(2)}(p_{n+1}, q_n) - V^{(2)}(p_n, q_n) \end{pmatrix} \in T_{(p_n,q_n)}(\mathcal{M} \times \mathcal{N}). \tag{3.72}$$

As it turns out, with all our assumptions in place, this bias is also negligible and vanishes when $n \to \infty$. Therefore, we get the following result for RRM Algorithms 3.1–3.4 with sequential updates:

Proposition 3.19. *Consider the alternating variant* (RRM-alt) *and assume that* $V^{(i)}$ *are bounded and* L*-Lipschitz in both of their arguments. Under the same hypotheses, the result of Proposition 3.15 holds as stated for the alternating variant of Algorithms 3.1–3.4.*

Proof. We only have to show that the extra bias term in (3.72) vanishes with probability 1. As $V^{(1)}$ is bounded, the same argument as in Lemma 3.9 implies that

$$\lim_{n\to\infty} \alpha_n |V^{(1)}(p_n, q_n) + Z_n^{(1)}| = 0, \quad \text{a.s.}$$

Therefore, for n large enough, p_{n+1} lies in the injectivity radius of p_n. Using Lipschitzness of $V^{(2)}$, we get

$$|V^{(2)}(p_{n+1}, q_n) - V^{(2)}(p_n, q_n)| \leq L\, d(p_{n+1}, p_n) = L \cdot \alpha_n |V^{(1)}(p_n, q_n) + Z_n^{(1)}|,$$

which vanishes as $n \to \infty$. The summability (3.22) of this bias term also follows by the boundedness of $V^{(1)}$ and summability of $Z_n^{(1)}$. □

3.9. APPLICATIONS TO LEARNING AND GAMES

We close this chapter by bringing some concrete implications of our general theory to the specific examples of Section 3.1. As seen below, the generality of (RRM) schemes allows us to not only unify several existing results, but also provide completely new ones.

3.9.1. Optimization on Manifolds

Recall the minimization problem

$$\min_{p\in\mathcal{M}} f(p)$$

where $f : \mathcal{M} \to \mathbb{R}$ is a smooth function which is not necessarily geodesically convex. Finding the minimum of f in many cases is an intractable problem and one is usually satisfied with a *critical point* of f, which is a point p such that $\nabla f(p) = 0$. This hints using the *gradient flow* of f; the flow of the vector field $V = -\nabla f$. This flow has the special property that the objective function f itself is a Lyapunov function for the set of its critical points.

Concretely, let Φ be the gradient flow of f and denote by Λ the set of critical points of f. It is clear that

$$\frac{d}{dt} f(\Phi_t(p)) = -|\nabla f(\Phi_t(p))|^2 \leq 0,$$

where the equality holds if and only if $\Phi_t(p) \in \Lambda$. This means that Λ is the set of equilibria of the flow Φ and f is a strict Lyapunov function for the set Λ. Moreover, by Sard's theorem [Ste99], the set of critical values of f, that is, $f(\Lambda)$, is of measure zero (and therefore, has empty interior). Therefore, Theorem 2.10 implies that every internally chain-transitive set of the gradient flow Φ is contained in a set of critical points of f on which f is constant. To use Sard's theorem, however, we have to assume that f and $\mathcal{M}$ are at least d times differentiable, where $d = \dim \mathcal{M}$.

Having characterized the internally chain-transitive sets of the gradient flow, we are now ready to apply our general asymptotic pseudo-trajectory theory to the optimization problem (3.10).

Proposition 3.20. *Let $\mathcal{M}$ be a d-dimensional Riemannian manifold and f be a smooth function, both of class C^d. Suppose that we run Algorithms 3.1–3.4 for the vector field $V = -\nabla f$ with an SFO satisfying (H4). Then, under Assumptions 3.1, 3.3 and 3.6, the induced sequence of iterates converges almost surely to a component of critical points of f on which f is constant. Additionally, if $\sup_p \mathbb{E}[|U(p;\omega)|^4] < \infty$, the conclusions above apply to all retraction-based variants of Algorithms 3.1–3.4.*

Let us note that several Euclidean algorithms are known to avoid undesirable solutions [see, e.g., CB19]. In [Hsi+23], we demonstrate that the general avoidance theory can be applied to Riemannian manifolds as well, indicating that many iterative Riemannian methods (including those based on retraction) converge with probability 1 solely to local minimizers.

3.9.2. Games on Manifolds

Recall the setup of Section 3.3.2 for games on manifolds: The space of all configurations of the game a product manifold $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_N$, where $\mathcal{M}_i$ is the strategy space of the ith player, and $u_i : \mathcal{M} \to \mathbb{R}$ is the payoff function of the ith player. Our first result below concerns the convergence of Algorithms 3.2–3.4 in a general class of (Riemannian) monotone games known as α-accretive games [Wan+10]. The proof is an immediate consequence of Propositions 3.15, 3.18 and 3.19, along with the main result of [Kri14].

Proposition 3.21. *Let $V = [\nabla_{p_i} u_i]$ be an α-*accretive game field *for some $\alpha > 0$.*

This means that for all $r \geq 0$,

$$(1+\alpha r)\, d(p,p') \leq d\big(\exp_p(rV(p)), \exp_{p'}(rV(p'))\big).$$

Then Algorithms 3.1–3.4, as well as their alternating or retraction-based variants, converge to the game's set of Nash–Stampacchia equilibria.

To the best of our knowledge, most of the algorithms we consider are new in the setting of Riemannian monotone games except for the *deterministic* gradient and extra-gradient methods [TH12; NSS16; FQT20; Kha+20; CLC21].

While being quite general, accretivity is a strong, convexity-like assumption about the games. In our next result, we prove general convergence for a class of non-convex potential games [KK02; ME05]. The proof is immediate from combining Propositions 3.15–3.19.

Proposition 3.22. *Let $V = [\nabla_{p_i} u_i]$ be a game field associated with a Riemannian potential game. Then Algorithms 3.1–3.4, as well as their alternating or retraction-based variants, converge to the critical points of the game potential.*

For Riemannian potential games, the convergence of the continuous-time dynamics (3.16) is well known, but we are not otherwise aware of a similar result for stochastic, discrete-time Riemannian Robbins–Monro methods. Our theory bridges this gap by showing that the same guarantees are in fact achieved by a wide array of RRM schemes.

3.9.3. Limit Cycles

We conclude this section by showing that, in complement to the convergence results above, our theory can also be used to derive convergence to limit cycles that arise in more general Riemannian settings.

▷ **Example 3.4.** The following example is taken from [DM21]: Consider the vector field on the 2-sphere $S^2 := \{p = (x,y,z) \in \mathbb{R}^3 : x^2+y^2+z^2 = 1\}$, defined by

$$V(p) = \begin{bmatrix} -y \\ x \\ 0 \end{bmatrix} + \left(z^2 - \frac{1}{4}\right) \begin{bmatrix} -xz \\ -yz \\ x^2+y^2 \end{bmatrix}. \tag{3.73}$$

By using spherical coordinates, one can see that the associated flow has two limit cycles and two equilibrium points; see Fig. 3.6. Therefore, the internally chain-transitive sets of V in (3.73) contain an attracting periodic orbit. Our Propositions 3.15 and 3.19 then imply that any RRM scheme driven by V also has a chance to get trapped in the limit cycle. To the best of our knowledge,

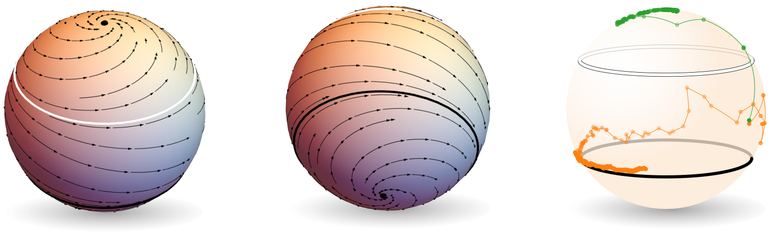


Figure 3.6. The flow induced by the game field (3.73). *Left and middle:* The two limit cycles are shown as two bands (white is repelling and black is absorbing). The equilibria are shown as black dots on the north and south pole. *Right:* Two trajectories starting from the same point, following a noisy estimate of V. The orange trajectory gets trapped in the limit cycle on the bottom, and the green trajectory goes to the absorbing equilibrium point at the north pole.

this is the first rigorous example of a cycling problem for Riemannian stochastic approximation in the literature. ◁

3.10. CONCLUSIONS AND DISCUSSIONS

To conclude, our theory offers a comprehensive framework for analyzing the convergence of Riemannian Robbins–Monro schemes that initially appear quite disparate. By checking simple conditions on the error terms Z_n as specified in Assumptions 3.4 and 3.5, our analysis allows us to infer the algorithm's long-term behavior by studying the deterministic dynamics of the flow (3.16). Despite the versatility of our results, they merely scratch the surface of the full potential of (RRM), leaving several vital research avenues open:

(1) In numerous applications, particularly in game theory and sequential online learning, direct access to V may not be feasible, necessitating the use of *zeroth-order* or *bandit* optimization methods. A significant question is whether a Riemannian Kiefer–Wolfowitz algorithm [KW52] can be analyzed within the (RRM) framework and whether there are fundamental differences compared to the Euclidean context.

(2) Our analysis relies heavily on the diminishing step-size assumption, which is applicable in many practical scenarios. However, many algorithms employ *constant* step-sizes, which our current theory does not address. It would be

valuable to explore if existing techniques for constant step-size stochastic approximation schemes [KH81; KY97] can be extended to a manifold setting.

(3) There is a dichotomy between the case where $\operatorname{inj}(\mathcal{M})$ is ∞ or not. This shows up specifically in our assumptions on noise and bias. While our analysis cannot tolerate unbounded noise or bias when $\operatorname{inj}(\mathcal{M}) < \infty$, it is interesting to see if this is merely an artifact of our analysis, or there could be some counterexamples.

(4) The retraction conditions of Assumption 3.7 impose uniform estimates for comparison with the exponential map. We believe that enough regularity of the retraction is sufficient for these estimates to hold, but the exact formulation of this regularity is an open question.

BIBLIOGRAPHIC NOTES

The foundational work of Bonnabel [Bon13] marked the inception of using retractions in Riemannian optimization by exploring scenarios in which the vector field V is the Riemannian gradient of a certain objective function. Subsequent studies [ZS16; Tri+18; BAC19; CB19; Lez20; Wan+21b] have elaborated on these initial findings, specifically for Riemannian stochastic gradient descent. Meanwhile, a parallel stream of research [FO02; LLM09; BFM17; HW21] obtain analogous results for the Riemannian proximal point method.

These works predominantly concern scenarios where V embodies a gradient field and thus, do not extend to non-gradient contexts. Nevertheless, an array of studies [FPN05; TH12; NSS16; FQT20; Kha+20; CLC21] has provided a limited generalization to non-gradient cases by examining Riemannian extragradient methods under geodesic monotonicity assumptions. This assumption, akin to convexity conditions, anticipates that V consistently orients towards its set of roots (which is a connected set in this case) in a suitable, geodesic sense, facilitating convergence through methodologies paralleling monotone operator theory in Hilbert spaces [BC17].

An alternative approach aiming to confirm the asymptotic pseudo-trajectory property within Riemannian stochastic approximation schemes is presented by Shah [Sha21]. However, there are several limitations in Shah's argument that diminish its overall applicability.

In [Sha21, p. 1131], the author asserts that proving the asymptotic pseudo-trajectory property involves demonstrating that

$$\lim_{t\to\infty} \sup_{h\in[0,T]} \|\bar{x}(t+h) - \hat{x}^{\tau_n}(h)\| \to 0.$$

Here, $\bar{x}$ denotes a linear interpolation in the coordinate space $\mathbb{R}^d$ between the coordinates of successive iterates of the algorithm, and $\hat{x}^{\tau_n}$ represents the coordinate representation of the flow orbit initiated at time τ_n.

While the primary objective is to demonstrate that the distance between the geodesic interpolation and the flow orbit diminishes, the distance under consideration is between the linear interpolation and the flow orbit within an arbitrary coordinate chart. This approach implicitly assumes that the distance in terms of the linear interpolation within Euclidean space sufficiently controls the geodesic distance, an assertion that remains unclear.

The existence of such coordinate charts throughout the manifold suggests its global flatness [Ili06], thus making the analysis excessively restrictive and unsuitable for authentic Riemannian contexts. This concern was a significant motivation for our effort to provide a more rigorous treatment of this problem.

Finally, recent works by Durmus et al. [Dur+20; Dur+21] consider a generic version of RRM schemes, incorporating both diminishing and constant step-sizes. However, the analysis of schemes with a constant step-size in [Dur+21] is unable to assert almost sure convergence and has an ergodic interpretation—a perspective that diverges from the main focus of this thesis. The setting of [Dur+20] is closer in spirit to our work, especially in considering bias impacts on V; still, their deductions are confined to dynamics permitting a Lyapunov function.

Durmus et al. [Dur+20] also delve into the sensitivity of general RRM schemes under bias influences. Central findings therein assert that an RRM scheme's error is bounded by the worst-case bias encountered. In contrast, our Theorem 3.4 demonstrates that the error magnitude can indeed be moderated by the asymptotic bias. Hence, while prior works assert mere boundedness of errors in algorithms like the Riemannian extra-gradient or proximal point method, our result substantiates their convergence to zero. Another notable difference lies in our methodological approach; unlike previous studies that presume the existence of a Lyapunov function, we derive one within our analysis of Theorem 3.6.

CHAPTER FOUR

STOCHASTIC APPROXIMATION FOR LANGEVIN-TYPE SDES

In this chapter, we explore a new set of stochastic approximation algorithms that arise from discretizing certain stochastic differential equations, and are often used for sampling from probability distributions and simulating stochastic models, even with noisy and incomplete information. Our main objective is to adapt these algorithms into a framework defined on the space of probability measures. We will show that, under certain conditions, these algorithms converge to compact, connected, attractor-free sets in Wasserstein distance. Specifically, we will prove that these algorithms form an asymptotic pseudo-trajectory in the space of probability measures equipped with Wasserstein metric, and identify the necessary conditions for their convergence.

Originality. Main results of this chapter are published in the conference proceedings [KHK23a]. There are considerable differences between this chapter and the mentioned publication, in notation, proofs, and content.

LIST OF IMPORTANT RESULTS

- **Corollary 4.7.** A set of probability measures with bounded second moments is compact in the $(2-\varepsilon)$-Wasserstein space for all $\varepsilon \in (0,1]$.
- **Theorem 4.9.** An SDE discretization with noise and bias in the evaluation of the drift constitutes an asymptotic pseudo-trajectory in the quadratic Wasserstein space for the flow corresponding to the Fokker–Planck equation of the SDE.
- **Lemma 4.10.** The only internally chain-transitive set for the flow of the Langevin diffusion is the singleton consisting of the target measure.
- **Theorem 4.13.** For dissipative and Lipschitz drifts and Lipschitz diffusion coefficients, the iterates of an SDE discretization have uniformly bounded second moments under some conditions on the diffusion coefficient, noise, and bias.
- **Lemma 4.14.** The only internally chain-transitive set for the flow of the dual mirror Langevin diffusion is the singleton consisting of the pushforward of the target measure under the mirror map.
- **Lemma 4.15.** The only internally chain-transitive set for the flow of the (primal) mirror Langevin diffusion is the singleton consisting of the target measure.

4.1. INTRODUCTION

In this chapter, we consider the time-homogeneous stochastic differential equation

$$dX_t = v(X_t)\,dt + \sigma(X_t)\,dW_t, \tag{4.1}$$

as well as various methods for its discretization. Here, v is a drift, σ is a diffusion coefficient, and W_t is a Brownian motion. Our primary objective is to understand the asymptotic behavior of these discretization schemes. By asymptotic behavior, we refer to the analysis of whether the probability law of the discretization at iteration n converges to some target distribution as $n \to \infty$.

One can think of two scenarios where the asymptotics of an SDE and a discretization thereof is important:

Sampling from a known distribution

The first scenario involves starting from an SDE which is known to have a desired and unique stationary distribution. This means that starting from an initial distribution, say standard Gaussian, the probability law of the solution to the SDE converges to the stationary distribution. This scenario is particularly relevant to the problem of sampling, where the goal is to produce a random variable with a given probability distribution.

A particularly efficient method for sampling is to use SDEs such as (overdamped or underdamped) Langevin diffusion and mirror Langevin diffusion. These SDEs share the property that the probability law of their solution converges towards a desired target distribution that is baked into the SDE itself. Notably, these SDEs depend on the target measure only through the score function, which is the gradient of the log-density of the target. This feature is significant because it eliminates the necessity of knowing the normalization constant for a density, a notoriously challenging problem in high-dimensional spaces.

The relevant question here is whether a discretization of an SDE demonstrates similar convergence behavior as the original SDE. While basic discretizations usually exhibit this behavior, it is desirable to determine if convergence holds for alternative algorithms used in practice. Notably, demonstrating that an SDE converges to the stationary distribution typically requires very mild conditions; see, for instance, the classical result of [RT96b, Thm. 2.1] for the Langevin diffusion. However, the difficulty increases for discretizations, especially those incorporating noise and bias, necessitating more structural assumptions about the target density, such as log-concavity or adherence to specific functional inequalities.

Modeling phenomena

In the second scenario, SDEs are used to model phenomena. Famous examples include modeling the movement of a molecule in a viscous fluid by the underdamped Langevin diffusion in physics, and dynamics of the price of a stock in the Black–Scholes options pricing model by the geometric Brownian motion in mathematical finance. Another class of examples are mean-field models, where the evolution of an entire population comprising an infinite number of "particles" is of interest. By considering one representative particle from the population, one can derive the *McKean–Vlasov process*, which is an SDE with both the drift and diffusion terms dependent on the probability density of the population at each time. These equations are particularly useful for modeling systems of interacting particles. An example in physics is the kinetic equation

$$dX_t = -\nabla f(X_t)\,dt - (\nabla g * \varrho_t)(X_t)\,dt + \sqrt{2}\,dW_t,$$

where ϱ_t denotes the population density at time t, f represents a potential energy, and g denotes an interaction energy; $g(x-y)$ captures the interaction between a particle located at x and another at y, and its gradient is the force one puts on the other. Another relevant example is the training of an infinitely wide two-layer neural network via stochastic gradient descent. By modeling each neuron as a particle, one can show that the resulting evolution adheres to a McKean–Vlasov process; see [KHK23b] and references therein for more examples.

In order to simulate a McKean–Vlasov process, a two-level discretization approach is required. Initially, one estimates the population with N particles, transforming the mean-field equation into a system of N coupled SDEs. Next, a step-sizing rule is applied to each SDE in this system, resulting in a system of recursive equations that involve noise and bias, depending on the chosen discretization algorithm.

The dynamics of this discrete-time system may markedly differ from the original McKean–Vlasov equation, and it is crucial to determine if there are any similarities between the two. This turns out to be a complex question and has spawned its own body of literature known as *propagation of chaos* results. Essentially, these results show that if N is sufficiently large, the distribution of the continuous-time N-particle approximation will be close to the mean-field solution. For time discretization, however, we have to see whether the probability law of a discretization converges to the same targets as the continuous-time SDE or not.

Both of these scenarios underscore our main objective in this chapter: understanding the asymptotics of an SDE discretization. In this chapter, we focus on studying SDEs of the form (4.1), which we call *Langevin-type SDEs*. For mean-field applications and systems of coupled SDEs, we refer the reader to [KHK23b].

Root-finding and SDEs

Our insight from previous chapters indicates that when dealing with noisy and biased update rules, we enter the realm of stochastic approximation. In stochastic approximation, the objective is to find the zeros of a vector field (or a function) by incorporating stochastic information. For the case of this chapter, we show that many of the examples mentioned before have a root-finding formulation under the hood. To gain intuition, we begin with the sampling problem and subsequently extend this intuition to general Langevin-type SDEs.

Analysis via the Fokker–Planck equation

Suppose that the goal is to sample from the target distribution $\mu \propto e^{-f}$. Many practical sampling algorithms are based on the overdamped Langevin diffusion

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t. \tag{4.2}$$

The main method for understanding the behavior of this SDE is to find the probability distribution function as a function of time using the Fokker–Planck equation, which is a deterministic PDE describing how the probability density ϱ_t of X_t evolves in time:

$$\partial_t \varrho_t = \nabla \cdot (\varrho_t \nabla f) + \Delta \varrho_t = \nabla \cdot \big(\varrho_t \nabla \log \frac{\varrho_t}{\mu}\big). \tag{4.3}$$

Many properties of the SDE can be read from the Fokker–Planck equation. For example, if $\varrho_t = \mu$, we see that the right-hand side becomes zero, implying $\partial_t \varrho_t = 0$; this shows that μ is a stationary distribution of the Langevin diffusion.

Geometry of the Fokker–Planck equation

A profound way to comprehend the Fokker–Planck equation (4.3) is through gradient flows in the Wasserstein space. Let us briefly mention how this is possible. In their influential paper, Jordan, Kinderlehrer, and Otto [JKO98] showed that by endowing the space of probability measures with the Wasserstein distance—a distance originating from optimal transport theory—and defining a notion of gradient flow of a functional defined on this space, the Fokker–Planck equation becomes the gradient flow of the relative entropy functional $\varrho \mapsto H(\varrho \,|\, \mu)$. Later, Otto [Ott01] studied similar PDEs and realized that one could impose a Riemannian structure on the Wasserstein space by defining tangent spaces, inner product, and so on, turning the Wasserstein space of distributions into an infinite-dimensional Riemannian manifold. This Riemannian structure facilitates defining gradients of functionals, transforming the gradient flow into the (formal) ODE[1]

$$\dot{\varrho}_t = -\nabla_{W_2} H(\varrho_t \,|\, \mu).$$

Here, ∇_{W_2} denotes the gradient in the sense of Otto. This formalization implies a trend towards (a unique) equilibrium: the relative entropy becomes a Lyapunov function, with its only critical point being μ.

Inspired by this observation, various authors studied Langevin diffusion discretizations through the lens of Fokker–Planck equation and gradient flows. These works use relative entropy as the Lyapunov function and assess the "dissipation of entropy" across iterations. Specifically, they measure the reduction in the Lyapunov function at each iteration by comparing it to the original Fokker–Planck equation. Works such as [VW19] derive a Fokker–Planck equation for a single iteration of the sampling algorithm and perform comparative analysis with the corresponding PDE of the Langevin diffusion. This approach facilitates deriving non-asymptotic bounds on the distance between Langevin diffusion and its discretization. Such analyses typically rely on functional inequalities (since they are conducted at PDE level), such as Log-Sobolev or Poincaré inequalities, and measure closeness of distributions via relative entropy.

Our take on the result of Otto is the following: An SDE effectively induces a semi-flow on the space of probability measures; starting from any initial distribution for X_0, the solution of the Fokker–Planck equation traces a curve in the Wasserstein space, corresponding to an orbit of the semi-flow. Furthermore, a discretization algorithm (such as a sampling algorithm) implicitly constructs a sequence of points within this space, i.e., the law of the consecutive iterates of the algorithm correspond to a sequence of points in the Wasserstein space. Given that the semi-flow is well-structured (for example, it exhibits a gradient flow structure in the case of Langevin diffusion) and converges to some target measure, our task is to demonstrate that the law of the iterates of the discretization algorithm forms an asymptotic pseudo-trajectory of the semi-flow. As established in previous chapters, showing this property ensures that noise and bias do not adversely affect the asymptotic behavior of the algorithm.

Recall that the notion of an asymptotic pseudo-trajectory makes sense for a continuous curve in a metric space. Otto's Riemannian structure on the space of probability distributions closely aligns with the Wasserstein geometry. Indeed, by endowing the space of probability distributions with the quadratic Wasserstein

[1] There are two ways of writing the gradient flow, depending on how the tangent space is defined. For example, in the sense of Ambrossio, Gigli, and Savaré [AGS05], the gradient flow writes as

$$\partial_t \varrho_t = -\nabla \cdot (\varrho_t \nabla_{W_2} H(\varrho_t \,|\, \mu)).$$

distance, one obtains the same distance function as the one derived from Otto's Riemannian structure. Therefore, within the Wasserstein space, we first show that any Langevin-type SDE induces a semi-flow. We then construct an interpolation between the iterates of the discretization algorithm in this space, obtaining a continuous curve. Finally, we verify the asymptotic pseudo-trajectory property of this curve.

Convergence conditions and compactness

To ensure convergence, we need to demonstrate that the law of the algorithm's iterates forms a precompact set. In the Wasserstein space, compact sets possess a structure that facilitates this. Specifically, well-known conditions such as dissipativity can lead to precompactness of the law of the iterates. Since dissipativity is a reasonable assumption for any relevant Langevin-type SDE, we can thus ensure that meaningful convergence criteria are satisfied.

Chapter Roadmap

In Section 4.2, we review the necessary background knowledge for this chapter. This includes parts of stochastic calculus, Fokker–Planck equations, and Wasserstein spaces. Section 4.3 sets up the stage for a stochastic approximation analysis of SDE discretization algorithms. This includes specifying the basic template for a discretization algorithm, describing how to interpolate between the iterates, as well as identifying the flow corresponding to the SDE. In Section 4.4, we bring the main theorem of this chapter: Theorem 4.9, which states that under some conditions, the law of the iterates of an SDE discretization algorithm forms an asymptotic pseudo-trajectory of the flow corresponding to the SDE. This section also identifies the internally chain-transitive sets of the Langevin diffusion. Section 4.5 is all about stability conditions; those that imply precompactness of the iterates of the algorithm. We show that dissipativity-type conditions are sufficient for stability. In Section 4.6, we go over six different sampling algorithms used in practice, and show that they all follow the template mentioned earlier. These examples include the mirror Langevin algorithm, that simulates the mirror Langevin diffusion instead of the Langevin diffusion. We prove there that the mirror Langevin diffusion also has similar internally chain-transitive sets as the Langevin diffusion. We conclude this chapter in Section 4.7 and bring extra pointers to the literature in the bibliographic notes at the end of the chapter.

4.2. A PRIMER ON SDES AND WASSERSTEIN GEOMETRY

In this short section, we review basic concepts related to SDEs and Wasserstein spaces. The interested reader is referred to the book of Le Gall [Le 16] for background in stochastic calculus, the book of Villani [Vil03] or Santambrogio [San15] for an introduction to optimal transport and Wasserstein spaces, and the book of Ambrossio, Gigli, and Savaré [AGS05] for a rigorous treatment of gradient flows in metric and Wasserstein spaces.

4.2.1. Itô Calculus

We work withing a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. For a process $(N_t)_{t\geq 0}$ and a stopping time τ, we define the *stopped process* N^τ to be $N^\tau_t = N_t$ if $t \leq \tau$ and $N^\tau_t = N_\tau$ if $t > \tau$. A continuous adapted process $(M_t)_{t\geq 0}$ is a *local martingale* if there exists a sequence $\tau_1, \tau_2, \ldots$ of stopping times such that $\tau_k \to \infty$ almost surely, and for all $k \geq 1$, M^{τ_k} is a continuous martingale. A process $(Z_t)_{t\geq 0}$ is a continuous *semimartingale* if it can be written as $Z_t = M_t + V_t$ where M is a local martingale in this filtered probability space, and V is the difference between two adapted continuous nondecreasing processes started from 0. The process V is usually called the *finite variation part* of the semimartingale Z.

For a local martingale $(M_t)_{t\geq 0}$, the *quadratic variation* process $\langle M\rangle$ is the a.s. unique adapted nondecreasing process such that $((M_t)^2 - \langle M\rangle_t)_{t\geq 0}$ is a local martingale. The quadratic variation of a semimartingale is defined to be the quadratic variation of its local martingale part. For two processes $(M_t)_{t\geq 0}$ and $(N_t)_{t\geq 0}$, the *cross-variation* $\langle M, N\rangle$ is defined via

$$\langle M, N\rangle_t = \frac{1}{4}(\langle M+N\rangle_t + \langle M-N\rangle_t).$$

One of the most important theorems in stochastic calculus is the Itô's formula:

Theorem 4.1 (Itô's Formula)**.** *Let Z be a continuous semimartingale with decomposition $Z = M + V$, and assume that $Z_t \in D \subseteq \mathbb{R}^d$ a.s. for all $t \geq 0$. Let F be a C^2 real-valued function defined on D. Then $F(Z_t)$ is a continuous semimartingale, and almost surely*

$$F(Z_t) = F(Z_0) + \sum_{i=1}^{d} \int_0^t \partial_i F(Z_s)\, dZ^i_s + \frac{1}{2} \sum_{1\leq i,j\leq d} \int_0^t \partial^2_{ij} F(Z_s)\, d\langle M^i, M^j\rangle_s.$$

Corollary 4.2. *Let X be a semimartingale satisfying the SDE*

$$dX_t = v_t(X_t)\,dt + \sigma_t(X_t)\,dW_t,$$

where $v : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d\times k}$, and W is a standard k-dimensional Brownian motion. Then, for $F \in C^2(\mathbb{R}^d)$ we have

$$\begin{aligned} F(X_t) = F(X_0) &+ \int_0^t \langle \nabla F(X_s), v_s(X_s)\rangle\,ds + \frac{1}{2}\int_0^t \langle \nabla^2 F(X_s), G_s(X_s)\rangle\,ds \\ &+ \int_0^t \nabla F(X_s)^\top \sigma_s(X_s)\,dW_s \end{aligned}$$

where $G_t(x) = \sigma_t(x)\sigma_t(x)^\top$.

Remark. If the local martingale part of $(F(X_t))_{t\geq 0}$ is a martingale, then

$$\mathbb{E}[F(X_t)] = \mathbb{E}\int_0^t \langle \nabla F(X_s), b_s(X_s)\rangle\,ds + \frac{1}{2}\,\mathbb{E}\int_0^t \langle \nabla^2 F(X_s), G_s(X_s)\rangle\,ds. \qquad \Diamond$$

An important property of Itô processes is the following lemma, known as *Itô isometry*:

Lemma 4.3. *Let $(W_t)_{t\geq 0}$ be a standard Brownian motion in $\mathbb{R}^k$. For any matrix-valued process $H \in \mathbb{R}^{d\times k}$ adapted to the same filtration as W, it holds*

$$\mathbb{E}\left[\left\|\int_0^t H_s\,dW_s\right\|^2\right] = \mathbb{E}\left[\int_0^t \|H_s\|_{\mathrm{F}}^2\,ds\right],$$

where $\|H_s\|_{\mathrm{F}}$ is the Frobenius norm of the matrix H_s.

4.2.2. The Fokker–Planck Equation

We continue with some basic notions regarding SDEs and their laws. We say that the SDE

$$dX_t = v(X_t)\,dt + \sigma(X_t)\,dW_t, \tag{4.4}$$

admits a *strong solution*, if, for a given standard k-dimensional Brownian motion $(W_t)_{t\geq 0}$ in some filtered probability space and $x_0 \in \mathbb{R}^d$, there exists a continuous semimartingale $(X_t)_{t\geq 0}$ adapted to the same filtration, such that for all $t \geq 0$,

$$X_t = x_0 + \int_0^t v(X_s)\,ds + \int_0^t \sigma(X_s)\,dW_s.$$

This means that a strong solution is fully determined by the Brownian motion in the equation and uses no other randomness. There are several sufficient conditions that imply existence and uniqueness of strong solutions for SDEs. Here, we mention a simple criterion that is the analogue of Picard–Lindelöf theorem for ODEs:

Theorem. *Given is a filtered probability space along a with a k-dimensional Brownian motion $(W_t)_{t\geq 0}$. Suppose that v and σ are Lipschitz functions with values in $\mathbb{R}^d$ and $\mathbb{R}^{d\times k}$, respectively. Then:*

(i) For all $x_0 \in \mathbb{R}^d$, there exists a continuous semimartingale X adapted to the same filtration as W such that for all $t \geq 0$,

$$X_t = x_0 + \int_0^t v(X_s)\,ds + \int_0^t \sigma(X_s)\,dW_s.$$

(ii) If Y is another semimartingale such that for all $t \geq 0$,

$$Y_t = x_0 + \int_0^t v(Y_s)\,ds + \int_0^t \sigma(Y_s)\,dW_s,$$

then $X = Y$ almost surely.

If the initialization x_0 is a random variable independent of the σ-algebra of the Brownian motion, the result of the theorem is still valid, with the difference that the strong solution is adapted to the larger filtration that contains both the σ-algebra of the Brownian motion and x_0.

The *Fokker–Planck equation* describes the evolution of the law of a solution of an SDE. Assume that x_0 has a density ϱ_0 with respect to the Lebesgue measure, and let $(X_t)_{t\geq 0}$ be the strong solution of (4.4) starting at x_0. It turns out that X_t admits a density ϱ_t for all $t \geq 0$, and these densities solve the PDE

$$\frac{\partial}{\partial t}\varrho_t(x) = -\sum_{i=1}^d \frac{\partial}{\partial x^i}(\varrho_t(x)v_i(x)) + \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d \frac{\partial^2}{\partial x^i \partial x^j}(\varrho_t(x)G_{ij}(x)),$$

where $G(x) := \sigma(x)\sigma(x)^\top \in \mathbb{R}^{d\times d}$. Letting $\nabla \cdot w$ denote the divergence of a vector field w, $\nabla^2 h$ denote the Hessian of h, and $\langle A, B\rangle := \operatorname{tr}(AB^\top)$, we can write the equation above in a more compact way:

$$\partial_t \varrho_t = -\nabla \cdot (\varrho_t v) + \frac{1}{2}\langle \nabla^2, \varrho_t\, G\rangle. \tag{4.5}$$

Remark. One can also consider *weak solutions* of the PDE (4.5), therefore removing the necessity of existence of smooth densities. We say $(\varrho_t)_{t\geq 0}$ is a weak solution of the Fokker–Planck PDE (4.5), if for all $T > 0$ and all test functions $\psi \in C_c^\infty((0,T)\times \mathbb{R}^d)$, it holds

$$\int_0^T \int_{\mathbb{R}^d} (\partial_t \psi)\, d\varrho_t\, dt = \int_0^T \int \langle \nabla\psi, v\rangle\, d\varrho_t\, dt + \frac{1}{2}\int_0^T \int \langle \nabla^2\psi, G\rangle\, d\varrho_t\, dt.$$

See [San15, Def. 4.1] for a rigorous treatment. ♢

While much of the theory we develop works for weak solutions, here and in the sequel, we tacitly assume that all measures have a density with respect to the Lebesgue measure, and we identify a measure with said density. We also denote by $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ the space of all probability measures that have second moments and are absolutely continuous with respect to the Lebesgue measure.

Remark. The Fokker–Planck equation can be seen as a special case of the more general continuity equation. For a velocity field $w : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$, the *continuity equation* is the PDE

$$\partial_t \varrho_t + \nabla\cdot(\varrho_t w_t) = 0$$

with no-flux boundary conditions. This equation encodes mass-preservation of ϱ_t:

$$\frac{d}{dt}\int_{\mathbb{R}^d} \varrho_t(x)\, dx = -\int_{\mathbb{R}^d} \nabla\cdot(\varrho_t(x) w_t(x))\, dx = 0.$$

The continuity equation is also known as the *transport equation* for the following reason: Define the flow of the non-autonomous ODE generated by w_t, that is,

$$\begin{cases} \dfrac{\partial \Psi}{\partial t}(t,x) = w_t(x) \\ \Psi(0,x) = x. \end{cases}$$

Then it turns out that $\varrho_t = \Psi_{t\#}\varrho_0$ [see, e.g., San15, Thm. 4.4]. This means that the mass cannot teleport, and no "source" or "sink" of mass exists; mass only moves continuously according to the flow Ψ. A rigorous treatment of the continuity equation and the Fokker–Planck equation can be found in [San15, Ch. 4]. ♢

Let us go back to the Fokker–Planck equation. For the matrix-valued function $G : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, define the divergence $\nabla\cdot G(x)$ to be the vector in $\mathbb{R}^d$ whose ith component is the divergence of the ith row of $G(x)$. With this notation,

$$\langle \nabla^2, G(x)\rangle = \nabla\cdot(\nabla\cdot G)(x) = \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2 G_{ij}(x)}{\partial x^i \partial x^j}.$$

Thus, $\nabla \cdot (\varrho_t G) = G\nabla \varrho_t + \varrho_t \nabla \cdot G = \varrho_t (G\nabla \log \varrho_t + \nabla \cdot G)$, and we can rewrite the Fokker–Planck equation (4.5) in the form of a continuity equation:

$$\partial_t \varrho_t + \nabla \cdot \bigl(\varrho_t \{v - \tfrac{1}{2} G\nabla \log \varrho_t - \tfrac{1}{2}\nabla \cdot G\}\bigr) = 0. \tag{4.6}$$

With the aid of this formulation, we will later define a semi-flow on the space of probability measures.

4.2.3. Wasserstein Spaces

Let μ and ν be two probability measures on $\mathbb{R}^d$. A probability distribution π on $\mathbb{R}^d \times \mathbb{R}^d$ is called a *coupling* between μ and ν if its first marginal is μ and its second marginal is ν; that is, for all bounded continuous functions f and g,

$$\iint f(x)\, d\pi(x,y) = \int f(x)\, d\mu(x), \quad \text{and} \quad \iint g(y)\, d\pi(x,y) = \int g(y)\, d\nu(y).$$

We denote by $\Pi(\mu,\nu)$ the set of all couplings between μ and ν. For any $p \geq 1$, define the *p-Wasserstein distance* of μ to ν as

$$W_p(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left(\int \|x-y\|^p \, d\pi(x,y) \right)^{1/p}.$$

This distance satisfies the axioms of a metric and is well-defined for all probability measures with finite p-moments, that is, the set

$$\mathcal{P}_p(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int \|x\|^p \, d\mu(x) < \infty \right\},$$

The *p-Wasserstein space* is the space $\mathcal{P}_p(\mathbb{R}^d)$ equipped with the p-Wasserstein distance. It is a remarkable fact that this space is a complete metric space, in the sense that any Cauchy sequence has a limit.

In the sequel, we mostly work with the 2-Wasserstein space, also called the *quadratic Wasserstein space*, or simply, the Wasserstein space.

4.2.4. Precompact Sets in the Wasserstein Space

An important building block in the analysis of stochastic approximation algorithms is the precompactness of the iterates. In this short section we bring sufficient conditions for a set to be precompact in the Wasserstein topology.

The following proposition, extracted from [AGS05, Prop. 7.1.5], gives a complete characterization of precompact sets in $\mathcal{P}_p(\mathbb{R}^d)$:

Proposition 4.4. *A set $\mathcal{K} \subset \mathcal{P}_p(\mathbb{R}^d)$ is precompact (in the metric topology of the p-Wasserstein distance) if and only if*

(i) it is tight, *that is, for each $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset \mathbb{R}^d$ such that*
$$\sup_{\mu \in \mathcal{K}} \mu(\mathbb{R}^d \setminus K_\varepsilon) \leq \varepsilon, \tag{4.7}$$

(ii) it has uniformly integrable p-moments, *that is, for each $\varepsilon > 0$, there is some $R > 0$ such that*
$$\sup_{\mu \in \mathcal{K}} \int_{\{\|x\| > R\}} \|x\|^p \, d\mu(x) < \varepsilon. \tag{4.8}$$

Let us remark that by Prokhorov's theorem [see, e.g., Bil99, Thm. 5.1], tightness of $\mathcal{K}$ implies its precompactness in the topology of narrow convergence (convergence against bounded continuous functions). The additional uniform integrability condition ensures precompactness in the p-Wasserstein topology, which is a stronger topology compared to the narrow topology.

It turns out that having bounded moments is a sufficient condition for tightness of a set of probability measures. This follows from the following lemma [see also AGS05, Rem. 5.1.5]:

Lemma 4.5. *If there is a function $g : \mathbb{R}^d \to \mathbb{R}_+$ such that $\{g \leq \lambda\}$ is compact for any $\lambda \in \mathbb{R}$ and $\int g \, d\mu \leq C$ for all $\mu \in \mathcal{K}$, then $\mathcal{K}$ is tight. Consequently, if $\sup_{\mu \in \mathcal{K}} \int \|x\|^p \, d\mu(x) < \infty$ for some $p \geq 1$, then $\mathcal{K}$ is tight.*

Proof. Fix $\varepsilon > 0$ and set $\lambda = C/\varepsilon$. Then, by Markov's inequality
$$\mu(\{g > \lambda\}) \leq \frac{1}{\lambda} \int g \, d\mu \leq \frac{C}{\lambda} = \varepsilon.$$

Since $K_\varepsilon := \{g \leq \lambda\}$ is compact, we have proven the tightness of $\mathcal{K}$. Letting $g(x) = \|x\|^p$ shows the second claim of the lemma. □

A useful criterion for checking p-uniform integrability condition (4.8) is the following [AGS05, Eqn. (5.1.20)]:

Lemma 4.6. *Let $\mathcal{K} \subset \mathcal{P}(\mathbb{R}^d)$. If $\sup_{\mu \in \mathcal{K}} \int \|x\|^q \, d\mu(x) < \infty$ for some $q > p$, then $\mathcal{K}$ is p-uniformly integrable.*

Proof. Let $M_q = \sup_{\mu \in \mathcal{K}} \int \|x\|^q \, d\mu(x)$ and notice that $\mathbf{1}_{\|x\| > R} < \|x\|^{q-p}/R^{q-p}$. Therefore, for any $\mu \in \mathcal{K}$,
$$\int \|x\|^p \mathbf{1}_{\|x\| > R} \, d\mu(x) < \frac{1}{R^{q-p}} \int \|x\|^q \, d\mu(x) \leq \frac{M_q}{R^{q-p}}.$$

Letting $R > (M_q/\varepsilon)^{p-q}$ makes the above less than ε. □

Combining these two lemmas, we can derive a sufficient condition for precompactness in Wasserstein topology as follows:

▶ **Corollary 4.7.** *If a sequence of measures $\{\mu_n\}_{n\in\mathbb{N}}$ has uniformly bounded second moments, it is precompact in the $(2-\varepsilon)$-Wasserstein space for any $0 < \varepsilon \leq 1$.*

Proof. Lemma 4.5 implies that the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ is tight. Lemma 4.6 shows that this sequence is p-uniformly integrable for any $p < 2$. Thus, Proposition 4.4 shows that the sequence is precompact in all $(2-\varepsilon)$-Wasserstein spaces with $0 < \varepsilon \leq 1$. □

It is instructive to give an example of a set of probability measures in $\mathcal{P}_p(\mathbb{R}^d)$ with uniformly bounded p-moments that is *not* precompact (i.e., it does not have uniformly integrable p-moments).

▷ **Example 4.1.** For $p \geq 1$, consider the sequence of probability measures defined on the real line: $\mu_n = (1-n^{-p})\delta_0 + n^{-p}\delta_n$. Firstly, it is evident that the p-moment of all these measures is equal to 1, and thus, is uniformly bounded. Additionally, the p-Wasserstein distance between μ_n and δ_0 is equal to 1 for all $n \in \mathbb{N}$. This means that this sequence is a subset of the closed ball $\overline{B_1(\delta_0)}$ in $\mathcal{P}_p(\mathbb{R}^d)$. However, despite these properties, this family does not have uniformly integrable p-moments:

$$\sup_{n\geq 1} \int_{|x|>R} |x|^p \, d\mu_n(x) = 1,$$

for any given $R > 0$. Consequently, the family $\{\mu_n\}$ is not precompact in the p-Wasserstein space. An alternative perspective on this is to note that although μ_n converges narrowly to δ_0, any subsequence converging in W_p would require the p-moment converge to the p-moment of δ_0, which is 0; this is impossible for the given family of measures, since the p-moment of all μ_n is 1. ◁

4.3. LANGEVIN–ROBBINS–MONRO SCHEMES

We consider discretizations of the time-homogeneous SDE

$$dX_t = v(X_t)\,dt + \sigma(X_t)\,dW_t, \tag{4.9}$$

where $v : \mathbb{R}^d \to \mathbb{R}^d$ is a vector field, called the *drift*, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times k}$ is the *diffusion matrix*, and $(W_t)_{t\geq 0}$ is a k-dimensional standard Brownian motion. We have already seen an example of this SDE: the Langevin diffusion (4.2) corresponds to setting $v = -\nabla f$ and $\sigma \equiv \sqrt{2}\, I_{d\times d}$ in (4.9).

A general template for discretizing the SDE (4.9) is as follows: Starting from an initial point $\boldsymbol{x}_0 \in \mathbb{R}^d$, the iterates $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ follow the recursion

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \gamma_{n+1}\{v(\boldsymbol{x}_n) + Z_{n+1}\} + \sqrt{\gamma_{n+1}}\,\sigma(\boldsymbol{x}_n)\,\xi_{n+1}, \tag{LRM}$$

where

(1) $\boldsymbol{x}_n \in \mathbb{R}^d$ denotes the state of the algorithm at iteration n,

(2) ξ_{n+1}'s are i.i.d. standard Gaussian random variables in $\mathbb{R}^k$,

(3) Z_{n+1} is a (random or deterministic) error term,

(4) and γ_{n+1} is the algorithm's step-size policy.

If $Z_n = 0$ for all $n \in \mathbb{N}$, the resulting scheme is known as the *Euler–Maruyama* discretization. In the above, we assume that the error terms and the standard Gaussians are generated after $\boldsymbol{x}_n$ and use a similar indexing convention as in Section 2.5. We will later decompose the error term into noise and bias. However, this needs a bit of measure-theoretic care, which we will discuss below.

4.3.1. Stochastic Interpolation

An essential step in a dynamical system analysis for a stochastic approximation algorithm is interpolating consecutive iterates to obtain a continuous curve. In Euclidean spaces, we did so by using straight line segments, and in Riemannian manifolds with geodesics. As in this chapter we focus on the law of the iterates, we need to devise a method to connect the laws of consecutive iterates using a continuous curve in the Wasserstein space. Our approach is to construct a stochastic process whose initial and terminal laws correspond to the law of two consecutive iterates. This construction facilitates comparison with the SDE (4.9).

We start by interpolating the terms $\sqrt{\gamma_{n+1}}\,\xi_{n+1}$ in (LRM) with a standard Brownian motion. Let $(W_t)_{t\geq 0}$ be a standard Brownian motion defined on a filtered probability space with the filtration $(\mathcal{F}_t^W)_{t\geq 0}$ satisfying the usual conditions,[2] and let $\tau_n = \sum_{k=1}^n \gamma_k$ be the effective time that has elapsed up to iteration n. As the Brownian motion has stationary and independent increments, if follows that the sequence of random variables $\{(W_{\tau_1} - W_0), (W_{\tau_2} - W_{\tau_1}), (W_{\tau_3} - W_{\tau_2}), \ldots\}$ are independent Gaussian random variables and

$$W_{\tau_{n+1}} - W_{\tau_n} \sim \mathcal{N}(0, (\tau_{n+1} - \tau_n)I) \overset{\text{law}}{=} \sqrt{\gamma_{n+1}}\,\xi_{n+1}.$$

[2] We say a filtration satisfies the usual conditions if it is *right-continuous* and *complete*, i.e., for all $t \geq 0$, $\mathcal{F}_t = \bigcap_{\varepsilon>0} \mathcal{F}_{t+\varepsilon}$, and each $\mathcal{F}_t$ contains all $\mathbb{P}$-null sets.

Interpolating the error terms Z_{n+1} across iterations is more intricate due to the extra randomness in Z_{n+1} that might not be adapted to the natural filtration of the Brownian motion, as well as possible dependency of the error term on the Brownian path. We will see shortly that if all the extra randomness (such as the random seeds used for generating stochastic gradients via a stochastic first-order oracle) is created at the *beginning* of each iteration and before the creation of ξ_{n+1}, then there exists a continuous interpolation for the error terms that has our desirable properties. Let us formalize this assumption, which is easily verified for all our examples in Section 4.6 under the condition of using a SFO; see Section 3.7 for a reminder about SFOs.

▷ **Assumption 4.1.** *For each iteration $n \geq 1$, all the extra noise involved in Z_{n+1} can be injected at the start of the iteration. This means that there exists a σ-algebra $\mathcal{G}_{n+1}$, independent of the Brownian motion after τ_n, so that Z_{n+1} is $(\mathcal{G}_{n+1} \vee \mathcal{F}^W_{\tau_{n+1}})$-measurable.*

An example of an error term that satisfies the assumption above is when Z_{n+1} is a measurable function of the current iterate $\boldsymbol{x}_n$, some extra randomness ω_{n+1} (which is $\mathcal{G}_{n+1}$-measurable), and the Brownian path in the interval $[\tau_n, \tau_{n+1}]$.

We construct the interpolation for the error terms satisfying the assumption above based on the classical martingale representation theorem, which we recall below [see, e.g., RW00, Thm. 36.1]:

Theorem 4.8. *Let $(W_t)_{t\geq 0}$ be a k-dimensional Brownian motion defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, and suppose that $(\mathcal{F}_t)_{t\geq 0}$ is the natural filtration generated by W and $\mathcal{F}_0$; i.e., $\mathcal{F}_t = \sigma(W_s : s \leq t) \vee \mathcal{F}_0$. Then, for any $T \geq 0$ and every square integrable random variable $Y \in \mathcal{F}_T$, there exists predictable processes $\beta^1, \dots, \beta^k$ satisfying $\mathbb{E}[\int_0^t (\beta^i_s)^2\, ds] < \infty$ for all $t \in [0, T]$, so that the following continuous representation holds:*

$$\mathbb{E}[Y \mid \mathcal{F}_t] = \mathbb{E}[Y \mid \mathcal{F}_0] + \sum_{i=1}^k \int_0^t \beta^i_s \, dW^i_s, \quad 0 \leq t \leq T.$$

Let us illustrate how to use this theorem for constructing an interpolation for the first iteration of (LRM). We are given Z_1, which by Assumption 4.1 is $(\mathcal{G}_1 \vee \mathcal{F}^W_{\gamma_1})$-measurable. By enlarging the natural filtration of the Brownian motion at time 0 with $\mathcal{G}_1$, we obtain the right-continuous filtration $(\mathcal{F}_t)$. Theorem 4.8 then implies that $\mathbb{E}[Z_1 \mid \mathcal{F}_t]$ interpolates between $\mathbb{E}[Z_1 \mid \mathcal{F}_0]$ and Z_1 by a continuous process, as t ranges over $[0, \gamma_1]$. Based on this, we can interpolate between the first two iterates $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ as

$$X_t = \boldsymbol{x}_0 + tv(\boldsymbol{x}_0) + t\,\mathbb{E}[Z_1 \mid \mathcal{F}_t] + \sigma(\boldsymbol{x}_0)W_t, \quad t \in [0, \gamma_1].$$

Similarly, by iteratively enlarging the filtration at times $\{\tau_n\}_{n\geq 0}$, we construct the *stochastic interpolation* of the iterates $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ as

$$X_t = \boldsymbol{x}_n + (t-\tau_n)(v(\boldsymbol{x}_n) + \mathbb{E}[Z_{n+1}\,|\,\mathcal{F}_t]) + \sigma(\boldsymbol{x}_n)\,(W_t - W_{\tau_n}), \tag{4.10}$$

for $t \in [\tau_n, \tau_{n+1}]$.

Remark. It is straightforward to see that the law of the stochastic interpolation (4.10) forms a continuous curve in the Wasserstein space. What we have to show is that $W_2(X_{t+\delta}, X_t)$ can get arbitrarily small provided that δ is chosen small. Suppose that $t, t+\delta \in [\tau_n, \tau_{n+1}]$. We have

$$\begin{aligned} W_2(X_{t+\delta}, X_t) &\leq \mathbb{E}\,\|X_{t+\delta} - X_t\|^2 \\ &\leq O(\delta) + \mathbb{E}\big[\|(t+\delta)\,\mathbb{E}[Z_{n+1}\,|\,\mathcal{F}_{t+\delta}] - t\,\mathbb{E}[Z_{n+1}\,|\,\mathcal{F}_t]\|^2\big] \end{aligned}$$

The second term on the right-hand side can be bounded using the properties of the interpolation. For brevity, let M_t be the jth coordinate of $\mathbb{E}[Z_{n+1}\,|\,\mathcal{F}_t]$. Theorem 4.8 and Itô isometry (Lemma 4.3) then imply that

$$\begin{aligned} \mathbb{E}[((t+\delta)M_{t+\delta} - tM_t)^2] &\leq 2(t+\delta)^2\,\mathbb{E}[(M_{t+\delta} - M_t)^2] + 2\delta^2\,\mathbb{E}[(M_t)^2] \\ &\leq 2(t+\delta)^2\,\mathbb{E}\left[\sum_{i=1}^{d}\int_t^{t+\delta} (\beta_s^i)^2\,ds\right] + 2\delta^2\,\mathbb{E}[(Z_{n+1}^j)^2]. \end{aligned}$$

By the continuity of the local martingales $\int_0^t \beta_s^i\,dW_s^i$ [see RW00, Thm. 36.5], this term can get as small as desired by choosing δ small enough, showing the continuity of the law of $(X_t)_{t\geq 0}$ in the Wasserstein space. $\lozenge$

Remark. One might think of using geodesic interpolation between the iterates in the following sense. Suppose μ_n and μ_{n+1} are the laws of $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n+1}$. Assuming that μ_n has a density, we can consider the optimal transport map T, such that $\mathrm{T}_{\#}\mu_n = \mu_{n+1}$. Then, the continuous curve $(\varrho_t)_{t\in[0,1]}$, defined as $\varrho_t = ((1-t)\mathrm{Id} + t\mathrm{T})_{\#}\mu_n$ is an interpolation (called the *displacement interpolation*). In a geometric sense, this interpolation is the minimizing geodesic connecting μ_n and μ_{n+1}, which is reminiscent of (but not the same as) the interpolation used in Chapter 3. The catch, however, is that this interpolation requires the knowledge of the optimal transport map, which is hard to guess given that the update in (LRM) involves noise and bias. Moreover, using this interpolation followed by a Riemannian argument in the Wasserstein space turns out to be very complicated and needs careful assessment of regularity. $\lozenge$

4.3.2. Flows in the Wasserstein Space

The ODE method for stochastic approximation requires identifying the mean dynamics (or the flow) associated with the algorithm (LRM). Recall that in the Euclidean case, the usual way to obtain this ODE is to "average away" all the noise and tend the step-size to zero. This can be done in this case by eliminating all the noise, as well as the Gaussians ξ_n. What we end up with is the ordinary differential equation $dX_t = v(X_t)\,dt$ defined on $\mathbb{R}^d$. While this ODE captures some behavioral aspects of the iterates in (LRM), it does not offer a complete dynamical picture. Specifically, one can show that the iterates of (LRM) do not form an asymptotic pseudo-trajectory for the flow of this ODE. The reason is that the order of the noise is roughly $\sqrt{\gamma_n}$ rather than γ_n, due to the presence of the Gaussian noise.

As our primary focus in this chapter is the asymptotic behavior of the law of $\boldsymbol{x}_n$ (rather than $\boldsymbol{x}_n$ as a point in $\mathbb{R}^d$), we demonstrate below that the SDE (4.9) indeed induces a *flow* within the Wasserstein space, thereby enabling the application of the ODE method.

With the aid of the Fokker–Planck equation (see Section 4.2.2), we can define the *flow* corresponding to the SDE (4.9) as the function $\Phi : [0, \infty) \times \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ with

$$\Phi_0 = \mathrm{Id}, \quad \partial_t \Phi_t(\varrho) = -\nabla \cdot (\Phi_t(\varrho)\, v) + \frac{1}{2}\langle \nabla^2, \Phi_t(\varrho)\, G \rangle. \tag{4.11}$$

In other words, $\Phi_t(\varrho)$ is the solution of the Fokker-Planck equation with initial datum ϱ at time t.

Working within the space of probability measures has its own intricacies, and one usually prefers doing the analysis at the level of random variables instead. It turns out that we only need some strong solution to the SDE (4.9) whose distribution evolves as what the flow prescribes. This process will have enough information to allow us giving a result at the level of distributions.

Using the same Brownian motion $(W_t)_{t \geq 0}$ used for interpolating the iterates $\{\boldsymbol{x}_n\}_{n \in \mathbb{N}}$, we construct the flow process as follows: For a fixed $t \geq 0$, let $(W_h^{(t)})_{h \geq 0}$ be the Brownian motion "restarted at t," i.e., $W_h^{(t)} = W_{t+h} - W_t$, and define the *flow process* $\phi^{(t)}$ starting at time t to be the strong solution of the SDE (4.9) started at X_t. In other words,

$$\phi_h^{(t)} = X_t + \int_0^h v(\phi_s^{(t)})\, ds + \int_0^h \sigma(\phi_s^{(t)})\, dW_s^{(t)}. \tag{4.12}$$

The flow process $\phi^{(t)}$ has the property that $\mathrm{law}(\phi_h^{(t)}) = \Phi_h(\mathrm{law}(X_t))$. Notice that by using the same Brownian motion, the interpolation and the flow processes become synchronously coupled. This property will be useful later, when we consider the difference of the interpolation and the flow process. If it is clear from the context, we drop the (t) from the flow process and simply write ϕ.

4.4. DYNAMICS OF LANGEVIN–ROBBINS–MONRO SCHEMES

We are now ready to state the main dynamical result of this chapter. What we show is that if the iterates of (LRM) have uniformly bounded second moments, that is, if $\sup_{n\geq 0} \mathbb{E}\,\|\boldsymbol{x}_n\|^2 < \infty$, then, under some assumptions, the law of the stochastic interpolation becomes an asymptotic pseudo-trajectory of the flow induced by the SDE (4.9). We see later that uniformly bounded second moments is a form of stability of the algorithm, and prove that it is implied by dissipativity-type conditions on the drift.

Let us state our blanket assumptions that underlie the rest of this chapter. These are as follows:

▷ **Assumption 4.2.** *The drift v is L_v-Lipschitz and the diffusion matrix σ is L_σ-Lipschitz in Frobenius norm, i.e., $\|\sigma(x) - \sigma(y)\|_{\mathrm{F}} \leq L_\sigma \|x - y\|$.*

▷ **Assumption 4.3.** *The Robbins–Monro summability conditions hold, i.e.,*

$$\sum \gamma_n = \infty \quad \text{and} \quad \sum \gamma_n^2 < \infty. \tag{4.13}$$

▷ **Assumption 4.4.** *The error terms Z_{n+1} satisfy Assumption 4.1. Moreover, there exists a decomposition $Z_{n+1} = U_{n+1} + B_{n+1}$ of the error into noise and bias terms, such that*

$$\mathbb{E}[U_{n+1} \,|\, \mathcal{F}_{\tau_n -}] = 0 \quad \text{and} \quad \sup_{n\in\mathbb{N}} \mathbb{E}\,\|U_{n+1}\|^2 < \infty, \tag{4.14}$$

and[3]

$$\mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] \lesssim \gamma_{n+1}^2 \|v(\boldsymbol{x}_n)\|^2 + \gamma_{n+1}. \tag{4.15}$$

Remark 4.1. A few remarks are in order:

(1) Global Lipschitzness of v and σ ensure that the SDE (4.9) has globally defined strong solutions. This is similar to Chapter 3, where we assumed Lipschitzness and completeness of the vector field to have a globally-defined flow. While this assumption can be restrictive in some scenarios, we remark

[3] We write $a \lesssim b$ if there exists some universal constant $C > 0$ such that $a \leq C \cdot b$.

that is a standard assumption in the context of stochastic approximation; see Section 2.5.

(2) The noise condition (4.14) simply states that the noise terms form a martingale difference sequence with bounded second moments. The bias assumption (4.15) prevents the bias B_{n+1} from overpowering $v(\boldsymbol{x}_n)$. ◊

The main theorem of this chapter is:

▶ **Theorem 4.9.** *Suppose that the drift v and diffusion matrix σ are Lipschitz continuous. Consider the sequence of iterates $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ of (LRM)*

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \gamma_{n+1}\{v(\boldsymbol{x}_n) + Z_{n+1}\} + \sqrt{\gamma_{n+1}}\,\sigma(\boldsymbol{x}_n)\,\xi_{n+1},$$

where the step-sizes γ_n satisfy Robbins–Monro conditions (4.13), and the error terms Z_n satisfy Assumption 4.4. Moreover, assume that the second moments of the iterates $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ are uniformly bounded. Then, almost surely, the stochastic interpolation $(X_t)_{t\geq 0}$ of $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ is an asymptotic pseudo-trajectory of the flow of the SDE

$$dY_t = v(Y_t)\,dt + \sigma(Y_t)\,dW_t$$

in the quadratic Wasserstein space.

We prove this theorem in Section 4.4.1 below. An important implication of Theorem 4.9 is almost sure last-iterate convergence in Wasserstein distance for many sampling algorithms. This follows from the structure of the internally chain-transitive sets of the corresponding SDE, the limit-set theorem (Theorem 2.6), and Corollary 4.7. Below, we show that for the Langevin diffusion, the only internally chain-transitive set is the singleton $\{\mu\}$, where μ is the stationary distribution of (4.2). Later, in Lemmas 4.14 and 4.15, we show a similar structure holds for the mirror Langevin diffusion—another SDE that is used for sampling from probability distributions.

▶ **Lemma 4.10.** *Consider the Langevin diffusion*

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t. \tag{4.16}$$

Then, the only internally chain-transitive set for the flow corresponding to this SDE is the singleton $\{\mu \propto e^{-f}\}$.

Proof. Define the functional $V(\cdot) = H(\cdot \mid \mu)$ to be the relative entropy with respect to μ. We show that V is a Lyapunov function for $\{\mu\}$ in the sense of Definition 2.9. Let μ_t be the density of the solution of the Langevin diffusion

(4.16) and observe that

$$\frac{d}{dt}V(\mu_t) = \frac{d}{dt}\int \log\frac{\mu_t(x)}{\mu(x)}\,\mu_t(x)\,dx = \int \log\frac{\mu_t(x)}{\mu(x)}\partial_t\mu_t(x)\,dx.$$

Using the Fokker-Planck equation, we can replace $\partial_t\mu_t(x)$:

$$\begin{aligned}\frac{d}{dt}V(\mu_t) &= \int \log\frac{\mu_t(x)}{\mu(x)}\,\nabla\cdot\big(\mu_t(x)(\nabla f(x)+\nabla\log\mu_t(x))\big)\,dx\\ &= \int \log\frac{\mu_t(x)}{\mu(x)}\,\nabla\cdot\left(\mu_t(x)\nabla\log\frac{\mu_t(x)}{\mu(x)}\right)dx\\ &= -\int\left\|\nabla\log\frac{\mu_t(x)}{\mu(x)}\right\|^2\mu_t(x)\,dx.\end{aligned}$$

The second equality follows from the fact that $\nabla f = -\nabla\log\mu$, and the last equality is due to the integration by parts. The last quantity, called the relative Fisher information, is strictly positive for all measures other than μ. Thus, V is Lyapunov for $\{\mu\}$. Theorem 2.10 then shows that the only point in the internally chain-transitive set of Φ is μ. As a bonus, this also shows the uniqueness of the stationary distribution of (4.16). □

4.4.1. Proof of Theorem 4.9

The proof of Theorem 4.9 employs similar ideas as that of Theorem 3.4 in Chapter 3. We construct the *Picard process*, similar to (3.26) in Chapter 3, to "lie in between" the stochastic interpolation and the flow process. Later, we decompose the distance of the interpolation to the flow process to the sum of their distances to the Picard process. While in Chapter 3, the construction of the Picard process was done by "integrating" the vector field along the interpolation's trajectory, here we construct an adapted process that is coupled to both the flow process and the stochastic interpolation. This coupling allows us to easily bound the distances from the Picard process to the flow process and the stochastic interpolation. The treatment of error terms are also similar to Chapter 3, with the upside that we do not need to deal with different coordinate systems.

Before diving into the proof of Theorem 4.9, let us populate some facts and inequalities that we use throughout the proof.

- Suppose $A \in \mathbb{R}^{d\times k}$ and ξ is a zero-mean Gaussian vector in $\mathbb{R}^k$ with covariance matrix aI_k. Then, $\mathbb{E}\,\|A\xi\|^2 = \mathbb{E}\,\mathrm{tr}(A^\top A\xi\xi^\top) = a\|A\|_{\mathrm{F}}^2$.

- For vectors $v_1,\dots,v_k$, it holds $\|v_1+\dots+v_k\|^2 \le k(\|v_1\|^2+\dots+\|v_k\|^2)$.

- For random vectors X and Y, it holds $\mathbb{E}\langle X, Y\rangle \leq (\mathbb{E}\|X\|^2\,\mathbb{E}\|Y\|^2)^{1/2}$.
- For positive numbers $a_1, \ldots, a_k$, it holds $\sqrt{a_1 + \cdots + a_k} \leq \sqrt{a_1} + \cdots + \sqrt{a_k}$.
- For any integrable vector-valued function a, it follows by Cauchy-Schwarz that
$$\left\| \int_0^t a(s)\, ds \right\|^2 \leq t \int_0^t \|a(s)\|^2\, ds.$$

Proof of Theorem 4.9. Recall the construction of the interpolation (4.10) and the flow process (4.12). Central to our analysis is the *Picard process started at time* t, defined as

$$Y_h^{(t)} = X_t + \int_0^h v(X_{t+s})\, ds + \int_0^h \sigma(X_{t+s})\, dW_s^{(t)}. \tag{4.17}$$

Similar to (4.12), $W_s^{(t)} = W_{t+s} - W_t$ is the Brownian motion restarted at time t. It is important to note that we are reusing the same Brownian motion, making the Picard process adapted and synchronously coupled to both the flow process and the stochastic interpolation. Intuitively, we think of the Picard process as one step of the Picard iteration for successive approximations to solve ODEs. It is thus expected that its trajectory is close to the original interpolation, as well as to that of the flow process, playing the role of a "bridge." In the sequel, if t is known from the context, we only write Y_h instead of $Y_h^{(t)}$.

Let us fix $t, T > 0$, and for $h \in [0, T]$ decompose the distance between the stochastic interpolation and the flow process as

$$\mathbb{E}\|X_{t+h} - \phi_h\|^2 \leq 2\,\mathbb{E}\|Y_h - \phi_h\|^2 + 2\,\mathbb{E}\|X_{t+h} - Y_h\|^2. \tag{4.18}$$

We now bound each term of this decomposition. Throughout the proof, we denote by $L = \max\{L_v, L_\sigma\}$.

The first term controls how close is the Picard process to the flow process. Using Itô isometry (Lemma 4.3), Lipschitzness of v and σ, and $h \leq T$, we have

$$\begin{aligned}
&\mathbb{E}\|Y_h - \phi_h\|^2 \\
&\quad= \mathbb{E}\left\| \int_0^h (v(\phi_s) - v(X_{t+s}))\, ds + \int_0^h (\sigma(\phi_s) - \sigma(X_{t+s}))\, dW_s^{(t)} \right\|^2 \\
&\quad\leq 2h \int_0^h \mathbb{E}[\|v(\phi_s) - v(X_{t+s})\|^2]\, ds + 2\,\mathbb{E}\left[\int_0^h \|\sigma(\phi_s) - \sigma(X_{t+s})\|_{\mathrm{F}}^2\, ds \right] \\
&\quad\leq 2(T+1)L^2 \int_0^h \mathbb{E}[\|\phi_s - X_{t+s}\|^2]\, ds.
\end{aligned} \tag{4.19}$$

The last term on the right-hand side gives us a hint for using Grönwall's lemma later.

Similar to Section 2.5, let us define the continuous-to-discrete counter $m(t) = \sup\{n \geq 1 : \tau_n \leq t\}$. Also, define the continuous-time, piecewise-constant, adapted processes

$$\overline{X}_t = \boldsymbol{x}_n, \quad \overline{\gamma}_t = \gamma_{n+1}, \quad \overline{Z}_t = Z_{n+1}, \quad \forall t \in [\tau_n, \tau_{n+1}).$$

To bound the distance between the Picard process and the stochastic interpolation (i.e., the second term in (4.18)), we can use the piecewise-constant processes above and write

$$X_{t+h} - Y_h = \int_t^{t+h} v(\overline{X}_s) - v(X_s)\, ds + \int_t^{t+h} (\sigma(\overline{X}_s) - \sigma(X_s))\, dW_s + \Delta_Z(t,h),$$

where $\Delta_Z(t,h)$ is the accumulation of error from time t to $t+h$ and is equal to

$$\begin{aligned} &\Delta_Z(t,h) \\ &\quad = -(t-\tau_k)\,\mathbb{E}[Z_{k+1} \,|\, \mathcal{F}_t] + \sum_{i=k}^{n-1} \gamma_{i+1} Z_{i+1} + (t+h-\tau_n)\,\mathbb{E}[Z_{n+1} \,|\, \mathcal{F}_{t+h}], \end{aligned} \tag{4.20}$$

with $k = m(t)$ and $n = m(t+h)$. As one might guess, $\|\Delta_Z(t,h)\|$ eventually becomes negligible as $t \to \infty$, since the step-size becomes small. The next lemma, whose proof can be found in Appendix B.1, confirms this intuition:

Lemma 4.11. *Suppose that Assumptions 4.1–4.4 hold and the iterates have uniformly bounded second moments. Then, for any fixed $T > 0$, it holds*

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \mathbb{E}\,\|\Delta_Z(t,h)\|^2 = 0.$$

Therefore, using the bounds

$$\mathbb{E}\left\|\int_t^{t+h} (v(X_s) - v(\overline{X}_s))\, ds\right\|^2 \leq h \int_t^{t+h} \mathbb{E}\,\|v(X_s) - v(\overline{X}_s)\|^2\, ds$$

and

$$\mathbb{E}\left\|\int_t^{t+h} (\sigma(X_s) - \sigma(\overline{X}_s))\, dW_s\right\| \leq \mathbb{E}\left[\int_t^{t+h} \|\sigma(X_s) - \sigma(\overline{X}_s)\|_{\mathrm{F}}^2\, ds\right],$$

we obtain

$$\mathbb{E}\,\|X_{t+h} - Y_h\|^2 \le 3(h+1)L^2 \int_t^{t+h} \mathbb{E}\,\|X_s - \overline{X}_s\|^2\, ds + 3\,\mathbb{E}\,\|\Delta_Z(t,h)\|^2. \quad (4.21)$$

Thus, it remains to control how far (on average) the stochastic interpolation gets from the iterates during one iteration. Let $n = m(s)$ so that $\overline{X}_s = \boldsymbol{x}_n$. We can then compute

$$\begin{aligned}
&\mathbb{E}\,\|X_s - \boldsymbol{x}_n\|^2 \\
&\quad = \mathbb{E}\big[\|(s-\tau_n)\{v(\boldsymbol{x}_n) + \mathbb{E}[Z_{n+1}\,|\,\mathcal{F}_s]\} + \sigma(\boldsymbol{x}_n)\,(W_s - W_{\tau_n})\|^2\big] \\
&\quad \le 3\gamma_{n+1}^2\,\mathbb{E}\,\|v(\boldsymbol{x}_n)\|^2 + 3\gamma_{n+1}^2\,\mathbb{E}\big[\|\mathbb{E}[Z_{n+1}\,|\,\mathcal{F}_s]\|^2\big] + 3\,\mathbb{E}\,\|\sigma(\boldsymbol{x}_n)\,(W_s - W_{\tau_n})\|^2
\end{aligned}$$

and since conditional expectation is a projection in L^2,

$$\le 3\gamma_{n+1}^2\,\mathbb{E}\,\|v(\boldsymbol{x}_n)\|^2 + 3\gamma_{n+1}^2\,\mathbb{E}\,\|Z_{n+1}\|^2 + 3\,\mathbb{E}\,\|\sigma(\boldsymbol{x}_n)\,(W_s - W_{\tau_n})\|^2,$$

and by Lemma B.1,

$$\le 3\gamma_{n+1}^2\,\mathbb{E}\,\|v(\boldsymbol{x}_n)\|^2 + 3\gamma_{n+1}^2\,\mathbb{E}\,\|Z_{n+1}\|^2 + 3\gamma_{n+1}\,\mathbb{E}\,\|\sigma(\boldsymbol{x}_n)\|_{\mathrm{F}}^2. \quad (4.22)$$

Moreover, as v and σ are Lipschitz and the iterates have bounded second moments, it follows that $\mathbb{E}\,\|v(\boldsymbol{x}_n)\|^2$ and $\mathbb{E}\,\|\sigma(\boldsymbol{x}_n)\|_{\mathrm{F}}^2$ are bounded by constants. This is because

$$\frac{1}{2}\,\mathbb{E}\,\|v(\boldsymbol{x}_n)\|^2 \le \mathbb{E}\,\|v(\boldsymbol{x}_n) - v(0)\|^2 + \|v(0)\|^2 \le L_v^2 \sup_{k\ge 0} \mathbb{E}\,\|\boldsymbol{x}_k\|^2 + \|v(0)\|^2,$$

and similarly for $\mathbb{E}\,\|\sigma(\boldsymbol{x}_n)\|_{\mathrm{F}}^2$. Thus, by Assumption 4.4 on the error terms,

$$\mathbb{E}\,\|Z_{n+1}\|^2 \le 2\,\mathbb{E}\,\|U_{n+1}\|^2 + 2\,\mathbb{E}\,\|B_{n+1}\|^2 \lesssim \gamma_{n+1}^2\,\mathbb{E}\,\|v(\boldsymbol{x}_n)\|^2 + \gamma_{n+1} + O(1).$$

Plugging this estimate into (4.22) shows that $\mathbb{E}\,\|X_s - \overline{X}_s\|^2 \le C\,\overline{\gamma}_s$ for some constant C not depending on s. Hence, continuing from (4.21), we have

$$\begin{aligned}
\mathbb{E}\,\|X_{t+h} - Y_h\|^2 &\le 3(h+1)L^2 C \int_t^{t+h} \overline{\gamma}_s\, ds + 3\,\mathbb{E}\,\|\Delta_Z(t,h)\|^2 \\
&\le 3(h+1)L^2 Ch \sup_{s\in[t,t+h]} \overline{\gamma}_s + 3\,\mathbb{E}\,\|\Delta_Z(t,h)\|^2 \\
&\le 3(T+1)^2 L^2 C \sup_{s\in[t,t+T]} \overline{\gamma}_s + 3 \sup_{s\in[0,T]} \mathbb{E}\,\|\Delta_Z(t,s)\|^2.
\end{aligned}$$

Taking supremum over $h \in [0, T]$ and noticing that the right-hand side is independent of h, together with Lemma 4.11 yields

$$A_t \coloneqq \sup_{0 \le h \le T} \mathbb{E}\, \|X_{t+h} - Y_h\|^2 \lesssim \sup_{s \in [t, t+T]} \overline{\gamma}_s + \sup_{h \in [0,T]} \mathbb{E}\, \|\Delta_Z(t, h)\|^2, \tag{4.23}$$

implying $A_t \to 0$ almost surely as $t \to \infty$. Therefore, the Picard process gets arbitrary close to the original interpolation.

Let us return to the decomposition (4.18). By taking expectation and using (4.19) and (4.23), an application of Grönwall's lemma gives

$$\begin{aligned} \mathbb{E}\, \|X_{t+h} - \phi_h\|^2 &\le 2(T+1)L^2 \int_0^h \mathbb{E}\, \|X_{t+s} - \phi_s\|^2 \, ds + 2A_t \\ &\le 2A_t \exp\bigl(s(T+1)L^2\bigr) \\ &\le 2A_t \exp(T(T+1)L^2), \end{aligned}$$

Thus,

$$\lim_{t \to \infty} \sup_{h \in [0,T]} \mathbb{E}\, \|X_{t+h} - \phi_h\|^2 = 0.$$

Recall that the quadratic Wasserstein distance between (the laws of) X_{t+h} and ϕ_h is the infimum of expected distance squared over the set of all of their couplings. As ϕ_h has the same marginal as the Langevin diffusion started from X_t at time h and the synchronous coupling of the interpolation and the flow process induces a specific coupling between them, we directly get

$$W_2(X_{t+h}, \phi_h) \le \mathbb{E}[\|X_{t+h} - \phi_h\|^2]^{\frac{1}{2}},$$

which implies

$$\lim_{t \to \infty} \sup_{s \in [0,T]} W_2(\operatorname{law}(X_{t+h}), \Phi_h(\operatorname{law}(X_t))) = 0. \qquad \square$$

4.5. STABILITY VIA DISSIPATIVITY

Theorem 4.9 and Corollary 4.7 in tandem show that under Assumptions 4.1–4.4 and uniformly bounded second moments, the desirable last-iterate convergence of a Langevin–Robbins–Monro scheme in any $(2-\varepsilon)$-Wasserstein space is immediately attained. Therefore, in this section, we turn our focus to establishing the bounded moment condition for Langevin–Robbins–Monro schemes.

There is a long history on conditions that ensure bounded moments for iterative algorithms, which has culminated in the so-called dissipativity properties. We consider one such example below.

Definition 4.12. The drift v is called (α, β)-*dissipative* for some constants $\alpha > 0$ and $\beta \geq 0$, if for all $x \in \mathbb{R}^d$,

$$\langle x, v(x) \rangle \leq -\alpha \|x\|^2 + \beta. \tag{4.24}$$

It is well established that dissipativity ensures the boundedness of the second moments for the basic Euler–Maruyama discretization of Langevin dynamics, whether employing deterministic or stochastic gradient oracles [Hal88; MT93; RT96a; LP02; Lem05; TTV16; RRT17]. However, these studies do not address the broader class of Langevin–Robbins–Monro schemes, particularly when a non-zero bias is present. As we will demonstrate in Section 4.6, this bias is vital for integrating more sophisticated discretization schemes. To this end, our next result shows that for a wide class of Langevin–Robbins–Monro schemes, the moment bounds essentially come for free under dissipativity.

Theorem 4.13. *Let v be an (α, β)-dissipative Lipschitz drift and σ be a Lipschitz diffusion matrix; let $\{\boldsymbol{x}_n\}_{n \in \mathbb{N}}$ be the iterates of a Langevin–Robbins–Monro scheme. Assume that $\lim_{n\to\infty} \gamma_n = 0$ and the bias satisfies the condition (4.15). Then either of the following conditions imply $\sup_n \mathbb{E}\, \|\boldsymbol{x}_n\|^2 < \infty$:*

(i) The Lipschitz constant L_σ of σ satisfies $L_\sigma^2 < \alpha$.

(ii) The diffusion coefficient σ is bounded in Frobenius norm.

At a first glance, condition (i) in Theorem 4.13 seems artificial. It turns out, however, that for SDEs with multiplicative noise, a bound on L_σ is necessary for contraction. We defer this discussion to Remark 4.2 in Section 4.6.

Proof of Theorem 4.13. For brevity, let us write $\mathcal{F}_n$ instead of $\mathcal{F}_{\tau_n}$ and $\mathbb{E}_n[\cdot]$ instead of $\mathbb{E}[\cdot \,|\, \mathcal{F}_n]$. Expanding $\boldsymbol{x}_{n+1}$ as

$$\|\boldsymbol{x}_{n+1}\|^2 = \|\boldsymbol{x}_n + \gamma_{n+1}\{v(\boldsymbol{x}_n) + Z_{n+1}\} + \sqrt{\gamma_{n+1}}\sigma(\boldsymbol{x}_n)\,\xi_{n+1}\|^2$$

and ignoring every term that has zero mean under $\mathbb{E}_n[\cdot]$, we get

$$\begin{aligned}\mathbb{E}_n[\|\boldsymbol{x}_{n+1}\|^2] = \|\boldsymbol{x}_n\|^2 &+ 2\gamma_{n+1}\,\mathbb{E}_n[\langle \boldsymbol{x}_n, v(\boldsymbol{x}_n) + B_{n+1}\rangle] \\ &+ \gamma_{n+1}^2\,\mathbb{E}_n[\|v(\boldsymbol{x}_n) + Z_{n+1}\|^2] + \gamma_{n+1}\,\mathbb{E}_n[\|\sigma(\boldsymbol{x}_n)\,\xi_{n+1}\|^2] \\ &+ 2\gamma_{n+1}^{3/2}\,\mathbb{E}_n[\langle \sigma(\boldsymbol{x}_n)\,\xi_{n+1}, B_{n+1}\rangle].\end{aligned}$$

Note that we used the fact that U_{n+1} is independent of ξ_{n+1} given $\mathcal{F}_n$.

Let us now focus on case (i), where σ is assumed to satisfy $L_\sigma < \alpha$. By repeatedly using the bias condition (4.15), the fact that

$$\mathbb{E}_n[\|B_{n+1}\|] \leq (\mathbb{E}_n[\|B_{n+1}\|^2])^{1/2} \lesssim \gamma_{n+1}\|v(\boldsymbol{x}_n)\| + \sqrt{\gamma_{n+1}},$$

the dissipativity of v, Lipschitzness of v in the sense that $\|v(x)\| \leq L_v\|x\| + \|v(0)\|$ and similarly for σ, and the Cauchy-Schwarz inequality, we obtain the following recursive bound on $a_n := \mathbb{E}\,\|\boldsymbol{x}_n\|^2$:

$$a_{n+1} \leq \left(1 - 2\gamma_{n+1}(\alpha - L_\sigma^2) + o(\gamma_{n+1})\right) a_n + O(\gamma_{n+1})\sqrt{a_n} + O(\gamma_{n+1}). \tag{4.25}$$

We now show that there is some positive $S > 0$ and $n_0 \in \mathbb{N}$ such that $a_n \leq S$ for all $n \geq n_0$. Let us rewrite the inequality above as

$$a_{n+1} \leq a_n(1 - e_n) + g_n\sqrt{a_n} + h_n.$$

For n_0 large enough, it holds for all $n \geq n_0$ that

$$c_1\gamma_{n+1} \leq e_n \leq C_1\gamma_{n+1}, \quad g_n \leq C_2\gamma_{n+1}, \quad h_n \leq C_3\gamma_{n+1}. \tag{4.26}$$

Now observe that it is sufficient for S to satisfy

$$(1 - e_n)S + g_n\sqrt{S} + h_n \leq S,$$

or equivalently, $-e_nS + g_n\sqrt{S} + h_n \leq 0$. As the left-hand side is a quadratic equation in terms of $\sqrt{S}$ with negative leading coefficient, it suffices to have a uniform (in n) upper bound on its larger root; taking S larger than that root gives us the desired upper bound on a_n. The larger root computes as

$$\frac{g_n + \sqrt{g_n^2 + 4h_ne_n}}{2e_n} \leq \frac{C_2 + \sqrt{C_2^2 + 4C_1C_3}}{2c_1},$$

where we used the bounds (4.26). Taking S larger than the right-hand side of the inequality above satisfies our desiderata.

For case (ii), where σ is assumed to be bounded, by a similar computation that leads to (4.25) in case (i), we can derive the recursion

$$a_{n+1} \leq (1 - 2\gamma_{n+1}\alpha + o(\gamma_{n+1}))\, a_n + O(\gamma_{n+1})\sqrt{a_n} + O(\gamma_{n+1}).$$

The rest of the argument is the same as in case (i). □

Remark. With the techniques developed by Durmus and Moulines [DM17], we are able to prove boundedness of the second moments of the iterates under milder conditions using specific Foster–Lyapunov drift conditions. Specifically, for the sampling problem via Langevin diffusion or dual mirror Langevin diffusion, if we assume that the drift satisfies *weak dissipativity*, in the sense that for some $\alpha, \beta > 0$ and $0 < \kappa \le 1$,

$$\langle x, v(x) \rangle \le -\alpha \|x\|^{1+\kappa} + \beta, \quad \forall x \in \mathbb{R}^d,$$

and if the noises involved are sub-Gaussian, we can get the desired boundedness property. As opposed to dissipativity, which requires quadratic growth of f outside a compact set when $v = -\nabla f$, weak dissipativity only entails super-linear growth and therefore, is considerably weaker. While the analysis for this case is interesting, it requires different machinery from those we discussed in this chapter; therefore, we choose not to mention them in this thesis, and refer the interested reader to Theorem 4 and its proof in [KHK23a]. $\Diamond$

4.6. SAMPLING ALGORITHMS

The generality of the Langevin–Robbins–Monro template allows us to capture many existing algorithms used for sampling from probability distributions, and suggests ways to design new ones. In the following subsections, we showcase instances of (LRM) that are typically used for sampling in practice.

Let $\mu \propto e^{-f}$ be the target distribution. Similar to Chapter 3, we will discuss algorithms that operate with indirect access to the gradient ∇f through a stochastic first-order oracle. In essence, when an SFO is invoked at a point $x \in \mathbb{R}^d$, given a random seed ω from a set of seeds Ω, it produces a random vector $\tilde{\nabla} f(x; \omega)$ that is of the form

$$\tilde{\nabla} f(x; \omega) = \nabla f(x) + U(x; \omega), \tag{4.27}$$

where the noise term $U(x; \omega)$ is assumed to be zero-mean and have bounded second moments:

$$\mathbb{E}_\omega[U(x; \omega)] = 0 \quad \text{and} \quad \mathbb{E}_\omega[\|U(x; \omega)\|^2] \le C_U, \quad \forall x \in \mathbb{R}^d. \tag{4.28}$$

4.6.1. Basic Discretizations

Similar to the stochastic gradient descent algorithm for optimization, the classic *Stochastic Gradient Langevin Dynamics* [WT11] uses the vanilla Euler–Maruyama discretization of (4.2), but employs stochastic gradients. This algorithm is mostly used in Bayesian learning and sampling from posterior distributions, as the log-likelihood function is in the form of a sum over all data points in a dataset; stochastic gradients are obtained by selecting a subset of data points uniformly at random in each iteration, leading to the algorithm

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_{n+1}\tilde{\nabla} f(\boldsymbol{x}_n) + \sqrt{2\gamma_{n+1}}\,\xi_{n+1}, \tag{SGLD}$$

where $\tilde{\nabla} f$ is the gradient of the negative log-likelihood of a random batch of the data. Setting $U_{n+1} \coloneqq \tilde{\nabla} f(\boldsymbol{x}_n) - \nabla f(\boldsymbol{x}_n)$ and $B_{n+1} = 0$ makes (SGLD) the first (and simplest) example of an Langevin–Robbins–Monro scheme.

If (SGLD) is the forward Euler–Maruyama method for discretizing (4.2), the *Proximal Langevin Algorithm* [Wib19] would be the backward one:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_{n+1}\nabla f(\boldsymbol{x}_{n+1}) + \sqrt{2\gamma_{n+1}}\,\xi_{n+1}. \tag{PLA}$$

In practice, this method is useful if one can solve (PLA) for $\boldsymbol{x}_{n+1}$; the proximal operator should be easy to compute. Setting $B_{n+1} \coloneqq \nabla f(\boldsymbol{x}_{n+1}) - \nabla f(\boldsymbol{x}_n)$, we see that this algorithm also follows the Langevin–Robbins–Monro template. Lemma B.2 in Appendix B.2 shows that the bias of (PLA) satisfies the bias condition (4.15).

4.6.2. Randomized Mid-point Method

The *Randomized Mid-point Method* [SL19] is an alternative discretization scheme to Euler–Maruyama and has been proposed for both underdamped and overdamped Langevin diffusions. Let us describe this method for the overdamped Langevin diffusion (4.2); in doing so we closely follow the argument of [HBE20].

Suppose the algorithm starts at the point $\boldsymbol{x}_0$. Having the step-size γ in mind, we explicitly solve the (overdamped) Langevin diffusion (4.2) initialized at $\boldsymbol{x}_0$ for γ amount of time and arrive at the point $\boldsymbol{x}^*(\gamma)$ which satisfies

$$\boldsymbol{x}^*(\gamma) = \boldsymbol{x}_0 - \int_0^\gamma \nabla f(\boldsymbol{x}^*(s))\,ds + \sqrt{2}\,W_\gamma. \tag{4.29}$$

Surely we cannot explicitly compute the integral above, and our goal is to estimate it. For this, we look at the integral from 0 to γ as an *expected value* of a uniformly distributed random variable and approximate it using one sample. Concretely, we let α to be a random variable uniformly distributed in $[0,1]$, independent of everything else, and use the unbiased estimate

$$\int_0^\gamma \nabla f(\boldsymbol{x}^*(s))\,ds = \gamma\,\mathbb{E}_\alpha[\nabla f(\boldsymbol{x}^*(\gamma\alpha))] \approx \gamma\nabla f(\boldsymbol{x}^*(\alpha\gamma))$$

as an approximation of the integral. The Randomized Mid-point method then proceeds by estimating $\boldsymbol{x}^*(\alpha\gamma)$ by the Euler–Maruyama discretization:

$$\boldsymbol{x}^*(\alpha\gamma) \approx \boldsymbol{x}_0 - \alpha\gamma\nabla f(\boldsymbol{x}_0) + \sqrt{2}\, W_{\alpha\gamma}.$$

It is important to mention that we have used the same Brownian motion as (4.29). By properties of the Brownian motion, given α, the covariance of W_γ and $W_{\alpha\gamma}$ is $\alpha\gamma I$. Therefore, writing $W_\gamma =: \sqrt{\gamma}\,\xi$ and $W_{\alpha\gamma} =: \sqrt{\alpha\gamma}\,\xi'$, it holds that ξ and ξ' are standard Gaussian, and given α, their cross-covariance is $\mathbb{E}[\xi'\xi^\top] = \sqrt{\alpha}\, I$. Using these variables, we can define the Randomized Mid-point Method's iterates

$$\begin{cases} \boldsymbol{x}_{n+1/2} = \boldsymbol{x}_n - \gamma_{n+1}\alpha_{n+1}\tilde{\nabla} f(\boldsymbol{x}_n) + \sqrt{2\gamma_{n+1}\alpha_{n+1}}\,\xi'_{n+1}, \\ \quad \boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_{n+1}\tilde{\nabla} f(\boldsymbol{x}_{n+1/2}) + \sqrt{2\gamma_{n+1}}\,\xi_{n+1}, \end{cases} \tag{RMM}$$

where $\{\alpha_n\}_{n\in\mathbb{N}}$ are independent and uniformly distributed in $[0,1]$, and are independent of everything else, and ξ_{n+1} and ξ'_{n+1} are standard Gaussian random variables with cross-covariance $\sqrt{\alpha_{n+1}} I$. As in the previous examples, we also allow noisy evaluation of the gradients.

To cast (RMM) into the Langevin–Robbins–Monro template, we set

$$\begin{aligned} B_{n+1} &:= \nabla f(\boldsymbol{x}_{n+1/2}) - \nabla f(\boldsymbol{x}_n), \quad \text{and} \\ U_{n+1} &:= \tilde{\nabla} f(\boldsymbol{x}_{n+1/2}) - \nabla f(\boldsymbol{x}_{n+1/2}). \end{aligned}$$

Lemma B.3 in Appendix B.2 shows that the bias of (RMM) satisfies the bias condition (4.15).

Inspecting the update rule of (RMM), we see that it requires *two* gradient oracle calls at each iteration. Inspired by the *optimistic gradient methods* in optimization and online learning, we propose to “recycle” the past gradients:

$$\begin{cases} \boldsymbol{x}_{n+1/2} = \boldsymbol{x}_n - \gamma_{n+1}\alpha_{n+1}\tilde{\nabla} f(\boldsymbol{x}_{n-1/2}) + \sqrt{2\gamma_{n+1}\alpha_{n+1}}\,\xi'_{n+1} \\ \quad \boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_{n+1}\tilde{\nabla} f(\boldsymbol{x}_{n+1/2}) + \sqrt{2\gamma_{n+1}}\,\xi_{n+1}, \end{cases} \tag{ORMM}$$

where α_{n+1}, ξ_{n+1}, ξ'_{n+1}, and $\tilde{\nabla} f$ are the same as in (RMM). This algorithm again falls into the category of Langevin–Robbins–Monro schemes with the same noise and bias as (RMM). Lemma B.4 in Appendix B.2 shows that the bias of (RMM) satisfies the bias condition (4.15). Let us remark that (ORMM) requires *one* gradient oracle, thereby halving the per-iteration cost of (RMM). To our knowledge, the scheme (ORMM) is new.

4.6.3. Stochastic Runge–Kutta Method

In addition to the simple (stochastic) Euler–Maruyama discretization, there is a class of more sophisticated discretization methods of (4.2) known as higher-order integrators. The *Stochastic Runge–Kutta method* [Li+20] is an example of an order 1.5 integrator, with iterates

$$\begin{cases} h_1 = \boldsymbol{x}_n + \sqrt{2\gamma_{n+1}}\left((\frac{1}{2} + \frac{1}{\sqrt{6}})\,\xi_{n+1} + \frac{1}{\sqrt{12}}\,\xi'_{n+1}\right) \\ h_2 = \boldsymbol{x}_n - \gamma_{n+1}\tilde{\nabla} f(\boldsymbol{x}_n) + \sqrt{2\gamma_{n+1}}\left((\frac{1}{2} - \frac{1}{\sqrt{6}})\,\xi_{n+1} + \frac{1}{\sqrt{12}}\,\xi'_{n+1}\right), \\ \boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{1}{2}\gamma_{n+1}(\tilde{\nabla} f(h_1) + \tilde{\nabla} f(h_2)) + \sqrt{2\gamma_{n+1}}\,\xi_{n+1}, \end{cases} \tag{SRK}$$

where ξ_{n+1} and ξ'_{n+1} are independent standard Gaussian random variables. This algorithm is a Langevin–Robbins–Monro scheme with

$$\begin{aligned} B_{n+1} &\coloneqq \tfrac{1}{2}(\nabla f(h_1) + \nabla f(h_2)) - \nabla f(\boldsymbol{x}_n), \quad \text{and} \\ U_{n+1} &\coloneqq \tfrac{1}{2}(\tilde{\nabla} f(h_1) - \nabla f(h_1)) + \tfrac{1}{2}(\tilde{\nabla} f(h_2) - \nabla f(h_2)). \end{aligned}$$

Lemma B.5 in Appendix B.2 shows that the bias of (RMM) satisfies the bias condition (4.15).

Finally, similar to (ORMM), we remark that one can recycle the past gradients of the Stochastic Runge–Kutta method to save oracle calls at each iteration. Since the idea is entirely the same, we omit the details.

4.6.4. Mirror Langevin Algorithm

The *Mirror Langevin algorithm* [Hsi+18; Zha+20; AC21], which is the sampling analogue of the celebrated mirror descent scheme in optimization [NJ83; BT03], is an example of using an SDE different from the Langevin diffusion for the task of sampling. This algorithm uses a strictly convex function φ, known as the Bregman potential, to change the Euclidean geometry to a favorable local geometry. This can be useful for sampling from bounded domains (where the support of the target distribution is not the whole $\mathbb{R}^d$), or in cases where a local change of variables improves the conditioning of the problem. In short, using the gradient of φ, one maps the original *primal* space into the *dual* space, and performs the desired operations (such as gradient steps or adding a Gaussian noise) in the dual. Below, we make this precise. A more general and complete picture of mirror descent is given in the next chapter, where we discuss a generic mirror descent scheme in the space of measures.

In the dual space, the iterates of the Mirror Langevin algorithm are

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_{n+1}\nabla f(\nabla\varphi^\star(\boldsymbol{x}_n)) + \sqrt{2\gamma_{n+1}}(\nabla^2\varphi^\star(\boldsymbol{x}_n)^{-1})^{1/2}\,\xi_{n+1}, \tag{ML}$$

where $\varphi^\star$ is the *Fenchel conjugate* of φ (see [Roc97] or (5.4) in Chapter 5 for a definition). This iteration is the Euler–Maruyama discretization of the *mirror Langevin diffusion*

$$dX_t = -\nabla f(\nabla\varphi^\star(X_t))\,dt + \sqrt{2}\,(\nabla^2\varphi^\star(X_t)^{-1})^{1/2}\,dW_t. \tag{4.30}$$

In our framework, (ML) fits into (LRM) by taking $v = -\nabla f \circ \nabla\varphi^\star$ and $\sigma = (\nabla^2\varphi^\star)^{-1/2}$, making it an example with state-dependent diffusion matrix.

It is worthwhile to translate our assumptions on the drift v and diffusion σ into conditions on f and σ. As it turns out, these conditions are the same as (or weaker than) those mentioned in the mirror Langevin literature [Li+22].

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an open convex set and suppose that $\varphi : \mathcal{X} \to \mathbb{R}^d$ is a twice-differentiable strictly convex function, so that $\nabla\varphi$ is a bijection between $\mathcal{X}$ and $\mathbb{R}^d$, and $\|\nabla\varphi(x)\| \to \infty$ as x approaches the boundary of $\mathcal{X}$. In this case, it holds for all $x \in \mathbb{R}^d$ that

$$\nabla\varphi(\nabla\varphi^\star(x)) = x, \quad \text{and} \tag{4.31}$$

$$(\nabla^2\varphi^\star)^{-1}(\nabla\varphi(x)) = \nabla^2\varphi(x). \tag{4.32}$$

(1) The L_v-Lipschitzness of the drift v corresponds to f being *L_v-smooth relative to φ* [Li+22, (A2)]:

$$\|\nabla f(x) - \nabla f(y)\| \le L_v\|\nabla\varphi(x) - \nabla\varphi(y)\|$$

This simply follows from (4.31) and surjectivity of $\nabla\varphi$; letting $x' = \nabla\varphi(x)$ gives

$$-v(x') = \nabla f(\nabla\varphi^\star(x')) = \nabla f(\nabla\varphi^\star(\nabla\varphi(x))) = \nabla f(x).$$

Invoking Lipschitzness of v gives the desired inequality.

(2) The L_σ-Lipschitzness of σ in Frobenius norm corresponds to *modified self-concordance with parameter L_σ^2* [Li+22, (A1)]:

$$\|\nabla^2\varphi(x)^{1/2} - \nabla^2\varphi(y)^{1/2}\|_{\mathrm{F}} \le L_\sigma\|\nabla\varphi(x) - \nabla\varphi(y)\|. \tag{4.33}$$

This follows from (4.32) and surjectivity of $\nabla\varphi$ similar to the previous item.

(3) As far as the knowledge of the author goes, α-dissipativity of the drift v does not correspond to a previously known condition on f and φ in the mirror

Langevin literature. However, dissipativity is implied by (and therefore, is weaker than) *α-strong convexity of f with respect to φ* [Li+22, (A3)]; that is, if

$$\langle \nabla f(x) - \nabla f(y), \nabla\varphi(x) - \nabla\varphi(y)\rangle \geq \alpha\|\nabla\varphi(x) - \nabla\varphi(y)\|^2. \tag{4.34}$$

Using (4.31), this condition can be equivalently written as

$$\langle v(x) - v(y), x - y\rangle \leq -\alpha\|x - y\|^2.$$

To see why this condition implies dissipativity, suppose that $v(x^*) = 0$ for some $x^* \in \mathbb{R}^d$; such a point exists, since $\|\nabla f(x)\| \to \infty$ as $x \to \partial\mathcal{X}$.

$$\langle v(x) - v(x^*), x - x^*\rangle = \langle v(x), x - x^*\rangle \leq -\alpha\|x - x^*\|^2.$$

Thus,

$$\langle v(x), x\rangle \leq -\alpha\|x\|^2 + 2\alpha\langle x, x^*\rangle - \alpha\|x^*\|^2 + \langle x^*, v(x)\rangle.$$

As v is Lipschitz, the right-hand side is $\leq -\alpha\|x\|^2 + g\|x\| + h$, for some $g, h \geq 0$. It is then evident that there is some $\alpha', \beta > 0$ such that the right-hand side is $\leq -\alpha'\|x\|^2 + \beta$; in other words, v is (α', β)-dissipative.

Thus far, we have shown that the mirror Langevin algorithm (ML) fits into the Langevin–Robbins–Monro scheme, and that we can interpret the assumptions on the drift and diffusion solely in terms of the target potential f and the Bregman potential φ. For the limits of the algorithm, however, we have to identify the internally chain-transitive sets of the flow corresponding to mirror Langevin diffusion (4.30). However, since the iterations of (ML) and the SDE evolve in the dual space, its stationary distribution $\tilde{\mu}$ is going to be the pushforward of the target measure $\mu \propto e^{-f}$ under the mirror map $\nabla\varphi$, that is, $\tilde{\mu} := (\nabla\varphi)_{\#}\mu$ [Li+22, Sec. 2.3]. In Lemma 4.14 below we show that the dual mirror Langevin diffusion indeed has $\tilde{\mu}$ as its stationary measure and that $\{\tilde{\mu}\}$ is the only internally chain-transitive set for the corresponding flow. Next, in Lemma 4.15 we show a similar result, but for the primal mirror Langevin diffusion, showing that its stationary distribution is μ and $\{\mu\}$ is its only internally chain-transitive set.

▶ **Lemma 4.14.** *The only internally chain-transitive set for the flow corresponding to the Mirror Langevin diffusion in its dual form* (4.30) *is the singleton* $\{\tilde{\mu}\}$, *where* $\tilde{\mu} = (\nabla\varphi)_{\#}\mu$.

Proof. Before proving this lemma, let us find the density of $\tilde{\mu}$. Using the change of variables formula, we can obtain [see Li+22, Sec. 2.3]

$$\log\tilde{\mu}(x) = -f(\nabla\varphi^\star(x)) - \log\det(\nabla^2\varphi^\star)^{-1}(x) + \text{constant}.$$

Therefore,

$$\nabla \log \tilde{\mu}(x) = -(\nabla^2 \varphi^\star)(x) \nabla f(\nabla \varphi^\star(x)) - \nabla^2 \varphi^\star(x) \nabla \cdot (\nabla^2 \varphi^\star)^{-1}(x). \tag{4.35}$$

For mirror Langevin diffusion, it is more convenient to work with the χ^2 divergence. Recall that for two probability densities ϱ and ν, the χ^2 divergence of ϱ from ν is defined as

$$\chi^2(\varrho \,|\, \nu) = \int \frac{\varrho(x)^2}{\nu(x)} \, dx - 1.$$

Define the functional $V(\cdot) = \chi^2(\cdot \,|\, \tilde{\mu})$. Similar to the case for Langevin diffusion, we show that V is a Lyapunov function for $\{\tilde{\mu}\}$ in the sense of Definition 2.9. Let μ_t be the density of the solution of the mirror Langevin diffusion (4.30). Then μ_t satisfies the Fokker–Planck equation $\partial_t \mu_t = -\nabla \cdot (\mu_t w_t)$ with

$$w_t = -\nabla f \circ \nabla \varphi^\star - (\nabla^2 \varphi^\star)^{-1} \nabla \log \mu_t - \nabla \cdot (\nabla^2 \varphi^\star)^{-1}.$$

Comparing with (4.35), we see that

$$w_t = (\nabla^2 \varphi^\star)^{-1} \nabla \log \tilde{\mu} - (\nabla^2 \varphi^\star)^{-1} \nabla \log \mu_t = -(\nabla^2 \varphi^\star)^{-1} \nabla \log \frac{\mu_t}{\tilde{\mu}}.$$

Therefore, for $\mu_t \neq \tilde{\mu}$, it holds

$$\begin{aligned} \frac{d}{dt} \chi^2(\mu_t \,|\, \tilde{\mu}) = \frac{d}{dt} \int \frac{\mu_t(x)^2}{\tilde{\mu}(x)} \, dx &= 2 \int \frac{\mu_t(x)}{\tilde{\mu}(x)} \partial_t \mu_t(x) \, dx \\ &= -2 \int \left\langle \nabla \log \frac{\mu_t(x)}{\mu(x)}, (\nabla^2 \varphi^\star)^{-1} \nabla \log \frac{\mu_t}{\tilde{\mu}} \right\rangle < 0, \end{aligned}$$

where we used integration by parts, and the inequality is due to the fact that φ is strictly convex. A similar use of Theorem 2.10 shows that V is a Lyapunov function for $\{\tilde{\mu}\}$, as well as that $\tilde{\mu}$ is the unique stationary distribution of the dual mirror Langevin dynamics. □

For illustrative purposes, let us do a similar proof for the mirror Langevin diffusion in the primal form. Firstly, this diffusion can be obtained by Itô's formula for the process $Y_t = \nabla \varphi^\star(X_t)$, where X_t follows the dual mirror Langevin diffusion; see also [Wib19, App. A].

▶ **Lemma 4.15.** *Consider the Mirror Langevin diffusion in its primal form:*

$$dY_t = (\nabla \cdot (\nabla^2\varphi(Y_t)^{-1}) - \nabla^2\varphi(Y_t)^{-1}\nabla f(Y_t))\,dt \\ + \sqrt{2}\,(\nabla^2\varphi(Y_t)^{-1})^{1/2}\,dW_t. \quad (4.36)$$

Then, the only internally chain-transitive set for the flow corresponding to this SDE is the singleton $\{\mu\}$, where $\mu \propto e^{-f}$.

Proof. Define the functional $V(\cdot) = \chi^2(\cdot \,|\, \mu)$. Similar to Lemma 4.14, we show that V is a Lyapunov function for $\{\mu\}$. Let μ_t be the density of the solution of the mirror Langevin diffusion (4.36). Then, μ_t satisfies the Fokker-Planck equation $\partial_t\mu_t = -\nabla \cdot (\mu_t w_t)$ with

$$\begin{aligned} w_t &= \nabla \cdot (\nabla^2\varphi)^{-1} - (\nabla^2\varphi)^{-1}\nabla f - (\nabla^2\varphi)^{-1}\nabla \log \mu_t - \nabla \cdot (\nabla^2\varphi)^{-1} \\ &= -(\nabla^2\varphi)^{-1}\nabla f - (\nabla^2\varphi)^{-1}\nabla \log \mu_t \\ &= -(\nabla^2\varphi)^{-1}\nabla \log \frac{\mu_t}{\mu}. \end{aligned}$$

See also [Wib19, Lem. 3] for a similar derivation. Therefore, for $\mu_t \neq \mu$, it holds

$$\begin{aligned} \frac{d}{dt}\chi^2(\mu_t \,|\, \mu) = \frac{d}{dt}\int \frac{\mu_t(x)^2}{\mu(x)}\,dx = 2\int \frac{\mu_t(x)}{\mu(x)}\partial_t\mu_t(x)\,dx \\ = -2\int \left\langle \nabla \log \frac{\mu_t(x)}{\mu(x)}, (\nabla^2\varphi)^{-1}\nabla \log \frac{\mu_t}{\mu} \right\rangle < 0, \end{aligned}$$

where we used integration by parts, and the inequality is due to the fact that φ is strictly convex. Similar to Lemma 4.14 we get the desired result. □

We finish this section with the following remark on the Lipschitz constant of the diffusion matrix which we postponed in the discussion of our assumptions in Theorem 4.13 for the stability theorem.

Remark 4.2. In the context of Mirror Langevin algorithm, we remark that the condition (i) in Theorem 4.13 (namely, $L_\sigma^2 < \alpha$) is merely asserting a positive contraction rate for mirror Langevin diffusion. Contraction is a powerful tool used in proving convergence (as well as obtaining non-asymptotic mixing times) of sampling algorithms. We say the SDE

$$dX_t = v(X_t)\,dt + \sigma(X_t)\,dW_t$$

is *contractive* with rate $r > 0$ if, for any two strong solutions X_t and X'_t that are synchronously coupled (i.e., they share the same Brownian motion), there is a

$t_0 > 0$ such that

$$\mathbb{E}\,\|X_t - X_t'\|^2 \leq e^{-2rt}\,\mathbb{E}\,\|X_0 - X_0'\|^2, \quad \forall t \in (0, t_0).$$

If an SDE is contractive, it has a stationary distribution. It turns out [see Li+22, Lem. 4] that the Mirror Langevin Diffusion in dual form (4.30) is contractive with rate $\alpha - L_\sigma^2$, where we recall that α is the dissipativity constant of the drift (or the strong-convexity parameter of f with respect to φ as in (4.34)) and L_σ is the Lipschitz constant of the diffusion matrix (or the square root of the modified self-concordance parameter of φ as in (4.33)). In general, a bound on L_σ is necessary for an SDE with multiplicative noise to contract; see [Li+22, App. C] for a worked-out example regarding the Geometric Brownian motion. ♢

4.7. CONCLUSIONS

In this chapter, we introduced a new, unified framework for analyzing a wide range of discretization schemes for SDEs, and in particular, sampling algorithms based on the Langevin diffusion, thus laying the theoretical ground for using them in practice, as well as motivating new and more efficient algorithms that enjoy rigorous guarantees. We built on the ideas from dynamical system theory present in the previous chapters, and gave a rather complete picture of the asymptotic behavior of many first-order discretization algorithms. In short, our results help with the following:

(1) **Validating existing methods:** Methods like mirror Langevin and randomized mid-point method currently lack even asymptotic guarantees in fully non-convex scenarios, such as sampling from neural network-defined distributions. Our work fills this gap by offering the first rigorous justification for these schemes, supporting practitioners in utilizing these methods confidently.

(2) **Facilitating new algorithm design:** Our work motivates novel sampling methods through a straightforward verification of Assumptions 4.1–4.4. An illustrative instance involves the randomized mid-point method and the Runge–Kutta integrators, wherein a substantial 50% reduction in computation per iteration can be achieved without compromising convergence by simply recycling past gradients, shown in (ORMM). The balance between the benefits of saving gradient oracles and potential drawbacks remains an open question, necessitating case-by-case practical evaluation. Nevertheless,

our theory provides a flexible algorithmic design template that extends beyond the current literature's scope.

While our asymptotic pseudo-trajectory result holds under very mild conditions, a severe limitation of our current framework is that it only applies to algorithms based on discretizing SDEs, whereas there are numerous practical sampling schemes, such as Metropolis–Hastings, that are not immediately linked to an SDE. Lifting such constraint is an interesting future work.

Another significant direction for future research is the analysis of discretizations with constant step-size. In practice, many of the algorithms discussed in this chapter utilize a constant step-size due to its robustness and potential for faster initial convergence, particularly with larger step-sizes. Such an approach is especially beneficial in data-scarce scenarios where maximizing the use of available data is critical, and small increments are less effective.

A promising theoretical approach for constant step-size is the averaging of iterates. By employing larger than usual gains and relying on offline averaging to mitigate the increased noise from the larger step-size, substantial overall improvement can be achieved. However, the tools and analysis required for this approach differ from those presented in this chapter. Therefore, we defer this intriguing line of research to the future.

BIBLIOGRAPHIC NOTES

The field of structured non-convex sampling is extensive. A *structured non-convex problem* is one that involves additional assumptions on the target density, such as strong convexity outside a ball or specific functional inequalities like the log-Sobolev or Poincaré inequalities. Under these conditions, researchers have derived non-asymptotic rates for sampling algorithms that are based on the Langevin diffusion [see, e.g., RRT17; Che+18; Xu+18; Li+19; VW19; ZXG19; MMS20; Che+21; Ma+21; Mou+22]. This chapter focuses on generic Langevin-type SDE discretization, with sampling as a special case. For the case of sampling, we essentially work with a generic non-convex sampling problem, which is NP-hard and its convergence is asymptotic at best.

Key related works [LP02; TTV16; BBC17; DM17; Bal+22] focus on the asymptotic convergence of sampling algorithms based on the Langevin diffusion, with minimal regularity assumptions on the potential function f. In comparison, our results either improve upon or are distinct from these studies, as discussed below.

Guarantees for discretization schemes. Lamberton and Pages [LP02] and Lemaire [Lem05] analyze the Euler–Maruyama discretization of the Langevin diffusion (4.2) with deterministic gradients (i.e., $Z_{n+1} \equiv 0$), proving weak convergence of the average iterates under a moment condition that is slightly weaker than bounded second moments. Although the moment condition in [LP02; Lem05] is stated in a weaker form than boundedness of the second moments, it is typically only verified on a special case that is equivalent to dissipativity, and thus implies the required boundedness of the moments [see e.g., LP02, Rem. 3]. Their analysis is further extended by Teh, Thiery, and Vollmer [TTV16] to incorporate stochastic gradients. Later, the last-iterate convergence of the simple Euler–Maruyama discretization of (4.2) is studied by Durmus and Moulines [DM17], who prove the convergence in the total variation distance under weak dissipativity. Another work on a similar setting as [DM17] is by Benaïm, Bouguet, and Cloez [BBC17], where the convergence criterion is given in terms of an integral probability metric of the form

$$d_{\mathcal{B}}(\mu, \nu) \coloneqq \sup_{\psi \in \mathcal{B}} |\mathbb{E}_{\mu}[\psi] - \mathbb{E}_{\nu}[\psi]|$$

for a certain class of test functions $\mathcal{B}$ that is known to imply weak convergence, but not convergence in total variation or Wasserstein distances.

Compared to these results, our guarantees possess the following desirable features:

- The convergence is always on the last iterates instead of the average iterates.
- As we tolerate biased algorithms, the class of discretization schemes we consider is significantly more general than the ones in existing work.

Finally, we note that our results are incomparable to the recent work of Balasubramanian et al. [Bal+22], who derive the same result as in [LP02; Lem05], i.e., average-iterate, weak convergence of deterministic Euler–Maruyama discretization. A remarkable feature of the analysis in [Bal+22] is that it does not require any bounded moments, and in particular, their bounds can be applied to target distributions with unbounded variance. However, the downside of [Bal+22] is that, in the presence of noise and bias, their analysis produces a bound that does not vanish as $n \to \infty$ [see Bal+22, Thm. 15]. In contrast, our framework can tolerate quite general noises and biases, and gives stronger asymptotic guarantees (Wasserstein vs. weak convergence; last-iterate vs. average-iterate).

Mean-Squared-Error analysis. A powerful framework for quantifying the global discretization error of a numerical algorithm is the mean-squared-error analysis framework [MT04]. This framework furnishes a general recipe for controlling short- and long-term integration errors. For sampling, this framework has been applied to prove convergence rates for Langevin Monte-Carlo (i.e., the Euler–Maruyama discretization of the Langevin diffusion (4.2)) in the strongly convex setting [Li+20; LZT22]. Similar to our work, the convergence obtained in these works is last-iterate and in Wasserstein distance. One of the essential ingredients in the latter work is the contraction property of the SDE, which is ensured by the strong convexity assumption. This, in turn, implies strong non-asymptotic convergence guarantees.

Sampling as optimization. One of the main themes in proving error bounds for sampling is the natural relation between sampling and optimization in the Wasserstein space. This point of view, when applied to strongly convex potentials, has produced numerous non-asymptotic guarantees; see [DK19; Che23] for a recent account and the references therein. Note that strong convexity is crucial for the analysis used in the aforementioned work. Moreover, the error bounds for biased and noisy discretizations do not decrease with the step-size or iteration count; see [DK19, Thm. 4, Eqn. (14)]. This means that while the bound is non-asymptotic, it does not automatically result in an asymptotic convergence. Finally, we stress that these approaches are orthogonal to our techniques: We view a sampling algorithm as a (noisy and biased) discretization of a dynamical system (and not necessarily a gradient flow), and use tools from dynamical system theory to provide asymptotic convergence results.

While, to our knowledge, our framework is significantly different from previous works on sampling, we acknowledge that similar ideas of creating an auxiliary process in-between the iterates and the continuous-time flow is not entirely new and has been touched upon in the literature [see, e.g., BEL18; Cha+21]. That being said, our specific approach in building the Picard process and its development into a wider array of algorithms, i.e., Langevin–Robbins–Monro schemes, undoubtedly plays a pivotal role in our analysis. Moreover, the integration of the Picard process with the theory of asymptotic pseudo-trajectories offers dual benefits to our study, and we view these as our unique contributions to this area of research.

CHAPTER FIVE

STOCHASTIC APPROXIMATION FOR ENTROPIC OPTIMAL TRANSPORT

In the previous chapters, we looked at stochastic approximation algorithms where the corresponding continuous-time flow was clear from the algorithm and straightforward to construct. For example, in Chapter 3, we examined algorithms that find the root of a vector field V on a Riemannian manifold, which the V itself created a flow on the manifold, allowing us to analyze the algorithm's behavior using stochastic approximation. Similarly, in Chapter 4, we used the Fokker–Planck equation of an SDE to construct a flow in Wasserstein space.

However, some practical algorithms do not come with an obvious continuous-time flow, which is essential for using methods like the ODE method or similar stochastic approximation techniques. In this chapter, we will explore two such algorithms: the Sinkhorn algorithm for solving Entropic Optimal Transport problems, and the iterative proportional fitting procedure for solving Schrödinger Bridge problems. These algorithms are widely used in practice, usually with noisy and incomplete data. To really understand how these algorithms work, we need a thorough stochastic approximation analysis.

Originality. Main results of this chapter are published in the conference proceedings [KHK24]. There are considerable differences, in notation and content, between this chapter and the mentioned publication.

LIST OF IMPORTANT RESULTS

- **Lemma 5.6.** A recursive formulation of the step-sized Sinkhorn iteration, similar to the original Sinkhorn algorithm in its primal form.
- **Lemma 5.7.** The dual formulation of the step-sized Sinkhorn iteration, giving an update rule for the Schrödinger potentials.
- **Proposition 5.9.** The dual step-sized Sinkhorn iteration is a dual mirror descent iteration.
- **Proposition 5.10.** The infinitesimal behavior of the step-sized Sinkhorn iteration in terms of both the coupling and the Schrödinger potentials. This leads to the definition of the Sinkhorn and dual Sinkhorn flows.
- **Theorem 5.12.** Convergence rate of $O(t^{-1})$ for the Sinkhorn flow and its dual flow in continuous-time.
- **Theorem 5.13.** Convergence of the step-sized Sinkhorn algorithm under noisy evaluations of the gradients. The proof is similar to (but not the same as) the one for stochastic mirror descent.
- **Theorem 5.14.** The iterates of the step-sized Sinkhorn algorithm using noisy and biased gradient evaluations form an asymptotic pseudo-trajectory of the Sinkhorn flow. Precompactness of the iterates (here, the Schrödinger potentials) ensure last-iterate convergence.
- **Theorem 5.20.** The explicit drift formula for the step-sized IPF iteration.

5.1. INTRODUCTION

Many contemporary challenges in machine learning can be reframed as an *entropic optimal transport* (EOT) problem in the space of probability measures. EOT involves adding an entropy regularization term to the classic *optimal transport* (OT) problem. This regularization not only stabilizes computational algorithms but also produces smoother and more interpretable transport plans.

One prominent application of EOT is the *Schrödinger bridge* (SB) problem, which aims to dynamically transform one probability measure into another over time. This concept has shown to be particularly useful in various fields that require an understanding of complex continuous-time stochastic systems.

In machine learning, EOT has found widespread applications. EOT is used to improve generative modeling algorithms [GPC17]. The addition of entropy regularization mitigates the issue of overfitting to noisy datasets and facilitates the learning of smoother probability distributions. This makes models like Generative Adversarial Networks and Variational Autoencoders more effective and robust. It is also used for aligning distributions from different domains to improve model generalizability [Cou+16].

Beyond machine learning, EOT has significant applications in the sciences. In molecular biology, it aids in modeling the dynamical behavior of molecular systems, providing insights into state transitions in complex biochemical processes. For example, EOT can model how proteins fold or how molecular reactions proceed over time, offering valuable information for drug development and biochemical research. In the study of single-cell dynamics, EOT helps in understanding the progression and differentiation of cells over time. By examining how probability distributions of cell states evolve, researchers can map out developmental pathways and identify critical regulatory mechanisms [Sch+19; Bun+23].

Traditionally, the entropic optimal transport problem is solved in practice using the Sinkhorn algorithm. This algorithm has been primarily viewed as an alternating projection method and have been extensively studied from this perspective. However, recent work has reinterpreted the Sinkhorn algorithm for discrete probability distributions as a mirror descent scheme—a well-known optimization algorithm—thereby providing new insights into understanding the Sinkhorn algorithm through classical optimization theory. This perspective has also been extended to continuous probability measures. These results show that Sinkhorn iterations can be regarded as mirror descent steps, specifically with step-size 1.

Building on this perspective, this chapter introduces a novel step-sized, as well as a continuous-time, variant of the Sinkhorn algorithm on the space of probability measures. The two main objectives are: First, to deepen the understanding of

Sinkhorn iterates by demonstrating that convergence relies on the choice of the mirror map, objective function, and constraints, rather than the previously emphasized step-size 1. Using stochastic mirror descent analysis, this new algorithm maintains convergence even in the presence of noise. Second, to employ the ODE method and perform stochastic approximation analysis for solving entropic optimal transport with noisy and biased data.

We frame our findings within the Schrödinger bridge context and provide a mirror descent interpretation of the Iterative Proportional Fitting procedure—a sibling of Sinkhorn algorithm commonly used for solving the Schrödinger bridge problem in the machine learning community. Additionally, we demonstrate that these new mirror descent iterations can be described via stochastic differential equations with explicit drift formulas. These contributions collectively advance both theoretical understanding and practical methodologies in the application of mirror descent and Sinkhorn algorithms to machine learning and related fields.

Outline of the Chapter

The mirror descent algorithm plays a pivotal role in this chapter; we review its essential ideas in Section 5.2. This section is intended for a reader without background knowledge in mirror descent.

In the previous chapter, we looked at the space of probability measures as a complete metric space by using the Wasserstein distance. This chapter, however, we work with the linear structure of the space of (signed) measures, and most importantly, convexity. In Section 5.3, we review the basics of (infinite-dimensional) topological vector spaces, dual pairs, and bits of convex analysis. We specifically bring essential properties of the relative entropy functional (or the Kullback–Leibler divergence), which plays a central role in this chapter.

In Section 5.4 we introduce the entropic optimal transport problem, some of the important properties of its optimal solution, as well as the Sinkhorn algorithm, a method to find the optimal solution.

In Section 5.5, a step-sized variant of the Sinkhorn algorithm is derived based on the discrete-time mirror descent scheme. This variant includes original Sinkhorn iterations as a special case when the step-size is set to be 1. Letting the step-sizes to zero, we derive a continuous-time flow in Section 5.6, which comprises *primal* and *dual Sinkhorn flows*. We prove convergence of these continuous-time flows to the optimal solution of EOT and derive a rate of convergence.

Section 5.7 is all about the convergence of the introduced schemes. We show strong asymptotic and non-asymptotic guarantees for the discrete-time schemes based on stochastic mirror descent analysis and stochastic approximation analysis.

We finish this chapter with mentioning the Schrödinger bridge problem in Section 5.8 and deliver a mirror descent interpretation of the iterative proportional fitting procedure. Further to this, we establish that the new mirror descent iterations can be expressed via SDEs with explicit drifts.

A short review of the related works, as well as extra pointers to the literature is provided in the bibliographic notes section at the end of the chapter.

5.2. MIRROR DESCENT ESSENTIALS

In this section, we revisit the fundamental components of the classical mirror descent scheme of Nemirovsky and Judin [NJ83]. For ease of understanding, we will focus on mirror descent in the Euclidean setting. In the next section, we mention how these concepts extend beyond this scope and show how they can be applied to the infinite-dimensional space of measures.

5.2.1. Minimizing Movement Interpretation

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a convex objective function that is differentiable on its domain, and $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set. The idea of the mirror descent algorithm is to change the underlying Euclidean metric to facilitate solving the constrained minimization problem

$$\text{minimize } F(x) \text{ subject to } x \in \mathcal{C}. \tag{5.1}$$

This change of metric is done via a so-called *Bregman potential*, which is assumed to be a strictly convex differentiable function $\varphi : \mathbb{R}^d \to \mathbb{R}$. Concretely, the *mirror descent* (MD) algorithm produces the iterates $x_0, x_1, x_2, \ldots$, where $x_0 \in \mathcal{C}$ is arbitrary, and

$$x_{n+1} = \operatorname*{arg\,min}_{x \in \mathcal{C}} \left\{ F(x_n) + \langle \nabla F(x_n), x - x_n \rangle + \frac{D_\varphi(x \,|\, x_n)}{\gamma_n} \right\}. \tag{5.2}$$

Here, γ_n is a sequence of step-sizes, and $D_\varphi(\cdot \,|\, \cdot)$ is the *Bregman divergence* associated with φ, defined as

$$D_\varphi(x' \,|\, x) \coloneqq \varphi(x') - \varphi(x) - \langle \nabla\varphi(x), x' - x \rangle. \tag{5.3}$$

The iteration (5.2) has a *minimizing movement* interpretation: At each step, we linearize the objective F at the current iterate and minimize it while staying “close” to the current iterate. The measure of closeness is determined by the Bregman divergence.

5.2.2. Mirror Descent Dual Iterations

Besides the minimizing movement interpretation, a particularly insightful approach for studying MD is through convex duality. Before getting into the details, let us recall the basics of convex duality in Euclidean setting.

Suppose $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is a proper closed convex function. The function f is *proper* if it never attains the value $-\infty$ and its domain $\operatorname{dom} f = \{x : f(x) < +\infty\}$ is nonempty, and it is *closed* if its sublevel sets $\{x : f(x) \leq a\}$ are closed subsets of $\mathbb{R}^d$ for all $a \in \mathbb{R}$. The *Fenchel conjugate* of f is the function $f^\star : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$

$$f^\star(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y\rangle - f(x)\}. \tag{5.4}$$

An important property of the Fenchel conjugate is that it is a proper closed convex function, and that it satisfies the *Fenchel–Young inequality*

$$f^\star(y) + f(x) \geq \langle x, y\rangle, \quad \forall x, y \in \mathbb{R}^d.$$

A vector v is a *subgradient* of f at $x \in \operatorname{dom} f$ if

$$f(y) \geq f(x) + \langle v, y - x\rangle, \quad \forall y \in \mathbb{R}^d.$$

We denote the set of all subgradients of f at x by $\partial f(x)$ and call it the *subdifferential of f at x*. The following theorem shows the importance of the Fenchel–Young inequality in identifying subgradients of both f and its Fenchel conjugate $f^\star$.

Theorem 5.1 (Roc97, Thm. 23.5)**.** *For any proper convex function f and any vector x, the following conditions on a vector y are equivalent to each other:*

(a) $y \in \partial f(x)$;

(b) $f(x) + f^\star(y) \leq \langle x, y\rangle$;

(c) $f(x) + f^\star(y) = \langle x, y\rangle$.

Moreover, if f is closed, the conditions above are equivalent to

(d) $x \in \partial f^\star(y)$.

As the optimization problem (5.1) that we are considering is constrained, it will be useful to define a "constrained version" of the Bregman potential φ. For this, let $\mathrm{I}_{\mathcal{C}}$ be the *convex indicator function* of the set $\mathcal{C}$, defined as

$$\mathrm{I}_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C}, \\ +\infty & \text{if } x \notin \mathcal{C}, \end{cases}$$

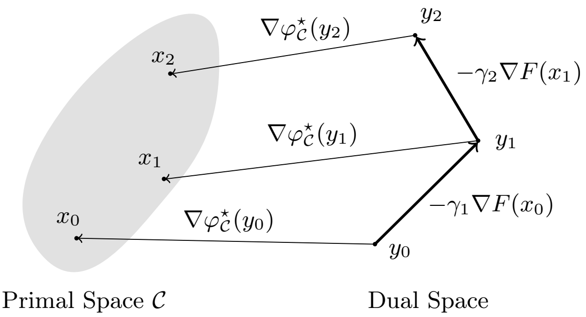

$\nabla\varphi_{\mathcal{C}}^{\star}(y_2)$
y_2
x_2
$-\gamma_2\nabla F(x_1)$
$\nabla\varphi_{\mathcal{C}}^{\star}(y_1)$
y_1
x_1
$-\gamma_1\nabla F(x_0)$
x_0
$\nabla\varphi_{\mathcal{C}}^{\star}(y_0)$
y_0
Primal Space $\mathcal{C}$
Dual Space

Figure 5.1. Mirror Descent dual iteration (5.7). The function $\nabla\varphi$, known as the *mirror map*, links the primal space $\mathcal{C}$ to the dual space, and $\nabla\varphi_{\mathcal{C}}^{\star}$, known as the *dual mirror map*, is the link from the dual to the primal space.

and define the constrained version of φ as $\varphi_{\mathcal{C}} := \varphi + \mathrm{I}_{\mathcal{C}}$. The Fenchel conjugate of $\varphi_{\mathcal{C}}$ satisfies

$$\varphi_{\mathcal{C}}^{\star}(y) = \sup_{x\in\mathbb{R}^d}\{\langle x, y\rangle - (\varphi + \mathrm{I}_{\mathcal{C}})(x)\} = \sup_{x\in\mathcal{C}}\{\langle x, y\rangle - \varphi(x)\}, \tag{5.5}$$

which is strictly convex because φ is strictly convex. Therefore, $\varphi_{\mathcal{C}}^{\star}$ is essentially differentiable[1] [Roc97, Thm. 26.3], and by Danskin's theorem [Dan67], its gradient is

$$\nabla\varphi_{\mathcal{C}}^{\star}(y) = \arg\max_{x\in\mathcal{C}}\{\langle x, y\rangle - \varphi(x)\}, \quad \forall y \in \operatorname{int}(\operatorname{dom}\varphi_{\mathcal{C}}^{\star}). \tag{5.6}$$

This is one of the key components that makes it possible to define the dual mirror descent iteration.

After this short digression, let us go back to the study of mirror descent. We shall call any $x \in \mathcal{C}$ a *primal point* and any $y \in \operatorname{dom}(\nabla\varphi_{\mathcal{C}}^{\star})$ a *dual point*. As it will shortly become clear, $\nabla\varphi_{\mathcal{C}}^{\star}$ and $\nabla\varphi$ act as links between the primal space $\mathcal{C}$ and dual space $\operatorname{dom}(\nabla\varphi_{\mathcal{C}}^{\star})$. Using these links, we derive an equivalent iteration in the dual space, made precise below.

Let $x_0 \in \mathcal{C}$ and $y_0 \in \operatorname{dom}(\nabla\varphi_{\mathcal{C}}^{\star})$ be a pair of primal and dual points satisfying $x_0 = \nabla\varphi_{\mathcal{C}}^{\star}(y_0)$. One can take, for example, $x_0 \in \mathcal{C}$ to be arbitrary and let $y_0 = \nabla\varphi(x_0)$; see (5.8) in Remark 5.1 below. The following recursion is called the

[1] A proper convex function f is *essentially differentiable* if the interior of its domain $\mathcal{D} := \operatorname{int}(\operatorname{dom} f)$ satisfies the conditions (a) $\mathcal{D}$ is nonempty, (b) f is differentiable in $\mathcal{D}$, (c) and for any sequence $x_1, x_2, \ldots$ converging to the boundary of $\mathcal{D}$, it holds $\lim_{n\to\infty}\|\nabla f(x_n)\| = +\infty$.

mirror descent dual iterations:

$$\begin{cases} y_{n+1} = y_n - \gamma_n \nabla F(x_n), \\ x_{n+1} = \nabla \varphi^\star_{\mathcal{C}}(y_{n+1}). \end{cases} \tag{5.7}$$

Let us spell out what this iteration does: First, the gradient of the objective F is evaluated at the current primal iterate x_n. This gradient is then used to update the current dual iterate y_n to y_{n+1}; this makes sense, as gradients essentially live in the dual space.[2] Finally, to obtain the next primal iterate x_{n+1}, one maps y_{n+1} back to $\mathcal{C}$ via the link $\nabla \varphi^\star_{\mathcal{C}}$, and therefore, closing the loop. Moreover, one can forget about the primal variables altogether and rewrite the dual iteration (5.7) solely in terms of the dual variables:

$$y_{n+1} = y_n - \gamma_n \nabla F(\nabla \varphi^\star_{\mathcal{C}}(y_n)).$$

Although the dual MD update seems simpler compared to the primal update (5.2), one can show that under certain conditions, the primal and dual iterations coincide; see [BT03]. Figure 5.1 illustrates the first iterations of the dual mirror descent update (5.7).

▷ **Example 5.1.** As a warm-up for the rest of this chapter, let us instantiate the theory developed above in a concrete and well-known example. We will see that the ubiquitous exponential weights algorithm is an instance of the mirror descent algorithm.

Suppose $\mathcal{C}$ is the $(d-1)$-dimensional probability simplex in $\mathbb{R}^d$, i.e.,

$$\mathcal{C} = \{x \in \mathbb{R}^d : x^1 + \cdots + x^d = 1 \text{ and } x^i \geq 0 \ \ \forall i = 1, \ldots, d\}.$$

Define the Bregman potential φ as $\varphi(x) = \sum_{i=1}^d x^i \log x^i$. Computing the Fenchel conjugate of the constrained version $\varphi^\star_{\mathcal{C}}$ is rather straightforward:

$$\varphi^\star_{\mathcal{C}}(y) = \sup_{x \in \mathcal{C}} \{\langle x, y \rangle - \varphi(x)\} = \log \sum_{i=1}^d \exp(y^i).$$

Moreover, the maximizer of the above maximization is

$$\nabla \varphi^\star_{\mathcal{C}}(y) = \left(\frac{\exp(y^1)}{Z}, \ldots, \frac{\exp(y^d)}{Z} \right), \quad Z = \sum_{i=1}^d \exp(y^i).$$

[2] $\nabla F(x_n)$ is the Riesz representative of the derivative of F at x_n, i.e., $DF(x_n)[v] = \langle \nabla F(x_n), v \rangle$ for all primal vectors v.

Let F be an arbitrary differentiable function we wish to minimize over the simplex and whose domain includes the simplex. Letting $g_n = \nabla F(x_n)$, the mirror descent dual update (5.7) updates the ith coordinate of x_n as $x^i_{n+1} \propto \exp(y^i_{n+1}) = \exp(y^i_n - \gamma_n g^i_n) \propto x^i_n \cdot \exp(-\gamma_n g^i_n)$, which means

$$x_{n+1} \propto x_n \odot \exp(-\gamma_n g_n) \propto x_0 \odot \exp(-\textstyle\sum_{k=1}^n \gamma_k g_k).$$

Here, $\odot$ is elementwise multiplication of vectors. This update is exactly the *exponential weights* algorithm: One keeps a weight vector (a vector of positive numbers), and updates it at each iteration by elementwise multiplication with the exponential of the loss vector $-\gamma \nabla F$. After re-normalizing, one obtains the current primal iterate on the simplex. $\triangleleft$

Remark 5.1. A dual point y uniquely identifies a primal point $x = \nabla \varphi^\star_{\mathcal{C}}(y)$, but several dual points might be associated to the same primal point x (i.e., the map $\nabla \varphi^\star_{\mathcal{C}}$ might not be injective). It is, however, always the case that $\nabla \varphi(x)$ is associated with x. To see this, notice that (5.6) implies

$$\nabla \varphi^\star_{\mathcal{C}}(\nabla \varphi(x)) = \arg\max_{x' \in \mathcal{C}} \{\langle x', \nabla \varphi(x)\rangle - \varphi(x')\} = \arg\min_{x' \in \mathcal{C}} D_\varphi(x' \,|\, x) = x. \tag{5.8}$$

In other words, $\nabla \varphi^\star$ is the left inverse of $\nabla \varphi$ on $\mathcal{C}$. An example for the case where $\nabla \varphi^\star_{\mathcal{C}}$ is not injective is the one in Example 5.1: For any $a \neq 0$ and $y \in \mathbb{R}^d$, it holds that $\nabla \varphi^\star_{\mathcal{C}}(y + a\mathbf{1}) = \nabla \varphi^\star_{\mathcal{C}}(y)$. $\diamondsuit$

In this section we discussed the mirror descent algorithm on Euclidean spaces for pedagogical reasons. Indeed, the majority of the concepts discussed above can also be directly applied to the space of measures, albeit with additional considerations. Specifically, one must treat differentiability with caution, as all the objective functions and Bregman potentials we will be considering in this chapter are not differentiable in the usual sense. This primarily has to do with topological nuances, given that the domain of these functions has an empty interior. Section 5.3 below explains how to deal with these issues and highlights the parts of the discussion above that are different from the Euclidean case.

5.3. CONVEX ANALYSIS IN THE SPACE OF MEASURES

In this section, we bring necessary elements from convex analysis in infinite dimensional spaces and specifically, topological vector spaces. This exposition is merely to bring all the necessary material in one place, and is not at all a reference to learn this subject. The reader is encouraged to consult a functional analysis book with a focus on convexity, such as the excellent book of Aliprantis and Border [AB06] or Attouch, Buttazzo, and Michaille [ABM14]. It is assumed that the reader is acquainted with basic notions of topology.

5.3.1. Dual Pairs and the Weak Topology

A *dual pair* of spaces is a pair $\langle X, X' \rangle$ of vector spaces together with a bilinear functional $(x, x') \mapsto \langle x, x' \rangle$ from $X \times X'$ to $\mathbb{R}$ that separates the points of X and X'; that is, if $\langle x, x' \rangle = 0$ for all $x' \in X'$, then $x = 0$, and similarly, if $\langle x, x' \rangle = 0$ for all $x \in X$, then $x' = 0$. The bilinear map $\langle \cdot, \cdot \rangle$ is sometimes called the *duality* of the pairing. In a dual pair, each space can be interpreted as a space of linear functionals on the other. For example, for each $x \in X$ we can associate the linear functional $x' \mapsto \langle x, x' \rangle$. Therefore, X can be identified with a vector subspace of $\mathbb{R}^{X'}$ and inherits its product topology.[3] We refer to this topology as the *weak topology* on X and denote it by $\sigma(X, X')$. Therefore, a sequence $\{x_n\} \subset X$ converges weakly (i.e., in the weak topology) to $x \in X$ if and only if $\langle x_n, x' \rangle \to \langle x, x' \rangle$ in $\mathbb{R}$ for all $x' \in X'$. An important property of dual pairs is that the topological dual[4] of $(X, \sigma(X, X'))$ is X'.

We now construct the main dual pair used in this chapter, namely $\langle L^1, L^\infty \rangle$. Fix a compact set $\mathcal{X} \subset \mathbb{R}^d$ and a finite, regular, positive measure μ on $\mathcal{X}$. Let $\mathcal{M}$ be the set of all finite (signed) measures on $\mathcal{X}$ that are absolutely continuous with respect to μ. By the Radon–Nikodym theorem, this space is nothing other than $L^1(\mu)$, in the sense that every measure $\nu \in \mathcal{M}$ is identified with its density with respect to μ:

$$\nu \in \mathcal{M} \longleftrightarrow \frac{d\nu}{d\mu} \in L^1(\mu).$$

Consider the dual pair $\langle L^1(\mu), L^\infty(\mu) \rangle$ with the duality

$$\langle f, g \rangle := \int_{\mathcal{X}} f g \, d\mu \quad \forall f \in L^1(\mu), g \in L^\infty(\mu).$$

This duality induces the weak topology $\sigma(L^1, L^\infty)$ on $L^1(\mu)$ which in the sequel we refer to as the weak topology. Note that with the identification $\mathcal{M} \cong L^1(\mu)$,

[3] Recall that the *product topology* on the space $\mathbb{R}^X$ of real-valued functions on X is the weakest topology that evaluation functions are continuous. This topology is also known as topology of pointwise convergence.

[4] The *topological dual* of a topological vector space X is the space of all *continuous* linear functionals on X, where continuity is meant with respect to the topology of X.

the duality above can also be written between measures and functions:

$$\langle \nu, g \rangle = \int \frac{d\nu}{d\mu}\, g\, d\mu = \int g\, d\nu.$$

5.3.2. Convex Functions and Derivatives

We now study convex functions and different notions of derivatives of functions on topological vector spaces. In Section 5.3.3, we instantiate these results on the relative entropy functional.

Let (X, τ) be a topological vector space with topology τ. A function $f : X \to \mathbb{R} \cup \{\pm\infty\}$ is called *τ-lower semi-continuous* (l.s.c.) if its sublevel sets $\{x : f(x) \leq a\}$ are τ-closed for all $a \in \mathbb{R}$. Equivalently, f is τ-l.s.c. if for any sequence (or more precisely, any net) $x_n \to x$, it holds that $\liminf f(x_n) \geq f(x)$. One of the main properties of l.s.c. functions is that *on a τ-compact set, a τ-lower semi-continuous function attains its minimum.* This is a generalization of the Weierstrass theorem to l.s.c. functions.

A function $f : X \to \mathbb{R} \cup \{\pm\infty\}$ is called *proper* if it never assumes the value $-\infty$ and its domain $\operatorname{dom} f = \{x \in X : f(x) < \infty\}$ is nonempty. A remarkable property of proper convex functions and convex sets in a normed vector space is that lower semi-continuity and closeness is the same with respect to the weak or norm topology:

Theorem 5.2 (ABM14, Thms. 3.3.2 and 3.3.3)**.** *A nonempty convex subset of a normed vector space is closed for the norm topology if and only if it is closed for the weak topology. Similarly, a proper convex function on a normed vector space is l.s.c. with respect to the weak topology if and only if it is l.s.c. with respect to the norm topology.*

As the norm topology is stronger than the weak topology (i.e., it has more closed sets), the main usage of this theorem is getting lower semi-continuity (resp. closedness) with respect to the weak topology for free when a convex function (resp. a convex set) is l.s.c. (resp. closed) in the norm topology; something that is usually easier to assess.

For a proper convex function $f : X \to \mathbb{R} \cup \{\pm\infty\}$, let us define the *one-sided directional derivative* at $x \in X$ and in the direction $v \in X$ as

$$d^+ f(x)(v) = \lim_{\lambda \downarrow 0} \frac{f(x + \lambda v) - f(x)}{\lambda} \in \mathbb{R} \cup \{\pm\infty\}.$$

Convexity of f implies that the difference quotients $\frac{f(x+\lambda v) - f(x)}{\lambda}$ are nonincreasing as λ decreases, so the limit above exists (though it might be $-\infty$). Moreover, it can be shown that the function $v \mapsto d^+ f(x)(v)$ is a positively homogeneous convex function. If the sublinear function $d^+ f(x)(\cdot)$ is actually linear (and finite valued), it is called the *Gâteaux derivative* of f at x, denoted by $\nabla f(x)$.

The directional derivative is closely related to subgradients of f, which we shall discuss next. In a dual pair $\langle X, X' \rangle$, we say $x' \in X'$ is a *subgradient* of a proper convex function f at the point x, if

$$f(y) \geq f(x) + \langle y - x, x' \rangle \quad \text{for all } y \in X.$$

The *subdifferential of f at x* is the set of all subgradients of f at x and is denoted by $\partial f(x)$. One can show [see AB06, Thms. 7.15-7.17] that $x' \in X'$ is a subgradient of f at x if and only if $x'(\cdot) \leq d^+ f(x)(\cdot)$. Moreover, the subdifferential of f at x is a singleton if and only if $d^+ f(x)(\cdot)$ is the Gâteaux derivative of f at x.

5.3.3. The Relative Entropy

We now instantiate the notions introduced above for the relative entropy functional. In what follows, we consider a compact set $\mathcal{X} \subset \mathbb{R}^d$ equipped with a positive regular reference measure μ, and consider the dual pair $\langle L^1(\mu), L^\infty(\mu) \rangle$. Recall that the space of signed measures that are absolutely continuous with respect to μ can be identified with $L^1(\mu)$. Define the *relative entropy* functional $H_\mu : L^1(\mu) \to \mathbb{R} \cup \{+\infty\}$ as

$$H_\mu(f) := \int_{\mathcal{X}} \{f(x) \log f(x) + 1 - f(x)\} \, d\mu(x) \tag{5.9}$$

if the integral is defined and otherwise, $H_\mu(f) := +\infty$. Note that if both μ and the measure $d\nu = f \, d\mu$ corresponding to f are probability measures, the definition above boils down to

$$H_\mu(f) = H(\nu \,|\, \mu) = \int_{\mathcal{X}} \log \frac{d\nu}{d\mu} \, d\nu, \tag{5.10}$$

and is called the *Kullback–Leibler divergence*.

It is well known that H_μ is a proper strictly convex function and is non-negative. Another important property of the relative entropy is that it is weakly l.s.c. and its level sets are weakly compact.

Theorem 5.3 (Egg93, Lem. 2.1, Cor. 2.2, Lem. 2.3)**.** *The relative entropy H_μ is lower semi-continuous in the weak topology of $L^1(\mu)$; that is, its sublevel sets $\{f \in L^1(\mu) : f \geq 0, H_\mu(f) \leq a\}$ are weakly closed subsets of $L^1(\mu)$ for all $a \geq 0$. Moreover, these sublevel sets are convex and weakly compact.*

We can also compute the directional derivative of the relative entropy. In special cases, it turns out that this functional has subgradients and is even Gâteaux differentiable. This is rather surprising, as in general, the interior of the domain of the relative entropy with respect to the norm topology of $L^1(\mu)$ is empty. The monotone convergence theorem allows us to compute the directional derivative of H_μ as

$$d^+ H_\mu(f)(v) = \int_{\mathcal{X}} v(x) \log f(x)\, d\mu(x) \tag{5.11}$$

whenever the integral is finite [Res05, Lem. 4.1]. Moreover, $\partial H_\mu(f)$ is nonempty if and only if $f \in L^\infty_+(\mu)$ and is bounded away from zero; in this case,

$$\partial H_\mu(f) = \{\log f\}. \tag{5.12}$$

Denote by $L^\infty_{++}(\mu)$ the set of all non-negative functions $f \in L^\infty_+(\mu)$ that are bounded away from zero. As the subdifferential $\partial H_\mu(f)$ is a singleton, $d^+ H_\mu(f)(\cdot)$ is indeed the Gâteaux derivative of H_μ at f, meaning that it is a linear functional with finite values. Since for $f \in L^\infty_{++}(\mu)$, it holds that $\log f \in L^\infty(\mu)$, (5.11) shows further that and for all $v \in L^1(\mu)$

$$\nabla H_\mu(f)(v) = d^+ H_\mu(f)(v) = \langle v, \log f\rangle, \tag{5.13}$$

making the Gâteaux derivative a continuous linear functional, where continuity is understood in the weak topology on $L^1(\mu)$.

Disintegration and the Chain Rule of Relative Entropy

In the sequel, we also use the relative entropy of a marginal of a joint distribution with respect to some reference measure. Before we mention the corresponding properties, it is helpful to review some measure-theoretic notions that help us work with joint distributions.

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces (complete separable metric spaces). The *disintegration of measure theorem* allows one to write a probability measure on $\mathcal{X} \times \mathcal{Y}$ as an average of probability measures on $\{x\} \times \mathcal{Y}$ for $x \in \mathcal{X}$. In particular, if π is a probability measure on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$-marginal μ, then there exists a measurable map $x \mapsto \pi(\cdot \mid x)$ from $\mathcal{X}$ to $\mathcal{P}(\mathcal{Y})$, uniquely determined $d\mu(x)$-almost everywhere, such that

$$\pi = \int_{\mathcal{X}} (\delta_x \otimes \pi(\cdot \mid x))\, d\mu(x).$$

Here, δ_x is a Dirac mass at x. In other words, for every $f \in C_b(X \times Y)$, it holds

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x,y)\,d\pi(x,y) = \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} f(x,y)\,d\pi(y \mid x)\right] d\mu(x).$$

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are equipped with their Borel σ-algebras. Let π, π' be two Borel probability measures on $\mathcal{X} \times \mathcal{Y}$, and suppose that π is absolutely continuous with respect to π'. By [Léo13b, Thm. 1.6], it holds that the $\mathcal{X}$-marginal of π, denoted by π_x, is also absolutely continuous with respect to the $\mathcal{X}$-marginal of π', and one has the following decomposition of the Radon–Nikodym derivative of π with respect to π':

$$\frac{d\pi}{d\pi'}(x,y) = \frac{d\pi_x}{d\pi'_x}(x)\,\frac{d\pi(\cdot \mid x)}{d\pi'(\cdot \mid x)}(y).$$

Moreover, one has the chain rule of the relative entropy [Léo13a, App. A]:

$$H(\pi \,|\, \pi') = H(\pi_x \,|\, \pi'_x) + \int_{\mathcal{X}} H(\pi(\cdot \mid x) \,|\, \pi'(\cdot \mid x))\,d\pi_x(x) \tag{5.14}$$

Therefore, $H(\pi \,|\, \pi') \geq H(\pi_x \,|\, \pi'_x)$ with equality if and only if $\pi(\cdot \mid x) = \pi'(\cdot \mid x)$ for $d\pi_x(x)$-almost every $x \in \mathcal{X}$.

After this short digression, let us go back to the properties of relative entropy of marginals of joint distributions. Specifically, let $\mathcal{X}$ and $\mathcal{Y}$ be compact sets in $\mathbb{R}^d$ and π be a positive measure on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$-marginal μ. Define the functional $F : L^1(\pi) \to \mathbb{R} \cup \{+\infty\}$ as

$$F(\pi') = H_\mu(\pi'_x).$$

Here, we abused the notation and used the same symbol for the measure π' and its density with respect to π. By a similar computation as in (5.11) and (5.12), we get for all $v \in L^1(\pi)$

$$d^+F(\pi')(v) = \int_{\mathcal{X}\times\mathcal{Y}} v(x,y) \log \frac{d\pi'_x}{d\mu}(x)\,d\pi(x,y) \tag{5.15}$$

if the integral is finite. Moreover, the subdifferential $\partial F(\pi')$ is nonempty if and only if $\pi'_x \in L^\infty_{++}(\mu)$, and in this case,

$$\partial F(\pi') = \left\{\log \frac{d\pi'_x}{d\mu}\right\},$$

which similarly implies that F is Gâteaux differentiable at π', with its Gâteaux derivative being a continuous linear functional on $L^1(\pi)$.

5.4. ENTROPIC OPTIMAL TRANSPORT AND THE SINKHORN ALGORITHM

In this section, we review the central properties of the entropic optimal transport, also known as entropy-regularized optimal transport, and the Sinkhorn algorithm. The results mentioned here are classical, and the interested reader is referred to the book of Peyré and Cuturi [PC20] and the survey of Nutz [Nut22] for further details on the subject.

5.4.1. Entropic Optimal Transport

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of $\mathbb{R}^d$, equipped with probability measures μ and ν, respectively. Consider the product space $\Omega := \mathcal{X} \times \mathcal{Y}$ with the product measure $\mu \otimes \nu$ and consider the space of measures $L^1(\mu \otimes \nu)$ along with the weak topology induced by the duality with $L^\infty(\mu \otimes \nu)$; see Section 5.3 for the related terminology. For brevity, we write $L^1(\Omega)$ instead of $L^1(\mu \otimes \nu)$. For a cost function $c \in L^\infty(\Omega)$ and a regularization parameter $\varepsilon > 0$, the *entropic optimal transport* (EOT) problem is the minimization

$$\mathrm{OT}^\varepsilon(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \left\{ \int_\Omega c\, d\pi + \varepsilon H(\pi \,|\, \mu \otimes \nu) \right\}, \tag{OT^ε}$$

where $\Pi(\mu, \nu) \subset L^1(\Omega)$ is the set of all probability measures on Ω with marginals μ and ν, and H is the relative entropy functional defined in (5.10). One reason for considering such regularized optimization is to solve (OT^ε) for small $\varepsilon > 0$ and obtain an approximation of the (unregularized) optimal transport problem that corresponds to setting $\varepsilon = 0$ in (OT^ε):

$$\mathrm{OT}(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_\Omega c\, d\pi. \tag{OT^0}$$

Besides, the problem (OT^ε) is of its own interest, and has connections to other important problems, such as the Schrödinger bridge problem, which we shall discuss next. See the introduction of this chapter and the bibliographic notes at the end for more applications of the entropic optimal transport problem.

Using the cost c and ε, let us define the reference measure R_ε on Ω as

$$dR_\varepsilon \propto e^{-c/\varepsilon}\, d(\mu \otimes \nu). \tag{5.16}$$

We can then rewrite the objective in (OT^ε) using R_ε as

$$\begin{aligned}\int c\,d\pi + \varepsilon H(\pi\,|\,\mu\otimes\nu) &= \varepsilon\int\left(\frac{c}{\varepsilon} + \log\frac{d\pi}{d(\mu\otimes\nu)}\right)d\pi\\ &= \varepsilon H(\pi\,|\,R_\varepsilon) - \varepsilon\log\int e^{-c/\varepsilon}\,d(\mu\otimes\nu).\end{aligned}$$

As the second term in the last equation is independent of π, (OT^ε) is equivalent to the following optimization problem, known as the *static Schrödinger problem*:

$$\min_{\pi\in\Pi(\mu,\nu)} H(\pi\,|\,R_\varepsilon). \tag{$\mathrm{S}_{\mathrm{static}}$}$$

This problem is referred to as "static" because it seeks a one-shot coupling, π, that transforms μ into ν. In Section 5.8, we explore the Schrödinger bridge problem in greater detail and also introduce a "dynamic" version. Note that the reference measure R_ε encodes all the data about the regularization parameter ε, as well as the cost c. Without loss of generality, we assume in the sequel that the cost c is normalized in such a way that equality holds in (5.16).

It is easy to see that $\Pi(\mu,\nu)$ is a strongly closed, convex subset of $L^1(\Omega)$, and by Theorem 5.2, it is also weakly closed. As H is weakly l.s.c. and has weakly compact sublevel sets (Theorem 5.3), the problem ($\mathrm{S}_{\mathrm{static}}$) admits a minimizer. Moreover, since H is strictly convex, this minimizer is unique, and we denote it by $\pi^{\varepsilon,\mathrm{opt}}$. This optimal solution admits the following "dual representation" [Nut22, Thm. 4.2]: There exists potential functions $f\in L^\infty(\mu)$ and $g\in L^\infty(\nu)$, unique up to constants, such that

$$d\pi^{\varepsilon,\mathrm{opt}} = \exp\Big(f\oplus g - \frac{c}{\varepsilon}\Big)\,d(\mu\otimes\nu) = \exp(f\oplus g)\,dR_\varepsilon. \tag{5.17}$$

Here, we use the notation $(f\oplus g)(x,y) = f(x)+g(y)$, and call f and g the *Schrödinger potentials*[5] of $\pi^{\varepsilon,\mathrm{opt}}$. Moreover, the reverse direction also holds: If a probability measure π on Ω has marginals μ and ν, i.e., if $\pi\in\Pi(\mu,\nu)$, and is of the form (5.17), then it is the optimal solution of $\mathrm{OT}^\varepsilon(\mu,\nu)$.

The Schrödinger potentials f and g together define the optimal coupling for ($\mathrm{S}_{\mathrm{static}}$). However, because of the special structure of the optimal coupling, it turns out that given one of f or g, we can derive the other. As the $\mathcal{Y}$-marginal of $\pi^{\varepsilon,\mathrm{opt}}$

[5] Beware that some authors (such as Nutz and Wiesel [NW21]) prefer writing the dual representation as $d\pi_\varepsilon^{\mathrm{opt}} = \exp\{(f\oplus g - c)/\varepsilon\}\,d(\mu\otimes\nu)$, and call these f and g Schrödinger potentials.

is ν, it holds for any test function $\phi \in C_b(\mathcal{Y})$ that

$$\begin{aligned}\int_{\mathcal{Y}} \phi(y)\, d\nu(y) &= \int_{\Omega} \phi(y)\, d\pi^{\varepsilon,\mathrm{opt}}(x,y) \\ &= \int_{\mathcal{Y}} \phi(y) \int_{\mathcal{X}} e^{f(x)+g(y)-c(x,y)/\varepsilon}\, d\mu(x)\, d\nu(y) \\ &= \int_{\mathcal{Y}} \phi(y)\, e^{g(y)} \left[\int_{\mathcal{X}} e^{f(x)-c(x,y)/\varepsilon}\, d\mu(x)\right] d\nu(y).\end{aligned}$$

This means that for $d\nu(y)$-almost every $y \in \mathcal{Y}$,

$$g(y) = -\log \int_{\mathcal{X}} e^{f(x)-c(x,y)/\varepsilon}\, d\mu(x). \tag{5.18a}$$

With a similar argument for the $\mathcal{X}$-marginal, it holds for $d\mu(x)$-almost every $x \in \mathcal{X}$,

$$f(x) = -\log \int_{\mathcal{Y}} e^{g(y)-c(x,y)/\varepsilon}\, d\nu(y). \tag{5.18b}$$

The pair of coupled equations (5.18) is called the *Schrödinger equations*. It is straightforward to see that if a pair of potentials (f, g) satisfy the Schrödinger equations, then they are the Schrödinger potentials corresponding to the optimal solution of (OT$^\varepsilon$).

Existence of a pair of Schrödinger potentials for the optimal solution of EOT is a consequence of a property of the optimal solution called cyclical invariance. While we do not use this concept in this chapter, it is instructive to see a different perspective to Schrödinger potentials. Following [Nut22], we call a probability measure π on Ω *cyclically invariant* with respect to R_ε, if it is absolutely continuous with respect to R_ε and its density satisfies for any set of pairs $\{(x_i, y_i) \in \Omega : i = 1, \ldots, k\}$,

$$\prod_{i=1}^{k} \frac{d\pi}{dR_\varepsilon}(x_i, y_i) = \prod_{i=1}^{k} \frac{d\pi}{dR_\varepsilon}(x_i, y_{i+1}).$$

By convention, we define y_{k+1} to be y_1. It turns out that a measure π is cyclically invariant if and only if its density with respect to R_ε has the form

$$\frac{d\pi}{dR_\varepsilon} = \exp(f \oplus g),$$

for some measurable functions f and g [Nut22, Lem. 2.7]. From the discussion above, it follows that a measure in $\Pi(\mu,\nu)$ is cyclically invariant if and only if it is the optimal solution of (OT^ε). In the sequel, we let $\Pi_{c,\varepsilon}$ to be the set of all cyclically invariant measures with respect to R_ε. In other words, $\Pi_{c,\varepsilon}$ are those measures that are the optimal solution of EOT for their own marginals:

$$\Pi_{c,\varepsilon} = \{\pi \in \mathcal{P}(\Omega) : \pi \ll R_\varepsilon \text{ and } \pi \text{ is optimal for } \mathrm{OT}^\varepsilon(\pi_x, \pi_y)\}. \tag{5.19}$$

Remark. There are strong parallels between EOT and OT. Let us recall the *Kantorovich duality* for OT. Assume for the moment that the cost c is continuous and bounded. Then, the dual of (OT^0) is the maximization

$$\sup\left\{\int_{\mathcal{X}} f\,d\mu + \int_{\mathcal{Y}} g\,d\nu \;:\; f \in L^1(\mathcal{X}),\, g \in L^1(\mathcal{Y}),\, f \oplus g \le c\right\}, \tag{5.20}$$

where f and g can be taken to be continuous functions. It turns out that strong duality holds and the value of the dual problem (5.20) is equal to the value of the primal problem (OT^0). The maximizers of (5.20) are called the *Kantorovich potentials*, denoted by f^0 and g^0. A special property of these potentials is that they satisfy the pair of equations

$$\begin{aligned} f^0(x) &= \inf_{y\in\mathcal{Y}}\{c(x,y) - g^0(y)\}, \\ g^0(y) &= \inf_{x\in\mathcal{X}}\{c(x,y) - f^0(x)\}. \end{aligned} \tag{5.21}$$

In the optimal transport jargon, they are a pair of conjugate c-concave functions [Vil03, Rem. 1.12].

Let us now mention an intuitive connection between (5.21) and the Schrödinger equations (5.18). To avoid confusion, denote the Schrödinger potentials of (OT^ε) by f^ε and g^ε. By looking at (5.18a) as a “Log-Sum-Exp” applied to $f^\varepsilon(\cdot)+c(\cdot,y)/\varepsilon$, we can expect that

$$\begin{aligned} \varepsilon g^\varepsilon(y) &= -\varepsilon \log \int \exp([\varepsilon f^\varepsilon(x) - c(x,y)]/\varepsilon)\,d\mu(x) \\ &\approx -\varepsilon \sup_x\{[\varepsilon f^\varepsilon(x) - c(x,y)]/\varepsilon\} = \inf_x\{c(x,y) - \varepsilon f^\varepsilon(x)\}. \end{aligned}$$

Thus, $(\varepsilon f^\varepsilon, \varepsilon g^\varepsilon)$ satisfy (5.21) *approximately*. Moreover, as ε gets smaller, the approximation of Log-Sum-Exp by maximum gets better, and one can expect that $\varepsilon f^\varepsilon \to f^0$ and $\varepsilon g^\varepsilon \to g^0$ as $\varepsilon \to 0$. While our argument here is merely intuitive and informal, this result is indeed true; see [Nut22, Sec. 5.4] for a rigorous statement and a proof. ◊

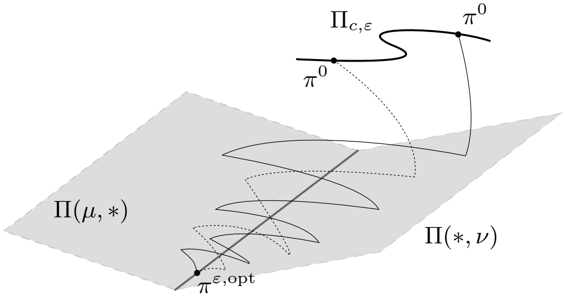


Figure 5.2. An intuitive depiction of the dependence of Sinkhorn iterates and their limit on the initialization. The planes represent the subspaces of the space of measures that have the $\mathcal{X}$- (resp. $\mathcal{Y}$-) marginals set to μ (resp. ν). The thick black curve is the set of all cyclically invariant couplings. When initialized on this set, the fate of the iterates is going to be $\pi^{\varepsilon,\mathrm{opt}}$. Two of such trajectories are depicted; one solid and one dashed. The entropic projection is a nonlinear projection, therefore it is depicted as curved lines.

5.4.2. The Sinkhorn Algorithm

A popular method for solving (OT^ε) is the *Sinkhorn algorithm* [SK67; Cut13]: Starting from $\pi^0 \in \Pi_{c,\varepsilon}$ defined in (5.19), the Sinkhorn algorithm in its primal formulation produces the iterates

$$\begin{aligned} \pi^{n+1/2} &:= \operatorname*{arg\,min}_{\pi\in\Pi(*,\nu)} H(\pi \mid \pi^n), \\ \pi^{n+1} &:= \operatorname*{arg\,min}_{\pi\in\Pi(\mu,*)} H(\pi \mid \pi^{n+1/2}). \end{aligned} \tag{Sink$_1$}$$

Here, $\Pi(\mu,*)$ denotes the set of all probability measures in $L^1(\Omega)$ whose $\mathcal{X}$-marginal is μ, and $\Pi(*,\nu)$ is defined similarly. Each iteration of (Sink_1) is an entropic projection onto $\Pi(*,\nu)$ followed by an entropic projection onto $\Pi(\mu,*)$, making it similar to alternating projection algorithms used in constrained convex optimization. We use the notation $\pi^{n+1} = \mathrm{Sin}_1(\pi^n)$ to represent a full Sinkhorn iteration. Later in Section 5.5, we will see that this iteration is an instance of mirror descent with constant step-size 1, hence the subscript 1 in Sin_1.

A key attribute of the Sinkhorn algorithm is that all the information concerning the cost c and the regularization parameter ε is encoded in the initialization π^0; the operator Sin_1 itself is independent of c and ε; see Fig. 5.2. Moreover, each step of (Sink$_1$) is of the form of a rescaling:

$$\frac{d\pi^{n+1/2}}{d\pi^n}(x,y) = \frac{d\nu}{d\pi^n_y}(y), \tag{5.22a}$$

$$\frac{d\pi^{n+1}}{d\pi^{n+1/2}}(x,y) = \frac{d\mu}{d\pi^{n+1/2}_x}(x), \tag{5.22b}$$

where π_x is the $\mathcal{X}$-marginal of the coupling π and π_y is the $\mathcal{Y}$-marginal. The rescaling property (5.22) is a simple consequence of the chain rule of relative entropy (5.14). For example, to show (5.22a), we write

$$H(\pi \,|\, \pi^n) = H(\pi_y \,|\, \pi^n_y) + \int_{\mathcal{Y}} H(\pi(\cdot \mid y) \,|\, \pi^n(\cdot \mid y))\, d\pi_y(y)$$

As $\pi^{n+1/2}$ is the minimizer of $H(\pi \,|\, \pi^n)$ among all couplings in $\Pi(*,\nu)$, we shall have $\pi^{n+1/2}(\cdot \mid y) = \pi^n(\cdot \mid y)$ for $d\nu(y)$-almost every y. This in turn implies for $d(\mu \otimes \nu)(x,y)$-almost every $(x,y) \in \Omega$,

$$\frac{d\pi^{n+1/2}}{d\pi^n}(x,y) = \frac{d\pi^{n+1/2}_y}{d\pi^n_y}(y)\, \frac{d\pi^{n+1/2}(\cdot \mid y)}{d\pi^n(\cdot \mid y)}(x) = \frac{d\nu}{d\pi^n_y}(y),$$

which is (5.22a). The other rescaling (5.22b) can be obtained similarly. The relations (5.22) readily translate into a relation between Schrödinger potentials, as described in the lemma below.

Lemma 5.4. *The iterates π^n, $n \geq 0$, of the Sinkhorn algorithm* (Sink$_1$) *are of the form*

$$d\pi^n = \exp(f_n \oplus g_n)\, dR_\varepsilon,$$

where $f_n \in L^\infty(\mu)$ and $g_n \in L^\infty(\nu)$. Furthermore, f_n can be obtained from g_n via (5.18b)*, and π^n can be recovered solely from g_n.*

Note that the same property also holds for the half-iterates $\pi^{n+1/2}$, $n \geq 0$, with the difference that $g_{n+1/2}$ can be obtained from $f_{n+1/2}$ via (5.18a), and that $\pi^{n+1/2}$ can be recovered from $f_{n+1/2}$.

Lemma 5.4, together with (5.22), gives a recursion for the Schrödinger potentials:

$$\exp\{f_{n+1/2}(x) + g_{n+1/2}(y) - f_n(x) - g_n(y)\} = \frac{d\nu}{d\pi^n_y}(y).$$

As the right-hand side is a function of y, it holds for $d\mu(x)$-almost every x that

$$f_{n+1/2}(x) = f_n(x),$$

and similarly, (5.22b) implies that for $d\nu(y)$-almost every y,

$$g_{n+1}(y) = g_{n+1/2}(y).$$

In short, the Sinkhorn algorithm updates g in half-iterations, and updates f in integer iterations. This is summarized in the *dual Sinkhorn iteration*, which describes the whole process only in terms of the Schrödinger potentials:

$$\begin{aligned} g_{n+1}(y) &= -\log \int_{\mathcal{X}} e^{f_n(x) - c(x,y)/\varepsilon} \, d\mu(x), \\ f_{n+1}(x) &= -\log \int_{\mathcal{Y}} e^{g_{n+1}(y) - c(x,y)/\varepsilon} \, d\nu(y). \end{aligned} \tag{Dual-Sink$_1$}$$

Remark 5.2. Let us remark briefly on how (Dual-Sink$_1$) is used in practice to solve EOT for measures with finite support. Suppose μ and ν are supported on $\{x_1, \ldots, x_k\}$ and $\{y_1, \ldots, y_\ell\}$, respectively. Let $K \in \mathbb{R}_+^{k \times \ell}$ be the matrix with entries

$$K_{i,j} = e^{-c(x_i,y_j)/\varepsilon} \, \mu(x_i) \, \nu(y_j);$$

let $\vec{\mu} = (\mu(x_1), \ldots, \mu(x_k))$ and $\vec{\nu} = (\nu(y_1), \ldots, \nu(y_\ell))$ be the probability vectors for the measures μ and ν. Then, the dual iteration can be equivalently described as follows: Define the sequence $u_n \in \mathbb{R}_+^k$ and $v_n \in \mathbb{R}_+^\ell$ as $u_n = e^{f_n}$ and $v_n = e^{g_n}$. Then, (Dual-Sink$_1$) becomes

$$v_{n+1} = \frac{\vec{\nu}}{K^\top u_n}, \quad u_{n+1} = \frac{\vec{\mu}}{K v_{n+1}}, \tag{5.23}$$

where division of vectors is meant to be elementwise. One can also recover each primal iteration of (Sink$_1$) as

$$\pi^n = \operatorname{diag}(u_n) \, K \operatorname{diag}(v_n). \qquad \Diamond$$

In Section 5.5.3 below, we demonstrate another way to think about the Schrödinger potentials and their relation to the primal iterates in Lemma 5.4. Specifically, in Proposition 5.8 we show that retrieving π^n from g_n is the same as applying the "dual mirror map" to g_n.

Remark 5.3. Another implication of Lemma 5.4 is that the iterates π^n of the Sinkhorn algorithm are all absolutely continuous with respect to R_ε (and therefore, with respect to $\mu \otimes \nu$), and their densities are *bounded, positive, and bounded away from zero.* This will be important later in Section 5.5, as the relative entropy functional is Gâteaux differentiable at points in $L^\infty_{++}(\Omega)$; see Section 5.3. $\Diamond$

5.5. STEP-SIZED SINKHORN

Besides the classic alternating projection viewpoint mentioned in the previous section, it has been recently pointed out by a series of works [Mis19; Lég20; AKL22] that the Sinkhorn iteration (Sink_1) has a mirror descent interpretation. Mishchenko [Mis19] made the connection between Sinkhorn and MD for the case of finitely supported measures. Later, Léger [Lég20] showed that the Sinkhorn algorithm can be viewed as a mirror descent iteration in the space of probability measures, and was able to show rates of convergence even in situations where classical results, such as [FL89], do not provide meaningful convergence rates. Later, Aubin-Frankowski, Korba, and Léger [AKL22] made this result more rigorous and studied, in a general context, mirror descent with relative smoothness in the space of probability measures.

In this section, we first bring the general idea of interpreting Sinkhorn as MD in the space of measures. We then generalize the Sinkhorn algorithm into a step-sized algorithm. Our generalized algorithm has both primal and dual forms. We heavily use the machinery of mirror descent, convex duality and analysis in the space of measures, and refer the reader to Sections 5.2 and 5.3 for the related background.

5.5.1. Sinkhorn as Mirror Descent

Let us start by mentioning the concrete framework of mirror descent in the space of probability measures. Recall that $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of $\mathbb{R}^d$, and $\Omega = \mathcal{X} \times \mathcal{Y}$. We work with the dual pair $\langle L^1(\Omega), L^\infty(\Omega)\rangle$, where we write $L^1(\Omega)$ as a shorthand for $L^1(\mu \otimes \nu)$ and similarly for $L^\infty(\Omega)$. For an objective function $F : L^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$, a Bregman potential $\varphi : L^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$, a constraint set $\mathcal{C} \subset L^1(\Omega)$, and a step-size sequence $\{\gamma_n\}_{n\in\mathbb{N}}$, the *Mirror Descent iteration* in its primal form is defined via the recursion

$$\pi^{n+1} = \operatorname*{arg\,min}_{\pi \in \mathcal{C}} \left\{ F(\pi^n) + d^+F(\pi^n)(\pi - \pi^n) + \frac{D_\varphi(\pi \,|\, \pi^n)}{\gamma_n} \right\}, \tag{5.24}$$

where the Bregman divergence $D_\varphi(\cdot \,|\, \cdot)$ is also defined with the help of directional derivatives:

$$D_\varphi(\pi' \,|\, \pi) = \varphi(\pi') - \varphi(\pi) - d^+\varphi(\pi)(\pi' - \pi).$$

Existence and uniqueness of the minimizers in this scheme shall be verified in a case-by-case basis; here we only do so for the specific case of the Sinkhorn algorithm and its step-sized version.

A specific choice of the objective function F, the Bregman potential φ, the constraint set $\mathcal{C}$, and the step-size sequence $\{\gamma_n\}$, renders the iteration (5.24) into the Sinkhorn algorithm (Sink_1). Specifically, if one sets

$$F(\pi) := H(\pi_y \,|\, \nu), \quad \varphi(\pi) := H(\pi \,|\, R_\varepsilon), \quad \mathcal{C} := \Pi(\mu, *), \quad \text{and} \quad \gamma_n \equiv 1, \tag{5.25}$$

it is shown in [AKL22, Prop. 5] that the full iteration (Sink_1) can be written as the mirror descent update (5.24). It is instructive and useful for later computations to show why this relation holds.

Consider the iterations $\{\pi^n\}$ of the Sinkhorn algorithm (Sink_1). What we show is that these iterations satisfy (5.24) for the specific choices made in (5.25). Recall from Remark 5.3 that $\pi^n \in L^\infty_{++}(\Omega)$ for all n. Therefore, as in (5.13), φ has a Gâteaux derivative at π^n and

$$d^+\varphi(\pi^n)(\pi - \pi^n) = \langle \nabla\varphi(\pi^n), \pi - \pi^n \rangle = \int_\Omega \log \frac{d\pi^n}{dR_\varepsilon}\, d(\pi - \pi^n).$$

Substituting this in the definition of the Bregman divergence gives

$$D_\varphi(\pi \,|\, \pi^n) = H(\pi \,|\, \pi^n).$$

Moreover, as mentioned in (5.15), the directional derivative of the objective F computes as

$$d^+F(\pi^n)(\pi - \pi^n) = \langle \nabla F(\pi^n), \pi - \pi^n \rangle = \int_\Omega \log \frac{d\pi^n_y}{d\nu}(y)\, d(\pi - \pi^n)(x, y).$$

Thus, the objective of the mirror descent update evaluates as

$$\begin{aligned}
&F(\pi^n) + \langle \nabla F(\pi^n), \pi - \pi^n \rangle + D_\varphi(\pi \,|\, \pi^n) \\
&\quad = \int_{\mathcal{Y}} \log \frac{d\pi^n_y}{d\nu}\, d\pi^n_y + \int_\Omega \log \frac{d\pi^n_y}{d\nu}\, d(\pi - \pi^n) + \int_\Omega \log \frac{d\pi}{d\pi^n}\, d\pi \\
&\quad = \int_\Omega \log \frac{d\pi^n_y}{d\nu}\, d\pi + \int_\Omega \log \frac{d\pi}{d\pi^n}\, d\pi \\
&\quad = \int_\Omega \log \frac{d\pi^n}{d\pi^{n+1/2}}\, d\pi + \int_\Omega \log \frac{d\pi}{d\pi^n}\, d\pi \\
&\quad = H(\pi \,|\, \pi^{n+1/2}),
\end{aligned}$$

where the penultimate equality follows from (5.22a). Taking $\arg\min$ over $\mathcal{C}$ gives the full iteration (Sink_1).

5.5.2. Sinkhorn with Arbitrary Step-sizes

Thus far, we have seen that the Sinkhorn iteration (Sink$_1$) is an instance of the general (primal) MD iteration (5.24) with constant step-size 1. Once this connection to MD is established, we can use arbitrary step-sizes to get a new Sin$_\gamma$-*iteration*, defined as follows:

Definition 5.5 (Sin$_\gamma$-iteration)**.** Let $F(\pi) = H(\pi_y \,|\, \nu)$, $\varphi(\pi) = H(\pi \,|\, R_\varepsilon)$, and $\mathcal{C} = \Pi(\mu, *)$ be the same as in (5.25), and $\{\gamma_n\}_{n\in\mathbb{N}}$ be a sequence of step-sizes satisfying $\gamma_n \leq 1$ for all n. Starting from $\pi^0 \in \Pi_{c,\varepsilon}$, the Sin$_\gamma$-iterates are defined as

$$\pi^{n+1} = \operatorname*{arg\,min}_{\pi\in\mathcal{C}} \left\{ \langle \nabla F(\pi^n), \pi - \pi^n \rangle + \frac{D_\varphi(\pi \,|\, \pi^n)}{\gamma_n} \right\}, \tag{Sink$_\gamma$}$$

and write $\pi^{n+1} = \mathrm{Sin}_{\gamma_n}[\pi^n]$.

Existence and uniqueness of the Sin$_\gamma$-iterates follow in the exact same way as for Sinkhorn iterates [AKL22]: As the directional derivative of F is linear and continuous at π^n and $D_\varphi(\cdot \,|\, \cdot)$ is the relative entropy, the function in the arg min of (Sink$_\gamma$) becomes l.s.c. and therefore, has weakly compact level-sets. This shows the existence of a minimizer. Uniqueness follows from strict convexity of the Bregman divergence.

It turns out that the Sin$_\gamma$-iteration has a simpler formulation that resembles the original Sinkhorn algorithm in its primal form (Sink$_1$):

▶ **Lemma 5.6.** *The* Sin$_\gamma$*-iterates* π^n *defined in* (Sink$_\gamma$) *satisfy the recursion*

$$\pi^{n+1/2} := \operatorname*{arg\,min}_{\pi\in\Pi(*,\nu)} H(\pi \,|\, \pi^n), \tag{5.26a}$$

$$\pi^{n+1} := \operatorname*{arg\,min}_{\pi\in\Pi(\mu,*)} \left\{ \gamma_n H(\pi \,|\, \pi^{n+1/2}) + (1-\gamma_n) H(\pi \,|\, \pi^n) \right\}. \tag{5.26b}$$

Proof. The proof follows by computing the objective of (Sink$_\gamma$):

$$\begin{aligned}
&F(\pi^n) + \langle \nabla F(\pi^n), \pi - \pi^n \rangle + \frac{D_\varphi(\pi \,|\, \pi^n)}{\gamma_n} \\
&= \int d\pi^n \log \frac{d\pi^n_y}{d\nu} + \int d(\pi - \pi^n) \log \frac{d\pi^n_y}{d\nu} + \frac{H(\pi \,|\, \pi^n)}{\gamma_n} \\
&= \int d\pi \log \frac{d\pi^n_y}{d\nu} + \frac{H(\pi \,|\, \pi^n)}{\gamma_n}.
\end{aligned}$$

Now define $\pi^{n+1/2}$ as in (5.26a). Since this step is shared with (Sink$_1$) iteration, we can use the rescaling property (5.22a) and find that the last equation is

$$\begin{aligned}
&= \int d\pi \log \frac{d\pi^n}{d\pi^{n+1/2}} + \frac{H(\pi \,|\, \pi^n)}{\gamma_n} \\
&= \int d\pi \log \left\{ \frac{d\pi^n}{d\pi^{n+1/2}} \cdot \left(\frac{d\pi}{d\pi^n} \right)^{1/\gamma_n} \right\} \\
&= \frac{1}{\gamma_n} \int d\pi \log \left\{ \left(\frac{d\pi^n}{d\pi^{n+1/2}} \right)^{\gamma_n} \cdot \left(\frac{d\pi}{d\pi^n} \right)^{\gamma_n} \cdot \left(\frac{d\pi}{d\pi^n} \right)^{1-\gamma_n} \right\} \\
&= \frac{1}{\gamma_n} \int d\pi \log \left\{ \left(\frac{d\pi}{d\pi^{n+1/2}} \right)^{\gamma_n} \cdot \left(\frac{d\pi}{d\pi^n} \right)^{1-\gamma_n} \right\} \\
&= \frac{1}{\gamma_n} \left\{ \gamma_n H(\pi \,|\, \pi^{n+1/2}) + (1-\gamma_n) H(\pi \,|\, \pi^n) \right\}.
\end{aligned}$$

As the factor $1/\gamma_n$ is irrelevant for the minimization, we get the desired result. □

The new iteration (5.26) becomes exactly (Sink$_1$) by setting $\gamma_n = 1$ for all $n \in \mathbb{N}$. For step-sizes $\gamma_n < 1$, however, the behavior of (5.26) becomes more interesting, as the second step becomes a regularized projection onto $\Pi(\mu, *)$, and the regularization gets more powerful when γ_n gets smaller. This introduces a form of stability in the algorithm, and enables us later to prove convergence of (5.26) under noisy updates; something that does not necessarily hold for the original Sinkhorn algorithm.

5.5.3. Dual Sinkhorn Iteration

Similar to the Sinkhorn algorithm, the Sin$_\gamma$-iteration admits a dual representation. This enables us to construct an efficient algorithm for this new iteration, and later in Section 5.6, allows us to define a novel flow in the space of probability measures.

To derive the dual iteration we first show that the step-sized iteration (Sink$_\gamma$) admits a rescaling interpretation, similar to (5.22). This will then result in an update for the corresponding Schrödinger potentials, which automatically leads us to the dual algorithm.

Let us start from the primal formulation (5.26). As the first step (5.26a) is shared with the original Sinkhorn algorithm, we readily know that the rescaling (5.22a) holds. For the other step (5.26b), after conditioning on x and using the chain rule of relative entropy, we obtain that the conditional distribution $\pi^{n+1}(\cdot \mid x)$

shall minimize

$$\gamma_n \int_{\mathcal{X}} H(\pi(\cdot \mid x) \,|\, \pi^{n+1/2}(\cdot \mid x))\, d\mu(x) + (1-\gamma_n) \int_{\mathcal{X}} H(\pi(\cdot \mid x) \,|\, \pi^{n}(\cdot \mid x))\, d\mu(x),$$

which is equivalent to minimizing

$$\int_{\mathcal{X}} d\mu(x) \int_{\mathcal{Y}} d\pi(y \mid x) \log\biggl\{ \left(\frac{d\pi^{n+1/2}(y \mid x)}{d\pi^{n}(y \mid x)} \right)^{1-\gamma_n} \frac{d\pi(y \mid x)}{d\pi^{n+1/2}(y \mid x)} \biggr\}.$$

For $d\mu(x)$-almost every x, define the probability measure $\varrho(\cdot \mid x)$ via

$$\frac{d\varrho(y \mid x)}{d\pi^{n+1/2}(y \mid x)} \propto \left(\frac{d\pi^{n}(y \mid x)}{d\pi^{n+1/2}(y \mid x)} \right)^{1-\gamma_n}.$$

Then, after removing terms not depending on $\pi(\cdot \mid x)$, the objective above becomes $\int_{\mathcal{X}} H(\pi(\cdot \mid x) \,|\, \varrho(\cdot \mid x))\, d\mu(x)$, whose minimizer is $\varrho(\cdot \mid x)$. Therefore, the rescaling form of (Sink$_\gamma$) is

$$\frac{d\pi^{n+1/2}}{d\pi^{n}}(x,y) = \frac{d\nu}{d\pi^{n}_{y}}(y), \tag{5.27a}$$

$$\frac{d\pi^{n+1}}{d\pi^{n+1/2}}(x,y) \propto \frac{d\mu}{d\pi^{n+1/2}_{x}}(x) \left(\frac{d\pi^{n}(y \mid x)}{d\pi^{n+1/2}(y \mid x)} \right)^{1-\gamma_n}. \tag{5.27b}$$

This readily implies the existence of Schrödinger potentials, as well as a dual formulation of the Sin$_\gamma$-iterates. The proof is straightforward and is omitted.

► **Lemma 5.7.** *The* Sin$_\gamma$*-iterates* π^n *defined in* (Sink$_\gamma$) *admit the representation*

$$d\pi^{n} = \exp(f_n \oplus g_n)\, dR_\varepsilon, \tag{5.28}$$

with $f_n \in L^\infty(\mu)$ *and* $g_n \in L^\infty(\nu)$*. Moreover, the potentials* g_n *satisfy the recursion*

$$g_{n+1} = g_n - \gamma_n \log \frac{d\pi^{n}_{y}}{d\nu}, \tag{5.29}$$

and f_{n+1} *is computed from* g_{n+1} *in the same way as in* (5.18b)*. Consequently, the* Sin$_\gamma$*-iteration admits the dual formulation:*

$$\begin{aligned} g_{n+1}(y) &= (1-\gamma_n)\, g_n(y) - \gamma_n \log \int_{\mathcal{X}} e^{f_n(x) - c(x,y)/\varepsilon}\, d\mu(x), \\ f_{n+1}(x) &= -\log \int_{\mathcal{Y}} e^{g_{n+1}(y) - c(x,y)/\varepsilon}\, d\nu(y). \end{aligned} \tag{Dual-Sink$_\gamma$}$$

Remark 5.4. The dual iteration for the step-sized Sinkhorn algorithm admits an efficient implementation for measures with finite support. Using the same notation as in Remark 5.2, it is straightforward to see that by defining $u_n = e^{f_n}$ and $v_n = e^{g_n}$, the iteration (Dual-Sink$_\gamma$) becomes

$$v_{n+1} = v_n^{1-\gamma_n} \odot \left(\frac{\vec{\nu}}{K^\top u_n} \right)^{\gamma_n}, \quad u_{n+1} = \frac{\vec{\mu}}{K v_{n+1}}, \tag{5.30}$$

where $\odot$ is elementwise product of vectors. ♢

5.5.4. Dual Mirror Descent Interpretation

We now take a step further and make a connection between the dual Sinkhorn iteration (Dual-Sink$_\gamma$) and the dual mirror descent iteration (5.7) in Section 5.2.

Let $L^\infty_{++}(\Omega) \cap \mathcal{C} \subset L^1(\Omega)$ be the *primal space* and $L^\infty(\Omega)$ be the *dual space*. The Gâteaux derivative of φ creates a link from the primal to the dual space, in the sense that $\nabla\varphi$ maps a measure π in the primal space to $\log \frac{d\pi}{dR_\varepsilon}$, which is, by construction, in $L^\infty(\Omega)$. The map from the dual space to the primal space is obtained formally by applying the Gâteaux derivative of $\varphi^\star_{\mathcal{C}}$ (if it exists) to a dual point $h \in L^\infty(\Omega)$.

To have a complete dual MD picture, we have to first verify that $\varphi^\star_{\mathcal{C}}$ admits a Gâteaux derivative, and show that plugging this derivative into the MD dual iteration (5.7) gives back the dual Sin$_\gamma$-iteration (Dual-Sink$_\gamma$). Proposition 5.8 below gives a formula for the Fenchel conjugate $\varphi^\star_{\mathcal{C}}$ and shows that it indeed Gâteaux differentiable.

Proposition 5.8. *The Fenchel conjugate of $\varphi_{\mathcal{C}} = \varphi + \mathrm{I}_{\mathcal{C}}$ evaluated at $h \in L^\infty(\Omega)$ is given by $\varphi^\star_{\mathcal{C}}(h) = \langle \hat{\pi}, h \rangle - H(\hat{\pi} \,|\, R_\varepsilon)$, where $\hat{\pi} \in \mathcal{C} \cap L^\infty_{++}(\Omega)$ satisfies*

$$\frac{d\hat{\pi}}{dR_\varepsilon}(x,y) = \frac{e^{h(x,y)}}{\int_{\mathcal{Y}} e^{h(x,y')} e^{-c(x,y')/\varepsilon}\, d\nu(y')}. \tag{5.31}$$

Moreover, $\varphi^\star_{\mathcal{C}}$ has a Gâteaux derivative at $h \in L^\infty(\Omega)$, which is given by

$$\nabla \varphi^\star_{\mathcal{C}}(h) = \hat{\pi}. \tag{5.32}$$

Proof. Recall that

$$\varphi^\star_{\mathcal{C}}(h) = \sup_{\pi \in \mathcal{C}} \{\langle \pi, h \rangle - H(\pi \,|\, R_\varepsilon)\} = -\inf_{\pi \in \mathcal{C}} \{H(\pi \,|\, R_\varepsilon) - \langle \pi, h \rangle\}.$$

As $h \in L^\infty(\Omega)$, define the probability measure $d\varrho \propto e^h \, dR_\varepsilon$ and write

$$H(\pi \,|\, R_\varepsilon) - \langle \pi, h \rangle = \int \left(\frac{d\pi}{dR_\varepsilon} - h \right) d\pi = H(\pi \,|\, \varrho) + \log \int e^h \, dR_\varepsilon.$$

The second term is independent of π, thus,

$$\varphi_{\mathcal{C}}^\star(h) = - \inf_{\pi \in \mathcal{C}} H(\pi \,|\, \varrho).$$

Since $\mathcal{C}$ is weakly closed and H is strictly convex with weakly compact sublevel sets, this minimization has a unique minimizer $\hat{\pi}$. By the chain rule of relative entropy and using $\hat{\pi} \in \mathcal{C}$, we obtain

$$H(\hat{\pi} \,|\, \varrho) = H(\mu \,|\, \varrho_x) + \int H(\hat{\pi}(\cdot \mid x) \,|\, \varrho(\cdot \mid x)) \, d\mu(x).$$

Consequently, for $d\mu(x)$-almost every x, it should hold $\hat{\pi}(\cdot \mid x) = \varrho(\cdot \mid x)$. However, by the properties of Radon–Nikodym derivative, we know

$$\frac{d\varrho(\cdot \mid x)}{d\nu}(y) = \frac{d\varrho}{d(\mu \otimes \nu)}(x, y) \Big/ \frac{d\varrho_x}{d\mu}(x)$$

and

$$\frac{d\varrho_x}{d\mu}(x) = \int_{\mathcal{Y}} \frac{d\varrho}{d(\mu \otimes \nu)}(x, y) \, d\nu(y) = \frac{\int_{\mathcal{Y}} e^h e^{-c/\varepsilon} \, d\nu}{\int_\Omega e^h \, dR_\varepsilon}.$$

Putting it all together, we obtain

$$\frac{d\hat{\pi}(\cdot \mid x)}{d\nu}(y) = \frac{e^{h(x,y)} e^{-c(x,y)/\varepsilon}}{\int_{\mathcal{Y}} e^{h(x,y')} e^{-c(x,y')/\varepsilon} \, d\nu(y')}.$$

Thus, as $\hat{\pi}_x = \mu$, it holds that $d\hat{\pi}_x / d\mu = 1$ and

$$\frac{d\hat{\pi}}{d(\mu \otimes \nu)}(x, y) = \frac{e^{h(x,y)} e^{-c(x,y)/\varepsilon}}{\int_{\mathcal{Y}} e^{h(x,y')} e^{-c(x,y')/\varepsilon} \, d\nu(y')},$$

proving (5.31).

We now show (5.32). As the set $\mathcal{C}$ is weakly closed, its convex indicator $\mathrm{I}_{\mathcal{C}}$ is weakly l.s.c., implying lower semi-continuity of $\varphi_{\mathcal{C}}$. By Theorem 5.1, $\pi \in \partial \varphi_{\mathcal{C}}^\star(h)$ if and only if

$$\varphi_{\mathcal{C}}(\pi) + \varphi_{\mathcal{C}}^\star(h) = \langle \pi, h \rangle.$$

As for any fixed $h \in L^\infty(\Omega)$ the maximizer of $\sup_{\pi \in \mathcal{C}} \{ \langle \pi, h \rangle - H(\pi \,|\, R_\varepsilon) \}$ is unique, it holds that $\partial\varphi^\star_{\mathcal{C}}(h) = \{\hat{\pi}\}$. As the subdifferential is a singleton, the subgradient is indeed the Gâteaux derivative of $\varphi^\star_{\mathcal{C}}$ at h. Moreover, as $\hat{\pi}$ is in $L^\infty(\Omega)$, this Gâteaux derivative is continuous. □

Remark 5.5. Let us make the important remark that the values of the Fenchel conjugate $\varphi^\star_{\mathcal{C}}(h)$ and its first variation $\nabla\varphi^\star_{\mathcal{C}}$ do not change if one adds a function of the x variable to h. Concretely, for any $u \in L^\infty(\mathcal{X})$, it holds that $\varphi^\star_{\mathcal{C}}(h+u) = \varphi^\star_{\mathcal{C}}(h)$ and $\nabla\varphi^\star_{\mathcal{C}}(h+u) = \nabla\varphi^\star_{\mathcal{C}}(h)$. This is because the additional u in (5.31) gets cancelled out from the numerator and denominator. Specifically, if $h = f \oplus g$, then

$$\varphi^\star_{\mathcal{C}}(f \oplus g) = \varphi^\star_{\mathcal{C}}(g), \quad \nabla\varphi^\star_{\mathcal{C}}(f \oplus g) = \nabla\varphi^\star_{\mathcal{C}}(g). \qquad \Diamond$$

We are now in the position to translate the dual Sin_γ-iteration (Dual-Sink$_\gamma$) into the language of mirror descent. More precisely, we show in Proposition 5.9 below that (Dual-Sink$_\gamma$) is precisely the dual mirror descent iteration (5.7) described in Section 5.2.

▶ **Proposition 5.9.** *Let $\{\pi^n\}_{n\geq 0}$ be the Sin_γ-iterates starting from $\pi^0 \in \Pi_{c,\varepsilon}$, and let f_n and g_n be the corresponding Schrödinger potentials, evolving as in (Dual-Sink$_\gamma$). Let $\tilde{\pi}^0 = \pi^0$ and $\tilde{g}_0 = g_0$, and define $\tilde{\pi}^n$ and $\tilde{g}_n$ for $n \geq 1$ via the recursion*

$$\begin{cases} \tilde{g}_{n+1} = \tilde{g}_n - \gamma_n \nabla F(\tilde{\pi}^n), \\ \tilde{\pi}^{n+1} = \nabla\varphi^\star_{\mathcal{C}}(\tilde{g}_{n+1}). \end{cases} \tag{5.33}$$

Then, for all $n \geq 0$, it holds that $\tilde{\pi}^n = \pi^n$ and $\tilde{g}_n = g_n$.

Proof. Consider a sequence of potentials $\tilde{f}_n \in L^\infty(\mathcal{X})$, which together with $\tilde{g}_n$ satisfy $d\tilde{\pi}^n = \exp(\tilde{f}_n \oplus \tilde{g}_n)\, dR_\varepsilon$. We show via induction that $\tilde{f}_n$ and $\tilde{g}_n$ are updated exactly in the same way as in (Dual-Sink$_\gamma$), thus proving the lemma.

Suppose by induction hypothesis that $\tilde{\pi}^n = \exp(\tilde{f}_n \oplus \tilde{g}_n)\, dR_\varepsilon$ and that $\tilde{\pi}^n = \pi^n$ and $\tilde{g}_n = g_n$ (and hence, $\tilde{f}_n = f_n$). The first step of (5.33) is

$$\begin{aligned} \tilde{g}_{n+1}(y) = g_n(y) - \gamma_n \nabla F(\pi^n)(y) &= g_n(y) - \gamma_n \log \frac{d\pi^n_y}{d\nu}(y) \\ &= g_n(y) - \gamma_n \log e^{g_n(y)} \int e^{f_n(x) - c(x,y)/\varepsilon}\, d\mu(x) \\ &= (1-\gamma_n) g_n(y) - \gamma_n \log \int e^{f_n(x) - c(x,y)/\varepsilon}\, d\mu(x), \end{aligned}$$

which is precisely the first step of dual Sin_γ-iteration (Dual-Sink$_\gamma$). This shows $\tilde{g}_{n+1} = g_{n+1}$. The next step of (5.33) is $\tilde{\pi}^{n+1} = \nabla\varphi^\star_{\mathcal{C}}(\tilde{g}_{n+1}) = \nabla\varphi^\star_{\mathcal{C}}(g_{n+1})$. By

Proposition 5.8, we have

$$\frac{d\tilde{\pi}^{n+1}}{dR_\varepsilon}(x,y) = \frac{e^{g_{n+1}(y)}}{\int_{\mathcal{Y}} e^{g_{n+1}(y')} e^{-c(x,y')/\varepsilon}\, d\nu(y')} =: \exp(\tilde{f}_{n+1} \oplus g_{n+1})(x,y),$$

where we defined

$$\tilde{f}_{n+1}(x) := -\log \int_{\mathcal{Y}} e^{g_{n+1}(y)} e^{-c(x,y)/\varepsilon}\, d\nu(y).$$

Observe that this is exactly the same as the second step in (Dual-Sink$_\gamma$), showing that $\tilde{f}_{n+1} = f_{n+1}$. We thus conclude that $\tilde{\pi}^{n+1} = \pi^{n+1}$. Therefore, the dual MD iteration gives the same iterates as the dual Sinkhorn iteration. □

5.6. SINKHORN FLOWS

We now study the limiting behavior of the operator Sin$_\gamma$ as $\gamma \to 0$. This allows us to define two continuous-time flows: one on the space of probability measures, and another one in the dual space. To avoid unnecessary measure-theoretic complications, we focus on the following two scenarios throughout this section:

(1) The probability measures μ and ν are absolutely continuous with respect to the Lebesgue measure with continuous and bounded densities, and the cost c is continuous and bounded. It is easy to see that in this case, all Sin$_\gamma$-iterates admit bounded and continuous densities; specifically, the Schrödinger potentials are also continuous and bounded.

(2) The probability measures μ and ν have finite support.

5.6.1. Sinkhorn Flow as a Mirror Flow

We start by computing the limit of the Sin$_\gamma$ operator defined in Definition 5.5. While the Sin$_\gamma$-iteration can be initialized at any measure in $\Pi_{c,\varepsilon}$, here we only consider those initial measures that have the correct $\mathcal{X}$-marginal μ. This will remove the disparity of the initialization and the rest of the iterates, which all have their $\mathcal{X}$-marginal equal to μ by construction. This helps us later when defining a continuous flow.

► **Proposition 5.10.** *Fix a coupling $\pi \in \Pi(\mu, *) \cap \Pi_{c,\varepsilon}$. For $\gamma > 0$, let $\pi^\gamma = \mathrm{Sin}_\gamma[\pi]$ be the result of one step of* (Sink$_\gamma$) *with step-size γ applied to π. Then,*

$$\frac{d}{d\gamma}\bigg|_{\gamma=0} \pi^\gamma(x,y) = -\pi(x,y)\log\frac{d\pi_y}{d\nu}(y) + \pi(x,y)\int_{\mathcal{Y}} \log\frac{d\pi_y}{d\nu}(y')\,d\pi(y' \mid x).$$

Moreover, if $d\pi = \exp(f \oplus g)\,dR_\varepsilon$ for some $f \in C_b(\mathcal{X})$ and $g \in C_b(\mathcal{Y})$, then $d\pi^\gamma = \exp(f_\gamma \oplus g_\gamma)\,dR_\varepsilon$, and f_γ and g_γ satisfy

$$\begin{aligned}\frac{d}{d\gamma}\bigg|_{\gamma=0} g_\gamma(y) &= -\log\frac{d\pi_y}{d\nu}(y),\\ \frac{d}{d\gamma}\bigg|_{\gamma=0} f_\gamma(x) &= \int_{\mathcal{Y}} \log\frac{d\pi_y}{d\nu}(y')\,d\pi(y' \mid x).\end{aligned}$$

Proof. Since $\pi \in \Pi_{c,\varepsilon}$, it is readily implied that $d\pi = \exp(f \oplus g)\,dR_\varepsilon$ for some $f \in C_b(\mathcal{X})$ and $g \in C_b(\mathcal{Y})$. Lemma 5.7 then implies that π^γ is also of the same form: $d\pi^\gamma = \exp(f_\gamma \oplus g_\gamma)\,dR_\varepsilon$.

By the recursion (5.29), we see that $\frac{1}{\gamma}(g_\gamma - g) = -\log\frac{d\pi_y}{d\nu}$. Taking the limit as $\gamma \to 0$ gives the desired formula for $\frac{d}{d\gamma}\big|_{\gamma=0} g_\gamma$.

From the second step of (Dual-Sink$_\gamma$) and taking derivative with respect to γ at $\gamma = 0$, we get

$$\begin{aligned}\frac{d}{d\gamma}\bigg|_{\gamma=0} f_\gamma(x) &= \int_{\mathcal{Y}} \log\frac{d\pi_y}{d\nu}(y')\,e^{g(y')-c(x,y')/\varepsilon}\,d\nu(y') \Big/ \int_{\mathcal{Y}} e^{g(y')-c(x,y')/\varepsilon}\,d\nu(y')\\ &= \int_{\mathcal{Y}} \log\frac{d\pi_y}{d\nu}(y')\,e^{f(x)+g(y')-c(x,y')/\varepsilon}\,d\nu(y')\\ &= \int_{\mathcal{Y}} \log\frac{d\pi_y}{d\nu}(y')\,d\pi(y' \mid x),\end{aligned}$$

where for the second equality we used the fact that $\pi \in \Pi(\mu, *)$ and (5.18b), and in the last equality we used the disintegration of π:

$$\frac{d\pi(\cdot \mid x)}{d\nu}(y') \cdot \frac{d\pi_x}{d\mu}(x) = \frac{d\pi}{d\mu \otimes \nu}(x, y').$$

Having the derivatives of f_γ and g_γ with respect to γ, it is straightforward to compute the derivative of $\log \pi^\gamma$:

$$\begin{aligned}\frac{d}{d\gamma}\bigg|_{\gamma=0} \log \frac{d\pi^\gamma}{dR_\varepsilon}(x,y) &= \frac{d}{d\gamma}\bigg|_{\gamma=0} (f_\gamma(x) + g_\gamma(y)) \\ &= -\log \frac{d\pi_y}{d\nu}(y) + \int_{\mathcal{Y}} \log \frac{d\pi_y}{d\nu}(y')\, d\pi(y' \mid x).\end{aligned}$$

Subsequently, using $\frac{d}{d\gamma}\big|_{\gamma=0} \pi^\gamma(x,y) = \pi(x,y) \frac{d}{d\gamma}\big|_{\gamma=0} \log \pi^\gamma(x,y)$, we obtain the result of the lemma. □

In view of Proposition 5.10, we are now ready to define the Sinkhorn and the dual Sinkhorn flows.

Definition 5.11. For any $\pi^0 \in \Pi(\mu, *) \cap \Pi_{c,\varepsilon}$, construct the curve $(\pi^t)_{t\geq 0}$ in $L^1(\Omega)$ whose velocity is determined by

$$\frac{d}{dt}\pi^t(x,y) = -\pi^t(x,y) \log \frac{d\pi^t_y}{d\nu}(y) + \pi^t(x,y)\, \mathbb{E}_{\pi^t(\cdot|x)}\left[\log \frac{d\pi^t_y}{d\nu}\right]. \tag{5.34}$$

We call the mapping $(t, \pi^0) \mapsto \pi^t$ the *Sinkhorn flow*. Similarly, if $\frac{d\pi^0}{dR_\varepsilon} = \exp(f_0 \oplus g_0)$, we call the mapping $(t, g_0) \mapsto g_t$ the *dual Sinkhorn flow*, which describes the evolution of the Schrödinger potential g_t corresponding to π^t:

$$\frac{d}{dt} g_t = -\log \frac{d\pi^t_y}{d\nu}. \tag{5.35}$$

As the Sinkhorn and dual Sinkhorn flows emerge from driving the step-size of an MD iteration to zero, it is natural to anticipate a *mirror flow* interpretation. Indeed, using the formulas for the mirror map $\nabla\varphi$ and its dual $\nabla\varphi^\star_{\mathcal{C}}$, we can define a mirror flow as follows. Fix $\pi^0 \in \Pi_{c,\varepsilon} \cap \Pi(\mu, *)$ and any $h_0 \in L^\infty(\Omega)$ that satisfies $\nabla\varphi^\star_{\mathcal{C}}(h_0) = \pi^0$. Consider the evolution

$$\begin{cases} \dfrac{d}{dt} h_t = -\nabla F(\pi^t), \\ \quad \pi^t = \nabla\varphi^\star_{\mathcal{C}}(h_t), \end{cases} \tag{5.36}$$

which can be equivalently written as

$$\frac{d}{dt} h_t = -(\nabla F \circ \nabla\varphi^\star_{\mathcal{C}})(h_t). \tag{5.37}$$

As mentioned in Remark 5.5, we can take h_0 to be a function of the y variable only; this is possible since $\pi^0 \in \Pi_{c,\varepsilon}$. The specific choice of $h_0 = g_0$ yields the evolution (5.35), therefore confirming our anticipation that dual Sinkhorn flow is a dual mirror flow.

Remark 5.6. Following Remarks 5.2 and 5.4, let us mention how the Sinkhorn and dual Sinkhorn flows look like for measures with finite support. We continue using the notation in the mentioned remarks. Define the vectors $u_t := e^{f_t}$ and $v_t := e^{g_t}$, where f_t and g_t are the Schrödinger potentials for the Sinkhorn flow. From (5.35), it follows that the dual Sinkhorn flow in this case is

$$\dot{v}_t = -v_t \odot \log \frac{\vec{\nu}}{v_t \odot (K^\top u_t)}.$$

This can also be verified by taking the limit $\gamma \to 0$ in (5.30). Similarly,

$$\dot{u}_t = \frac{u_t}{K v_t} \odot K\left(v_t \odot \log \frac{\vec{\nu}}{v_t \odot (K^\top u_t)}\right).$$

Given u_t and v_t, we have the evolution for π^t as

$$\dot{\pi}_t = \operatorname{diag}(\dot{u}_t) K \operatorname{diag}(v_t) + \operatorname{diag}(u_t) K \operatorname{diag}(\dot{v}_t). \tag{5.38}$$

See Fig. 5.3 on page 180 for an example depicting an entropic optimal transport problem, along with the flow values at different settings. ◊

We end this section with convergence properties of the continuous-time Sinkhorn and dual Sinkhorn flows. In the next section, we see how these flows enable us to use the stochastic approximation machinery employed in previous chapters.

5.6.2. Convergence of Sinkhorn Flows

Let us now establish the convergence rate of the Sinkhorn flows. Theorem 5.12 below shows that the continuous-time Sinkhorn flow (and similarly its dual flow) converge at a rate of $O(1/t)$ to the optimal solution of (OT$^\varepsilon$). While our proof is guided by the mirror flow formalism presented in the previous section, it does not follow directly from existing results for mirror flows such as [KBB15; Tze+23]. This is primarily due to the presence of the additional constraint set $\mathcal{C}$, which is absent in conventional mirror flow analyses.

▶ **Theorem 5.12.** *Starting from $\pi^0 \in \Pi_{c,\varepsilon} \cap \Pi(\mu, *)$, consider the Sinkhorn flow $(\pi^t)_{t\geq 0}$ and the corresponding dual flow $(g_t)_{t\geq 0}$, as defined in (5.34) and (5.35). Then, for any $t > 0$,*

$$F(\pi^t) \leq \frac{D_{\varphi^\star_{\mathcal{C}}}(g_0 \,|\, g_{\mathrm{opt}})}{t} = O\big(t^{-1}\big),$$

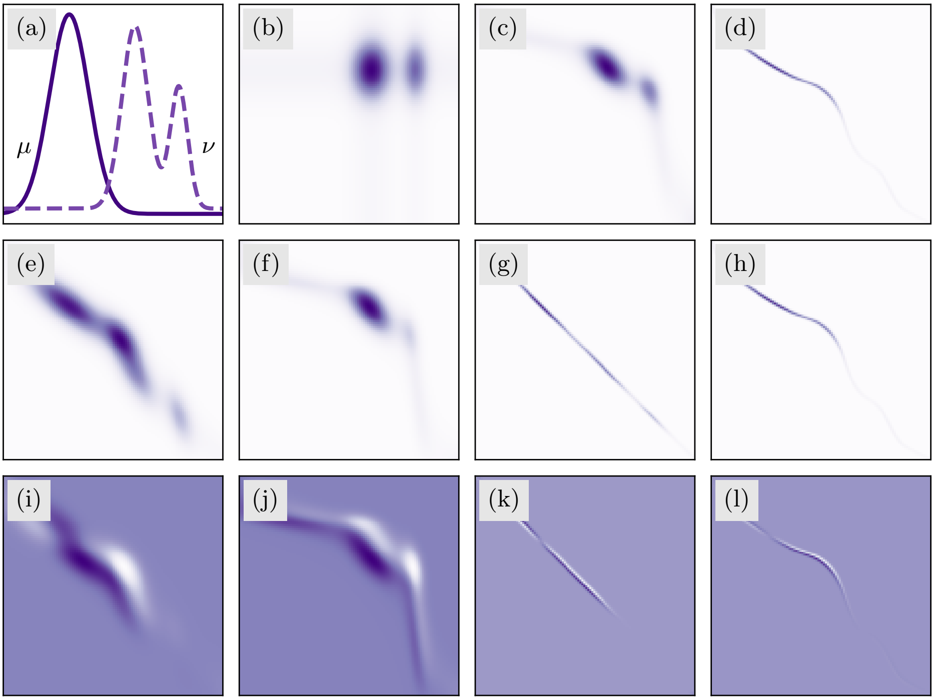


Figure 5.3. An entropic optimal transport problem for measures μ and ν shown in (a). The measure μ is a Gaussian and ν is a mixture of two Gaussians. The product measure $\mu \otimes \nu$ is shown in (b). Figures (c) and (d) show the optimal coupling for $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-4}$, respectively. Observe that for larger ε, the optimal solution is "blurry", while for smaller ε, the optimal solution looks more like an optimal transport map. The next two rows show a coupling π on top and the corresponding time derivative $\dot{\pi}$, as in (5.38), underneath it. (e) and (i): For $\varepsilon = 10^{-2}$ at initialization. (f) and (j): For $\varepsilon = 10^{-2}$ after 10 iterations of Sinkhorn. (g) and (k): For $\varepsilon = 10^{-4}$ at initialization. (h) and (l): For $\varepsilon = 10^{-4}$ after 100 iterations of Sinkhorn. In all the figures, except (a), the vertical axis is $\mathcal{X}$ and the horizontal is $\mathcal{Y}$.

where g_{opt} is a Schrödinger potential of the optimal coupling for $(\text{OT}^{\varepsilon})$*. In other words, the $\mathcal{Y}$-marginal of π^t converges (in relative entropy) to ν with a rate of $1/t$.*

Proof. First, we show that the objective function F is decreasing along the flow:

$$\begin{aligned}\frac{d}{dt}F(\pi^t) &= \left\langle \nabla F(\pi^t), \frac{d}{dt}\pi^t \right\rangle = \left\langle \nabla F(\pi^t), \pi^t \frac{d}{dt}\log \pi^t \right\rangle \\ &= \int_{\Omega} \log \frac{d\pi^t_y}{d\nu}(y) \left[\int_{\mathcal{Y}} \log \frac{d\pi^t_y}{d\nu}(y')\, d\pi^t(y' \mid x) \right] d\pi^t(x,y) \\ &\qquad - \int_{\Omega} \left(\log \frac{d\pi^t_y}{d\nu}(y) \right)^2 d\pi^t(x,y). \end{aligned} \tag{5.39}$$

Defining $k(x) = \int_{\mathcal{Y}} \log \frac{d\pi^t_y}{d\nu}(y')\, d\pi^t(y' \mid x)$, we see that the first term above is

$$\begin{aligned}\int_{\Omega} \log \frac{d\pi^t_y}{d\nu}(y)\, k(x)\, d\pi^t(x,y) &= \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} \log \frac{d\pi^t_y}{d\nu}(y)\, d\pi^t(y \mid x) \right] k(x)\, d\mu(x) \\ &= \int_{\mathcal{X}} k(x)^2\, d\mu(x).\end{aligned}$$

Now, by Jensen inequality, we have

$$\begin{aligned}\int_{\mathcal{X}} k(x)^2\, d\mu(x) &= \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} \log \frac{d\pi^t_y}{d\nu}(y)\, d\pi^t(y \mid x) \right]^2 d\mu(x) \\ &\le \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} \left(\log \frac{d\pi^t_y}{d\nu}(y) \right)^2 d\pi^t(y \mid x) \right] d\mu(x) \\ &= \int_{\Omega} \left(\log \frac{d\pi^t_y}{d\nu}(y) \right)^2 d\pi^t(x,y).\end{aligned}$$

Plugging this back into (5.39) shows that

$$\frac{d}{dt}F(\pi^t) \le 0. \tag{5.40}$$

For a Schrödinger potential g and its corresponding coupling $\pi = \nabla \varphi^{\star}_{\mathcal{C}}(g)$, define

$$\begin{aligned}V(g) := D_{\varphi^{\star}_{\mathcal{C}}}(g \,|\, g_{\text{opt}}) &= \varphi^{\star}_{\mathcal{C}}(g) - \varphi^{\star}_{\mathcal{C}}(g_{\text{opt}}) - \langle \nabla \varphi^{\star}_{\mathcal{C}}(g_{\text{opt}}), g - g_{\text{opt}} \rangle \\ &= \varphi^{\star}_{\mathcal{C}}(g) - \varphi^{\star}_{\mathcal{C}}(g_{\text{opt}}) - \langle \pi^{\text{opt}}, g - g_{\text{opt}} \rangle\end{aligned} \tag{5.41}$$

and observe that $\nabla V(g) = \nabla\varphi^\star_{\mathcal{C}}(g) - \nabla\varphi^\star_{\mathcal{C}}(g_{\mathrm{opt}}) = \pi - \pi^{\mathrm{opt}}$. We treat V as a Lyapunov function for the dual mirror flow. For that, we compute

$$\frac{d}{dt}V(g_t) = \langle \nabla V(g_t), \frac{d}{dt}g_t\rangle = -\langle \pi^t - \pi^{\mathrm{opt}}, \nabla F(\pi^t)\rangle \leq F(\pi^{\mathrm{opt}}) - F(\pi^t),$$

where the inequality is due to convexity of F. Thus,

$$V(\pi^t) - V(\pi^0) = \int_0^t \frac{d}{ds}V(\pi^s)\,ds \leq \int_0^t F(\pi^{\mathrm{opt}}) - F(\pi^s)\,ds \leq t(F(\pi^{\mathrm{opt}}) - F(\pi^t)),$$

where the last inequality is because $F(\pi^t)$ is nonincreasing. Since V is a Bregman divergence, $V \geq 0$, and we obtain the result. □

Let us remark that this shows the convergence of the $\mathcal{Y}$-marginal of the flow to the correct distribution ν (the $\mathcal{X}$-marginal is always μ by construction). Given that the cost c is bounded, this convergence can be translated into a stronger convergence of the Schrödinger potentials g_t and the coupling π^t itself. While we do not consider this type of convergence in this thesis, the reader is referred to [Nut22, Thm 6.15] for general ideas regarding proving this type of result.

5.7. STOCHASTIC APPROXIMATION FOR EOT

In this section, we illustrate how to leverage our theory to enhance the convergence of Sinkhorn schemes in scenarios involving noisy and biased gradients. In entropic optimal transport, neural networks are commonly used to parameterize transport plans. The Sinkhorn iterations (Sink$_1$) require solving an (infinite-dimensional) optimization problem that is approximated via multiple stochastic gradient steps over neural networks. However, inherent stochasticity in computations can prevent convergence. This necessitates a remedy to overcome the convergence issues, as proposed by Hanzely and Richtárik [HR21].

In this section, we introduce two improvements offered by our framework to address these challenges. First, Theorem 5.13 shows that our step-sized method (Sink$_\gamma$) with constant step-size $\gamma = O(n^{-1/2})$, results in a convergence rate of $O(n^{-1/2})$, when the "stochastic gradients" remain unbiased with finite variance. Second, Theorem 5.14 establishes asymptotic last-iterate convergence when one employs noisy and biased gradient estimates. We remark that these results are theoretical, and one has to adapt them to the desired practical scenario.

5.7.1. Convergence of Sinkhorn under Noise

We start with stochastic unbiased gradients. In this case, the Sin_γ-iteration becomes

$$\pi^{n+1} = \arg\min_{\pi \in \mathcal{C}} \left\{ \langle \tilde{\nabla} F(\pi^n), \pi - \pi^n \rangle + \frac{D_\varphi(\pi \,|\, \pi^n)}{\gamma_n} \right\}, \tag{5.42}$$

where $\tilde{\nabla} F$ is an unbiased noisy estimate of the Gâteaux derivative ∇F. Theorem 5.13 shows that the ergodic averages of this sequence converges to the optimal coupling for (OT^ε), in expectation.

▶ **Theorem 5.13.** *Suppose that $\tilde{\nabla} F$ is a stochastic estimate of ∇F such that $\mathbb{E}[\tilde{\nabla} F(\pi)] = \nabla F(\pi)$ and $\mathbb{E}\,\|\tilde{\nabla} F(\pi)\|_\infty^2 \leq \sigma^2 < \infty$ for all π. Consider the iterations π^n generated by (5.42) using a fixed step-size γ. Then, with $\bar{\pi}^n := \frac{1}{n}\sum_{k=0}^{n-1} \pi^k$, we have*

$$\mathbb{E}[H(\bar{\pi}_y^n \,|\, \nu)] \leq \frac{H(\pi^{\varepsilon,\mathrm{opt}} \,|\, R_\varepsilon)}{\gamma n} + \gamma\sigma^2. \tag{5.43}$$

The proof of Theorem 5.13 is established by combining our framework with the smoothness result of [AKL22] and a classical analysis of stochastic Bregman schemes [DEH21; HR21].

Proof. Since F is convex and 1-smooth relative to φ [AKL22, Lem. 6], by [HR21, Thm. 4.5] we have

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}[F(\pi^k) - F(\pi^{\mathrm{opt}})] \leq \frac{1}{\gamma n} D_\varphi(\pi^{\mathrm{opt}} \,|\, R_\varepsilon) + \sigma^2\gamma.$$

Since $\pi^{\varepsilon,\mathrm{opt}} \in \Pi(\mu,\nu)$, it holds that $F(\pi^{\mathrm{opt}}) = H(\pi_y^{\mathrm{opt}} \,|\, \nu) = 0$. Moreover, $D_\varphi(\pi^{\mathrm{opt}} \,|\, R_\varepsilon) = H(\pi^{\mathrm{opt}} \,|\, R_\varepsilon)$. The proof follows by the convexity of F:

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}[F(\pi^k)] \leq \mathbb{E}[F(\tfrac{1}{n}\textstyle\sum_{k=0}^{n-1} \pi^k)] = \mathbb{E}[H(\bar{\pi}_y^n \,|\, \nu)]. \qquad \square$$

Suppose we intend to run (5.42) for n iterations. Then, (5.43) immediately yields an $O(n^{-1/2})$ convergence rate by choosing the step-size $\gamma = O(n^{-1/2})$.

5.7.2. Convergence of Sinkhorn under Noise and Bias

There are two significant drawbacks in Theorem 5.13. First, since the stochastic estimate $\tilde{\nabla} F$ aims to capture noise introduced during the intermediate optimization procedures for neural networks, the unbiasedness assumption is rather restrictive. Second, even if $\tilde{\nabla} F$ is unbiased, we are still required to produce an ergodic iterate

$\bar{\pi}^n$, whereas in practice, the last iterate π^n is often the most utilized. To address these issues, we leverage stochastic approximation analysis, which relies on the continuous-time convergence in Theorem 5.12.

Let $\{\pi^n\}_{n\geq 0}$ be the sequence of measures generated by (5.42) with step-sizes γ_n and a noisy and biased oracle $\tilde{\nabla} F$, and let $\{g_n\}_{n\geq 0}$ be the sequence of corresponding Schrödinger potentials. As in previous chapters, define the effective time τ_n to be $\tau_n \coloneqq \sum_{k=0}^{n} \gamma_k$, which is the time that has elapsed up to the nth iteration of the discrete-time process g_n. Using τ_n, we consider the continuous-time piecewise-linear interpolation $g(t)$ of g_n:

$$g(t) \coloneqq g_n + \frac{t - \tau_n}{\tau_{n+1} - \tau_n}(g_{n+1} - g_n), \quad t \in [\tau_n, \tau_{n+1}].$$

Note that each $g(t)$ is a function in $L^\infty(\mathcal{Y})$, and by our considerations in the beginning of Section 5.6, we can take $g(t) \in C_b(\mathcal{Y})$. We make the following standard assumptions:

▷ **Assumption 5.1.** *Let π^n and $g(t)$ be given as above. We assume that (a) ∇F is Lipschitz and bounded on a neighborhood of $(\pi^n)_{n\in\mathbb{N}}$, and (b) $(g(t))_{t\geq 0}$ is a precompact set in the topology of uniform convergence of $C_b(\mathcal{Y})$.*

It is worth highlighting that Assumption 5.1 is a relatively mild technical condition that finds applicability in a wide range of practical scenarios. For example, it remains satisfied when employing bounded and Hölder continuous neural networks to parameterize distributions with compact support as a result of Arzelà–Ascoli theorem; see, for example, [WG24].

▶ **Theorem 5.14.** *Let π^n be given as above. Suppose Assumption 5.1 holds and the step-size rule γ_n satisfies the Robbins–Monro conditions $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$. Denote by $\mathcal{F}_n$ the filtration generated by the stochastic algorithm up to iteration n, and the noise and bias by*

$$U_n \coloneqq \tilde{\nabla} F(\pi^n) - \mathbb{E}[\tilde{\nabla} F(\pi^n) \,|\, \mathcal{F}_n],$$
$$B_n \coloneqq \mathbb{E}[\tilde{\nabla} F(\pi^n) \,|\, \mathcal{F}_n] - \nabla F(\pi^n).$$

Then, the Schrödinger potential g_n converges to g_{opt} if the following holds almost surely:

$$\lim_{n\to\infty} \|B_n\|_\infty = 0 \quad \textit{and} \quad \sup_n \mathbb{E}\, \|U_n\|_\infty^2 \leq \sigma^2 < \infty. \tag{5.44}$$

Theorem 5.14 offers two advantages over Theorem 5.13. First, it replaces ergodic convergence with the more desirable last-iterate convergence. Secondly, if we consider the bias as the error during the optimization of the neural network at each step of (5.42), then Theorem 5.14 allows for a level of flexibility where the precision of the intermediate steps may progressively improve, instead of always requiring perfect optimization as stipulated by the unbiased assumption in Theorem 5.13. However, we acknowledge that these advantages come at the cost of losing a non-asymptotic rate.

Proof. Assumption 5.1 and (5.44) ensure that $g(\cdot)$ is a precompact asymptotic pseudo-trajectory of the associated continuous-time dual flow given in (5.35); the proof is the same as the Euclidean case in Section 2.5. It follows from this association that the iterates $(g_n)_{n\geq 0}$ converge almost surely to an internally chain-transitive set of the dual flow. On the other hand, within the course of our proof for Theorem 5.12, we have established the existence of a Lyapunov function for the dual flow $V(g) := D_{\varphi^\star_C}(g \mid g_{\mathrm{opt}})$; see (5.41). Consequently, Theorem 2.10 implies that the only possible internally chain-transitive set is the set $\{g_{\mathrm{opt}}\}$. This, in turn, implies that almost surely, $g_n \to g_{\mathrm{opt}}$ in $C_b(\mathcal{Y})$. This immediately implies the weak convergence of π^n_y to ν, almost surely. □

5.8. SCHRÖDINGER BRIDGES

In Sections 5.5 and 5.6, we established the step-sized and continuous-time variants of the Sinkhorn algorithm for solving (OT^ε), which pertains to the "static" entropic optimal transport. In this section, we broaden our scope to encompass the dynamic scenario, often referred to as the Schrödinger bridge problem. Beyond adapting the results in Sections 5.5 and 5.6 to Schrödinger bridges, we provide additional insights by demonstrating that each time point in both the step-sized algorithm and the continuous-time flow can be characterized as a stochastic differential equation with a well-defined drift formula.

5.8.1. Review of Schrödinger Bridges

We start by reviewing some basic properties of the SB problem. Most of the material here are borrowed from the survey of Léonard [Léo13a] and the interested reader is referred to that survey for a more in-depth exposition.

By a *path measure* we mean a positive measure on the space of continuous functions. For instance, consider the space $\Omega = C([0,1];\mathbb{R}^d)$ of all continuous $\mathbb{R}^d$-valued functions on the interval $[0,1]$. Then, a stochastic process with almost sure continuous sample paths (such as a Brownian motion or any Langevin-type

SDE) induces a probability measure on Ω. The σ-algebra on Ω is generated by time projections:

$$X_t(\omega) := \omega_t, \quad \omega = (\omega_s)_{s\in[0,1]} \in \Omega, \quad t \in [0,1].$$

The mapping $X = (X_t)_{t\in[0,1]}$ is sometimes called the *canonical process*. Moreover, the topology of uniform convergence turns Ω into a Polish space. We denote by $\mathcal{P}(\Omega)$ the space of probability measures on Ω.

Given two probability measures μ_0, μ_1 on $\mathbb{R}^d$, the *Schrödinger bridge* (SB) problem refers to the following entropy minimization over the space of all path measures over $[0,1]$:

$$\min_{P\in\mathcal{P}(\Omega)} \{H(P \,|\, R) : P_0 = \mu_0,\, P_1 = \mu_1\}. \tag{$\mathrm{S_{dyn}}$}$$

Here, $P_t(\cdot) := P(X_t \in \cdot)$ is the marginal of P at time t, and R is a given path measure, referred to in the sequel as the *reference measure*. It is noteworthy to mention that ($\mathrm{S_{dyn}}$) is a convex optimization problem with convex constraints. Moreover, since $H(\cdot \,|\, R)$ is strictly convex, the solution of ($\mathrm{S_{dyn}}$) is unique, if it exists.

It turns out that solving ($\mathrm{S_{dyn}}$) is intimately related to solving the static Schrödinger problem ($\mathrm{S_{static}}$). Concretely, if $\hat{P}$ is the optimal solution of ($\mathrm{S_{dyn}}$), then $\hat{\pi} := \hat{P}_{01}$ is the optimal solution of

$$\min_{\pi\in\Pi(\mu_0,\mu_1)} H(\pi \,|\, R_{01}), \tag{5.45}$$

which is the same as ($\mathrm{S_{static}}$). Notice the notation $\hat{P}_{01}(\cdot) := \hat{P}((X_0, X_1) \in \cdot)$. Moreover, $\hat{P}$ disintegrated over its marginals at times 0 and 1 has the form:

$$\hat{P}(\cdot) = \int_{\mathbb{R}^d\times\mathbb{R}^d} R^{xy}(\cdot)\, d\hat{\pi}(x,y),$$

where $R^{xy}(\cdot) := R(\cdot \mid X_0 = x, X_1 = y)$. Note that this disintegration also shows how to construct the optimal solution to ($\mathrm{S_{dyn}}$) via a solution of ($\mathrm{S_{static}}$): The optimal solution to the ($\mathrm{S_{dyn}}$) has the same "bridges" (i.e., R^{xy}) as the reference measure, and the bridges are mixed via the optimal solution of the static Schrödinger problem.

A common choice for R is the law of a *reversible Brownian motion*. It is the Brownian motion whose forward dynamics is as usual, but its random initial position is "uniformly distributed on $\mathbb{R}^d$." In other words, $R(\cdot) = \int_{\mathbb{R}^d} \mathcal{W}^x(\cdot)\, dx$, where $\mathcal{W}^x$ is the law of the (usual) Brownian motion started at x. In some sense, the reversible Brownian motion can model a Brownian motion for which we lack knowledge about the starting position. Note that R has infinite mass and needs special treatment; we do not deal with the reversible Brownian motion in this thesis and only bring it as an example. We refer the interested reader to [Léo13a; Léo13b] for more details regarding unbounded path measures.

▷ **Example.** Let R be a reversible Brownian motion with diffusion parameter σ; the initial law is the Lebesgue measure on $\mathbb{R}^d$ and the forward dynamics is given by the law of a Brownian motion with diffusion matrix σI. It is straightforward to see that the joint distribution of times 0 and 1 satisfies

$$dR_{01}(x, y) \propto \exp(-\|x-y\|^2/2\sigma^2)\, dx\, dy.$$

Therefore, (5.45) becomes an instance of EOT with marginals μ_0 and μ_1, cost $c(x,y) = \frac{1}{2}\|x-y\|^2$, and $\varepsilon = \sigma^2$. This shows that ($\mathrm{S}_{\mathrm{dyn}}$) can be viewed as the dynamic formulation of (OT^ε) where, instead of merely seeking an optimal coupling, one solves for an entire stochastic process that transforms μ_0 into μ_1. ◁

5.8.2. IPF and its Interpretation as Mirror Descent

The classical algorithm for solving ($\mathrm{S}_{\mathrm{dyn}}$) is the *iterative proportional fitting* (IPF) procedure, which is the dynamic version of the Sinkhorn scheme: Starting from $P^{(0)} = R$, define for $n \geq 0$,

$$\begin{aligned} P^{(n+1/2)} &= \arg\min\{H(P \,|\, P^{(n)}) : P_1 = \mu_1\} \\ P^{(n+1)} &= \arg\min\{H(P \,|\, P^{(n+1/2)}) : P_0 = \mu_0\}. \end{aligned} \tag{IPF$_1$}$$

Similar to the Sinkhorn algorithm, we show that IPF can be interpreted through the lens of mirror descent. Specifically, we show that it is equivalent to an MD iteration with constant step-size 1. This finding serves as the dynamic counterpart to [AKL22, Prop. 5]. The proof is similar to the case of Sinkhorn and is omitted; see Section 5.5.1.

Proposition 5.15. *The iterations $P^{(n)}$ of* (IPF$_1$) *satisfy*

$$P^{(n+1)} = \arg\min_{P \in \mathcal{C}}\{\langle \nabla F(P^{(n)}), P - P^{(n)}\rangle + D_\varphi(P \,|\, P^{(n)})\}, \tag{5.46}$$

with $F(P) := H(P_1 \,|\, \mu_1)$, $\varphi(P) := H(P \,|\, R)$, and $\mathcal{C} := \{P : P_0 = \mu_0\}$.

Upon recognizing that IPF can be interpreted as MD iterations with a step-size of 1, we can proceed to investigate the MD iteration (5.46) with an arbitrary step-size γ_n:

$$P^{(n+1)} = \arg\min_{P \in \mathcal{C}} \left\{ \langle \nabla F(P^{(n)}), P - P^{(n)} \rangle + \frac{D_\varphi(P \,|\, P^{(n)})}{\gamma_n} \right\}. \tag{5.47}$$

A similar calculation to that of Lemma 5.6 reveals that (5.47) can be equivalently expressed as:

$$\begin{aligned} P^{(n+1/2)} &= \arg\min_{P_1 = \mu_1} \{ H(P \,|\, P^{(n)}) \}, \\ P^{(n+1)} &= \arg\min_{P_0 = \mu_0} \{ \gamma_n \, H(P \,|\, P^{(n+1/2)}) + (1 - \gamma_n) \, H(P \,|\, P^{(n)}) \}. \end{aligned} \tag{IPF$_\gamma$}$$

In analogy to the Sin$_\gamma$-iteration, we call the update rule above the IPF$_\gamma$*-iteration*.

5.8.3. The SDE Representation of IPF$_\gamma$-iterates

In this section, we assume that the reference measure R is induced by the law of the SDE

$$dZ_t = b_t(Z_t)\, dt + \sigma \, dW_t. \tag{5.48}$$

We first recall the important fact that for this reference measure, the IPF iterates can be expressed in terms of the *time-reversal* of SDEs, whose drift terms can be computed in practice via score-matching techniques. Next, we show a similar property for IPF$_\gamma$-iterates, making them amenable to computations. For this, we draw connections to stochastic optimal control, resulting in SDEs killed at random times. Before diving deep into the results of this section, let us make a short digression, discussing three important properties of diffusions.

Digression: Time-Reversal and Relative Entropy of Diffusions

Suppose $(Z_t)_{t \in [0,1]}$ is a *diffusion process* in $\mathbb{R}^d$, that is, a solution of the SDE

$$dZ_t = v_t(Z_t)\, dt + \sigma_t(Z_t)\, dW_t,$$

where $(W_t)_{t \in [0,1]}$ is a standard Brownian motion in $\mathbb{R}^d$. Let $\overline{Z}_t := Z_{1-t}$ be the time-reversed process. It turns out that under mild conditions, the reverse process $\overline{Z}_t$ is still a diffusion process with explicit drift and diffusion coefficients. Below, we bring a simplified version of this result for the case of constant diffusion $\sigma_t(z) \equiv \sigma$.

Theorem 5.16 (HP86, Thm. 2.1)**.** *Let $(Z_t)_{t \in [0,1]}$ be the strong solution of $dZ_t = v_t(Z_t)\, dt + \sigma \, dW_t$, and assume Z_t has a density ϱ_t for all $t \in [0,1]$. Then, under some regularity conditions, the reverse process $\overline{Z}_t := Z_{1-t}$ is a Markov diffusion* *process: Defining the drift*

$$w_t(x) = -v_{1-t}(x) + \sigma^2 \nabla \log \varrho_{1-t}(x),$$

there exists a Brownian motion $\overline{W}_t$ in some probability space such that $\overline{Z}_t$ is a solution to

$$d\overline{Z}_t = w_t(\overline{Z}_t)\, dt + \sigma\, d\overline{W}_t, \quad \overline{Z}_0 \sim \varrho_1. \tag{5.49}$$

See [HP86, (A)] for sufficient regularity conditions. Let us remark further that the reverse diffusion is a solution to a martingale problem, and hence, is a weak solution; it lives in a (possibly) different probability space. We refer the reader to [BGL14, Ch. 1] for the notions of the martingale problem and weak solutions.

Remark. A rigorous proof for Theorem 5.16 is not straightforward. We note that the time-reversal of a Markov process remains a Markov process. The proof involves demonstrating that the infinitesimal generator of this Markov process coincides with that of the time-reversed SDE (5.49). However, in this context, we will not follow this argument. Instead, we provide some intuition through heuristic computations.

Let $\delta \ll 1$. We compute the conditional law of Z_t given $Z_{t+\delta} = z_{t+\delta}$ for time-homogeneous drift v and constant diffusion σ. By the Bayes theorem, we have

$$\mathbb{P}(Z_t \in dz \mid z_{t+\delta}) \propto \varrho_t(z) \exp\Bigl\{-\frac{\|z_{t+\delta} - (z + \delta v(z))\|^2}{2\sigma^2\delta}\Bigr\},$$

where we approximated $v(z_{t+s}) = v(z)$ for $0 \le s \le \delta$. Using the Taylor approximation of $\log \varrho_t$ around $z_{t+\delta}$, we have

$$\varrho_t(z) \approx \varrho_t(z_{t+\delta}) \exp \langle \nabla \log \varrho_t(z_{t+\delta}), z - z_{t+\delta} \rangle.$$

Therefore,

$$\mathbb{P}(Z_t \in dz \mid z_{t+\delta}) \propto \exp\Bigl\{-\frac{\|z - (z_{t+\delta} - \delta v(z_{t+\delta}) + \sigma^2\delta \nabla \log \varrho_t(z_{t+\delta}))\|^2}{2\sigma^2\delta}\Bigr\},$$

which corresponds to the transition of an SDE with the same diffusion σ and drift

$$-v(z) + \sigma^2 \nabla \log \varrho_t(z). \qquad \diamondsuit$$

Next, we focus on absolute continuity of laws of diffusions. Let us begin with an illuminating example:

▷ **Example.** Suppose W_t is a standard Brownian motion on $\mathbb{R}$, and let P and Q be the laws of $(W_t)_{0 \le t \le 1}$ and $(2W_t)_{0 \le t \le 1}$, respectively. We claim that Q is not absolutely continuous with respect to P. To see this, consider the event A that consists of continuous functions f satisfying $\limsup_{t\downarrow 0} f(t)/(2t\log\log t)^{1/2} = 1$. Then, by the law of the iterated logarithm of Lévy, it holds that $Q(A) = 0$ and $P(A) = 1$. In other words, by just looking at one sample path (i.e., one trajectory), we can decide whether it is from a Brownian motion or of the twice a Brownian motion by checking if it is in A or not. ◁

The Cameron–Martin–Girsanov theorem states under which conditions the law of two semimartingales are absolutely continuous with respect to each other, and provides a formula for their Radon–Nikodym derivative. We bring here a simplified version of a general Girsanov formula in [Léo12, Thm. 1].

Theorem 5.17 (Girsanov)**.** *Let P be a path measure, under which the canonical process $(X_t)_{t\geq 0}$ has the semimartingale decomposition*

$$X_t = X_0 + \int_0^t b_s \, ds + \sigma W_t,$$

with W_t being a Brownian motion. Suppose the path measure Q is absolutely continuous with respect to P and $H(Q \,|\, P) < \infty$. Then there exists an $\mathbb{R}^d$-valued adapted process β_t with $\mathbb{E}_Q[\int_0^1 \|\beta_t\|^2 \, dt] < \infty$, such that X has the semimartingale decomposition

$$X_t = X_0 + \int_0^t (b_s + \beta_s) \, ds + \sigma W_t^Q, \quad Q\text{-a.s.},$$

where W^Q is a Q-Brownian motion. Moreover,

$$\frac{dQ}{dP}(\omega) = \frac{dQ_0}{dP_0}(\omega_0) \cdot \exp\left\{ \frac{1}{\sigma} \int_0^1 \langle \beta_t, dW_t \rangle - \frac{1}{2\sigma^2} \int_0^1 \|\beta_t\|^2 \, dt \right\}.$$

As a corollary, the last result gives an expression for the relative entropy of the law of two Markov diffusion processes:

Corollary 5.18. *Let P and Q be two path measures with $H(Q \,|\, P) < \infty$. Moreover, assume that under P, the canonical process X has the semimartingale decomposition*

$$X_t = X_0 + \int_0^t b_s \, ds + \sigma W_t^P,$$

where W^P is a P-Brownian motion, and under Q,

$$X_t = X_0 + \int_0^t c_s \, ds + \sigma W_t^Q,$$

where W^Q is a Q-Brownian motion. Then,

$$H(Q \,|\, P) = \mathbb{E}_Q\left[\log \frac{dQ}{dP}\right] = H(Q_0 \,|\, P_0) + \frac{1}{2\sigma^2}\, \mathbb{E}_Q\left[\int_0^1 \|c_t - b_t\|^2 \, dt\right].$$

After this brief digression, let us return to our original problem of representing the iterations of IPF and IPF$_\gamma$ as the laws of SDEs with explicit drifts. We begin by the first step of (IPF$_1$) (which is shared between IPF and IPF$_\gamma$). This result is already established in [De +21, Prop. 6]; we provide a proof for completeness.

Theorem 5.19. *Suppose $P^{(n)}$ is the law of the SDE*

$$dZ_t = v_t(Z_t)\, dt + \sigma \, dW_t, \quad Z_0 \sim \mu_0, \tag{5.50}$$

and the time-reversal of $P^{(n+1/2)}$ is given by the SDE

$$dY_t = w_{1-t}(Y_t)\, dt + \sigma \, d\overline{W}_t, \quad Y_0 \sim \mu_1. \tag{5.51}$$

Then the drift w_t satisfies

$$w_t(x) = -v_t(x) + \sigma^2 \nabla \log \varrho_t^n(x), \tag{5.52}$$

where ϱ_t^n is the density of $P_t^{(n)}$.

Proof. Let $\bar{P}^{(n)}$ be the law of the time-reversal of $P^{(n)}$. Since the time-reversal of $P^{(n+1/2)}$ solves $\arg\min\{H(P \,|\, \bar{P}^{(n)}) : P_0 = \mu_1\}$, its SDE representation is the same as the one for $\bar{P}^{(n)}$ with its initial distribution set to μ_1. By the time-reversal formula (Theorem 5.16), $\bar{P}^{(n)}$ corresponds to

$$dY_t = \big(-v_{1-t}(Y_t) + \sigma^2 \nabla \log \varrho_{1-t}^n(Y_t)\big)\, dt + \sigma \, dW_t, \qquad Y_1 \sim \mu_0,$$

where ϱ_t^n is the density of $P_t^{(n)}$. This means that this should coincide with the SDE for time reversal of $P^{(n+1/2)}$, giving the result of the theorem. □

We are now ready to state the main result of this section: the drift formula of the SDE representation for the second step of γ-IPF. For pedagogical reasons, we walk through the proof and state the final result in the end. The impatient reader is referred to Theorem 5.20 for the statement of the theorem we prove below.

Let $P^{(n)}$ be given by the scheme (IPF_γ), and let v_t^n be the (forward) drift corresponding to the SDE representation of $P^{(n)}$. Theorem 5.19 together with Theorem 5.16 show that the path measure $P^{(n+1/2)}$ corresponds to the time-reversal of the SDE (5.51), which is a process with the drift

$$v_t^{n+1/2} := -w_t + \sigma^2 \nabla \log \varrho_t^{n+1/2} = v_t^n + \sigma^2 \nabla \log \frac{\varrho_t^{n+1/2}}{\varrho_t^n} = v_t^n + \sigma^2 \nabla \log \ell_t^n, \tag{5.53}$$

with $\varrho_t^{n+1/2}$ being the density of $P_t^{(n+1/2)}$ and $\ell_t^n := \varrho_t^{n+1/2}/\varrho_t^n$.

The next step of (IPF_γ) is given by

$$P^{(n+1)} = \underset{P_0=\mu_0}{\arg\min}\{\gamma_n \, H(P \,|\, P^{(n+1/2)}) + (1-\gamma_n)\, H(P \,|\, P^{(n)})\}.$$

Let us make the ansatz that $P^{(n+1)}$ corresponds to an SDE of the form

$$dX_t^u = (b_t^\gamma(X_t^u) + u_t)\, dt + \sigma \, dW_t, \quad X_0^u \sim \mu_0, \tag{5.54}$$

where we define the drift b_t^γ as

$$b_t^\gamma := \gamma v_t^{n+1/2} + (1-\gamma) v_t^n,$$

which by (5.53) is equal to

$$b_t^\gamma = v_t^n + \gamma \cdot \sigma^2 \nabla \log \ell_t^n.$$

The reason that we take such SDE representation for $P^{(n+1)}$ is that, firstly, it should be a diffusion with the same diffusion coefficient to be absolutely continuous with respect to $P^{(n)}$, and its drift shall be a "weighted average" of the drifts of $P^{(n+1/2)}$ and $P^{(n)}$ with some correction u_t. Notice that this ansatz only helps us to embed our intuition in the formulation of the drift and imposes no restrictions; u_t can be any adapted process. Our goal in what follows is to find u_t.

By the relative entropy formula for diffusions (Corollary 5.18), we have for $P = P^{(n+1)}$,

$$H(P \,|\, P^{(n)}) = \frac{1}{2\sigma^2} \mathbb{E}_P \left[\int_0^1 \| b_t^\gamma (X_t^u) + u_t - v_t^n(X_t^u) \|^2 \, dt \right],$$

since $P_0 = P_0^{(n)} = \mu_0$. Likewise,

$$H(P \mid P^{(n+1/2)}) = H(\mu_0 \mid P_0^{(n+1/2)}) \\ + \frac{1}{2\sigma^2}\,\mathbb{E}_P\left[\int_0^1 \|b_t^\gamma(X_t^u) + u_t - v_t^{n+1/2}(X_t^u)\|^2\,dt\right].$$

Taking the weighted average of the two equations above and noticing that $H(\mu_0 \mid P_0^{(n+1/2)})$ is merely a constant not depending on P, we see that

$$\begin{aligned}&\gamma H(P \mid P^{(n+1/2)}) + (1-\gamma)H(P \mid P^{(n)}) \\ &\quad\doteq \frac{1}{\sigma^2}\,\mathbb{E}_P\left[\int_0^1 \frac{1}{2}\|u_t\|^2\,dt + \frac{\gamma(1-\gamma)}{2}\int_0^1 \|v_t^{n+1/2}(X_t^u) - v_t^n(X_t^u)\|^2\,dt\right] \\ &\quad= \frac{1}{\sigma^2}\,\mathbb{E}_P\left[\int_0^1 \frac{1}{2}\|u_t\|^2\,dt + \frac{\sigma^4\gamma(1-\gamma)}{2}\int_0^1 \|\nabla \log \ell_t^n(X_t^u)\|^2\,dt\right].\end{aligned}$$

Therefore, the minimization problem in (IPF$_\gamma$) reduces to the following: Find an adapted process $(u_t)_{t\in[0,1]}$ such that the following *cost functional* is minimized:

$$J[u] \coloneqq \mathbb{E}_P\left[\int_0^1 \frac{1}{2}\|u_t\|^2 + c_t(X_t^u)\,dt\right], \tag{5.55}$$

where $c_t(x) \coloneqq \sigma^4\gamma(1-\gamma)\|\nabla \log \ell_t^n(x)\|^2/2$. It is not hard to see that this is an instance of a stochastic optimal control problem with zero terminal cost, where u is "controlling" the stochastic process X_t^u such that the Stein score is minimized while spending the least amount of energy. The structure of this control problem suggests that the optimal control is of feedback type and is equal to the negative gradient of the value function. It therefore remains to find the value function.

The value function $V_t(x)$ of the optimal control problem shall satisfy the Hamilton–Jacobi–Bellman (HJB) equation, which writes as

$$\partial_t V_t(x) + \min_{u\in\mathbb{R}^d}\left\{\langle b_t^\gamma(x) + u, \nabla V_t(x)\rangle + \frac{\sigma^2}{2}\Delta V_t(x) + \frac{1}{2}\|u\|^2 + c_t(x)\right\} = 0,$$

along with $V_1(x) = 0$ for all $x \in \mathbb{R}^d$. The optimal value u^* of the inner optimization problem is $u^*(t,x) = -\nabla V_t(x)$, asserting that the optimal control is of feedback type. Plugging this value in the HJB equation gives

$$\partial_t V_t(x) - \frac{1}{2}\|\nabla V_t(x)\|^2 + \frac{\sigma^2}{2}\Delta V_t(x) + \langle b_t^\gamma(x), \nabla V_t(x)\rangle + c_t(x) = 0, \quad V_1(x) = 0.$$

Inspired by Fleming's logarithmic transformation [Fle77, Sec. 2], let us make the change of variables $V_t(x) = -\sigma^2 \log E_t(x)$ in the equation to get

$$\partial_t E_t(x) + \frac{\sigma^2}{2}\Delta E_t(x) + \langle b_t(x), \nabla E_t(x)\rangle = \frac{1}{\sigma^2} E_t(x)\, c_t(x), \quad E_1(x) = 1. \tag{5.56}$$

This PDE admits a probabilistic representation similar to Feynman–Kac formula. Following Pra and Pavon [PP90], we see that

$$E_t(x) = \mathbb{E}^{t,x}\left[\exp\left(-\frac{1}{\sigma^2}\int_t^1 c_s(Y_s)\, ds\right)\right],$$

where $(Y_s)_{s\in[t,1]}$ is the solution to the uncontrolled SDE $dY_t = b_t^\gamma(Y_t)\, dt + \sigma\, dW_t$ and $\mathbb{E}^{t,x}$ is expectation with respect to the law of Y started at time t from x.

Remark. One has to invoke an appropriate verification theorem to ensure the optimality of the mentioned control. This turns out to be straightforward for the stochastic optimal control problem (5.55); we refer the reader to [ØS19]. ◊

Putting everything together, we thus have proved the main theorem of this section, stated below:

▶ **Theorem 5.20.** *Let $P^{(n)}$ be given by the scheme (IPF$_\gamma$), and let v_t^n be the (forward) drift corresponding to its SDE representation. Then v_t^n satisfies the following recursion:*

$$v_t^{n+1}(x) = v_t^n(x) + \gamma\cdot\sigma^2\nabla\log\ell_t^n(x) - \nabla V_t^n(x), \tag{SDE$_\gamma$}$$

where

$$V_t^n(x) = -\sigma^2\log\mathbb{E}^{t,x}\left[\exp\left\{-\frac{\sigma^2\gamma(1-\gamma)}{2}\int_t^1 \left\|\nabla\log\ell_s^n(Y_s)\right\|^2 ds\right\}\right], \tag{5.57}$$

and the expectation is with respect to the law of the process $(Y_s)_{s\in[t,1]}$ starting at $Y_t = x$ and following the uncontrolled SDE

$$dY_s = \big(v_s^n(Y_s) + \gamma\cdot\sigma^2\nabla\log\ell_s^n(Y_s)\big)\, ds + \sigma\, dW_s. \tag{5.58}$$

When $\gamma_n \equiv 1$, the ∇V_t term in (SDE$_\gamma$) disappears, and the result of Theorem 5.20 becomes Theorem 5.19 applied twice, recovering the iterative formula for the SDE representation of IPF [De +21, Prop. 6].

We close this section by mentioning a few remarks about the proof, computational aspects of Theorem 5.20, as well as a formal flow corresponding to the IPF$_\gamma$-iterations.

Computational aspects

Although this thesis focuses on the theoretical understanding of the Sinkhorn and IPF iterates, let us briefly remark how the formula in Theorem 5.20 admits a practical implementation. To see this, notice that the $\nabla \log \ell_t^n$ term in (SDE_γ) is the standard *Stein score* ratio that can be estimated by various diffusion models and is present in most practical training procedures of SB. On the other hand, the computation of the additional term involving V_t needs some extra treatment.

A common practice is to connect the value function via the Feynman–Kac formula to SDEs with *killing*. Concretely, since the cost c_t in (5.56) is non-negative, one can simulate the uncontrolled SDE (5.58), and kill it at a rate $c_t/\sigma^2 = \frac{\gamma(1-\gamma)}{2}\|\nabla \log \ell_t^n\|^2$, that is,

$$\mathbb{P}[Y_{t+h} \text{ is killed} \mid Y_t] = \frac{\gamma(1-\gamma)}{2}\|\nabla \log \ell_t^n(Y_t)\|^2 + o(h).$$

These procedures are already employed in the SB community in other contexts [Liu+22; Par+23].

Another way is to use the Girsanov theorem (Theorem 5.17) and formulate V_t in terms of expectations with respect to a standard Brownian motion. That is,

$$\begin{aligned} V_t(x) = -\log \mathbb{E}\Bigg[\exp\Bigg(\frac{1}{\sigma}\int_t^1 \langle b_s^\gamma(x+\sigma W_{s-t}), dW_{s-t}\rangle \\ -\frac{1}{2\sigma^2}\int_t^1 \|b_s^\gamma(x+\sigma W_{s-t})\|^2 + c_s(x+\sigma W_{s-t})\, ds\Bigg)\Bigg], \end{aligned}$$

where the expectation is with respect to a standard Brownian motion $(W_t)_{t\geq 0}$. Given $\nabla \log \ell_t^n$, which is given by the usual score matching procedure in SB training, one can compute the value function using approximation techniques in control theory for integration with respect to standard Brownian motion [ZC22].

Schrödinger flows

Let us remark on how the results in this section naturally lead to a flow of SDEs, that is, an evolution of path measures $(P^s)_{s\geq 0}$ where each P^s is the law of an SDE with a certain drift v_t^s and diffusion coefficient σ.

To streamline the exposition, we make the simplifying assumption that $\sigma = 1$ and the reference measure R is given by the law of the reversible Brownian motion. Our conclusions remain applicable in the general case, with the cost of more involved notation.

Consider the static SB problem in (5.45), which is nothing but (OT^ε) with cost function $c(x,y) = \frac{1}{2}\|x-y\|^2$ and $\varepsilon = 1$. Let f^s and g^s be the Schrödinger potentials of the Sinkhorn flow (5.34). For each $s \geq 0$, define the path measures P^s on $\Omega = C([0,1];\mathbb{R}^d)$ by

$$\frac{dP^s}{dR}(\omega) = \exp(f^s(\omega_0) + g^s(\omega_1)), \quad \omega \in \Omega. \tag{5.59}$$

Similar to the static case, these path measures are known to solve the SB problem for their corresponding marginals μ_0^s, μ_1^s [Léo13a, Thm. 2.5] and, since f^s and g^s come from the Sinkhorn flow, $\mu_0^s = \mu_0$ for all $s \geq 0$.

We can now formally define an evolution of the path measures P^s, where at each time s, P^s admits an SDE representation which can be described using the Schrödinger potentials f^s, g^s: For each s, define the function G^s on $[0,1] \times \mathbb{R}^d$ by

$$G^s(t,z) := \log \mathbb{E}[\exp(g^s(R_1)) \,|\, R_t = z] \tag{5.60}$$

so that $G^s(1,\cdot) = g^s(\cdot)$. Then Léonard [Léo13a, Prop. 6] implies that P^s is the law of the SDE:

$$dX_t^s = \nabla G^s(t, X_t^s)\, dt + dW_t, \quad X_0^s \sim \mu_0. \tag{5.61}$$

As a result, the mapping $s \mapsto (G^s(t,\cdot))_{t\in[0,1]}$ can be regarded as the dynamic dual Sinkhorn flow associated with $(g^s)_{s\geq 0}$, while (5.61) can be considered as the continuous-time limit of the SDE representation of (IPF_γ), as $\gamma \to 0$.

5.9. CONCLUSIONS

In summary, this chapter introduced the continuous-time Sinkhorn algorithm as a novel approach to design schemes that maintain convergence in the presence of noise and bias. We extend these insights to Schrödinger bridges and the Iterative Proportional Fitting procedure. Our work paves the way for several promising avenues for future research. For instance, one direction involves exploring connections with other existing dynamics, such as those introduced by Conforti, Lacker, and Pal [CLP23], the Wasserstein mirror flow of Deb et al. [Deb+23], and mean-field Schrödinger dynamics of Claisse et al. [Cla+23]. The connection to the mirror flow also opens up the possibility of introducing momentum terms to achieve *acceleration* [KBB15; WWJ16]. These questions offer rich opportunities for further investigation.

BIBLIOGRAPHIC NOTES

The Sinkhorn algorithm was introduced by Sinkhorn and Knopp [SK67]. A modern account with additional historical notes can be found in the book of Peyré and Cuturi [PC20]. Traces of the iterative proportional fitting procedure can be found in the works of Fortet [For40] and Kullback [Kul68]. Classical analysis of the Sinkhorn algorithm—seen as an alternating projection method—can be found in, e.g., [Cut13; CGP16; PC20; GN22].

Mirror Descent was introduced by Nemirovsky and Judin [NJ83]. Later, Beck and Teboulle [BT03] showed that it is essentially a nonlinear projected subgradient method; this is the viewpoint we chose in the chapter.

For applications of Schrödinger Bridges in sampling, see [Ber+19; Hua+21]. For application in generative modeling see [Bor+21; CLT21; Wan+21a]. For molecular biology applications see, e.g., [Hol+23], and for single-cell dynamics see, e.g., [Bun+23] and references therein. For applications in mean-field games and deep reinforcement learning, see [Liu+22].

CHAPTER SIX

CONCLUSION AND OUTLOOK

This thesis has presented a comprehensive investigation into stochastic approximation algorithms through the lens of dynamical systems. By drawing on foundational theories and expanding into new contexts, this work offers both theoretical insights and practical algorithms. Below, we provide a summary of the key contributions from each chapter and outline directions for future research.

In Chapter 3, we extended the theoretical framework of Benaïm and Hirsch to Riemannian manifolds by adapting stochastic approximation algorithms to the manifold settings. Through examples in machine learning and game theory, we highlighted the practical relevance and challenges associated with nonlinear root-finding. A rather complete picture of the asymptotics of root-finding algorithms is given in the context of two main theorems: one showing the asymptotic pseudo-trajectory property, and other proving stability of the iterates in non-compact Hadamard manifolds. We also studied practical variations, namely retractions and alternations, and showed that these variations do not change the asymptotics of the algorithm.

Chapter 4 explored stochastic approximation algorithms in the Wasserstein space for analyzing discretizations of stochastic differential equations. We presented a unified framework showing that SDE discretization algorithms converge to the same limits as continuous SDEs. The results of this chapter have direct implications for a wide range of Langevin-based sampling algorithms, as well as those based on the mirror Langevin diffusion, namely last-iterate asymptotic convergence of the law of the iterates to the target distribution in Wasserstein distance.

In Chapter 5, we delved into the linear structure and convexity in the space of signed measures, focusing on the relative entropy functional. The step-sized Sinkhorn algorithm and its continuous-time counterpart were introduced, illustrating convergence to optimal solutions is possible amid noise and bias. The chapter also connected the Schrödinger Bridge problem and Iterative Proportional Fitting procedure to mirror descent techniques, offering rigorous convergence guarantees for their practical implementations.

Overall, our unified framework underscores that the convergence of various seemingly different stochastic approximation schemes can be analyzed through the deterministic dynamics of flows, provided certain criteria are met regarding noise and bias terms. This approach validates existing algorithms and facilitates the design of new ones.

Future Research Directions

While this thesis provides crucial insights and lays a solid theoretical foundation, several open questions and research directions remain:

(1) **Zeroth-Order Optimization:** Kiefer–Wolfowitz algorithms are essential in situations where it is not possible to access vector fields directly. This is particularly important in game theory and sequential online learning contexts. Although there is an extensive literature on Euclidean domains and some focused specifically on Riemannian optimization, a comprehensive theory like the one developed in Chapter 3 is lacking.

(2) **Constant Step-Size Algorithms:** Our analysis currently does not cover constant step-size stochastic approximation schemes, which are frequently used in practical applications. This is a significant limitation, especially for real-world implementations of these algorithms. Although the theory underlying constant step-size stochastic approximation differs substantially from that of diminishing step-size methods, we believe there may be potential to extend the existing Euclidean theory to the domains discussed in this thesis.

(3) **Beyond SDE-based Algorithms:** Practical sampling schemes such as Metropolis–Hastings do not immediately link to an SDE discretization. The main challenge is the “accept-reject” step, which is essentially a projection onto the set of reversible Markov processes. Incorporating these algorithms into the framework of stochastic approximation within the Wasserstein space is a promising future research direction. This theoretical approach can provide a more comprehensive understanding of sampling problems.

(4) **Continuous-Time Sinkhorn Extensions:** In our study [KHK24], we established preliminary links between the continuous-time Sinkhorn algorithm and significant dynamics including the Wasserstein mirror flow and mean-field Schrödinger dynamics. Future research could explore these connections

further and incorporate momentum terms for acceleration, thereby enhancing our comprehension of various flows in Wasserstein space. This could also aid in developing new, more efficient algorithms for solving Schrödinger bridges and entropic optimal transport problems.

In conclusion, this thesis advances the understanding and application of stochastic approximation algorithms across various complex settings. The insights gained herein suggest numerous promising avenues for future research, aiming to solve more intricate problems with robust, theoretically-backed methods.

APPENDIX A

PROOFS FOR CHAPTER 3

In this appendix, we bring missing proofs in Chapter 3. Appendix A.1 includes those results that are of a general geometric nature, and are used in the proofs and arguments in the chapter. The rest of the sections are organized by the corresponding theorems.

A.1. GENERAL GEOMETRIC RESULTS

Lemma A.1. *Let $\mathcal{U} \subset \mathcal{M}$ be a normal neighborhood around $p \in \mathcal{M}$; let $q \in \mathcal{U}$. Then, the parallel transport of a tangent vector $v \in T_p\mathcal{M}$ from p to q along the minimizing geodesic depends smoothly on p, q, and v.*

Proof. This lemma is folklore, and can possibly be found in some Riemannian geometry textbooks. We give a proof here for the sake of completeness.

We consider the normal coordinate system $\varphi : \mathcal{V} \subseteq \mathbb{R}^d \to \mathcal{U}$ centered at p. In this chart, the minimizing geodesic between p and q is the line segment $t \mapsto (tw^1, \ldots, tw^d)$, where $w = (w^1, \ldots, w^d)$ is the coordinate expression of $\exp_p^{-1}(q)$. Consider the parallel vector field $V(t)$ along this geodesic with $V(0) = v$. When expressed in φ, $V(t) = \sum_k V^k(t)\partial_i\big|_{\gamma(t)}$ satisfies the system of ODEs

$$\dot{V}^k(t) = -\sum_{i,j} \Gamma^k_{ij}(sw^1, \ldots, sw^d)\, w^i\, V^j(t), \quad k = 1, \ldots, d, \tag{A.1}$$

with initial conditions $V^k(0) = v^k$. Now define a set of new auxiliary functions W^k, $k = 1, \ldots, d$ and consider the parallel transport as the system of ODEs in $2d$

functions $(V^1, \dots, V^d, W^1, \dots, W^d)$

$$\begin{cases} \dot{V}^k(t) = -\sum_{i,j} \Gamma^k_{ij}(tW^1(t), \dots, tW^d(t))\, W^i(t)\, V^j(t) \\ \dot{W}^k(t) = 0 \end{cases} \tag{A.2}$$

with initial conditions

$$\begin{cases} W^k(0) = w^k, \\ V^k(0) = v^k. \end{cases}$$

As the solutions of smooth ODEs depend smoothly on initial conditions, as well as time, the solutions to (A.2) can be written as smooth functions $V^k(t, w, v)$ and $W^k(t, w, v)$. It follows immediately from the form of the equations that $W^k(t) \equiv w^k$, and therefore, V^k coincides with the solution of (A.1). Therefore, we obtain that $V^k(1)$, which is the parallel transport of v along the minimizing geodesic connecting p to q depends smoothly on both q and v. Smoothness in p follows by considering the parallel transport from q to p. □

Lemma A.2. *Let r be the radial distance function from a fixed point $p \in \mathcal{M}$, i.e., $r(q) = d(p, q)$. Then, for any absolutely continuous curve $\gamma : [a, b] \to \mathcal{M}$ with metric derivative $|\dot{\gamma}| \in L^1(a, b)$, it holds*

$$|r(\gamma(b)) - r(\gamma(a))| \le \int_a^b |\dot{\gamma}|(t)\, dt.$$

Proof. The result follows from [AGS05, Thm. 1.2.5] by noticing that r is 1-Lipschitz continuous. □

Lemma A.3 ([Lez20, Thm 3.12] or [CB21, Prop. A.3])**.** *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[\kappa_{\text{low}}, \kappa_{\text{up}}]$, and let $\kappa_{\max} = \max(|\kappa_{\text{up}}|, |\kappa_{\text{low}}|)$. For $v \in T_p\mathcal{M}$, consider the geodesic $\gamma(t) = \exp_p(tv)$. If γ is defined and has no interior conjugate point on the interval $[0, 1]$, then*

$$\forall w \in T_p\mathcal{M}, \qquad |(d\exp_p)_v(w) - \mathrm{P}_{p \to \gamma(1)}[w]| \le \kappa_{\max} \cdot f_{\kappa_{\text{low}}}(|v|) \cdot |w_\perp| \tag{A.3}$$

where $w_\perp := w - \frac{\langle v, w\rangle}{\langle v, v\rangle} v$ is the component of w orthogonal to v. The function $f_{\kappa_{\text{low}}}$ in (A.3) is defined as

$$f_{\kappa_{\text{low}}}(a) = \begin{cases} \frac{a^2}{6} & \text{if } \kappa_{\text{low}} = 0, \\ r^2 \left(1 - \frac{\sin(a/r)}{a/r}\right) & \text{if } \kappa_{\text{low}} = \frac{1}{r^2} > 0, \\ r^2 \left(\frac{\sinh(a/r)}{a/r} - 1\right) & \text{if } \kappa_{\text{low}} = -\frac{1}{r^2} < 0. \end{cases}$$

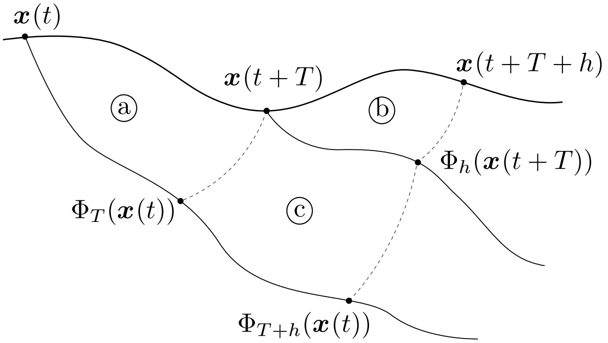

$\boldsymbol{x}(t)$
$\boldsymbol{x}(t+T)$
$\boldsymbol{x}(t+T+h)$
ⓐ
ⓑ
ⓒ
$\Phi_h(\boldsymbol{x}(t+T))$
$\Phi_T(\boldsymbol{x}(t))$
$\Phi_{T+h}(\boldsymbol{x}(t))$

Figure A.1. Induction argument for proving the asymptotic pseudo-trajectory property (see Lemma 3.7).

Moreover, the function $f_{\kappa_{\text{low}}}$ *is dominated by the case* $\kappa_{\text{low}} < 0$*; for all* $a \in \mathbb{R}_+$*,*

$$f_{\kappa_{\text{low}}}(a) \leq f_{-\kappa_{\max}}(a). \tag{A.4}$$

A.2. AUXILIARY RESULTS FOR THEOREM 3.4

Lemma 3.7. *Let* V *be a* C^1 *vector field and* Φ *be its corresponding flow. If a continuous piecewise-smooth curve* $\boldsymbol{x}$ *satisfies* (3.23) *for some* $T > 0$*, then it is an asymptotic pseudo-trajectory of the flow* Φ*.*

Proof. It is clear that (3.23) holds for all $0 < T' \leq T$. We show that it holds for $2T$, and thus concluding the lemma. First, observe that

$$\begin{aligned}\sup_{h\in[0,2T]} d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t))) \leq &\sup_{h\in[0,T]} d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t))) \\ &+ \sup_{h\in[T,2T]} d(\boldsymbol{x}(t+h), \Phi_h(\boldsymbol{x}(t))).\end{aligned}$$

See Fig. A.1 for an illustration. By the induction hypothesis, the first term vanishes as $t \to \infty$ (part ⓐ in the figure). So we only deal with the second term. We have by the triangle inequality

$$\begin{aligned}\sup_{h\in[T,2T]} d(\boldsymbol{x}(t+h),\Phi_h(\boldsymbol{x}(t))) &= \sup_{h\in[0,T]} d(\boldsymbol{x}(t+T+h),\Phi_{T+h}(\boldsymbol{x}(t)))\\ &\le \sup_{h\in[0,T]} d(\boldsymbol{x}(t+T+h),\Phi_h(\boldsymbol{x}(t+T)))\\ &\quad+\sup_{h\in[0,T]} d(\Phi_h(\boldsymbol{x}(t+T)),\Phi_{T+h}(\boldsymbol{x}(t))).\end{aligned}$$

Again, by the induction hypothesis, the first term (corresponding to part ⓑ of the figure) vanishes as $t\to\infty$. For the second term, notice that by the semigroup property of the flow,

$$d(\Phi_h(\boldsymbol{x}(t+T)),\Phi_{T+h}(\boldsymbol{x}(t))) = d(\Phi_h(\boldsymbol{x}(t+T)),\Phi_h(\Phi_T(\boldsymbol{x}(t)))).$$

The term on the right-hand side (corresponding to part ⓒ in the figure) can be bounded using Lemma A.4 by

$$\sup_{h\in[0,T]} d(\Phi_h(\boldsymbol{x}(t+T)),\Phi_{T+h}(\boldsymbol{x}(t))) \le e^{LT} d(\boldsymbol{x}(t+T),\Phi_T(\boldsymbol{x}(t)))$$

which also vanishes as $t\to\infty$ by assumption. We thus have shown the desired property. □

Lemma A.4. *Let $p,q\in\mathcal{M}$ and consider two integral curves $\Phi_s(p)$ and $\Phi_s(q)$, $s\in[0,T]$, of the flow Φ of a C^1, L-Lipschitz, and complete vector field V. Then, one has the estimate*

$$\sup_{s\in[0,T]} d(\Phi_s(p),\Phi_s(q)) \le e^{TL} d(p,q).$$

Proof. First, we recall the fact that the flow Φ, as a function of both t and p, is smooth. This follows from the fundamental theorem of flows on Riemannian manifolds [Lee12, Thm. 9.12], which shows the existence of a unique maximal smooth global flow (here, smoothness is both in time and space variables). Since for a complete vector field the flow is global, this implies that Φ is smooth everywhere.

Connect p and q by a minimizing geodesic $\gamma : [0,1]\to\mathcal{M}$. Consider the one-parameter family of curves $c : [0,T]\times[0,1]\to M$, defined as

$$c(s,t) = \Phi_s(\gamma(t)).$$

See Fig. A.2 for an illustration. As the flow Φ is globally smooth, one has that c is a smooth mapping. We denote $\partial_s c(s,t) := (dc)(\frac{\partial}{\partial s})$ and likewise for $\partial_t c$. By

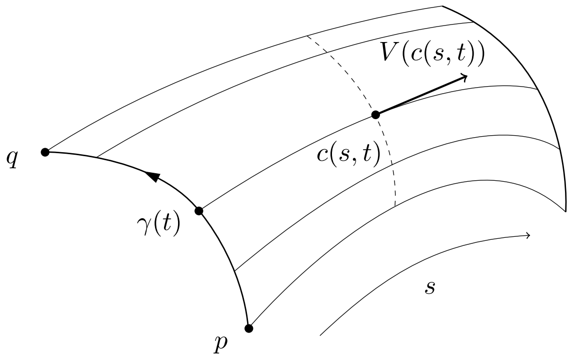


Figure A.2. Construction of a one-parameter family of curves using the flow of the vector field V starting at the geodesic γ (see Lemma A.4).

construction, we have $\partial_s c(s,t) = V(c(s,t))$. Now compute

$$\begin{aligned}
\frac{d}{ds}\frac{1}{2}|\partial_t c(s,t)|^2 &= \langle D_s \partial_t c(s,t), \partial_t c(s,t)\rangle \\
&= \langle D_t \partial_s c(s,t), \partial_t c(s,t)\rangle \qquad \text{(torsion-free)} \\
&= \langle D_t V(c(s,t)), \partial_t c(s,t)\rangle \\
&\leq |D_t V(c(s,t))| \cdot |\partial_t c(s,t)| \\
&\leq L|\partial_t c(s,t)|^2,
\end{aligned}$$

where in the last line, we used the Lipschitzness of V, in the sense that for any tangent vector v, we have $|\nabla_v V| \leq L|v|$; see the remark after Assumption 3.2. Integrating the above equation and using Grönwall inequality gives

$$|\partial_t c(s,t)|^2 \leq e^{2LT}|\partial_t c(0,t)|^2 = e^{2LT} d(p,q)^2.$$

Now, for each $s \in [0,T]$, define the energy

$$E(s) = \frac{1}{2}\int_0^1 |\partial_t c(s,t)|^2 \, dt.$$

Note that for any smooth curve $\beta : [0,1] \to \mathcal{M}$, we have $L(\beta)^2 \leq E(\beta)$ by the Cauchy-Schwarz inequality. Thus, for each $s \in [0,T]$,

$$d(\Phi_s(p), \Phi_s(q)) \leq L(c(s,\cdot)) \leq E(c(s,\cdot))^{\frac{1}{2}} \leq e^{TL} d(p,q). \qquad \square$$

A.3. AUXILIARY RESULTS FOR THEOREM 3.6

Lemma 3.14. *Let E be defined as in (3.59). Then E is negatively correlated with V everywhere, in the sense that*

$$\langle \nabla E(p), V(p)\rangle \leq 0, \quad \forall p \in \mathcal{M}. \tag{3.62}$$

Moreover, there exists a constant $C > 0$ such that $(\operatorname{Hess} E)_p(v,v) \leq C|v|^2$ and

$$E(p') \leq E(p) + \langle \nabla E(p), \exp_p^{-1}(p')\rangle + \frac{C}{2} d^2(p,p'), \quad \forall p, p' \in \mathcal{M}. \tag{3.63}$$

Proof. We begin by recalling that the gradient of E is given by

$$\nabla E(p) = \begin{cases} 0 & \text{if } r(p) \leq R, \\ \frac{f'(r(p))}{r(p)} \nabla k(p) & \text{if } r(p) > R. \end{cases}$$

By assumption, $f'(r(p))/r(p) \geq 0$ so $\langle \nabla E(p), V(p)\rangle$ and $\langle \nabla k(p), V(p)\rangle$ have the same sign if $r(p) > R$ and otherwise $\langle \nabla E(p), V(p)\rangle = 0$ if $r(p) \leq R$. We thus conclude that E and V are negatively correlated, as claimed.

Now, to compute the Hessian of E, notice that

$$\operatorname{Hess} E(p)[v,v] = \langle \nabla_v \nabla E(p), v\rangle.$$

Hence,

$$\begin{aligned} \operatorname{Hess} E(p)[v,v] &= \nabla_v \frac{f'(r(p))}{r(p)} \cdot \langle \nabla k(p), v\rangle + \frac{f'(r(p))}{r(p)} \langle \nabla_v \nabla k(p), v\rangle \\ &= \underbrace{\left\langle \nabla \frac{f'(r(p))}{r(p)}, v \right\rangle \cdot \langle \nabla k(p), v\rangle}_{\text{ⓐ}} + \underbrace{\frac{f'(r(p))}{r(p)} \operatorname{Hess} k(p)[v,v]}_{\text{ⓑ}}. \end{aligned} \tag{A.5}$$

Here we use the same notation for directional derivative of a scalar function and the covariant derivative. With this in mind, the first step in computing ⓐ is the observation that

$$\nabla \frac{f'(r(p))}{r(p)} = \left(f''(r(p)) - \frac{f'(r(p))}{r(p)} \right) \frac{1}{r^2(p)} \nabla k(p), \tag{A.6}$$

and hence

$$\begin{aligned}\textcircled{a} &= \left(f''(r(p)) - \frac{f'(r(p))}{r(p)}\right)\frac{1}{r^2(p)}\langle \nabla k(p), v\rangle^2 \\ &\leq \frac{C_2}{r^2(p)}|\nabla k(p)|^2|v|^2 = C_2\,|v|^2.\end{aligned}$$

For $\textcircled{b}$, as $x \coth x \leq 1 + x$ for $x \geq 0$, we obtain

$$\textcircled{b} \leq \frac{f'(r(p))}{r(p)}(1 + \kappa r(p))|v|^2 \leq C_1\,(1/R + \kappa)\,|v|^2.$$

Summing up everything, we obtain

$$\operatorname{Hess} E(p)[v,v] \leq (C_2 + C_1/R + C_1\kappa)|v|^2 =: C|v|^2, \tag{A.7}$$

that is, E has bounded Hessian. Moreover, E is smooth as a composition of smooth functions. Let $p, p' \in \mathcal{M}$ be arbitrary, and let $\gamma : [0,1] \to \mathcal{M}$ be a geodesic connecting the two. By Taylor's remainder theorem, there exists some $t \in (0,1)$ such that

$$E(p') = E(p) + \langle \nabla E(p), \dot\gamma(0)\rangle + \frac{1}{2}\operatorname{Hess} E(\gamma(t))[\dot\gamma, \dot\gamma].$$

Thus, invoking (A.7) and noting that $|\dot\gamma| = d(p,p')$ and $\dot\gamma(0) = \exp_p^{-1}(p')$, we obtain (3.63) and our proof is complete. □

Lemma A.5. *Let $h : \mathbb{R} \to \mathbb{R}$ be the function*

$$h(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \dfrac{e^{-1/x}}{e^{-1/x} + e^{-1/(1-x)}} & \text{if } x \in (0,1) \\ 1 & \text{if } x \geq 1, \end{cases} \tag{A.8}$$

and, for $R > 0$, let

$$f(x) = \int_0^x h(s - R)\,ds.$$

Then f is C^∞ and it satisfies the conditions (3.60) with $C_1 = 1$ and $C_2 = 2$. In addition, one has $f(x) \geq x - (R+1)$, and hence $f(x) = \Omega(x)$.

Proof. As $h(x) \in [0,1]$, we obtain that $f'(x) \in [0,1]$. By a straightforward computation, one observes that the first derivative of h is bounded as $0 \leq h'(x) \leq 2$, so

$$f''(x) = h'(x - R) \le 2.$$

To complete our proof, simply notice that, for $x \ge R + 1$, we have $f(x) = \int_0^x h(s - R)\, ds \ge \int_{R+1}^x 1\, ds = x - (R + 1) = \Omega(x)$, as claimed. □

A.4. PROOF OF PROPOSITION 3.15

Proposition 3.15. *Suppose that*

(H1) $\mathcal{M}$ is either a compact or a Hadamard manifold satisfying Assumption 3.1,

(H2) the vector field V is bounded and satisfies Assumption 3.2,

(H3) V is weakly coercive (3.58) in case $\mathcal{M}$ is not compact,

(H4) and the errors U of the SFO for V are zero-mean and have bounded second moments (3.66). If $\mathcal{M}$ is compact, the errors are further assumed to be a.s. uniformly bounded in norm.

Then, with probability 1, the iterates of Algorithms 3.1–3.4 converge to an internally chain-transitive set of the flow (3.16).

Proof. The proof boils down to verifying noise and bias conditions in Assumptions 3.4 and 3.5, which we do so in a case-by-case basis.

Algorithm 3.1 (RSGM). As $U_n = U(\boldsymbol{x}_n; \omega_n)$ and $B_n = 0$, given (H4) we are done.

Algorithm 3.2 (RPPM). We only have to deal with the bias since $U_n = 0$. First, observe that for n large enough (so that $\gamma_n < \frac{\operatorname{inj}\mathcal{M}}{V^*}$), $\boldsymbol{x}_{n+1}$ falls in the injectivity radius of $\boldsymbol{x}_n$. Using Lipschitzness of V, we have

$$|B_n| = |\mathrm{P}_{\boldsymbol{x}_{n+1}\to\boldsymbol{x}_n}[V(\boldsymbol{x}_{n+1})] - V(\boldsymbol{x}_n)| \le L \cdot d(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n).$$

It follows from (RPPM) that

$$|B_n| \le L\gamma_n |V(\boldsymbol{x}_{n+1})| = O(\gamma_n),$$

which is sufficient to ensure summability requirement of Assumption 3.5.

Algorithm 3.3 (RSEG). Notice that

$$\begin{aligned}|U_n| &= |\mathrm{P}_{\boldsymbol{x}_{n+1/2}\to\boldsymbol{x}_n}[U(\boldsymbol{x}_{n+1/2};\omega_{n+1/2})]| \\ &= |U(\boldsymbol{x}_{n+1/2};\omega_{n+1/2})|,\end{aligned}$$

which satisfies Assumption 3.4 by (H4). For the bias, an argument identical to the proof of Lemma 3.9 implies that $\gamma_n|\widetilde{V}(\boldsymbol{x}_n;\omega_n)| \to 0$ almost surely. Thus, $\boldsymbol{x}_{n+1/2}$ falls in the injectivity radius of $\boldsymbol{x}_n$ for large enough n and we have

$$\begin{aligned}|B_n| &= |\mathrm{P}_{\boldsymbol{x}_{n+1/2}\to\boldsymbol{x}_n}[V(\boldsymbol{x}_{n+1/2})] - V(\boldsymbol{x}_n)| \\ &\leq L\cdot d(\boldsymbol{x}_{n+1/2},\boldsymbol{x}_n) \\ &= L\cdot\gamma_n\cdot|\widetilde{V}(\boldsymbol{x}_n;\omega_n)|.\end{aligned}$$

Hence, $|B_n| \to 0$ with probability 1. For the summability, we see that by (H2) and (H4),

$$\mathbb{E}[|B_n|^2 \,|\, \mathcal{F}_n] \leq 2L^2\gamma_n^2((V^*)^2+\sigma^2) = O(\gamma_n^2) =: (B_n^*)^2,$$

and therefore, $B_n^* \to 0$ as $n\to\infty$ and $\sum\gamma_n\,\mathbb{E}[(B_n^*)^2]^{1/2} = O(\sum\gamma_n^2) < \infty$.

Algorithm 3.4 (ROG). The proof is exactly the same as for RSEG. □

APPENDIX B

PROOFS FOR CHAPTER 4

B.1. AUXILIARY RESULTS FOR THEOREM 4.9

Lemma 4.11. *Suppose that Assumptions 4.1–4.4 hold and the iterates have uniformly bounded second moments. Then, for any fixed $T > 0$, it holds*

$$\lim_{t\to\infty} \sup_{0\le h\le T} \mathbb{E}\,\|\Delta_Z(t,h)\|^2 = 0.$$

Proof. Let $k = m(t)$ and $n = m(t+h)$, and by Assumption 4.4, decompose the error terms into noise and bias. This implies that $\Delta_Z(t,h) = \Delta_U(t,h) + \Delta_B(t,h)$, where Δ_B is the same as Δ_Z with Z replaced by B:

$$\begin{aligned}\Delta_B&(t,h)\\ &= -(t-\tau_k)\,\mathbb{E}[B_{k+1}\,|\,\mathcal{F}_t] + \sum_{i=k}^{n-1}\gamma_{i+1}B_{i+1} + (t+h-\tau_n)\,\mathbb{E}[B_{n+1}\,|\,\mathcal{F}_{t+h}],\end{aligned}$$

and

$$\Delta_U(t,h) = -(t-\tau_k)U_{k+1} + \sum_{i=k}^{n-1}\gamma_{i+1}U_{i+1} + (t+h-\tau_n)U_{n+1}. \tag{B.1}$$

The reason for this simpler formulation is that U_{n+1} is $\mathcal{F}_{\tau_n+}$-measurable; we can therefore remove all the conditional expectations.

Let us first show that $\Delta_B(t,h)$ vanishes as $t\to\infty$. Without loss of generality, suppose that t is large enough so that $\overline{\gamma}_s \le 1$ for all $s \ge t$. Let us also write $\tilde{B}_t := \mathbb{E}[\overline{B}_t\,|\,\mathcal{F}_t]$. By the triangle inequality applied to the definition of $\Delta_B(t,h)$, we obtain

$$\|\Delta_B(t,h)\|^2 \leq \left(\sum_{i=k}^{n-1} \gamma_{i+1}\|B_{i+1}\| + (t+h-\tau_n)\|\tilde{B}_{t+h}\| + (t-\tau_k)\|\tilde{B}_t\| \right)^2$$

which, by the Cauchy-Schwarz inequality, as well as $\sum_{i=k}^{n-1} \gamma_{i+1} \leq h+1$ and $\gamma_{k+1}, \gamma_{n+1} < 1$,

$$\leq (h+3)\left(\sum_{i=k}^{n-1} \gamma_{i+1}\|B_{i+1}\|^2 + \gamma_{n+1}\|\tilde{B}_{t+h}\|^2 + \gamma_{k+1}\|\tilde{B}_t\|^2 \right),$$

Since conditional expectation is a projection in L^2, we have $\mathbb{E}\,\|\tilde{B}_{t+h}\|^2 \leq \mathbb{E}\,\|B_{n+1}\|^2$ and $\mathbb{E}\,\|\tilde{B}_t\|^2 \leq \mathbb{E}\,\|B_{k+1}\|^2$. Letting $l = m(t+T)$, we get

$$\sup_{h\in[0,T]} \mathbb{E}\,\|\Delta_B(t,h)\|^2 \leq (3+T)\cdot$$
$$\left(\sum_{i=n}^{l-1} \gamma_{i+1}\,\mathbb{E}\,\|B_{i+1}\|^2 + \sup_{k\leq j\leq l+1} \gamma_{j+1}\,\mathbb{E}\,\|B_{j+1}\|^2 + \gamma_{k+1}\,\mathbb{E}\,\|B_{k+1}\|^2 \right)$$

Since the second moment of the iterates are assumed to be bounded, the bias condition (4.15) along with Lipschitzness of v implies that $\mathbb{E}\,\|B_{n+1}\|^2 = O(\gamma_{n+1})$. Thus,

$$\sup_{h\in[0,T]} \mathbb{E}\,\|\Delta_B(t,h)\|^2 \lesssim (3+T)\left(\sum_{i=k}^{l-1} \gamma_{i+1}^2 + \sup_{k\leq j\leq l+1} \gamma_{j+1}^2 + \gamma_{k+1}^2 \right). \tag{B.2}$$

Observe that

$$\sum_{i=k}^{l-1} \gamma_{i+1}^2 \leq \left(\sup_{k\leq i\leq l-1} \gamma_{i+1} \right) \sum_{i=k}^{l-1} \gamma_{i+1} \leq T \sup_{k\leq i\leq l-1} \gamma_{i+1}.$$

Since the step-sizes vanish as $t \to \infty$, all the three terms in (B.2) vanish. Therefore,

$$\lim_{t\to\infty} \sup_{h\in[0,T]} \mathbb{E}\,\|\Delta_B(t,h)\|^2 = 0.$$

We now show that the same property holds for $\Delta_U(t,h)$. Recall that by Assumption 4.4, the noise terms have bounded second moments, which we call

C_U. We first decompose (B.1) in L^2:

$$\|\Delta_U(t,h)\|^2 \leq 3\left\|\sum_{i=k}^{n-1} \gamma_{i+1} U_{i+1}\right\|^2 + 3\gamma_{n+1}^2 \|U_{n+1}\|^2 + 3\gamma_{k+1}^2 \|U_{k+1}\|^2.$$

Letting $l = m(t+T)$, taking expectations and supremum over $h \in [0,T]$ gives

$$\sup_{h\in[0,T]} \mathbb{E}\,\|\Delta_U(t,s)\|^2 \leq 3 \sup_{k<n\leq l} \left\{\mathbb{E}\left\|\sum_{i=k}^{n-1} \gamma_{i+1} U_{i+1}\right\|^2 + 3(\gamma_{n+1}^2 + \gamma_{k+1}^2) C_U\right\}.$$

Since $\{U_n\}$ is a martingale difference sequence, $\{\sum_{i=k}^{n-1} \gamma_{i+1} U_{i+1}\}_{n>k}$ is a martingale. Thus, by the martingale property and boundedness of the second moments of U_n, we get

$$\mathbb{E}\left\|\sum_{i=k}^{n-1} \gamma_{i+1} U_{i+1}\right\|^2 = \sum_{i=k}^{n-1} \gamma_{i+1}^2\, \mathbb{E}\,\|U_{i+1}\|^2 \leq C_U \sum_{i=k}^{n-1} \gamma_{i+1}^2.$$

Hence,

$$\lim_{n\to\infty} \sup\left\{\mathbb{E}\left\|\sum_{i=k}^{n-1} \gamma_{i+1} U_{i+1}\right\|^2 : k < n \leq l\right\} \leq \lim_{n\to\infty} C_U \sum_{i=k}^{\infty} \gamma_{i+1}^2 = 0,$$

as the step-size sequence is square-summable. This shows that

$$\lim_{t\to\infty} \sup_{h\in[0,T]} \mathbb{E}\,\|\Delta_U(t,h)\|^2 = 0. \qquad \square$$

Lemma B.1. *Let A be an $\mathcal{F}_s$-measurable matrix and $t \geq s$. Then, it holds*

$$\mathbb{E}\,\|A\,(W_t - W_s)\|^2 = (t-s)\,\mathbb{E}[\operatorname{tr}(A^\top A)].$$

Proof.

$$\begin{aligned}\mathbb{E}\,\|A\,(W_t - W_s)\|^2 &= \mathbb{E}[(W_t - W_s)^\top A^\top A\,(W_t - W_s)]\\ &= \mathbb{E}[\operatorname{tr}\big(A^\top A\,(W_t - W_s)(W_t - W_s)^\top\big)]\\ &= \mathbb{E}[\mathbb{E}[\operatorname{tr}(A^\top A\,(W_t - W_s)(W_t - W_s)^\top)\,|\,\mathcal{F}_s]]\\ &= (t-s)\,\mathbb{E}[\operatorname{tr}(A^\top A)]. \end{aligned} \qquad \square$$

B.2. AUXILIARY LEMMAS FOR SAMPLING ALGORITHMS

Lemma B.2. *The bias of the Proximal Langevin algorithm* (PLA) *satisfies the bias condition* (4.15)*.*

Proof. Using Lipschitzness of ∇f we can write

$$\begin{aligned}\mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] &= \mathbb{E}[\|\nabla f(\boldsymbol{x}_{n+1}) - \nabla f(\boldsymbol{x}_n)\|^2 \,|\, \mathcal{F}_n] \\ &\leq L^2\, \mathbb{E}[\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_n\|^2 \,|\, \mathcal{F}_n] \\ &= L^2\, \mathbb{E}[\|-\gamma_{n+1}\nabla f(\boldsymbol{x}_{n+1}) + \sqrt{2\gamma_{n+1}}\,\xi_{n+1}\|^2 \,|\, \mathcal{F}_n] \\ &\leq 2L^2\gamma_{n+1}^2\, \mathbb{E}[\|\nabla f(\boldsymbol{x}_{n+1})\|^2 \,|\, \mathcal{F}_n] + 4L^2 d\gamma_{n+1}.\end{aligned}$$

Now, notice that $\|\nabla f(\boldsymbol{x}_{n+1})\|^2 \leq 2\|\nabla f(\boldsymbol{x}_{n+1}) - \nabla f(\boldsymbol{x}_n)\|^2 + 2\|\nabla f(\boldsymbol{x}_n)\|^2$. Since $\gamma_{n+1} \to 0$ as $n \to \infty$, we can assume that $4L^2\gamma_{n+1}^2 < \frac{1}{2}$. This gives

$$\mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] \leq \frac{1}{2}\, \mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] + 4L^2\gamma_{n+1}^2\|\nabla f(\boldsymbol{x}_n)\|^2 + 4L^2 d\gamma_{n+1},$$

which implies

$$\begin{aligned}\mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] &\leq 8L^2\gamma_{n+1}^2\|\nabla f(\boldsymbol{x}_n)\|^2 + 8L^2 d\gamma_{n+1} \\ &\lesssim \gamma_{n+1}^2\|\nabla f(\boldsymbol{x}_n)\|^2 + \gamma_{n+1}.\end{aligned}$$

□

Lemma B.3. *The bias of the randomized mid-point method* (RMM) *satisfies the condition* (4.15)*.*

Proof. Let $\tilde{\nabla} f(\boldsymbol{x}_n) = \nabla f(\boldsymbol{x}_n) + U(\boldsymbol{x}_n; \omega_n)$ and $\tilde{\nabla} f(\boldsymbol{x}_{n+1/2}) = \nabla f(\boldsymbol{x}_{n+1/2}) + U(\boldsymbol{x}_{n+1/2}; \omega_{n+1/2})$. Using the Lipschitzness of ∇f and $\alpha_{n+1} \leq 1$, we get

$$\begin{aligned}\mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] &= \mathbb{E}[\|\nabla f(\boldsymbol{x}_{n+1/2}) - \nabla f(\boldsymbol{x}_n)\|^2 \,|\, \mathcal{F}_n] \\ &\leq L^2\, \mathbb{E}[\|\boldsymbol{x}_{n+1/2} - \boldsymbol{x}_n\|^2 \,|\, \mathcal{F}_n] \\ &\leq 2L^2\big(\gamma_{n+1}^2\, \mathbb{E}[\|\nabla f(\boldsymbol{x}_n) + U(\boldsymbol{x}_n; \omega_n)\|^2 \,|\, \mathcal{F}_n] + 2\gamma_{n+1} d\big) \\ &\leq 4L^2\gamma_{n+1}^2\, \|\nabla f(\boldsymbol{x}_n)\|^2 + 2L^2\gamma_{n+1}^2 C_U^2 + 4L^2 d\gamma_{n+1} \\ &\lesssim \gamma_{n+1}^2\|\nabla f(\boldsymbol{x}_n)\|^2 + \gamma_{n+1}.\end{aligned}$$

□

Lemma B.4. *The bias of the optimistic Randomized Mid-point method* (ORMM) *satisfies the condition* (4.15)*.*

Proof. We have

$$\begin{aligned}\mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] &= \mathbb{E}[\|\nabla f(\boldsymbol{x}_{n+1/2}) - \nabla f(\boldsymbol{x}_n)\|^2 \,|\, \mathcal{F}_n] \\ &\leq L^2\, \mathbb{E}[\|\boldsymbol{x}_{n+1/2} - \boldsymbol{x}_n\|^2 \,|\, \mathcal{F}_n] \\ &= L^2\, \mathbb{E}[\|-\gamma_{n+1}\alpha_{n+1}\tilde{\nabla} f(\boldsymbol{x}_{n-1/2}) + \sqrt{2\gamma_{n+1}\alpha_{n+1}}\xi'_{n+1}\|^2 \,|\, \mathcal{F}_n] \\ &\leq 2L^2\gamma_{n+1}^2\, \mathbb{E}[\|\nabla f(\boldsymbol{x}_{n-1/2})\|^2 \,|\, \mathcal{F}_n] + 2L^2\gamma_{n+1}^2 C_U^2 + 4L^2 d\gamma_{n+1},\end{aligned}$$

where we used $\alpha_{n+1} \leq 1$. Similar to the proof of Lemma B.2 for (PLA), notice that $\|\nabla f(\boldsymbol{x}_{n-1/2})\|^2 \leq 2\|\nabla f(\boldsymbol{x}_{n-1/2}) - \nabla f(\boldsymbol{x}_n)\|^2 + 2\|\nabla f(\boldsymbol{x}_n)\|^2$. Since $\gamma_{n+1} \to 0$ as $n \to \infty$, we can assume that $4L^2\gamma_{n+1}^2 < \frac{1}{2}$, and we get

$$\begin{aligned}\mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] &\leq 8L^2\gamma_{n+1}^2\|\nabla f(\boldsymbol{x}_n)\|^2 + 8L^2\gamma_{n+1}^2 C_U^2 + 8L^2 d\gamma_{n+1} \\ &\lesssim \gamma_{n+1}^2\|\nabla f(\boldsymbol{x}_n)\|^2 + \gamma_{n+1}.\end{aligned}$$

□

Lemma B.5. *The bias of* (SRK) *satisfies the condition* (4.15).

Proof. We have

$$\mathbb{E}[\|\nabla f(h_1) - \nabla f(\boldsymbol{x}_n)\|^2 \,|\, \mathcal{F}_n] \leq 4L^2 d\gamma_{n+1}\Big((\tfrac{1}{2} + \tfrac{1}{\sqrt{6}})^2 + \tfrac{1}{12}\Big) = O(\gamma_{n+1}),$$

and similarly,

$$\mathbb{E}[\|\nabla f(h_2) - \nabla f(\boldsymbol{x}_n)\|^2 \,|\, \mathcal{F}_n] \lesssim \gamma_{n+1}^2\|\nabla f(\boldsymbol{x}_n)\|^2 + \gamma_{n+1}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[\|B_{n+1}\|^2 \,|\, \mathcal{F}_n] &\leq \mathbb{E}[\tfrac{1}{2}\|\nabla f(h_1) - \nabla f(\boldsymbol{x}_n)\|^2 + \tfrac{1}{2}\|\nabla f(h_2) - \nabla f(\boldsymbol{x}_n)\|^2 \,|\, \mathcal{F}_n] \\ &\lesssim \gamma_{n+1}^2\|\nabla f(\boldsymbol{x}_n)\|^2 + \gamma_{n+1}.\end{aligned}$$

□

BIBLIOGRAPHY

[AMS08] Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2008.

[AC21] Kwangjun Ahn and Sinho Chewi. “Efficient constrained sampling via the mirror-Langevin algorithm”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 28405–28418.

[AB06] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker’s Guide.* eng. 3rd ed. 2006. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[Ama83] Shun-Ichi Amari. “A Foundation of Information Geometry”. In: *Electronics and Communications in Japan (Part I: Communications)* 66.6 (1983), pp. 1–10.

[AGS05] Luigi Ambrossio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Lectures in Mathematics ETH Zürich. Basel: Birkhäuser-Verlag, 2005.

[ABM14] Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille. *Variational Analysis in Sobolev and BV Spaces.* Philadelphia, PA: Society for Industrial and Applied Mathematics, 2014. eprint: `https://epubs.siam.org/doi/pdf/10.1137/1.9781611973488`.

[AKL22] Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. “Mirror Descent with Relative Smoothness in Measure Spaces, with Application to Sinkhorn and EM”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17263–17275.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators.* Vol. 103. Springer, 2014.

[Bal+22] Krishna Balasubramanian, Sinho Chewi, Murat A. Erdogdu, Adil Salim, and Shunshi Zhang. “Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo”. In: *Conference on Learning Theory*. PMLR. 2022, pp. 2896–2923.

[BC17] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd ed. New York, NY, USA: Springer, 2017.

[BT03] Amir Beck and Marc Teboulle. “Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization”. In: *Operations Research Letters* 31.3 (2003), pp. 167–175.

[Ben99] Michel Benaïm. “Dynamics of Stochastic Approximation Algorithms”. In: *Séminaire de Probabilités XXXIII*. Ed. by Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer, 1999, pp. 1–68.

[BBC17] Michel Benaïm, Florian Bouguet, and Bertrand Cloez. “Ergodicity of inhomogeneous Markov chains through asymptotic pseudotrajectories”. In: *The Annals of Applied Probability* 27.5 (2017), pp. 3004–3049.

[BH96] Michel Benaïm and Morris W. Hirsch. “Asymptotic Pseudotrajectories and Chain Recurrent Flows, with Applications”. In: *Journal of Dynamics and Differential Equations* 8.1 (1996), pp. 141–176.

[BFM17] Glaydston Bento, Orizon Ferreira, and Jefferson Melo. “Iteration-complexity of gradient, subgradient and proximal point methods on Riemannnian manifolds”. In: *Journal of Optimization Theory and Applications* 173.2 (2017), pp. 548–562.

[BMP90] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

[Ber+19] Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E. Jacob. *Schrödinger Bridge Samplers*. 2019. arXiv: `1912.13170 [stat]`. preprint.

[Ber96] Dimitri P Bertsekas. *Neuro-Dynamic Programming*. Anthropological Field Studies. Athena Scientific, 1996.

[Bil99] Patrick Billingsley. *Convergence of probability measures*. 2nd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, 1999.

[Bon13] Silvere Bonnabel. “Stochastic Gradient Descent on Riemannian Manifolds”. In: *IEEE Transactions on Automatic Control* 58.9 (2013), pp. 2217–2229. arXiv: `1111.5280 [cs, math, stat]`.

[Bor08] Vivek S. Borkar. *Stochastic Approximation.* Hindustan Book Agency, 2008.

[Bor+21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. "Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling". In: Advances in Neural Information Processing Systems. 2021.

[Bou23] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds.* 1st ed. Cambridge University Press, 2023.

[BAC19] Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. "Global rates of convergence for nonconvex optimization on manifolds". In: *IMA Journal of Numerical Analysis* 39.1 (2019), pp. 1–33.

[BEL18] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. "Sampling from a log-concave distribution with projected Langevin Monte Carlo". In: *Discrete & Computational Geometry* 59 (2018), pp. 757–783.

[Bun+23] Charlotte Bunne, Ya-Ping Hsieh, Marco Cuturi, and Andreas Krause. "The Schrödinger Bridge between Gaussian Measures Has a Closed Form". In: International Conference on Artificial Intelligence and Statistics. 2023.

[BK81] Peter Buser and Hermann Karcher. *Gromov's almost flat manifolds.* Astérisque 81. Société mathématique de France, 1981.

[Car92] Manfredo Perdigão do Carmo. *Riemannian geometry / Manfredo do Carmo ; translated by Francis Flaherty.* eng. Mathematics. Theory and applications. Boston: Birkhäuser, 1992.

[Cha+21] Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. *On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case.* 2021. arXiv: `1905.13142 [math.ST]`.

[CE08] Jeff Cheeger and David Ebin. *Comparison Theorems in Riemannian Geometry.* Vol. 365. AMS Chelsea Publishing. American Mathematical Society, 2008.

[CLC21] Junfeng Chen, Sanyang Liu, and Xiaokai Chang. "Modified Tseng's extragradient methods for variational inequality on Hadamard manifolds". In: *Applicable Analysis* 100.12 (2021), pp. 2627–2640.

[CLT21] Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. "Likelihood Training of Schrödinger Bridge Using Forward-Backward SDEs Theory". In: International Conference on Learning Representations. 2021.

[CGP16] Yongxin Chen, Tryphon Georgiou, and Michele Pavon. "Entropic and displacement interpolation: a computational approach using the Hilbert metric". In: *SIAM Journal on Applied Mathematics* 76.6 (2016), pp. 2375–2396.

[Che+18] Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. "Sharp convergence rates for Langevin dynamics in the nonconvex setting". In: *arXiv preprint arXiv:1805.01648* (2018).

[Che23] Sinho Chewi. *Log-Concave Sampling.* 2023.

[Che+21] Sinho Chewi, Murat A. Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. "Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev". In: *arXiv preprint arXiv:2112.12662* (2021).

[Cla+23] Julien Claisse, Giovanni Conforti, Zhenjie Ren, and Songbo Wang. *Mean Field Optimization Problem Regularized by Fisher Information.* 2023. arXiv: `2302.05938 [math.PR]`.

[CLP23] Giovanni Conforti, Daniel Lacker, and Soumik Pal. *Projected Langevin dynamics and a gradient flow for entropic optimal transport.* 2023. arXiv: `2309.08598 [math.PR]`.

[CG24] Dario Corona and Roberto Giambò. "Global Models of Collapsing Scalar Field: Endstate". In: *Symmetry* 16.5 (2024), p. 583.

[Cou+16] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. *Optimal Transport for Domain Adaptation.* 2016. arXiv: `1507.00504 [cs.LG]`.

[CB19] Christopher Criscitiello and Nicolas Boumal. "Efficiently escaping saddle points on manifolds". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 5987–5997.

[CB21] Christopher Criscitiello and Nicolas Boumal. *An Accelerated First-Order Method for Non-Convex Optimization on Manifolds.* 2021. arXiv: `2008.02252 [cs, math]`. preprint.

[Cut13] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: Neural Information Processing Systems. 2013.

[DK19] Arnak S. Dalalyan and Avetik Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". In: *Stochastic Processes and their Applications* 129.12 (2019), pp. 5278–5311.

[Dan67] John M. Danskin. *The Theory of Max-Min and its Application to Weapons Allocation Problems.* Springer Berlin Heidelberg, 1967.

[De +21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. "Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling". In: *Advances in Neural Information Processing Systems.* Vol. 34. 2021, pp. 17695–17709.

[Deb+23] Nabarun Deb, Young-Heon Kim, Soumik Pal, and Geoffrey Schiebinger. *Wasserstein Mirror Gradient Flow as the Limit of the Sinkhorn Algorithm.* 2023. arXiv: `2307.16421 [math, stat]`. preprint.

[DEH21] Radu Alexandru Dragomir, Mathieu Even, and Hadrien Hendrikx. "Fast stochastic bregman gradient methods: Sharp analysis and variance reduction". In: *International Conference on Machine Learning.* PMLR. 2021, pp. 2815–2825.

[DM21] Theodore D. Drivas and Alexei A. Mailybaev. "'Life after Death' in Ordinary Differential Equations with a Non-Lipschitz Singularity". In: *Nonlinearity* 34.4 (2021), p. 2296.

[Dur+21] Alain Durmus, Pablo Jiménez, Éric Moulines, and Salem Said. "On Riemannnian Stochastic Approximation Schemes with Fixed Step-Size". In: *International Conference on Artificial Intelligence and Statistics.* PMLR. 2021, pp. 1018–1026.

[Dur+20] Alain Durmus, Pablo Jiménez, Éric Moulines, Salem Said, and Hoi-To Wai. *Convergence analysis of Riemannnian stochastic approximation schemes.* 2020. arXiv: `2005.13284`. preprint.

[DM17] Alain Durmus and Eric Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". In: *The Annals of Applied Probability* 27.3 (2017), pp. 1551–1587.

[Egg93] P. P. B. Eggermont. "Maximum Entropy Regularization for Fredholm Integral Equations of the First Kind". In: *SIAM Journal on Mathematical Analysis* 24.6 (1993), pp. 1557–1576.

[FP03] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems.* Springer Series in Operations Research. Springer, 2003.

[FQT20] Jingjing Fan, Xiaolong Qin, and Bing Tan. "Tseng's extragradient algorithm for pseudomonotone variational inequalities on Hadamard manifolds". In: *Applicable Analysis* (2020), pp. 1–14.

[FO02] OP Ferreira and PR Oliveira. "Proximal point algorithm on Riemannnian manifolds". In: *Optimization* 51.2 (2002), pp. 257–270.

[FPN05] Orizon Pereira Ferreira, LR Lucambio Pérez, and Sándor Zoltán Németh. "Singularities of monotone vector fields and an extragradient-type algorithm". In: *Journal of Global Optimization* 31.1 (2005), pp. 133–151.

[FV08] Alessio Figalli and Cedric Villani. "An Approximation Lemma about the Cut Locus, with Applications in Optimal Transport Theory". In: *Methods and Applications of Analysis* 15.2 (2008), pp. 149–154.

[Fle77] Wendell H. Fleming. "Exit Probabilities and Optimal Stochastic Control". In: *Applied Mathematics and Optimization* 4.1 (1977), pp. 329–346.

[For40] Robert Fortet. "Résolution d'un système d'équations de M. Schrödinger". In: Journal de mathématiques pures et appliquées (1940).

[FL89] Joel Franklin and Jens Lorenz. "On the Scaling of Multidimensional Matrices". In: *Linear Algebra and its Applications* (1989).

[FK82] Takahiko Fujita and Shin-ichi Kotani. "The Onsager-Machlup Function for Diffusion Processes". In: *Journal of Mathematics of Kyoto University* 22.1 (1982), pp. 115–130.

[GPC17] Aude Genevay, Gabriel Peyré, and Marco Cuturi. *Learning Generative Models with Sinkhorn Divergences*. 2017. arXiv: `1706.00292 [stat.ML]`.

[GN22] Promit Ghosal and Marcel Nutz. *On the Convergence Rate of Sinkhorn's Algorithm*. 2022. arXiv: `2212.06000 [math]`. preprint.

[Hal88] Jack K. Hale. "Asymptotic Behavior of Dissipative Systems". In: *Aerican Mathematical Society*. 1988.

[HR21] Filip Hanzely and Peter Richtárik. "Fastest Rates for Stochastic Mirror Descent Methods". In: *Computational Optimization and Applications* 79.3 (2021), pp. 717–766.

[HP86] U. G. Haussmann and E. Pardoux. "Time Reversal of Diffusions". In: *The Annals of Probability* 14.4 (1986), pp. 1188–1205.

[Hay86] Simon S. Haykin. *Adaptive filter theory*. Prentice-Hall information and system sciences series. Prentice-Hall, 1986.

[HBE20] Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. "On the Ergodicity, Bias and Asymptotic Normality of Randomized Midpoint Sampling Method". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 7366–7376.

[Hol+23] Lars Holdijk, Yuanqi Du, Ferry Hooft, Priyank Jaini, Bernd Ensing, and Max Welling. *Stochastic Optimal Control for Collective Variable Free Sampling of Molecular Transition Paths.* 2023. arXiv: `2207.02149 [physics, q-bio]`. preprint.

[Hsi+23] Ya-Ping Hsieh, Mohammad Reza Karimi, Andreas Krause, and Panayotis Mertikopoulos. “Riemannian Stochastic Optimization Methods Avoid Strict Saddle Points”. In: *Advances in Neural Information Processing Systems (NeurIPS).* 2023.

[Hsi+18] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. “Mirrored Langevin dynamics”. In: *Advances in Neural Information Processing Systems* 31 (2018).

[Hua+21] Jian Huang, Yuling Jiao, Lican Kang, Xu Liao, Jin Liu, and Yanyan Liu. *Schrödinger-Föllmer Sampler: Sampling without Ergodicity.* 2021. arXiv: `2106.10880 [stat]`. preprint.

[HAG15] Wen Huang, P.-A. Absil, and K. A. Gallivan. “A Riemannian Symmetric Rank-One Trust-Region Method”. In: *Mathematical Programming* 150.2 (2015), pp. 179–216.

[HW21] Wen Huang and Ke Wei. “Riemannnian proximal gradient methods”. In: *Mathematical Programming* (2021), pp. 1–43.

[Ili06] Bozhidar Z Iliev. *Handbook of normal frames and coordinates.* Vol. 42. Springer Science & Business Media, 2006.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. “The Variational Formulation of the Fokker–Planck Equation”. In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17.

[Jos17] Jürgen Jost. *Riemannian Geometry and Geometric Analysis.* Universitext. Cham: Springer International Publishing, 2017.

[Kak01] Sham M Kakade. “A Natural Policy Gradient”. In: *Advances in Neural Information Processing Systems.* Vol. 14. MIT Press, 2001.

[KHK23a] Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. “A Dynamical System View of Langevin-Based Non-Convex Sampling”. In: *Advances in Neural Information Processing Systems (NeurIPS).* 2023.

[KHK23b] Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. “Stochastic Approximation Algorithms for Systems of Interacting Particles”. In: *Advances in Neural Information Processing Systems (NeurIPS).* 2023.

[KHK24] Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. "Sinkhorn Flow as Mirror Flow: a Continuous-Time Framework for Generalizing the Sinkhorn Algorithm". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2024.

[Kar+22] Mohammad Reza Karimi, Ya-Ping Hsieh, Panayotis Mertikopoulos, and Andreas Krause. "The Dynamics of Riemannian Robbins-Monro Algorithms". In: *Proceedings of 35th Conference on Learning Theory (COLT)*. 2022.

[Kha+20] Konrawut Khammahawong, Poom Kumam, Parin Chaipunya, Jen-Chih Yao, Ching-Feng Wen, and Wachirapong Jirakitpuwapat. "An extragradient algorithm for strongly pseudomonotone equilibrium problems on Hadamard manifolds". In: *Thai Journal of Mathematics* 18.1 (2020), pp. 350–371.

[KW52] J. Kiefer and J. Wolfowitz. "Stochastic Estimation of the Maximum of a Regression Function". In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 462–466.

[KK02] Eric Klavins and Daniel E Koditschek. "Phase regulation of decentralized cyclic robotic systems". In: *The International Journal of Robotics Research* 21.3 (2002), pp. 257–275.

[Kor76] G. M. Korpelevich. "The extragradient method for finding saddle points and other problems". In: *Èkonom. i Mat. Metody* 12 (1976), pp. 747–756.

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. "Accelerated Mirror Descent in Continuous and Discrete Time". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc., 2015.

[Kri14] Alexandru Kristály. "Nash-type equilibria on Riemannian manifolds: A variational approach". In: *Journal de Mathématiques Pures et Appliquées* 101.5 (2014), pp. 660–688.

[Kul68] S. Kullback. "Probability Densities with Given Marginals". In: *The Annals of Mathematical Statistics* 39.4 (1968), pp. 1236–1243. JSTOR: 2239692.

[KC78] Harold J. Kushner and Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer New York, 1978.

[KH81] Harold J. Kushner and Hai Huang. "Asymptotic Properties of Stochastic Approximations with Constant Coefficients". In: *SIAM Journal on Control and Optimization* 19.1 (1981), pp. 87–105. eprint: `https://doi.org/10.1137/0319007`.

[KY97] Harold J. Kushner and G. G. Yin. *Stochastic approximation algorithms and applications*. Springer-Verlag, 1997.

[LP02] Damien Lamberton and Gilles Pages. "Recursive computation of the invariant distribution of a diffusion". In: *Bernoulli* (2002), pp. 367–405.

[Le 16] Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*. Springer International Publishing, 2016.

[Lee12] John M. Lee. *Introduction to Smooth Manifolds*. Vol. 218. Graduate Texts in Mathematics. New York, NY: Springer, 2012.

[Lee18] John M. Lee. *Introduction to Riemannian Manifolds*. Vol. 176. Graduate Texts in Mathematics. Cham: Springer International Publishing, 2018.

[Lég20] Flavien Léger. *A Gradient Descent Perspective on Sinkhorn*. 2020. arXiv: `2002.03758 [math]`. preprint.

[Lem05] Vincent Lemaire. "Estimation récursive de la mesure invariante d'un processus de diffusion." PhD thesis. Université de Marne la Vallée, 2005.

[Léo12] Christian Léonard. "Girsanov Theory Under a Finite Entropy Condition". In: *Séminaire de Probabilités XLIV*. Ed. by Catherine Donati-Martin, Antoine Lejay, and Alain Rouault. Springer, 2012, pp. 429–465.

[Léo13a] Christian Léonard. "A Survey of the Schrödinger Problem and Some of Its Connections with Optimal Transport". In: *Discrete and Continuous Dynamical Systems* 34.4 (2013), pp. 1533–1574.

[Léo13b] Christian Léonard. *Some Properties of Path Measures*. 2013. arXiv: `1308.0217 [math]`. preprint.

[Lez20] Mario Lezcano-Casado. *Curvature-dependant global convergence rates for optimization on manifolds of bounded geometry*. 2020. arXiv: `2008.02517 [math.OC]`. preprint.

[LLM09] Chong Li, Genaro López, and Victoria Martín-Márquez. "Monotone vector fields and the proximal point algorithm on Hadamard manifolds". In: *Journal of the London Mathematical Society* 79.3 (2009), pp. 663–683.

[Li+22] Ruilin Li, Molei Tao, Santosh S. Vempala, and Andre Wibisono. "The Mirror Langevin Algorithm Converges with Vanishing Bias". In: *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*. International Conference on Algorithmic Learning Theory. PMLR, 2022, pp. 718–742.

[LZT22] Ruilin Li, Hongyuan Zha, and Molei Tao. "Sqrt(d) Dimension Dependence of Langevin Monte Carlo". In: *The International Conference on Learning Representations*. 2022.

[Li+20] Xuechen Li, Denny Wu, Lester Mackey, and Murat A. Erdogdu. *Stochastic Runge-Kutta Accelerates Langevin Monte Carlo and Beyond*. 2020. arXiv: `1906.07868 [cs, stat]`. preprint.

[Li+19] Xuechen Li, Yi Wu, Lester Mackey, and Murat A. Erdogdu. "Stochastic runge-kutta accelerates langevin monte carlo and beyond". In: *Advances in neural information processing systems* 32 (2019).

[Liu+22] Guan-Horng Liu, Tianrong Chen, Oswin So, and Evangelos Theodorou. "Deep Generalized Schrödinger Bridge". In: *Advances in Neural Information Processing Systems*. 2022.

[Lju77] Lennart Ljung. "Analysis of recursive stochastic algorithms". In: *IEEE Transactions on Automatic Control* 22.4 (1977), pp. 551–575.

[LS83] Lennart Ljung and Torsten Söderström. *Theory and practice of recursive identification*. MIT press, 1983.

[Ma+21] Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. "Is there an analog of Nesterov acceleration for gradient-based MCMC?" In: *Bernoulli* 27.3 (2021), pp. 1942–1992.

[MMS20] Mateusz B. Majka, Aleksandar Mijatović, and Łukasz Szpruch. "Nonasymptotic bounds for sampling algorithms without log-concavity". In: *The Annals of Applied Probability* 30.4 (2020), pp. 1534–1581.

[MM63] FK Manasse and Charles W Misner. "Fermi normal coordinates and some basic concepts in differential geometry". In: *Journal of mathematical physics* 4.6 (1963), pp. 735–745.

[Mar70] B. Martinet. "Brève communication. Régularisation d'inéquations variationnelles par approximations successives". In: *Revue française d'informatique et de recherche opérationnelle. Série rouge* 4.R3 (1970), pp. 154–158.

[MS18] Panayotis Mertikopoulos and William H. Sandholm. "Riemannian game dynamics". In: *Journal of Economic Theory* 177 (2018), pp. 315–364.

[MT93] Sean P. Meyn and Richard L. Tweedie. "Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes". In: *Advances in Applied Probability* 25.3 (1993), pp. 518–548.

[MT04] Grigori N Milstein and Michael V Tretyakov. *Stochastic numerics for mathematical physics.* Vol. 39. Springer, 2004.

[Mis19] Konstantin Mishchenko. *Sinkhorn Algorithm as a Special Case of Stochastic Mirror Descent.* 2019. arXiv: `1909.06918 [cs, math, stat]`. preprint.

[Mou+22] Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. "Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity". In: *Bernoulli* 28.3 (2022), pp. 1577–1601.

[ME05] Abubakr Muhammad and Magnus Egerstedt. "Decentralized coordination with local interactions: Some new directions". In: *Cooperative Control.* Springer, 2005, pp. 153–170.

[NJ83] Arkadii Semenovich Nemirovsky and David Borisovich Judin. *Problem complexity and method efficiency in optimization.* eng. Wiley-Interscience series in discrete mathematics. J. Wiley, 1983.

[NSS16] JX Cruz Neto, PSM Santos, and PA Soares. "An extragradient method for equilibrium problems on Hadamard manifolds". In: *Optimization Letters* 10.6 (2016), pp. 1327–1336.

[Nut22] Marcel Nutz. "Introduction to Entropic Optimal Transport". Lecture Notes. 2022.

[NW21] Marcel Nutz and Johannes Wiesel. *Entropic Optimal Transport: Convergence of Potentials.* 2021. arXiv: `2104.11720 [math]`. preprint.

[ØS19] Bernt Øksendal and Agnès Sulem. *Applied Stochastic Control of Jump Diffusions.* Springer International Publishing, 2019.

[Ott01] Felix Otto. "The Geometry of Dissipative Evolution Equations: The Porous Medium Equation". In: *Communications in Partial Differential Equations* 26.1-2 (2001), pp. 101–174.

[Par+23] Matteo Pariset, Ya-Ping Hsieh, Charlotte Bunne, Andreas Krause, and Valentin De Bortoli. "Unbalanced Diffusion Schrödinger Bridge". In: *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems.* 2023.

[Pet16] Peter Petersen. *Riemannian Geometry.* Vol. 171. Graduate Texts in Mathematics. Cham: Springer International Publishing, 2016.

[PC20] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. 2020. arXiv: `1803.00567 [stat]`. preprint.

[Phe93] Robert Ralph Phelps. *Convex Functions, Monotone Operators and Differentiability*. 2nd ed. Lecture Notes in Mathematics. Springer-Verlag, 1993.

[Pop80] Leonid Denisovich Popov. "A modification of the Arrow–Hurwicz method for search of saddle points". In: *Mathematical Notes of the Academy of Sciences of the USSR* 28.5 (1980), pp. 845–848.

[PP90] Paolo Dai Pra and Michele Pavon. "On the Markov Processes of Schrödinger, the Feynman-Kac Formula and Stochastic Control". In: *Realization and Modelling in System Theory: Proceedings of the International Symposium MTNS-89, Volume I*. Ed. by M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran. Boston, MA: Birkhäuser, 1990, pp. 497–504.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. "Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis". In: *Conference on Learning Theory*. PMLR. 2017, pp. 1674–1703.

[RBS14] Lillian J. Ratliff, Samuel A. Burden, and S. Shankar Sastry. *On the Characterization of Local Nash Equilibria in Continuous Games*. 2014. arXiv: `1411.2168 [math.OC]`.

[Res05] Elena Resmerita. "Regularization of Ill-Posed Problems in Banach Spaces: Convergence Rates". In: *Inverse Problems* 21.4 (2005), p. 1303.

[RM51] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[RT96a] Gareth O. Roberts and Richard L. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* (1996), pp. 341–363.

[RT96b] Gareth O. Roberts and Richard L. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* 2.4 (1996), pp. 341–363.

[Roc97] R Tyrrell Rockafellar. *Convex analysis*. Vol. 11. Princeton university press, 1997.

[Roc76] R. Tyrrell Rockafellar. "Monotone Operators and the Proximal Point Algorithm". In: *SIAM Journal on Control and Optimization* 14.5 (1976), pp. 877–898.

[RW00] L. C. G. Rogers and David Williams. *Diffusions, Markov Processes and Martingales.* Cambridge University Press, 2000.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.* Springer International Publishing, 2015.

[Sch+19] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. "Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming". In: *Cell* 176.6 (2019), p. 1517.

[Sha21] Suhail M. Shah. "Stochastic Approximation on Riemannian Manifolds". In: *Applied Mathematics & Optimization* 83.2 (2021), pp. 1123–1151.

[SL19] Ruoqi Shen and Yin Tat Lee. *The Randomized Midpoint Method for Log-Concave Sampling.* 2019. arXiv: `1909.05503 [cs, math, stat]`. preprint.

[SK67] Richard Sinkhorn and Paul Knopp. "Concerning Nonnegative Matrices and Doubly Stochastic Matrices". In: *Pacific Journal of Mathematics* 21.2 (1967), pp. 343–348.

[Ste99] Shlomo Sternberg. *Lectures on differential geometry.* Vol. 316. American Mathematical Society, 1999.

[Str10] Daniel W Stroock. *Probability theory: An Analytic View.* 2nd ed. Cambridge University Press, 2010.

[SB98] Richard S Sutton and Andrew G Barto. "Reinforcement Learning: An Introduction". In: (1998).

[TH12] Guo-ji Tang and Nan-jing Huang. "Korpelevich's method for variational inequality problems on Hadamard manifolds". In: *Journal of Global Optimization* 54.3 (2012), pp. 493–509.

[TTV16] Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. "Consistency and fluctuations for stochastic gradient Langevin dynamics". In: *Journal of Machine Learning Research* 17 (2016).

[Tri+18] Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I Jordan. "Averaging stochastic gradient descent on Riemannnian manifolds". In: *Conference On Learning Theory.* PMLR. 2018, pp. 650–687.

[Tze+23] Belinda Tzen, Anant Raj, Maxim Raginsky, and Francis Bach. *Variational Principles for Mirror Descent and Mirror Langevin Dynamics.* 2023. arXiv: `2303.09532 [math.OC]`. preprint.

[VW19] Santosh Vempala and Andre Wibisono. “Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices”. In: *Advances in neural information processing systems* 32 (2019).

[Vil03] Cédric Villani. *Topics in optimal transportation.* American Mathematical Society, 2003.

[Wan+21a] Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. “Deep Generative Learning via Schrödinger Bridge”. In: *Proceedings of the 38th International Conference on Machine Learning.* International Conference on Machine Learning. PMLR, 2021, pp. 10794–10804.

[Wan+10] JH Wang, G López, Victoria Martín-Márquez, and Chong Li. “Monotone and accretive vector fields on Riemannian manifolds”. In: *Journal of optimization theory and applications* 146.3 (2010), pp. 691–708.

[WG24] Tao Wang and Ziv Goldfeld. *Neural Estimation Of Entropic Optimal Transport.* 2024. arXiv: `2405.06734 [math.ST]`.

[Wan+21b] Xi Wang, Zhipeng Tu, Yiguang Hong, Yingyi Wu, and Guodong Shi. “No-regret Online Learning over Riemannnian Manifolds”. In: *Thirty-Fifth Conference on Neural Information Processing Systems.* 2021.

[WT11] Max Welling and Yee Whye Teh. “Bayesian Learning via Stochastic Gradient Langevin Dynamics”. In: *Internation Conference on Machine Learning.* 2011, p. 8.

[Wib19] Andre Wibisono. *Proximal Langevin Algorithm: Rapid Convergence Under Isoperimetry.* 2019. arXiv: `1911.01469 [cs, math, stat]`. preprint.

[WWJ16] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. “A variational perspective on accelerated methods in optimization”. In: *Proceedings of the National Academy of Sciences* 113.47 (2016), E7351–E7358.

[Xu+18] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. “Global convergence of Langevin dynamics based algorithms for nonconvex optimization”. In: *Advances in Neural Information Processing Systems* 31 (2018).

[ZS16] Hongyi Zhang and Suvrit Sra. "First-order methods for geodesically convex optimization". In: *Conference on Learning Theory*. PMLR. 2016, pp. 1617–1638.

[Zha+20] Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. "Wasserstein control of mirror Langevin Monte Carlo". In: *Conference on Learning Theory*. PMLR. 2020, pp. 3814–3841.

[ZC22] Qinsheng Zhang and Yongxin Chen. "Path Integral Sampler: A Stochastic Control Approach For Sampling". In: *International Conference on Learning Representations*. 2022.

[ZS20] Xiaojing Zhu and Hiroyuki Sato. "Riemannian Conjugate Gradient Methods with Inverse Retraction". In: *Computational Optimization and Applications* 77.3 (2020), pp. 779–810.

[ZXG19] Difan Zou, Pan Xu, and Quanquan Gu. "Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 2936–2945.

CURRICULUM VITAE

Personal Data

Name	Mohammad Reza Karimi Jaghargh
Date of Birth	December 31, 1992
Place of Birth	Tehran, Iran
Citizen of	Iran

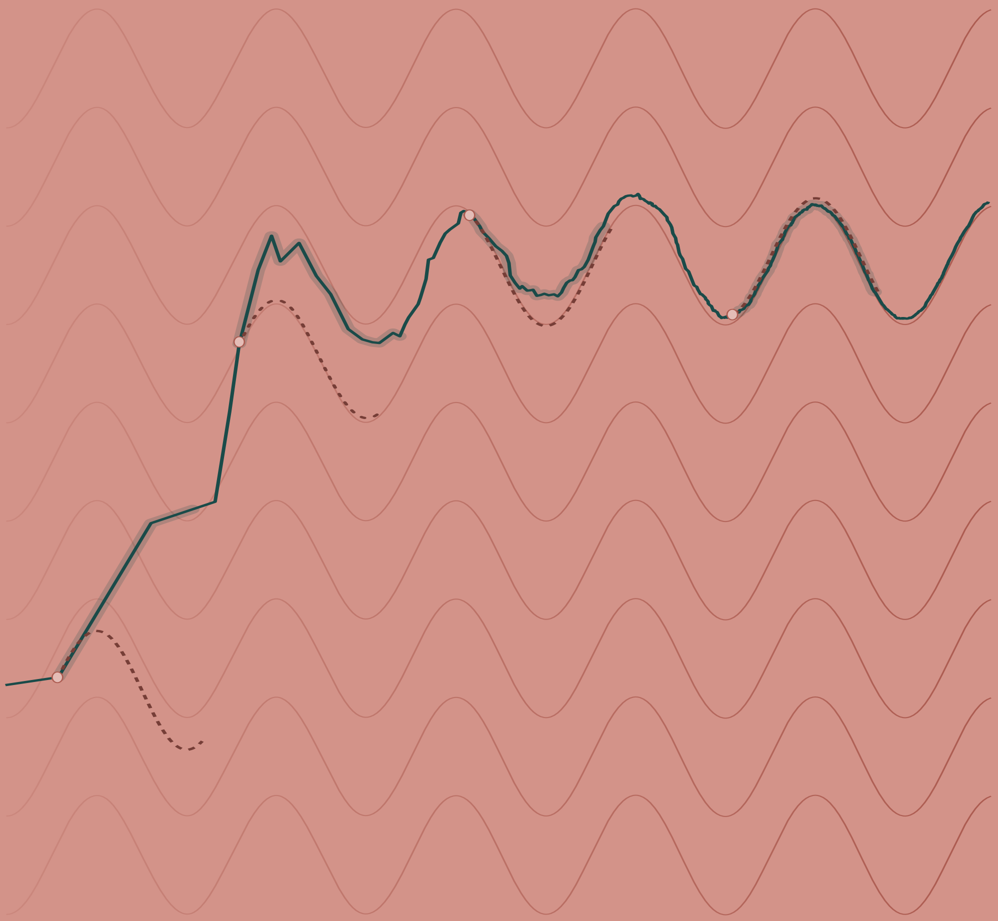
Education

2019–2024	Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland *Final degree:* Doctor of Science
2016–2019	Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland *Final degree:* Master of Science in Computer Science
2011–2016	Sharif University of Technology, Tehran, Iran *Final degree:* Bachelor of Science in Computer Engineering and Pure Mathematics

Employment

2016–2019	Research assistant *Learning and Adaptive Systems group*, Zürich, Switzerland
Summer 2015	Research intern *Max Planck Institute for Software Systems*, Kaiserslautern, Germany

INDEX



$$\lim_{t\to\infty}\sup_{h\in[0,T]} d(X(t+h),\Phi_h(X(t)))=0$$