Log-Concave Sampling

(unfinished draft)

Sinho Chewi

February 15, 2024

Contents

Preface **i**

I Diffusions in Continuous Time **1**

1 The Langevin Diffusion in Continuous Time **3**
1.1 A Primer on Stochastic Calculus . . . 3
1.2 Markov Semigroup Theory . . . 19
1.3 The Geometry of Optimal Transport . . . 30
1.4 The Langevin SDE as a Wasserstein Gradient Flow . . . 45
1.5 Overview of the Convergence Results . . . 51
Bibliographical Notes . . . 55
Exercises . . . 56

2 Functional Inequalities **65**
2.1 Overview of the Inequalities . . . 65
2.2 Proofs via Markov Semigroup Theory . . . 68
2.3 Operations Preserving Functional Inequalities . . . 77
2.4 Concentration of Measure . . . 88
2.5 Isoperimetric Inequalities . . . 97
2.6 Metric Measure Spaces . . . 108
2.7 Discrete Space and Time . . . 122
Bibliographical Notes . . . 124

Exercises . . . 126

3 Additional Topics in Stochastic Analysis **137**
3.1 Quadratic Variation . . . 137
3.2 Change of Measure in Path Space . . . 141
3.3 Doob's Transform . . . 147
3.4 Föllmer Drift . . . 151
3.5 Schrödinger Bridge . . . 155
Bibliographical Notes . . . 160
Exercises . . . 161

II Complexity of Sampling **163**

4 Analysis of Langevin Monte Carlo **165**
4.1 Proof via Wasserstein Coupling . . . 165
4.2 Proof via Interpolation Argument . . . 171
4.3 Proof via Convex Optimization . . . 175
4.4 Proof via Girsanov's Theorem . . . 180
Bibliographical Notes . . . 184
Exercises . . . 184

5 Faster Low-Accuracy Samplers **187**
5.1 Randomized Midpoint Discretization . . . 187
5.2 Hamiltonian Monte Carlo . . . 192
5.3 The Underdamped Langevin Diffusion . . . 196
Bibliographical Notes . . . 203
Exercises . . . 203

6 Convergence in Rényi Divergence **207**
6.1 Proof under LSI via Interpolation Argument . . . 207
Bibliographical Notes . . . 213
Exercises . . . 213

7 High-Accuracy Samplers **215**
7.1 Rejection Sampling . . . 215
7.2 The Metropolis–Hastings Filter . . . 217
7.3 An Overview of High-Accuracy Samplers . . . 220
7.4 Markov Chains in Discrete Time . . . 225

7.5 Analysis of MALA for a Feasible Start . . . 232
7.6 Analysis of MALA for a Warm Start . . . 236
Bibliographical Notes . . . 239
Exercises . . . 239

8 The Proximal Sampler 243
8.1 Introduction to the Proximal Sampler . . . 243
8.2 Convergence under Strong Log-Concavity . . . 245
8.3 Simultaneous Heat Flow and Time Reversal . . . 248
8.4 Convergence under Log-Concavity . . . 251
8.5 Convergence under Functional Inequalities . . . 253
8.6 Applications . . . 254
Bibliographical Notes . . . 257
Exercises . . . 257

9 Lower Bounds for Sampling 261
9.1 A Query Complexity Result in One Dimension . . . 261
9.2 Other Approaches . . . 270
Bibliographical Notes . . . 272
Exercises . . . 272

10 Structured Sampling 273
10.1 Coordinate Langevin . . . 273
10.2 Mirror Langevin . . . 273
10.3 Proximal Langevin . . . 287
10.4 Stochastic Gradient Langevin . . . 287
Bibliographical Notes . . . 287
Exercises . . . 287

11 Non-Log-Concave Sampling 291
11.1 Approximate First-Order Stationarity via Fisher Information . . . 291
11.2 Fisher Information Bound . . . 294
11.3 Applications of the Fisher Information Bound . . . 296
Bibliographical Notes . . . 299
Exercises . . . 299

References 301

Index 317

Preface

What This Book Contains

As discussed in the next section, a large portion of this book is dedicated to a systematic and unified treatment of recent developments in the complexity theory for log-concave sampling, with a particular emphasis on connections with the field of optimization. Many of these developments appear here in textbook form for the first time. Although this is still an active area of research, at this time there is enough beautiful mathematics and canonical theory that it seemed a shame not to have available an exposition which is accessible to, say, an ambitious graduate student.

From a broader view, however, it is not the specific applications to log-concave sampling, but rather the general perspective and techniques used, that will have the largest impact on the reader. With this in mind, the book includes several topics which are not directly related to sampling, but loosely illustrate the general theme of "modern applications of stochastic analysis to probability and statistics". The applications range from classical mathematical questions, such as concentration of measure and geometry, to instances in which the philosophy of diffusion processes has inspired recent algorithms for machine learning tasks. Although tastes change, the overall importance of this perspective only seems to grow with time.

The subject matter of this book touches upon many fields, such as geometry, PDE, stochastic calculus, etc., and a primary goal of the exposition here is to make the material accessible without extensive background knowledge in these topics. This means that at several places we have sacrificed full mathematical rigor in favor of (hopefully) more lucid explanations, referring to the original sources for details. As such, these subjects are not prerequisites for this book, although more background knowledge on the reader's part naturally translates into an healthier understanding of the context of the material. The main exceptions to this statement are: (1) we assume that the reader is familiar

with graduate-level analysis and probability; (2) since much of the theory of sampling is inspired by ideas from optimization, we highly recommend that the reader is familiar with the latter, as treated in, e.g., [Bub15; Nes18].

The Complexity of Sampling

In this book, we consider the following canonical sampling problem:

> Given query access to a smooth function $V : \mathbb{R}^d \to \mathbb{R}$, what is the minimum number of queries required to output an approximate sample from the probability density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$?

The problem formulation is chosen due to the following considerations. In many applications (some described below), we wish to sample from a probability density π, and we have an explicit function $V : \mathbb{R}^d \to \mathbb{R}$ such that $\pi \propto \exp(-V)$. In other words, since π is a probability density, then $\pi = \frac{1}{Z}\exp(-V)$, where $Z := \int \exp(-V)$ is called the *normalizing factor* (or the *partition function* in statistical physics). Although Z (and thus π) are explicitly given in terms of the known function V, a naïve evaluation of Z as a high-dimensional integral is intractable. Indeed, the usual approach of approximating an integral by a sum requires discretizing space via a fine grid whose size scales exponentially in the dimension d. Moreover, even if we had access to Z, it is still not clear how we could use this to sample from π. Therefore, the focus here is to develop direct methods for the sampling task which bypass the computation of Z.[1]

Not only do we want to develop fast algorithms, we also want to understand the inherent complexity of the sampling task, which in turn allows us to identify *optimal* algorithms. By *complexity*, we do not mean *computational* complexity, since proving lower bounds in that context is out of reach (besides, we would have to spend too much time worrying about the bit representation of V). Instead, following the well-trodden path of optimization, we adopt a model in which we only have access to V through queries made to an oracle, and our notion of complexity is the number of queries made. This is known as *oracle complexity* or *query complexity*; see [NY83, §1] for a detailed discussion. We will usually consider a first-order oracle, i.e. given a point $x \in \mathbb{R}^d$, the oracle returns $(V(x), \nabla V(x))$. Since V is only well-defined up to an additive constant, we can equivalently imagine that the oracle returns $(V(x) - V(0), \nabla V(x))$.

Before considering the problem further, here are some important applications.

[1] In fact, it goes the other way around: the state-of-the-art methods for approximately computing Z are based on the sampling methods we develop here.

1. (**Bayesian statistics**) Suppose that we wish to make inferences about a parameter ϑ of interest, which lies in a space $\Theta \subseteq \mathbb{R}^d$. As a Bayesian, we have a *prior* density p_ϑ over Θ which encodes our subjective beliefs about the value of ϑ prior to seeing any data. Next, we collect some data X which, conditionally on the value of ϑ, is drawn from a density $p_{X|\vartheta}(\cdot \mid \vartheta)$. According to Bayesian statistics, we should then compute the *posterior distribution*

$$p_{\vartheta|X}(\theta \mid X) \propto \frac{p_\vartheta(\theta)\, p_{X|\vartheta}(X \mid \theta)}{\int_\Theta p_\vartheta(\mathrm{d}\theta')\, p_{X|\vartheta}(X \mid \theta')},$$

   which encodes our new beliefs about ϑ after seeing the data.

   Typically, we have access to the functional forms of p_ϑ and $\{p_{X|\vartheta}(\cdot \mid \theta)\}_{\theta\in\Theta}$, so we can evaluate these densities and compute gradients. However, the denominator of $p_{\vartheta|X}$ is precisely the normalizing constant described previously and cannot be naïvely evaluated. Moreover, even if we had the functional form of $p_{\vartheta|X}(\cdot \mid X)$, we would still not be able to compute expectations $\mathbb{E}[\varphi(\vartheta) \mid X]$ of test functions w.r.t. the posterior without evaluating another high-dimensional integral. Instead, the sampling methods we discuss in this book can output random variables $\vartheta_1, \ldots, \vartheta_n$ whose distributions are approximately $p_{\vartheta|X}(\cdot \mid X)$, and the expectation can be approximated to arbitrary accuracy via the averages $n^{-1}\sum_{i=1}^n \varphi(\vartheta_i)$.

2. (**high-dimensional integration**) More generally, computing integrals of functions against a known density π is a fundamental task in scientific computing. In many high-dimensional applications, the strategy of drawing samples from π and then approximating integrals via Monte Carlo averages is in fact the only known way to efficiently tackle this problem.

3. (**privacy**) As machine learning algorithms are continually deployed in application domains with personal and sensitive information, there is growing concern about maintaining the privacy of the data on which the machine learning models are trained. One way to address this issue is to require that the algorithm be *differentially private*, which loosely speaking requires the output of the model to not depend too much on the presence or absence of a single data point. The most common method to achieve this goal is via the careful addition of noise to the algorithm. Readers who are interested in the mathematics of privacy will benefit from a healthy understanding of the analysis of sampling algorithms, and vice versa.

4. (**statistical physics**) In a physical system, $V(x)$ represents the energy of a state x. In this situation, thermodynamics predicts that the equilibrium distribution

over states is the *Boltzmann* (or *Gibbs*) distribution whose density is proportional to $\exp(-V/T)$ (where T is the temperature of the system). Naturally, sampling provides a method for probing properties of the equilibrium distribution. More subtly, the mixing time of specific sampling algorithms also provides information about the system such as metastability phenomena; we revisit this in Chapter 11.

Due to this physical interpretation, we will often refer to V as the *potential energy*.

5. (**uncertainty quantification**) In order to better understand the risks inherent in any given system, it is important to quantify how much uncertainty is present in any given prediction. This application is closely related to the discussion on Bayesian statistics, since a Bayesian framework is a natural approach for performing uncertainty quantification. More generally, the choice to use sampling rather than optimization reflects a desire to understand typical outcomes of a procedure rather than choosing a single fitted model which may fall victim to model misspecification.

Besides these examples, it is no surprise that sampling arises in many other applications, since sampling is a fundamental algorithmic primitive. As such, sampling methods are employed daily in applied domains such as biology, climatology, and cosmology.

Example 0.E.1 (Bayesian logistic regression)**.** For concreteness, let us consider the application of sampling to a Bayesian logistic regression problem. Suppose we have collected data in the form of pairs (X_i, Y_i), $i = 1, \dots, n$, where $X_i \in \mathbb{R}^d$ is a vector of covariates and $Y_i \in \{0, 1\}$ is a binary outcome. For example, Y_i might represent whether or not a certain drug is effective on the i-th patient in a clinical study. Here, we regard the covariates $\{X_i,\ i = 1, \dots, n\}$ to be deterministic and fixed, and we posit that the outcomes $\{Y_i,\ i = 1, \dots, n\}$ are independent with distributions

$$Y_i \sim \mathsf{Bernoulli}\Big(\frac{\exp\langle \vartheta, X_i\rangle}{1 + \exp\langle \vartheta, X_i\rangle}\Big),$$

Moreover, we take a Gaussian prior $\mathsf{normal}(0, \lambda^{-1} I_d)$ for ϑ, where $\lambda > 0$. The likelihood is given by

$$p_{Y_i|\vartheta}(y_i \mid \theta) = \Big(\frac{1}{1 + \exp\langle \theta, X_i\rangle}\Big)^{1-y_i} \Big(\frac{\exp\langle \theta, X_i\rangle}{1 + \exp\langle \theta, X_i\rangle}\Big)^{y_i}, \qquad y_i \in \{0, 1\},$$

and by independence, $p_{Y_1,\dots,Y_n|\vartheta}(y_1, \dots, y_n \mid \theta) = \prod_{i=1}^n p_{Y_i|\vartheta}(y_i \mid \theta)$. A computation via

Bayes rule yields the posterior $p_{\vartheta|Y_1,\dots,Y_n} \propto \exp(-V)$, where

$$V(\theta) = \sum_{i=1}^{n} \left(\ln(1 + \exp\langle \theta, X_i \rangle) - Y_i \langle \theta, X_i \rangle \right) + \frac{\lambda}{2} \|\theta\|^2 .$$

Note that V is λ-strongly convex. It is straightforward to find the minimizer of V via standard optimization methods (e.g., gradient descent), and this corresponds to finding the mode or *maximum a posteriori* (MAP) estimate of the parameter ϑ. On the other hand, it is less obvious how to obtain (approximate) samples from the posterior. In this book, we study algorithms which can solve this task accompanied by non-asymptotic complexity estimates.

Next, we turn towards the *how* rather than the *why*. A key theme of this book is the surprising and close connection between methods in *optimization* and methods in *sampling*. To illustrate, we introduce our first sampling method, which is the sampling analogue of the well-known gradient descent algorithm from optimization. The **Langevin diffusion** is the solution $(Z_t)_{t\geq 0}$ to the stochastic differential equation (SDE)

$$\mathrm{d}Z_t = \underbrace{-\nabla V(Z_t)\,\mathrm{d}t}_{\text{gradient flow}} + \underbrace{\sqrt{2}\,\mathrm{d}B_t}_{\text{Brownian motion}} .$$

With a pure gradient flow $\mathrm{d}Z_t = -\nabla V(Z_t)\,\mathrm{d}t$, we would expect the dynamics to converge to stationary points of V. The Brownian motion ensures that we fully explore the distribution π, as is required in sampling. Under mild conditions, the unique stationary distribution of the Langevin diffusion is indeed $\pi \propto \exp(-V)$, which makes this diffusion a good candidate upon which to base a sampling algorithm.

Since the Langevin diffusion is "a gradient flow + noise", it is no wonder that researchers have drawn parallels between this diffusion and the gradient flow from optimization. However, the connection actually lies much deeper than this superficial observation would suggest. There is a natural geometry on the space of probability measures with finite second moment, $\mathcal{P}_2(\mathbb{R}^d)$, namely the 2-Wasserstein distance W_2 from the theory of optimal transport. The space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ turns out to be much richer than a metric space; in fact, it is *almost* a Riemannian manifold. In turn, the Riemannian structure allows us to define *gradient flows* on this space. The punchline here is that if π_t denotes the law of Z_t, then *the curve of measures $t \mapsto \pi_t$ is the gradient flow of the Kullback–Leibler (KL) divergence* $\mathsf{KL}(\cdot \,\|\, \pi)$ *with respect to the W_2 geometry*. Hence, at the level of the trajectory $(Z_t)_{t\geq 0}$, the Langevin diffusion is a noisy gradient flow, but at the level of measures $(\pi_t)_{t\geq 0}$, it is *precisely* a gradient flow! This remarkable connection was introduced in the seminal work of Jordan, Kinderlehrer, and Otto [JKO98].

This perspective suggests that we can study the convergence of the Langevin diffusion using tools from optimization. For example, a standard assumption in the optimization literature which allows for fast rates of convergence is that of *strong convexity* of the objective function. Hence, we can ask under what conditions the functional $\mathsf{KL}(\cdot \,\|\, \pi)$ on the space of measures is strongly convex along W_2 geodesics. Quite pleasingly, this is equivalent to the (Euclidean) strong convexity of the potential V. Consequently, the assumption of strong convexity of V, which is natural in optimization for studying gradient flows, turns out to be natural in the sampling context as well.

Much of this book is devoted to the case when V is strongly convex; we refer to π as being *strongly log-concave*. Besides its naturality and simplicity, it is also a practical assumption. For example, in the application to Bayesian statistics, the Bernstein–von Mises theorem states that as the number of data points tends to infinity, the posterior distribution closely resembles a Gaussian distribution and is thus (almost) strongly log-concave, a fact which has already been exploited to give sampling guarantees in the context of Bayesian inverse problems (see, e.g., [NW20]). However, some of the results also apply to restricted classes of non-log-concave measures, and in Chapter 11 we will see what can be said about non-log-concave sampling in general.

Before using the Langevin diffusion for sampling, however, it is first necessary to discretize the process in time. The simplest discretization, known as the *Euler–Maruyama discretization*, proceeds by fixing a step size $h > 0$ and following the iteration

$$X_{(k+1)h} \coloneqq X_{kh} - h\,\nabla V(X_{kh}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,.$$

Since the Brownian increment $B_{(k+1)h} - B_{kh}$ has the $\mathsf{normal}(0, hI_d)$ distribution, this iteration can be easily implemented once we have access to a gradient oracle for V and the ability to draw standard Gaussian variables. This iteration is commonly known as the *Langevin Monte Carlo* (LMC) algorithm, or the *unadjusted Langevin algorithm* (ULA); in this book, we stick to the former acronym.

The LMC algorithm is the starting point of our study. As a result of research in the last decade, we now have the following guarantee. For any of the common divergences d between probability measures, e.g., $\mathsf{d}(\mu, \pi) = W_2(\mu, \pi)$ or $\mathsf{d}(\mu, \pi) = \sqrt{\mathsf{KL}(\mu \,\|\, \pi)}$, and with an appropriate choice of initialization and step size, the law μ_{Nh} of the N-th iterate of LMC satisfies $\mathsf{d}(\mu_{Nh}, \pi) \le \varepsilon$ with a number of iterations N which is polynomial in the problem parameters (the dimension d, the condition number κ of V, and the inverse accuracy $1/\varepsilon$). For example, when $\mathsf{d} = \sqrt{\mathsf{KL}}$, the state-of-the-art guarantee reads $N = \widetilde{O}(\kappa d/\varepsilon^2)$. Whereas the convergence of the continuous-time diffusion is classical and typically proven via abstract calculus, the quantitative non-asymptotic convergence of the discretized algorithm necessitates the development of a new toolbox of analysis techniques. A primary goal of this book to make this toolbox more accessible to researchers who are

not yet acquainted with the field.

Beyond the standard LMC algorithm, there is now a rich arsenal of algorithms in the sampling literature. Some algorithms are directly inspired by other optimization algorithms (e.g., mirror descent), whereas other algorithms have their roots in the classical theory of Markov processes (e.g., the use of a Metropolis–Hastings filter). We will also explore some of these more sophisticated algorithms in detail, as in many cases they represent substantial improvements over standard LMC.

Finally, although we began this introduction by discussing the goal of understanding the *complexity* of sampling, in fact the complexity is not yet fully understood. The issue here is that there are currently very few *lower bounds* on the complexity of sampling. This is in contrast with the field of optimization, in which oracle complexity lower bounds have in most situations identified nearly optimal algorithms for optimizing various function classes. In Chapter 9, we will explain the current progress towards achieving this goal for sampling, but much work remains to be done.

For example, here is the precise statement for a fundamental open question about the complexity of sampling.

> Let $\pi \propto \exp(-V)$ be a probability density on $\mathbb{R}^d$. Determine, up to a universal constant, the minimum number of queries to a first-order oracle for V required to output a sample whose law μ satisfies $\|\mu - \pi\|_{\mathrm{TV}} \leq \varepsilon$, uniformly over the following class of potentials: V is twice continuously differentiable, satisfying the conditions $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$.

What This Book Does Not Contain

At the risk of offending researchers who are omitted even from the list of omissions, here we point out a few egregious exclusions. First, as mentioned previously, the price we paid for a succinct exposition of a variety of fields is a lack of rigorous development of the fundamentals of said fields, which we leave to the reader to pursue more thoroughly.

The field of sampling has a rich literature spanning decades, and although we have made an effort to cite the works most relevant to the modern perspective, it was not possible to cite even a vanishing fraction of the applied and/or classical literature. This extends to even recent theoretical works on log-concave sampling, for which we have omitted any discussion of sampling from convex bodies or polytopes. Although these works constitute fundamental developments in the field, here we chose to limit our focus to the part of the literature which is more strongly inspired by optimization algorithms.

Naturally, the other topics we explore in the book are far from comprehensive.

Notational Conventions

The symbols $\wedge$ and $\vee$ mean "minimum" and "maximum" respectively. We write $a \lesssim b$ or $a = O(b)$ to mean that $a \leq Cb$ for a universal constant $C > 0$. Similarly, $a \gtrsim b$ or $a = \Omega(b)$ mean that $a \geq cb$ for a universal constant $c > 0$, and $a \asymp b$ or $a = \Theta(b)$ mean that both $a \lesssim b$ and $a \gtrsim b$.

For a function $f : \mathbb{R}^d \to \mathbb{R}$, we write $\partial_i f$ to denote the i-th partial derivative of f. The gradient ∇f is the vector of partial derivatives $(\partial_1 f, \ldots, \partial_d f)$, and the Hessian $\nabla^2 f$ is the matrix $(\partial_i \partial_j f)_{i,j \in [d]}$. For a vector field $v : \mathbb{R}^d \to \mathbb{R}^d$, we also use the notation ∇v to denote the Jacobian matrix of v. The divergence of a vector field v is $\operatorname{div} v = \nabla \cdot v = \sum_{i=1}^d \partial_i v_i$, and the Laplacian of f is $\Delta f = \operatorname{tr} \nabla^2 f = \sum_{i=1}^d \partial_i^2 f$.

Acknowledgements

This book started off as a series of lecture notes during my visit to New York University in Spring 2022, and owes much to the audience and hospitality there. They were further developed during my time as a member at the Institute for Advanced Study, from 2023 through 2024, during which I was supported by the Eric and Wendy Schmidt Fund.

Before its inception, however, the idea of writing a book was first conceived during the Simons Institute for the Theory of Computing program on Geometric Methods in Sampling and Optimization (GMOS) in Fall 2021. There, I met a lot of my long-term collaborators and learned much of what I currently know about sampling. Overall, I found the program to be incredibly intellectually stimulating and I am grateful.

I am indebted to collaborators with whom I have had many fruitful conversations, including (but not limited to): Kwangjun Ahn, Jason M. Altschuler, Francis Bach, Krishnakumar Balasubramanian, Silvère Bonnabel, Joan Bruna, Sébastien Bubeck, Sitan Chen, Yongxin Chen, Yuansi Chen, Xiang Cheng, Jaume de Dios Pont, Alain Durmus, Ronen Eldan, Murat A. Erdogdu, Patrik R. Gerber, Ramon van Handel, Ye He, Marc Lambert, Thibaut Le Gouic, Holden Lee, Yin Tat Lee, Jerry Li, Mufan (Bill) Li, Yuanzhi Li, Chen Lu, Jianfeng Lu, Yian Ma, Tyler Maunu, Shyam Narayanan, Philippe Rigollet, Lionel Riou-Durand, Adil Salim, Ruoqi Shen, Austin J. Stromme, Kevin Tian, Jure Vogrinc, Lihan Wang, Andre Wibisono, Anru R. Zhang, and Matthew S. Zhang.

I am also appreciative of the encouragement and suggestions from Sébastien Bubeck, Jonathan Niles-Weed, Martin Wainwright, and my advisor Philippe Rigollet, as well as the careful reading of Michael Diao, Aram-Alexandre Pooladian, and Yandi Shen.

Thank you to all of the friends who kept me sane throughout.

Part I

Diffusions in Continuous Time

CHAPTER 1

The Langevin Diffusion in Continuous Time

In this chapter, we study the continuous-time **Langevin diffusion** with potential V, which is the solution to the following stochastic differential equation (SDE):

$$\mathrm{d}Z_t = -\nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,. \tag{1.E.1}$$

We begin with a quick introduction to stochastic calculus in order to make sense of this equation. Then, we introduce two powerful frameworks for analyzing the Langevin diffusion: *Markov semigroup theory*, and *the calculus of optimal transport*. These frameworks are two perspectives on the same diffusion, and the abstract calculus rules we develop within each framework streamline important computations.

A rigorous mathematical treatment of the theory in this chapter requires addressing substantial analytical technicalities, such as checking that the various partial differential equations (PDEs) are well-posed and that the calculations are carefully justified. We will not attempt to do so here and instead refer to bibliography for detailed treatments. In particular, the "proofs" in this section are more like "proof sketches" which are meant to convey the main intuition.

1.1 A Primer on Stochastic Calculus

In this section, we introduce just enough stochastic calculus to understand the meaning of the SDE (1.E.1). See [Ste01; Le 16] for thorough expositions. We treat further topics in stochastic calculus in Chapter 3.

In Section 1.1.4, we discuss the rather technical construction of the Itô integral. For the remainder of the book, the details of the construction are less important than the calculation rules that follow. The reader who is unfamiliar with stochastic calculus is encouraged to skip Section 1.1.4 upon first reading.

1.1.1 The Itô Integral

Definition 1.1.1. (Standard) **Brownian motion** is a stochastic process $(B_t)_{t\geq 0}$ in $\mathbb{R}^d$ satisfying the following properties:

1. $B_0 = 0$.
2. (independence of increments) For all $0 < t_1 < \cdots < t_k$, the random variables $(B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_k} - B_{t_{k-1}})$ are mutually independent.
3. (law of the increments) For all $0 \leq s < t < \infty$,
$$B_t - B_s \sim \mathsf{normal}(0, t-s)\,.$$
4. (continuity of the paths) Almost surely, $t \mapsto B_t$ is continuous.

Brownian motion was originally introduced over a century ago as a model for the jittery path of a particle which is constantly colliding with surrounding molecules. Since its inception, Brownian motion has been used to model the flow of heat, to price options at the financial market, to solve partial differential equations, to tease out the geometry of manifolds, and of course, to sample from probability distributions. It is perhaps not clear at first sight that such a process even exists,[1] but it would take us too far afield to give a construction here.

Instead, our goal is to compute integrals involving Brownian motion: given a stochastic process $(\eta_t)_{t\geq 0}$, how do we make sense of an expression such as $\int_0^T \eta_t \, \mathrm{d}B_t$? Once we have stochastic integration in hand, we can then formulate and solve stochastic *differential equations*. The solution to such an equation is a diffusion process, no less jittery than the Brownian motion which drives it, and yet in the right hands it becomes an incredible tool for solving a plethora of disparate problems.

The main technical difficulty in defining the stochastic integral is that Brownian motion is an irregular process: for small $t > 0$, by definition $B_t \sim \mathsf{normal}(0, t)$, which means that $|B_t|$ is typically of size $\asymp \sqrt{t}$. In particular, this prevents Brownian motion from being

[1] Of course, this did not stop Einstein from using it to probe the microscopic structure of matter.

differentiable at 0, or indeed, anywhere. Nevertheless, such stochastic integrals can be meaningfully defined and used to build a far-reaching calculus.

We give the high-level idea of the construction of the Itô integral here, deferring details to Section 1.1.4. We work on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ which is *complete*, *filtered*, and *right-continuous*, meaning that there is an increasing family $(\mathscr{F}_t)_{t\geq 0}$ of σ-algebras with $\bigcup_{t=0}^{\infty} \mathscr{F}_t \subseteq \mathscr{F}$, with $\bigcap_{t>s} \mathscr{F}_t = \mathscr{F}_s$ for all $s \geq 0$, and such that $\mathscr{F}_0$ contains all subsets of null sets. We assume that Brownian motion is adapted to the filtration: B_t is $\mathscr{F}_t$-measurable for each $t \geq 0$.

Defining the Itô integral at a single time T. Suppose first that $(\eta_t)_{t\geq 0}$ is a process of the form

$$\eta_t = \sum_{i=0}^{k-1} H_i \, \mathbb{1}\{t \in (t_i, t_{i+1}]\} \tag{1.1.2}$$

for some $0 \leq t_0 < t_1 < \cdots < t_k$, where H_i is bounded and $\mathscr{F}_{t_i}$-measurable. We call η an **elementary process**. In this case, perhaps the only reasonable definition of the stochastic integral is to take

$$\int_0^T \eta_t \, \mathrm{d}B_t \coloneqq \sum_{i=0}^{k-1} H_i \left(B_{t_{i+1}\wedge T} - B_{t_i \wedge T}\right). \tag{1.1.3}$$

This is indeed what we shall do, but for the moment we will refrain from using the integral symbol and write this as $\mathcal{I}_{[0,T]}(\eta)$ to avoid confusion.

We would like to extend this definition to more general processes, but before doing so we record two key properties of the stochastic integral. The first is that $t \mapsto \mathcal{I}_{[0,t]}(\eta)$ is a continuous martingale, i.e., it is continuous and satisfies the following definition.

Definition 1.1.4. A process $(M_t)_{t\geq 0}$ is a **martingale** w.r.t. the filtration $(\mathscr{F}_t)_{t\geq 0}$ if for all $t \geq 0$, M_t is $\mathscr{F}_t$-measurable and integrable, and

$$\mathbb{E}[M_t \mid \mathscr{F}_s] = M_s, \qquad \text{for all } 0 \leq s < t.$$

Indeed, we deduce that $t \mapsto \mathcal{I}_{[0,t]}(\eta)$ is a martingale from the fact that H_i is $\mathscr{F}_{t_i}$-measurable for each i, and because $(B_t)_{t\geq 0}$ is a martingale.

The second key property is that we can compute the variance:

$$\mathbb{E}[\mathcal{I}_{[0,T]}(\eta)^2] = \mathbb{E}\Big[\Big|\sum_{i=0}^{k-1} H_i \left(B_{t_{i+1}\wedge T} - B_{t_i\wedge T}\right)\Big|^2\Big] = \sum_{k=0}^{k-1} \mathbb{E}[|H_i \left(B_{t_{i+1}\wedge T} - B_{t_i \wedge T}\right)|^2] \tag{1.1.5}$$

$$= \sum_{k=0}^{k-1} \mathbb{E}[H_i^2] \left((t_{i+1} \wedge T) - (t_i \wedge T) \right) = \mathbb{E} \int_0^T \eta_t^2 \, \mathrm{d}t \,. \tag{1.1.6}$$

Here, we used the basic properties listed in the definition of Brownian motion, such as independence of increments. This equation shows that if $\mathbb{P}_T := \mathbb{P} \otimes \mathfrak{m}|_{[0,T]}$, where $\mathfrak{m}|_{[0,T]}$ is the Lebesgue measure on $[0, T]$, then the mapping $\eta \mapsto \mathcal{I}_{[0,T]}(\eta)$ is an isometry from $L^2(\mathbb{P}_T)$ to $L^2(\mathbb{P})$. We use this isometry to extend the definition of the stochastic integral as follows.

For a more general process $(\eta_t)_{t \geq 0}$, assume that it is *progressive*[2] and satisfies the integrability condition

$$\|\eta\|_{L^2(\mathbb{P}_T)}^2 = \mathbb{E} \int_0^T \eta_t^2 \, \mathrm{d}t < \infty \,.$$

One shows that $(\eta_t)_{t \geq 0}$ can be approximated by elementary processes $\{(\eta_t^{(k)})_{t \geq 0} : k \in \mathbb{N}\}$ of the form (1.1.2) in the $L^2(\mathbb{P}_T)$ norm. For each k, the stochastic integral $\mathcal{I}_{[0,T]}(\eta^{(k)})$ is defined via (1.1.3), and $\lim_{k \to \infty} \mathcal{I}_{[0,T]}(\eta^{(k)})$ exists in $L^2(\mathbb{P})$ thanks to the isometry. We can then take the limit to be the definition of the stochastic integral $\mathcal{I}_{[0,T]}(\eta)$.

Defining the Itô integral as a stochastic process. Although the procedure above successfully defines $\mathcal{I}_{[0,t]}(\eta)$ for a *fixed* time $t > 0$, there is no guarantee of coherence between different times t. The trouble arises because $\mathcal{I}_{[0,t]}(\eta)$ is defined as a limit, but this limit is only well-specified up to an event of measure zero, and these measure zero events for different times t might conceivably accumulate into something more. This is undesirable because the true power of stochastic calculus comes from viewing the stochastic integral as a time-indexed stochastic process in its own right.

The key insight is to go back to the approximating sequence $\{(\eta_t^{(k)})_{t \geq 0} : k \in \mathbb{N}\}$. For each k, the Itô integral is defined as an entire process $t \mapsto \mathcal{I}_{[0,t]}(\eta^{(k)})$ via (1.1.28), and moreover this process is a continuous martingale. We can then apply powerful results on martingale convergence, which are developed in Section 1.1.4, in order to prove the following theorem.

Theorem 1.1.7. *Suppose that $(\eta_t)_{t \geq 0}$ is progressive and satisfies $\mathbb{E} \int_0^T \eta_t^2 \, \mathrm{d}t < \infty$. Then, there exists a continuous martingale, denoted $(\int_0^t \eta_s \, \mathrm{d}B_s)_{t \geq 0}$, which is adapted to $(\mathcal{F}_t)_{t \geq 0}$*

[2] The process $(\eta_t)_{t \geq 0}$ is progressive if for all $T \geq 0$, the mapping $(\omega, t) \mapsto \eta_t(\omega)$ is measurable w.r.t. $\mathcal{F}_T \otimes \mathcal{B}_{[0,T]}$, where $\mathcal{B}_{[0,T]}$ is the Borel σ-algebra on $[0, T]$.

and satisfies

$$\mathbb{E}\Big[\Big|\int_0^t \eta_s \, \mathrm{d}B_s\Big|^2\Big] = \mathbb{E}\int_0^t \eta_s^2 \, \mathrm{d}s\,, \qquad \textit{for all } t \in [0,T]\,. \tag{1.1.8}$$

The formula (1.1.8) *is called the* ***Itô isometry****.*

Also, for each $t \in [0,T]$*, it holds that* $\int_0^t \eta_s \, \mathrm{d}B_s = \mathcal{I}_{[0,t]}(\eta)$ *a.s.*

Extending the definition via localization. There is one final step which is traditionally taken, namely to expand the class of allowable integrands to progressive processes η with

$$\int_0^T \eta_s^2 \, \mathrm{d}s < \infty \qquad \text{almost surely}\,. \tag{1.1.9}$$

Note that this condition is weaker than the condition $\mathbb{E}\int_0^T \eta_s^2 \, \mathrm{d}s < \infty$. Such an extension is evidently mathematically interesting, as we would like our definitions to be as broad as possible. However, equally important is that it introduces the device of **localization**. On the whole, localization actually serves to *reduce* the number of technicalities in the subject: once introduced, it allows us to always work with a stopping time up to which the process is as nice as one desires (e.g., bounded). The flexibility and utility that localization thus brings cements its place as the natural mathematical framework for stochastic calculus. However, this is not the focus of the book, and in what follows we will usually brush over such localization arguments. For now, we simply introduce the basic definitions in order to show the reader that the idea is actually fairly straightforward.

Definition 1.1.10. A **stopping time** τ is a random variable such that for each $t \geq 0$, the event $\{\tau \leq t\}$ is $\mathscr{F}_t$-measurable.

Definition 1.1.11. An increasing sequence of stopping times $(\tau_n)_{n\in\mathbb{N}}$ is called a **localizing sequence** for η on $[0,T]$ if:

1. for all $n \in \mathbb{N}$, $(\eta_t \, \mathbb{1}\{t \leq \tau_n\})_{t\geq 0}$ has finite $\|\cdot\|_{L^2(\mathbb{P}_T)}$ norm, and
2. $\tau_n \to T$ almost surely.

The good news is that localizing sequences are easy to find, and the following proposition barely needs a proof (and so we omit it).

Proposition 1.1.12. *If η is a progressive process satisfying the condition (1.1.9), then the sequence $(\tau_n)_{n\in\mathbb{N}}$ defined by*

$$\tau_n := \inf\left\{t \geq 0 \;\middle|\; \int_0^t \eta_s^2 \,\mathrm{d}s \geq n\right\} \wedge T$$

is a localizing sequence for η on $[0, T]$.

The idea now is simple: for each progressive process η satisfying (1.1.9), let $(\tau_n)_{n\in\mathbb{N}}$ be a localizing sequence for η on $[0, T]$. For each $n \in \mathbb{N}$, by the definition of a localizing sequence, we can apply our existing definition of the Itô integral, which gives us a continuous martingale

$$t \mapsto \int_0^t \eta_s \,\mathbb{1}\{s \leq \tau_n\} \,\mathrm{d}B_s \,. \tag{1.1.13}$$

Then, we can define the Itô integral of η to be a limit of the processes (1.1.13). The details are straightforward, and omitted.

We can also define an analogue of martingales using localizing sequences; these are almost martingales, but lack the required integrability.

Definition 1.1.14. A process $(M_t)_{t\geq 0}$ is a **local martingale** if it is adapted to the filtration $(\mathscr{F}_t)_{t\geq 0}$ and there is an increasing sequence $(\tau_n)_{n\in\mathbb{N}}$ of stopping times such that $\tau_n \to \infty$ and for each n, the process $t \mapsto M_{t\wedge\tau_n} - M_0$ is a martingale w.r.t. $(\mathscr{F}_t)_{t\geq 0}$.

Proposition 1.1.15. *If η is a progressive process satisfying (1.1.9), then the Itô integral $t \mapsto \int_0^t \eta_s \,\mathrm{d}B_s$ is a continuous local martingale.*

Looking forward. The construction of the Itô integral may seem quite abstract; indeed, we are sorely lacking in examples. At this juncture, it is common to work out simple exercises such as computing $\int_0^t B_s \,\mathrm{d}B_s$, and while this is pedagogically natural it is also liable to mislead the reader into thinking that the main use of Itô integration is to solve synthetic problems with no apparent purpose. As counterintuitive as it may seem, our solution to the heavy amount of abstraction will be *more abstraction*. In the next section, we will develop the single most important computation rule in stochastic calculus (along with the Itô isometry (1.1.8)), called Itô's formula, after which we will hardly need to

return to the definition of a stochastic integral ever again. And even Itô's formula will be abstracted out into the language of Markov semigroups in Section 1.2. The upshot is that we introduced the Itô integral because it is the foundation of our field, but most of what we have developed thus far is not necessary for the remainder of the book.

1.1.2 Itô's Formula

With the Itô integral in hand, we consider the following class of processes.

Definition 1.1.16. A stochastic process $(X_t)_{t\geq 0}$ is an **Itô process** if it is of the form

$$X_t = X_0 + \int_0^t b_s \, \mathrm{d}s + \int_0^t \sigma_s \, \mathrm{d}B_s \,, \qquad \text{for } t \geq 0 \,,$$

where $(b_t)_{t\geq 0}$ takes values in $\mathbb{R}^d$, $(\sigma_t)_{t\geq 0}$ takes values in $\mathbb{R}^{d\times N}$, and $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^N$.

Implicit in the above definition is that the process should be well-defined: the coefficients $(b_t)_{t\geq 0}$ and $(\sigma_t)_{t\geq 0}$ should be progressive processes for which the integrals exist. Also, the random variable X_0 should be $\mathscr{F}_0$-measurable, in which case the process $(X_t)_{t\geq 0}$ is also progressive.

We refer to $(b_t)_{t\geq 0}$ as the **drift coefficient** and $(\sigma_t)_{t\geq 0}$ as the **diffusion coefficient**. When the drift coefficient is zero, then $(X_t)_{t\geq 0}$ is simply an Itô integral, and thus a continuous local martingale (Proposition 1.1.15). Otherwise, for a non-zero drift coefficient, the process $(X_t)_{t\geq 0}$ is no longer necessarily a local martingale. As a shorthand, we often write the Itô process in differential form:

$$\mathrm{d}X_t = b_t \, \mathrm{d}t + \sigma_t \, \mathrm{d}B_t \,. \tag{1.1.17}$$

Our goal is to understand how the Itô process transforms when we compose it with a smooth function $f : \mathbb{R}^d \to \mathbb{R}$. This leads to *Itô's formula*, which is the bread and butter of stochastic calculus computations.

Although the notation (1.1.17) is informal, it conveys the main intuition. For $h > 0$ small, we can approximate $X_{t+h} \approx X_t + h\, b_t + \sqrt{h}\, \sigma_t \xi$, where $\xi \sim \mathrm{normal}(0, I_N)$. Note that the $\sqrt{h}$ scaling comes from the fact that the Brownian increment $B_{t+h} - B_t$ has the $\mathrm{normal}(0, hI_d)$ distribution. Now suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Normally, to compute $f(X_{t+h}) - f(X_t)$ up to order $o(h)$, a *first-order* Taylor expansion of f suffices, but in stochastic calculus this would miss important terms arising

from the Brownian motion: indeed, second-order terms in $B_{t+h} - B_t$ are of order h and hence not negligible.

Therefore, we carry out the Taylor expansion to an extra term:

$$\begin{aligned} f(X_{t+h}) - f(X_t) &\approx \langle \nabla f(X_t), h\, b_t + \sqrt{h}\, \sigma_t \xi \rangle + \frac{1}{2} \langle h\, b_t + \sqrt{h}\, \sigma_t \xi, \nabla^2 f(X_t)\, (h\, b_t + \sqrt{h}\, \sigma_t \xi) \rangle \\ &= h \left\{ \langle \nabla f(X_t), b_t \rangle + \frac{1}{2} \langle \sigma_t \xi, \nabla^2 f(X_t)\, \sigma_t \xi \rangle \right\} + \sqrt{h}\, \langle \sigma_t^\mathsf{T} \nabla f(X_t), \xi \rangle + o(h)\,. \end{aligned}$$

This expression suggests that $(f(X_t))_{t \geq 0}$ is also an Itô process. The third term, which is of order $\sqrt{h}$, turns into an Itô integral once integrated. Perhaps the most interesting term is the second term, $\frac{h}{2} \langle \sigma_t^\mathsf{T}\, \nabla^2 f(X_t)\, \sigma_t, \xi \xi^\mathsf{T} \rangle$, which is a genuinely new feature of stochastic calculus. If we sum up many of these increments, we end up with an expression like $\frac{1}{2} \sum_{k=0}^{K} \langle \sigma_{t+kh}^\mathsf{T}\, \nabla^2 f(X_{t+kh})\, \sigma_{t+kh}, h\, \xi_k \xi_k^\mathsf{T} \rangle$. If we replace each $h\, \xi_k \xi_k^\mathsf{T}$ by its expectation $h\, I_N$ (which must be carefully justified; see the calculation of quadratic variation in Section 3.1), then this resembles a Riemann sum, which converges to the integral of $\frac{1}{2} \langle \nabla^2 f(X_t), \sigma_t \sigma_t^\mathsf{T} \rangle$. This is formalized in the following theorem.

Theorem 1.1.18 (Itô's formula). *Let $(X_t)_{t \geq 0}$ be an Itô process, $\mathrm{d}X_t = b_t\, \mathrm{d}t + \sigma_t\, \mathrm{d}B_t$, and let $f \in C^2(\mathbb{R}^d)$. Then, $(f(X_t))_{t \geq 0}$ is also an Itô process which satisfies, for $t \geq 0$:*

$$f(X_t) - f(X_0) = \int_0^t \left\{ \langle \nabla f(X_s), b_s \rangle + \frac{1}{2} \langle \nabla^2 f(X_s), \sigma_s \sigma_s^\mathsf{T} \rangle \right\} \mathrm{d}s + \int_0^t \langle \sigma_s^\mathsf{T} \nabla f(X_s), \mathrm{d}B_s \rangle\,.$$

We omit the proof, since the bulk of the intuition is carried in the informal Taylor series argument described above. Observe that since Itô integrals are (under appropriate integrability conditions) continuous martingales, the expectation of the last term in Itô's formula is typically zero. Therefore,[3]

$$\mathbb{E}\, f(X_t) - \mathbb{E}\, f(X_0) = \int_0^t \mathbb{E}\Big[\langle \nabla f(X_s), b_s \rangle + \frac{1}{2} \langle \nabla^2 f(X_s), \sigma_s \sigma_s^\mathsf{T} \rangle \Big]\, \mathrm{d}s\,, \tag{1.1.19}$$

or in differential form,

$$\partial_t\, \mathbb{E}\, f(X_t) = \mathbb{E}\Big[\langle \nabla f(X_t), b_t \rangle + \frac{1}{2} \langle \nabla^2 f(X_t), \sigma_t \sigma_t^\mathsf{T} \rangle \Big]\,.$$

Itô's formula can also be extended to time-dependent functions via

$$f(t, X_t) - f(0, X_0) = \int_0^t \left\{ \partial_s f(s, X_s) + \langle \nabla f(s, X_s), b_s \rangle + \frac{1}{2} \langle \nabla^2 f(s, X_s), \sigma_s \sigma_s^\mathsf{T} \rangle \right\} \mathrm{d}s$$

[3] Even when the Itô integral is only a local martingale, one can usually justify the formula (1.1.19) anyway using the localizing sequence and the dominated convergence theorem.

$$+ \int_0^t \langle \sigma_s^{\mathsf{T}} \nabla f(s, X_s), \mathrm{d}B_s \rangle \,.$$

We revisit and streamline Itô's formula in Section 3.1.

1.1.3 Existence and Uniqueness of SDEs

Let $b : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d \times N}$. We now consider the stochastic differential equation (SDE)

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}B_t \,. \tag{1.1.20}$$

Suppose we are given a complete filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ which supports a standard N-dimensional adapted Brownian motion B. A **solution** to the SDE is an adapted $\mathbb{R}^d$-valued process X such that

$$X_t = X_0 + \int_0^t b(s, X_s)\,\mathrm{d}s + \int_0^t \sigma(s, X_s)\,\mathrm{d}B_s \,.$$

The question we address in this section is under what conditions there exists a unique solution to the SDE. This is a question of a technical nature, but the answer is instructive because the proof introduces standard arguments that recur frequently in stochastic calculus. The main result is that if the coefficients b, σ are Lipschitz in space uniformly in time, then the SDE admits a unique solution.

Before proceeding, we need the following lemma, which is used throughout the book.

Lemma 1.1.21 (Grönwall's lemma). *Let $T > 0$ and let $g : [0, T] \to [0, \infty)$ be bounded and measurable. Assume there exists $C_1, C_2 \geq 0$ such that*

$$g(t) \leq C_1 + C_2 \int_0^t g\,, \qquad \forall t \in [0, T] \,.$$

Then,

$$g(t) \leq C_1 \exp(C_2 t)\,, \qquad \forall t \in [0, T] \,.$$

Proof. By iterating the assumption, for each $n \in \mathbb{N}$,

$$g(t) \leq C_1 + C_2 \int_0^t g(s_1)\,\mathrm{d}s_1 \leq C_1 + C_1 C_2 t + C_2^2 \int_0^t \int_0^{s_1} g(s_2)\,\mathrm{d}s_2\,\mathrm{d}s_1 \leq \cdots$$

$$\leq C_1 \sum_{k=0}^{n} \frac{(C_2 t)^k}{k!} + C_2^{n+1} \int_0^t \cdots \int_0^{s_n} g(s_{n+1}) \, \mathrm{d}s_{n+1} \cdots \mathrm{d}s_1 \, .$$

The remainder term is bounded by

$$\left| C_2^{n+1} \int_0^t \cdots \int_0^{s_n} g(s_{n+1}) \, \mathrm{d}s_{n+1} \cdots \mathrm{d}s_1 \right| \leq \frac{\|g\|_{\sup} \, (C_2 t)^{n+1}}{(n+1)!} \xrightarrow{n \to \infty} 0 \, ,$$

which gives the result. □

In the following theorem, for a matrix M,

$$\|M\|_{\mathrm{HS}}^2 := \mathrm{tr}(MM^{\mathsf{T}}) = \mathrm{tr}(M^{\mathsf{T}}M)$$

denotes the Hilbert–Schmidt (or Frobenius) norm of M. The uniqueness result states that if there are two solutions $X, \widetilde{X}$ to SDE on the same probability space, driven by the same Brownian motion, then $X = \widetilde{X}$.

Theorem 1.1.22 (existence and uniqueness of SDE solutions)**.** *Assume that b and σ are continuous, and there exists $C > 0$ such that for all $t \geq 0$ and $x, y \in \mathbb{R}^d$,*

$$\|b(t,x) - b(t,y)\| \vee \|\sigma(t,x) - \sigma(t,y)\|_{\mathrm{HS}} \leq C \, \|x - y\| \, .$$

Then, for any complete filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ and $x \in \mathbb{R}^d$, there exists a unique solution $(X_t)_{t \in [0,T]}$ for the SDE (1.1.20) with $X_0 = x$. Moreover, the solution $(X_t)_{t \in [0,T]}$ is a Markov process.

Proof. Uniqueness. Fix a time $T > 0$ and suppose there exist two solutions $(X_t)_{t \in [0,T]}$ and $(\widetilde{X}_t)_{t \in [0,T]}$ of the SDE on $[0,T]$ with $X_0 = \widetilde{X}_0$. For $t \geq 0$, we compute the difference between X_t and $\widetilde{X}_t$ using the Itô isometry (1.1.8), the Cauchy–Schwarz inequality, and the Lipschitz assumption:

$$\begin{aligned}
\mathbb{E}[\|X_t - \widetilde{X}_t\|^2] &\leq 2 \, \mathbb{E}\left[\left\| \int_0^t \{b(s, X_s) - b(s, \widetilde{X}_s)\} \, \mathrm{d}s \right\|^2 + \left\| \int_0^t \{\sigma(s, X_s) - \sigma(s, \widetilde{X}_s)\} \, \mathrm{d}B_s \right\|^2 \right] \\
&\leq 2 \, \mathbb{E}\left[T \int_0^t \|b(s, X_s) - b(s, \widetilde{X}_s)\|^2 \, \mathrm{d}s + \int_0^t \|\sigma(s, X_s) - \sigma(s, \widetilde{X}_s)\|_{\mathrm{HS}}^2 \, \mathrm{d}s \right] \\
&\leq 2C^2 \, (1 + T) \, \mathbb{E} \int_0^t \|X_s - \widetilde{X}_s\|^2 \, \mathrm{d}s \, .
\end{aligned}$$

From Gronwall's lemma (Lemma 1.1.21), we obtain $X_t = \widetilde{X}_t$ almost surely. Actually, this proof has a gap: we have not fully justified why we can apply the Itô isometry. To get around this, one may use technique of localization, the details of which are omitted.

Existence. We use a method of establishing existence of solutions to ODEs, known as **Picard iteration**. We start by defining the process $X^{(0)}$ to be the constant process with value x, and for $n \in \mathbb{N}^+$ we let

$$X_t^{(n)} := x + \int_0^t b(s, X_s^{(n-1)})\,\mathrm{d}s + \int_0^t \sigma(s, X_s^{(n-1)})\,\mathrm{d}B_s\,. \tag{1.1.23}$$

In other words, we "freeze" the coefficients of the SDE using the process from the previous stage of the iteration. The stochastic integrals make sense because inductively, each $X^{(n)}$ is adapted and has continuous sample paths. Thus, since

$$X_t^{(n+1)} - X_t^{(n)} = \int_0^t \{b(s, X_s^{(n)}) - b(s, X_s^{(n-1)})\}\,\mathrm{d}s + \int_0^t \{\sigma(s, X_s^{(n)}) - \sigma(s, X_s^{(n-1)})\}\,\mathrm{d}B_s$$

we bound

$$\begin{aligned}\mathbb{E}\sup_{[0,t]} \|X^{(n+1)} - X^{(n)}\|^2 \le 2\,\mathbb{E}\Big[&\sup_{u\in[0,t]} \Big\| \int_0^u \{b(s, X_s^{(n)}) - b(s, X_s^{(n-1)})\}\,\mathrm{d}s \Big\|^2 \\ &+ \sup_{u\in[0,t]} \Big\| \int_0^u \{\sigma(s, X_s^{(n)}) - \sigma(s, X_s^{(n-1)})\}\,\mathrm{d}B_s \Big\|^2 \Big] \\ &=: \mathrm{I} + \mathrm{II}\,.\end{aligned}$$

The first term is handled via Cauchy–Schwarz:

$$\mathrm{I} \le T\,\mathbb{E}\int_0^t \|b(s, X_s^{(n)}) - b(s, X_s^{(n-1)})\|^2\,\mathrm{d}s \le C^2 T \int_0^t \mathbb{E}\sup_{[0,s]} \|X^{(n)} - X^{(n-1)}\|^2\,\mathrm{d}s\,.$$

For the second term, recall that $t \mapsto \int_0^t \{\sigma(s, X_s^{(n)}) - \sigma(s, X_s^{(n-1)})\}\,\mathrm{d}B_s$ is a martingale. By Doob's L^2 maximal inequality (Corollary 1.1.31 in Section 1.1.4) and the Itô isometry (1.1.8), we can bound

$$\begin{aligned}\mathrm{II} &\le 4\,\mathbb{E}\Big[\Big\| \int_0^t \{\sigma(s, X_s^{(n)}) - \sigma(s, X_s^{(n-1)})\}\,\mathrm{d}B_s \Big\|^2 \Big] = 4\,\mathbb{E}\int_0^t \|\sigma(s, X_s^{(n)}) - \sigma(s, X_s^{(n-1)})\|_{\mathrm{HS}}^2\,\mathrm{d}s \\ &\le 4C^2 \int_0^t \mathbb{E}\sup_{[0,s]} \|X^{(n)} - X^{(n-1)}\|^2\,\mathrm{d}s\,.\end{aligned}$$

Putting these two bounds together,

$$\mathbb{E} \sup_{[0,t]} \|X^{(n+1)} - X^{(n)}\|^2 \leq 2C^2 (4+T) \int_0^t \mathbb{E} \sup_{[0,s]} \|X^{(n)} - X^{(n-1)}\|^2 \, \mathrm{d}s \, .$$

By induction (and using the fact that $\mathbb{E} \sup_{[0,T]} \|X^{(1)} - X^{(0)}\|^2 \leq C_T$ for some constant $C_T < \infty$, which follows from similar arguments), we get

$$\mathbb{E} \sup_{[0,T]} \|X^{(n)} - X^{(n-1)}\|^2 \leq C_T \frac{\{2C^2 T (4+T)\}^{n-1}}{(n-1)!} \, .$$

If we sum this, then we obtain

$$\mathbb{E} \sum_{n=1}^{\infty} \sup_{[0,T]} \|X^{(n)} - X^{(n-1)}\|^2 < \infty \implies \sum_{n=1}^{\infty} \sup_{[0,T]} \|X^{(n)} - X^{(n-1)}\|^2 < \infty \text{ almost surely} \, .$$

By completeness of $C([0,T])$, it implies that $X^{(n)}$ converges uniformly on $[0,T]$ to a continuous process X as $n \to \infty$.

Using again the Lipschitz assumption on the coefficients of the SDE, we have

$$\int_0^t \sigma(s, X_s^{(n)}) \, \mathrm{d}B_s - \int_0^t \sigma(s, X_s) \, \mathrm{d}B_s \to 0 \, ,$$
$$\int_0^t b(s, X_s^{(n)}) \, \mathrm{d}s - \int_0^t b(s, X_s) \, \mathrm{d}s \to 0 \, ,$$

almost surely. Passing to the limit in (1.1.23) shows that X solves the SDE on $[0,T]$.

There is one last argument to make, which is to find a solution for the SDE on the entire time interval $[0,\infty)$. For each $t > 0$, let $T > t$ and define X_t to be the solution of the SDE on $[0,T]$ at time t. The uniqueness assertion in the first half of the theorem shows that this is well-defined (regardless of the choice of T), and it also shows that the resulting process X is adapted, has continuous sample paths, and solves the SDE on $[0,\infty)$.

We omit the proof that $(X_t)_{t\in[0,T]}$ is Markov. □

Reflecting on the proof, the basic strategy is to coupling together two diffusions (with the same driving Brownian motion), use Lipschitz bounds on the coefficients, and apply Gronwall's inequality. The same strategy will also be used to obtain convergence bounds for sampling algorithms, albeit with a more quantitative goal in mind.

A theorem similar in spirit to Theorem 1.1.22 can be established under the assumption that b and σ are only *locally* Lipschitz, but in this case the solution to the SDE is not guaranteed to last for all time. The issue is that when the coefficients grow faster than linearly, there can be a finite (random) time $\mathfrak{e}$, called the *explosion time*, such that $\|X_t\| \to \infty$ as $t \to \mathfrak{e}$. This phenomenon is already present for ODEs (see Exercise 1.4). For the purposes of this book, the assumption of Lipschitz coefficients suffices.

1.1.4 Appendix: Construction of the Itô Integral

To appreciate the construction, it may be helpful to first discuss some technical subtleties. First, suppose that $(\eta_t)_{t\geq 0}$ is deterministic (non-random). Since $(B_t)_{t\geq 0}$ has continuous paths, one could try to define $\int_0^T \eta_t \, dB_t$ as a Riemann–Stieltjes integral, but due to the irregularity of Brownian motion this only works for integrands $(\eta_t)_{t\geq 0}$ which are "*locally of bounded variation*"[4]; in particular not all continuous integrands are allowed. To get around this, a standard idea is to *approximate* a continuous integrand $(\eta_t)_{t\geq 0}$ with a sequence of integrands $\{(\eta_t^{(k)})_{t\geq 0} : k \in \mathbb{N}\}$ which have bounded variation. For each k, it is no trouble to define $\int_0^T \eta_t^{(k)} \, dB_t$, and then we can define $\int_0^T \eta_t \, dB_t$ as a suitable limit. This idea can be carried out, but the notion of limit that is used is L^2. In particular, the limit may not exist in an almost sure sense.

Now suppose that $(\eta_t)_{t\geq 0}$ is a stochastic process. Then, it becomes technically challenging to carry out the requisite approximations; another idea is required. The new insight is that if we consider integrands which are adapted to a filtration, then the stochastic integrals $t \mapsto \int_0^t \eta_s \, dB_s$ are martingales, and we can leverage their powerful convergence theory to streamline the construction. We now proceed to implement this plan.

Throughout, we work on a complete, filtered, and right-continuous probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. We also recall the various classes of processes that we introduced in Section 1.1.1.

Definition 1.1.24. Let $(\eta_t)_{t\geq 0}$ be a stochastic process.

1. $(\eta_t)_{t\geq 0}$ is a **progressive process** if for all $T \geq 0$, the mapping $(\omega, t) \mapsto \eta_t(\omega)$ is measurable w.r.t. $\mathcal{F}_T \otimes \mathcal{B}_{[0,T]}$, where $\mathcal{B}_{[0,T]}$ is the Borel σ-algebra on $[0, T]$.

2. $(\eta_t)_{t\geq 0}$ is an **elementary process** if it is of the form

$$\eta_t = \sum_{i=0}^{k-1} H_i \, \mathbb{1}\{t \in (t_i, t_{i+1}]\}, \tag{1.1.25}$$

   where $0 \leq t_0 < t_1 < \cdots < t_k$ and for each i, H_i is bounded and $\mathcal{F}_{t_i}$-measurable.

3. $(\eta_t)_{t\in[0,T]}$ is a **square integrable process** if it is a progressive process and moreover $\mathbb{E}\int_0^T \eta_t^2 \, dt < \infty$.

The proof of the following technical result is omitted.

[4] See Section 3.1 for further discussion.

Lemma 1.1.26 (approximation via elementary processes). *For any square integrable process $(\eta_t)_{t\in[0,T]}$, there is a sequence $\{(\eta_t^{(k)})_{t\geq 0} : k \in \mathbb{N}\}$ of elementary processes with*

$$\|\eta - \eta^{(k)}\|^2_{L^2(\mathbb{P}_T)} = \mathbb{E}\int_0^T |\eta_t - \eta_t^{(k)}|^2 \,\mathrm{d}t \to 0 \qquad \text{as } k \to \infty\,.$$

Definition 1.1.27. Let $(\eta_t)_{t\geq 0}$ be a progressive process.

1. If $(\eta_t)_{t\geq 0}$ is an elementary process of the form (1.1.25), we define the **Itô integral** of $(\eta_t)_{t\geq 0}$ on $[0,T]$ via

$$\mathcal{I}_{[0,T]}(\eta) := \sum_{i=0}^{k-1} H_i \left(B_{t_{i+1}\wedge T} - B_{t_i\wedge T}\right). \tag{1.1.28}$$

2. If $(\eta_t)_{t\geq 0}$ is a square integrable process, then let $\{(\eta_t^{(k)})_{t\geq 0} : k \in \mathbb{N}\}$ be the approximating sequence furnished by Lemma 1.1.26. In this case, we define the **Itô integral** of $(\eta_t)_{t\geq 0}$ via

$$\mathcal{I}_{[0,T]}(\eta) := \lim_{k\to\infty} \mathcal{I}_{[0,T]}(\eta^{(k)})$$

where the limit is taken in $L^2(\mathbb{P})$.

The second part of the definition requires some justification. We checked in (1.1.5) and (1.1.6) that the Itô isometry holds for elementary processes: the mapping $\mathcal{I}_{[0,T]}$ is an isometry from elementary processes equipped with the $L^2(\mathbb{P}_T)$ norm, to the space $L^2(\mathbb{P})$ of square integrable random variables. Through this isometry, we deduce from the fact that $\{\eta^{(k)} : k \in \mathbb{N}\}$ is Cauchy that $\{\mathcal{I}_{[0,T]}(\eta^{(k)}) : k \in \mathbb{N}\}$ is also Cauchy. Since $L^2(\mathbb{P})$ is a complete metric space, there exists a limit $\mathcal{I}_{[0,T]}(\eta)$ of the latter sequence. We can then deduce that the Itô isometry (1.1.8) holds for square integrable processes as well.

Upon trying to view $t \mapsto \mathcal{I}_{[0,t]}(\eta)$ as a stochastic process, we encounter the usual measure-theoretic difficulty: for fixed t, $\mathcal{I}_{[0,t]}(\eta)$ is well-defined outside of a measure zero event, but we have to contend with uncountably many values of t and the measure zero events may accumulate. Overcoming this issue requires some principle that holds uniformly over $t \in [0,T]$; in our case, this principle is Doob's maximal inequality from the theory of martingales.

To set the stage, we can verify from the explicit formula (1.1.28) that $t \mapsto \mathcal{I}_{[0,t]}(\eta)$ is a continuous martingale when $(\eta_t)_{t\geq 0}$ is an elementary process. Before proceeding onwards, we need to further develop the theory of martingales.

Definition 1.1.29. A process $(M_t)_{t\geq 0}$ is a **submartingale** w.r.t. the filtration $(\mathcal{F}_t)_{t\geq 0}$ if for all $t \geq 0$, M_t is $\mathcal{F}_t$-measurable and integrable, and

$$\mathbb{E}[M_t \mid \mathcal{F}_s] \geq M_s\,, \qquad \text{for all } 0 \leq s < t\,.$$

The class of submartingales is far broader and more useful than the class of martingales. For example, if $(M_t)_{t\geq 0}$ is a martingale and $\varphi : \mathbb{R} \to \mathbb{R}$ is any convex function with $\mathbb{E}|\varphi(M_t)| < \infty$ for all $t \geq 0$, then Jensen's inequality for conditional expectations implies that $(\varphi(M_t))_{t\geq 0}$ is a submartingale.

One of the key facts about submartingales is that they easily converge, which is often deduced from Doob's maximal inequality.

Theorem 1.1.30 (Doob's maximal inequality)**.** *Let $(M_t)_{t\geq 0}$ be a continuous and non-negative submartingale. Then, for all $\lambda, T > 0$,*

$$\mathbb{P}\Big(\sup_{t\in[0,T]} M_t \geq \lambda\Big) \leq \frac{\mathbb{E}\, M_T}{\lambda}\,.$$

Proof. We prove the theorem for discrete-time submartingales $(M_n)_{n\in\mathbb{N}}$. The result for continuous-time submartingales can then be obtained via approximation.

Let $\tau := \min\{k \in \mathbb{N} : M_k \geq \lambda\}$. On the event $\{\tau \leq N\}$, we have $M_\tau \geq \lambda$, so

$$\lambda\,\mathbb{P}\Big(\max_{k=0,1,\dots,N} M_k \geq \lambda\Big) = \lambda\,\mathbb{P}(\tau \leq N) \leq \mathbb{E}[M_\tau\,\mathbb{1}\{\tau \leq N\}] = \sum_{k=0}^{N} \mathbb{E}[M_k\,\mathbb{1}\{\tau = k\}]\,.$$

Next, since $\{\tau = k\}$ is $\mathcal{F}_k$-measurable, the submartingale property yields

$$\mathbb{E}[M_k\,\mathbb{1}\{\tau = k\}] \leq \mathbb{E}\big[\mathbb{E}[M_N \mid \mathcal{F}_k]\,\mathbb{1}\{\tau = k\}\big] = \mathbb{E}[M_N\,\mathbb{1}\{\tau = k\}]\,.$$

Hence,

$$\lambda\,\mathbb{P}\Big(\max_{k=0,1,\dots,N} M_k \geq \lambda\Big) \leq \mathbb{E}\Big[M_N \sum_{k=0}^{N} \mathbb{1}\{\tau = k\}\Big] = \mathbb{E}[M_N\,\mathbb{1}\{\tau \leq N\}] \leq \mathbb{E}\, M_N\,,$$

where we used the assumption that the submartingale is non-negative. □

It yields the following corollary, which we leave as Exercise 1.1.

Corollary 1.1.31 (Doob's L^p maximal inequality). *Let $(M_t)_{t\geq 0}$ be a continuous and non-negative submartingale. Then, for all $p > 1$ and $T > 0$,*

$$\Big\| \sup_{t\in[0,T]} M_t \Big\|_{L^p(\mathbb{P})} \leq \frac{p}{p-1} \|M_T\|_{L^p(\mathbb{P})} .$$

We are now ready to finish the construction of the Itô integral.

Proof of Theorem 1.1.7. Let $(\eta_t)_{t\geq 0}$ be a square integrable process, with approximating sequence $\{(\eta_t^{(k)})_{t\geq 0} : k \in \mathbb{N}\}$ obtained from Lemma 1.1.26. We define

$$X_t^{(k)} := \mathcal{I}_{[0,t]}(\eta^{(k)}) ,$$

where $\mathcal{I}_{[0,t]}(\eta^{(k)})$ is defined by the explicit formula (1.1.28). For any $k, \ell \in \mathbb{N}$, the process $t \mapsto |X_t^{(k)} - X_t^{(\ell)}|^2$ is a continuous non-negative submartingale, so Doob's maximal inequality (Theorem 1.1.30) and the Itô isometry (1.1.8) yield

$$\begin{aligned}\mathbb{P}\Big(\sup_{t\in[0,T]} |X_t^{(k)} - X_t^{(\ell)}| \geq \lambda \Big) &= \mathbb{P}\Big(\sup_{t\in[0,T]} |X_t^{(k)} - X_t^{(\ell)}|^2 \geq \lambda^2 \Big) \\ &\leq \frac{\mathbb{E}[|X_T^{(k)} - X_T^{(\ell)}|^2]}{\lambda^2} = \frac{\|\eta^{(k)} - \eta^{(\ell)}\|^2_{L^2(\mathbb{P}_T)}}{\lambda^2} .\end{aligned}$$

As $(\eta^{(k)})_{k\in\mathbb{N}}$ is Cauchy, we can pick a sequence $n_0 < n_1 < n_2 < \cdots$ of integers such that $k, \ell \geq n_j$ implies $\|\eta^{(k)} - \eta^{(\ell)}\|^2_{L^2(\mathbb{P}_T)} \leq 2^{-3j}$. Take $\lambda = 2^{-j}$ to obtain

$$\mathbb{P}\Big(\sup_{t\in[0,T]} |X_t^{(n_j)} - X_t^{(n_{j+1})}| \geq 2^{-j} \Big) \leq 2^{-j} .$$

These probabilities are summable, so the Borel–Cantelli lemma implies

$$\text{almost surely,} \quad \sup_{t\in[0,T]} |X_t^{(n_j)} - X_t^{(n_{j+1})}| \leq 2^{-j} \quad \text{for all but finitely many } j .$$

In particular, the paths $\{(X_t^{(n_j)})_{t\geq 0} : j \in \mathbb{N}\}$ form a Cauchy sequence in $C([0,T])$ (equipped with the supremum norm). As $C([0,T])$ is complete, there is a limit $(X_t)_{t\geq 0}$ which belongs to $C([0,T])$. A similar argument, this time using Doob's L^2 maximal inequality (Corollary 1.1.31), shows that $X_t^{(n_j)} \to X_t$ in $L^2(\mathbb{P})$ as well. Since each $t \mapsto X_t^{(n_j)}$ is a continuous martingale, so is $t \mapsto X_t$.

Finally, for any fixed $t \in [0,1]$, on one hand we have $\mathcal{I}_{[0,t]}(\eta^{(n_j)}) = X_t^{(n_j)} \to X_t$ in $L^2(\mathbb{P})$ as noted above. On the other hand, the Itô isometry implies $\mathcal{I}_{[0,t]}(\eta^{(n_j)}) \to \mathcal{I}_{[0,t]}(\eta)$ in $L^2(\mathbb{P})$. Hence, almost surely, $X_t = \mathcal{I}_{[0,t]}(\eta)$. □

1.2 Markov Semigroup Theory

More thorough treatments of Markov semigroup theory can be found in [BGL14; Han16]. We will revisit and expand upon many of the topics introduced here in Chapter 2.

1.2.1 Basic Definitions and Kolmogorov's Equations

The core idea of Markov semigroup theory is to encode the behavior of a Markov process $(X_t)_{t\geq 0}$ via operators which act on functions. We will then develop calculus rules for working with these operators, and we can study the operators via functional analysis. This is analogous to how the linear algebraic study of the transition matrix of a discrete-time Markov chain reveals properties (e.g., ergodicity, convergence) of the chain.

Definition 1.2.1. For a time-homogeneous Markov process $(X_t)_{t\geq 0}$, its associated **Markov semigroup** $(P_t)_{t\geq 0}$ is the family of operators acting on functions via

$$P_t f(x) := \mathbb{E}[f(X_t) \mid X_0 = x] .$$

The Markov property and iterated conditioning yields the following lemma (exercise).

Lemma 1.2.2. *The Markov semigroup $(P_t)_{t\geq 0}$ satisfies $P_0 = \mathrm{id}$ and $P_s P_t = P_t P_s = P_{s+t}$ for all $s, t \geq 0$.*

In order to do calculus, we want to differentiate the semigroup $t \mapsto P_t$, which is accomplished via the following definition.

Definition 1.2.3. The **infinitesimal generator** $\mathscr{L}$ associated with a Markov semigroup $(P_t)_{t\geq 0}$ is the operator defined by

$$\mathscr{L} f := \lim_{t \searrow 0} \frac{P_t f - f}{t} ,$$

for all functions f for which the above limit exists.

Here we pause to warn the reader of some technical issues. The mathematical difficulties of Markov semigroup theory arise in trying to answer the following questions: on what space of functions is the generator defined, and in what sense is the above limit taken? As we shall see, a natural space of functions to consider is $L^2(\pi)$, with π denoting

the stationary distribution of the diffusion. However, the generator is usually a differential operator, and not all functions in $L^2(\pi)$ have enough regularity to lie in the domain of the generator. The theory of *unbounded* linear operators on a Hilbert space was developed to handle this situation, but it is rife with subtle distinctions such as the difference between symmetric and self-adjoint operators. We will brush over these issues and focus on the calculation rules.

Example 1.2.4. Let us compute the generator of the Langevin diffusion (1.E.1). In fact, the following computation is simply a consequence of Itô's formula (Theorem 1.1.18), but it does not hurt to derive this result from scratch. We approximate

$$Z_t = Z_0 - \int_0^t \nabla V(Z_s)\,\mathrm{d}s + \sqrt{2}\,B_t = Z_0 - t\,\nabla V(Z_0) + \sqrt{2}\,B_t + o(t)\,.$$

Assuming that $f \in C^2(\mathbb{R}^d)$ with bounded derivatives, we perform a Taylor expansion of f to second order.

$$\mathbb{E}\,f(Z_t) = \mathbb{E}[f(Z_0) + \langle \nabla f(Z_0), -t\,\nabla V(Z_0) + \sqrt{2}\,B_t\rangle + \langle \nabla^2 f(Z_0)\,B_t, B_t\rangle] + o(t)\,.$$

Since B_t is mean zero and independent of Z_0, with $\mathbb{E}[B_t B_t^\mathsf{T}] = tI_d$,

$$\begin{aligned}\mathbb{E}[f(Z_t) \mid Z_0 = z] &= f(z) - t\,\langle \nabla f(z), \nabla V(z)\rangle + t\,\mathrm{tr}\,\nabla^2 f(z) + o(t)\\ &= f(z) - t\,\langle \nabla f(z), \nabla V(z)\rangle + t\,\Delta f(z) + o(t)\,.\end{aligned}$$

Hence,

$$\mathscr{L} f(z) = \lim_{t \searrow 0} \frac{\mathbb{E}[f(Z_t) \mid Z_0 = z] - f(z)}{t} = \Delta f(z) - \langle \nabla V(z), \nabla f(z)\rangle\,.$$

The Markov semigroup and dynamics. As promised, the Markov semigroup captures the information that was contained in the original Markov process. One way to demonstrate this is to prove theorems which show that the Markov process can be completely recovered from its Markov semigroup. Another approach, which we now take up, is to show that the dynamics of the Markov process are captured via calculation rules involving the Markov semigroup.

Proposition 1.2.5 (Kolmogorov's backward equation)**.** *For all $t \geq 0$, it holds that*

$$\partial_t P_t f = \mathscr{L} P_t f = P_t \mathscr{L} f \,.$$

In particular, $\mathscr{L}$ commutes with the semigroup $(P_t)_{t\geq 0}$.

Proof. Observe that

$$\lim_{h \searrow 0} \frac{P_{t+h} f - P_t f}{h} = \lim_{h \searrow 0} \frac{P_h - \mathrm{id}}{h} P_t f = \mathscr{L} P_t f \,.$$

Repeating the computation but factoring out P_t on the left yields the second equality. □

There is a dual to this equation: let π_0 denote the density of X_0. Formally, we can write $\mathbb{E} f(X_t) = \int P_t f \, \mathrm{d}\pi_0 = \int f \, \mathrm{d}P_t^* \pi_0$, where P_t^* is the adjoint of P_t. This says that the law π_t of X_t is formally given by $P_t^* \pi_0$. Moreover, by Kolmogorov's forward equation,

$$\partial_t \int f \, \mathrm{d}P_t^* \pi_0 = \partial_t \int P_t f \, \mathrm{d}\pi_0 = \int P_t \mathscr{L} f \, \mathrm{d}\pi_0 = \int \mathscr{L} f \, \mathrm{d}P_t^* \pi_0 = \int f \, \mathrm{d}\mathscr{L}^* P_t^* \pi_0 \,.$$

Since this has to hold for all functions f, we conclude the following.

Proposition 1.2.6 (Kolmogorov's forward equation)**.** *For all $t \geq 0$,*

$$\partial_t P_t^* \pi_0 = \mathscr{L}^* P_t^* \pi_0 = P_t^* \mathscr{L}^* \pi_0 \,.$$

Here is another illuminating way to express these equations. Let $u_t := P_t f$, and let $\pi_t = P_t^* \pi_0$. Then:

$$\partial_t u_t = \mathscr{L} u_t \,, \qquad \text{(Kolmogorov's backward equation)}$$
$$\partial_t \pi_t = \mathscr{L}^* \pi_t \,. \qquad \text{(Kolmogorov's forward equation)}$$

The terms "backward" and "forward" are rather confusing, so we will not use them. Instead, we will refer to the evolution equation for the density (Kolmogorov's forward equation) as the **Fokker–Planck equation**.

Consequently, we obtain characterizations of stationarity. Recall that π is stationary for the Markov process if, when $X_0 \sim \pi$, then $X_t \sim \pi$ for all $t \geq 0$.

Proposition 1.2.7 (stationarity)**.** *The following are equivalent.*

1. *π is a stationary distribution for the Markov process.*
2. *$\mathscr{L}^* \pi = 0$.*

3. $\mathbb{E}_\pi \mathscr{L} f = 0$ *for all functions* f.

Proof. The equivalence between the first two statements is the Fokker–Planck equation. The third statement is the dual of the second statement. □

Example 1.2.8. Consider again the Langevin diffusion. For functions $f, g : \mathbb{R}^d \to \mathbb{R}$,

$$\int \mathscr{L} f \, g = \int \{\Delta f - \langle \nabla V, \nabla f \rangle\} \, g = \int f \, \{\Delta g + \operatorname{div}(g \, \nabla V)\}$$

where the second equality is integration by parts. This shows that

$$\mathscr{L}^* g = \Delta g + \operatorname{div}(g \, \nabla V) \, .$$

From here, we can solve for the stationary distribution. Write

$$0 = \mathscr{L}^* \pi = \Delta \pi + \operatorname{div}(\pi \, \nabla V) = \operatorname{div}\big(\pi \, (\nabla \ln \pi + \nabla V)\big) \, .$$

This can be solved by setting $\ln \pi = -V + \text{constant}$, i.e. $\pi \propto \exp(-V)$.

Corollary 1.2.9. *The stationary distribution of the Langevin diffusion* (1.E.1) *with potential* V *is* $\pi \propto \exp(-V)$.

1.2.2 Reversibility and the Spectrum

Consider a Markov semigroup $(P_t)_{t \geq 0}$ with generator $\mathscr{L}$ and stationary distribution π. Then, the natural space of functions to study is the Hilbert space $L^2(\pi)$. The analysis of the Markov process is particularly simple if the following condition holds.

Definition 1.2.10. The Markov semigroup $(P_t)_{t \geq 0}$ is **reversible** w.r.t. π if for all $f, g \in L^2(\pi)$ and all $t \geq 0$,

$$\int P_t f \, g \, \mathrm{d}\pi = \int f \, P_t g \, \mathrm{d}\pi \, .$$

Equivalently, for all f and g for which $\mathscr{L}f$ and $\mathscr{L}g$ are defined,

$$\int \mathscr{L}f\, g\, \mathrm{d}\pi = \int f\, \mathscr{L}g\, \mathrm{d}\pi\,.$$

If $X_0 \sim \pi$ and we take $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$ for events A and B, then it implies

$$\mathbb{P}\{X_t \in A,\ X_0 \in B\} = \mathbb{P}\{X_0 \in A,\ X_t \in B\}\,,$$

i.e., (X_0, X_t) has the same distribution as (X_t, X_0). This is the sense in which the associated Markov process is time-reversible.

The definition says that P_t and $\mathscr{L}$ are symmetric operators on $L^2(\pi)$, and thus we expect that P_t and $\mathscr{L}$ have real spectra. Also, since $\partial_t P_t = \mathscr{L}P_t$, we can formally write $P_t = \exp(t\mathscr{L})$, and so we expect P_t to be a positive operator, meaning that $\int f\, P_t f\, \mathrm{d}\pi \geq 0$ for all $f \in L^2(\pi)$, which can be checked from reversibility (write $\int f\, P_t f\, \mathrm{d}\pi = \int (P_{t/2}f)^2\, \mathrm{d}\pi$). Moreover, from the definition of P_t and Jensen's inequality,

$$\{P_t f(x)\}^2 = \mathbb{E}[f(X_t) \mid X_0 = x]^2 \leq \mathbb{E}[f(X_t)^2 \mid X_0 = x] = P_t(f^2)(x) \tag{1.2.11}$$

and integrating this yields $\int (P_t f)^2\, \mathrm{d}\pi \leq \int P_t(f^2)\, \mathrm{d}\pi = \int f^2\, \mathrm{d}\pi$, where the equality follows from stationarity of π. This shows that P_t is a contraction on $L^2(\pi)$ (in fact, on any $L^p(\pi)$, $p \in [1, \infty]$). Combining this with $P_t = \exp(t\mathscr{L})$ leads us to predict that $\mathscr{L}$ is a *negative* operator. Below, we will give a direct proof of this fact; unsurprisingly, the proof of negativity of $\mathscr{L}$ still relies on the crucial fact (1.2.11).

Definition 1.2.12. The **carré du champ** is the bilinear operator Γ defined via

$$\Gamma(f, g) := \frac{1}{2}\{\mathscr{L}(fg) - f\,\mathscr{L}g - g\,\mathscr{L}f\}\,.$$

The **Dirichlet energy** is the functional $\mathscr{E}(f, g) := \int \Gamma(f, g)\, \mathrm{d}\pi$.

Lemma 1.2.13. *For any function f, $\Gamma(f, f) \geq 0$.*

Proof. Recall from (1.2.11) that $P_t(f^2) \geq (P_t f)^2$ for all $t > 0$. In terms of $\mathscr{L}$,

$$f^2 + t\,\mathscr{L}(f^2) + o(t) \geq [f + t\,\mathscr{L}f + o(t)]^2 = f^2 + 2t\, f\,\mathscr{L}f + o(t)$$

and sending $t \searrow 0$ yields $\mathscr{L}(f^2) \geq 2f\,\mathscr{L}f$. (This proof does not require reversibility.) □

Theorem 1.2.14 (fundamental integration by parts identity)**.** *Suppose that the generator $\mathscr{L}$ and carré du champ Γ are associated with a Markov semigroup which is reversible w.r.t. π. Then, for any functions f and g,*

$$\int f\,(-\mathscr{L})g\,\mathrm{d}\pi = \int (-\mathscr{L})f\,g\,\mathrm{d}\pi = \int \Gamma(f,g)\,\mathrm{d}\pi = \mathscr{E}(f,g)\,.$$

Since the identity implies that $\mathscr{L}$ is symmetric, the integration by parts identity is in fact *equivalent* to reversibility. We can reformulate this identity in an illuminating way as

$$-\mathscr{L} = \nabla^{*,\pi}\,\nabla \tag{1.2.15}$$

where $(\cdot)^{*,\pi}$ denotes the adjoint in $L^2(\pi)$. This expression brings out the symmetry of $\mathscr{L}$ and also makes the following corollary clear.

Corollary 1.2.16. *For a reversible Markov semigroup, $-\mathscr{L} \geq 0$.*

Proof of Theorem 1.2.14. Since $\int \mathscr{L}h\,\mathrm{d}\pi = 0$ for all functions h (due to stationarity of π), the definition of Γ yields

$$\int \Gamma(f,g)\,\mathrm{d}\pi = \frac{1}{2}\int f\,(-\mathscr{L})g\,\mathrm{d}\pi + \frac{1}{2}\int g\,(-\mathscr{L})f\,\mathrm{d}\pi\,.$$

The two terms are equal due to reversibility. □

It is usually convenient for our operators to be positive, so from now on we will instead refer to the negative generator $-\mathscr{L}$.

When we introduced Kolmogorov's equations, we ended up with two PDEs, one involving the generator $\mathscr{L}$ and one involving its $L^2(\mathfrak{m})$ adjoint $\mathscr{L}^*$, where $\mathfrak{m}$ is the Lebesgue measure on $\mathbb{R}^d$. The issue is that we used the "wrong" inner product; $\mathscr{L}$ is not symmetric in $L^2(\mathfrak{m})$. If we now switch to $L^2(\pi)$, then instead of considering the density π_t with respect to $\mathfrak{m}$ we should consider the density $\rho_t := \pi_t/\pi$ with respect to π. Then, the Fokker–Planck equation becomes

$$\partial_t \rho_t = \mathscr{L}\rho_t\,. \tag{1.2.17}$$

For the rest of the section, the Markov semigroup is assumed reversible.

Example 1.2.18. Returning to the fundamental example of the Langevin diffusion, a computation shows that

$$\Gamma(f,f) = \frac{1}{2}\left\{\Delta(f^2) - \langle \nabla V, \nabla(f^2)\rangle - 2f\left(\Delta f - \langle \nabla V, \nabla f\rangle\right)\right\} = \|\nabla f\|^2 .$$

Incidentally, *carré du champ* means "square of the field" in French, and it is this expression which gives it its name. More generally, $\Gamma(f,g) = \langle \nabla f, \nabla g\rangle$.

The identity in Theorem 1.2.14 reads

$$\int f\left(-\Delta g + \langle \nabla V, \nabla g\rangle\right) \mathrm{d}\pi = \int g\left(-\Delta f + \langle \nabla V, \nabla f\rangle\right) \mathrm{d}\pi = \int \langle \nabla f, \nabla g\rangle \,\mathrm{d}\pi$$

which can be checked (using $\pi \propto \exp(-V)$) via integration by parts (naturally!), showing that the Langevin diffusion is indeed reversible w.r.t. π.

Gradient flow of the Dirichlet energy. It turns out that a reversible Markov process follows the *steepest descent* of the Dirichlet energy with respect to $L^2(\pi)$. To justify this, for a curve $t \mapsto u_t$ in $L^2(\pi)$, write $\dot{u}_t := \partial_t u_t$ for the time derivative. The $L^2(\pi)$ gradient of the functional $f \mapsto \mathscr{E}(f) := \mathscr{E}(f,f)$ at f is defined to be the element $\nabla_{L^2(\pi)}\mathscr{E}(f) \in L^2(\pi)$ such that for all curves $t \mapsto u_t$ with $u_0 = f$, it holds that

$$\partial_t\big|_{t=0}\, \mathscr{E}(u_t, u_t) = \int \dot{u}_0\, \nabla_{L^2(\pi)}\mathscr{E}(f)\, \mathrm{d}\pi .$$

From the integration by parts identity,

$$\partial_t\big|_{t=0}\, \mathscr{E}(u_t, u_t) = \partial_t\big|_{t=0} \int u_t\, (-\mathscr{L})u_t \,\mathrm{d}\pi = 2\int \dot{u}_0\, (-\mathscr{L})f \,\mathrm{d}\pi .$$

Therefore, $\nabla_{L^2(\pi)}\mathscr{E}(f) = -2\mathscr{L}f$.

The steepest descent of $\mathscr{E}$ is the curve $t \mapsto u_t$ such that $\dot{u}_t = -\nabla_{L^2(\pi)}\mathscr{E}(u_t) = 2\mathscr{L}u_t$. This is, up to a rescaling of time, precisely the equation satisfied by $t \mapsto P_t f$.

Spectral gap and convergence. Consider a reversible Markov semigroup $(P_t)_{t\geq 0}$ and recall that $P_t f(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. We are interested in the long-term behavior of $P_t f$. If the process mixes, then by definition it forgets its initial condition, so that $P_t f$ converges to a constant; moreover, this constant should be the average value $\int f \,\mathrm{d}\pi$ at stationarity. How do we establish a rate of convergence for $P_t f \to \int f \,\mathrm{d}\pi$?

We may assume that $\int f \, \mathrm{d}\pi = 0$, so we wish to prove $P_t f \to 0$. Also recall that, formally, $P_t = \exp(t\mathscr{L})$ with $\mathscr{L} \leq 0$. If we have a *spectral gap*

$$-\mathscr{L} \geq \lambda_{\min} > 0\,,$$

then we would expect that

$$\|P_t f\|^2_{L^2(\pi)} \leq \exp(-2\lambda_{\min} t) \, \|f\|^2_{L^2(\pi)}\,.$$

This is indeed the case. However, observe that since $P_t 1 = 1$ for all $t \geq 0$ and hence $\mathscr{L}1 = 0$, the spectral gap condition is only supposed to hold on the subspace of $L^2(\pi)$ which is orthogonal to constants.

Definition 1.2.19. The Markov process is said to satisfy a **Poincaré inequality** (PI) with constant C_{PI} if for all functions $f \in L^2(\pi)$,

$$\int f\,(-\mathscr{L})f \, \mathrm{d}\pi = \mathscr{E}(f,f) \geq \frac{1}{C_{\mathsf{PI}}} \, \|f - \mathrm{proj}_{1^\perp} f\|^2_{L^2(\pi)} = \frac{1}{C_{\mathsf{PI}}} \operatorname{var}_\pi f\,.$$

The Poincaré constant C_{PI} corresponds to the inverse of the spectral gap. Based on the calculus we have developed so far, it is not too difficult to prove the following result (differentiate $t \mapsto \|P_t f\|^2_{L^2(\pi)}$), so we leave it as Exercise 1.7.

Theorem 1.2.20. *The following are equivalent.*

1. *The Markov process satisfies a Poincaré inequality with constant C_{PI}.*
2. *For all $f \in L^2(\pi)$ with $\int f \, \mathrm{d}\pi = 0$ and all $t \geq 0$,*

$$\|P_t f\|^2_{L^2(\pi)} \leq \exp\Big(-\frac{2t}{C_{\mathsf{PI}}}\Big) \, \|f\|^2_{L^2(\pi)}\,.$$

In particular, we can apply this result to the semigroup corresponding to the Langevin diffusion (1.E.1) to obtain a spectral gap criterion for quantitative convergence. However, this result is mainly of use when we are interested in a specific test function f. More generally, it is useful to obtain bounds on the rate of convergence of the law π_t of X_t to the stationary distribution π. Recall (from (1.2.17)) that the relative density $\rho_t := \pi_t/\pi$ solves the equation $\partial_t \rho_t = \mathscr{L}\rho_t$, i.e., ρ_t is given by $\rho_t = P_t \rho_0$. We can therefore apply the preceding result to $f := \rho_0 - 1$.

For a probability measure μ, we define the **chi-squared divergence**

$$\chi^2(\mu \,\|\, \pi) := \Big\| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \Big\|^2_{L^2(\pi)} = \mathrm{var}_\pi \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \qquad \text{if } \mu \ll \pi\,,$$

with $\chi^2(\mu \,\|\, \pi) := \infty$ otherwise. The result can be formulated as follows.

Theorem 1.2.21. *The following are equivalent.*

1. *The Markov process satisfies a Poincaré inequality with constant C_{PI}.*
2. *For any initial distribution π_0 and all $t \geq 0$,*

$$\chi^2(\pi_t \,\|\, \pi) \leq \exp\Big(-\frac{2t}{C_{\mathsf{PI}}}\Big)\, \chi^2(\pi_0 \,\|\, \pi)\,.$$

Example 1.2.22. For the Langevin diffusion, the Poincaré inequality reads

$$\mathrm{var}_\pi f \leq C_{\mathsf{PI}}\, \mathbb{E}_\pi[\|\nabla f\|^2]$$

for all functions $f : \mathbb{R}^d \to \mathbb{R}$, where $\pi \propto \exp(-V)$.

1.2.3 The Log-Sobolev Inequality and Bakry–Èmery Theory

For sampling applications, the convergence result under a Poincaré inequality is not fully satisfactory because the chi-squared divergence at initialization is typically large, scaling exponentially in the dimension. The approach we explore next is to use the **Kullback–Leibler (KL) divergence** $\mathsf{KL}(\cdot \,\|\, \pi)$ as our objective functional, defined via

$$\mathsf{KL}(\mu \,\|\, \pi) := \int \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\, \mathrm{d}\pi = \int \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\, \mathrm{d}\mu \qquad \text{if } \mu \ll \pi\,,$$

and $\mathsf{KL}(\mu \,\|\, \pi) := \infty$ otherwise.

Recall the notation $\rho_t := \pi_t/\pi$ for the relative density of the Markov process w.r.t. π. Since $\partial_t \rho_t = \mathscr{L}\rho_t$, we can calculate via the integration by parts identity that

$$\partial_t\, \mathsf{KL}(\pi_t \,\|\, \pi) = \partial_t \int \rho_t \ln \rho_t\, \mathrm{d}\pi = \int (\ln \rho_t + 1)\, \mathscr{L}\rho_t\, \mathrm{d}\pi = -\mathscr{E}(\rho_t, \ln \rho_t)\,. \tag{1.2.23}$$

Hence, if $\mathscr{E}(\rho_t, \ln \rho_t) \gtrsim \mathsf{KL}(\pi_t \,\|\, \pi)$, then we obtain convergence to equilibrium for the diffusion in KL divergence, at an exponential rate.

Definition 1.2.24. The Markov process is said to satisfy a **log-Sobolev inequality** (LSI) with constant C_{LSI} if for all densities ρ w.r.t. π,

$$\mathsf{KL}(\rho\pi \parallel \pi) \le \frac{C_{\mathsf{LSI}}}{2} \mathscr{E}(\rho, \ln \rho) \,.$$

Theorem 1.2.25. *The following are equivalent.*

1. *The Markov process satisfies a log-Sobolev inequality with constant C_{LSI}.*
2. *For any initial distribution π_0 and all $t \ge 0$,*

$$\mathsf{KL}(\pi_t \parallel \pi) \le \exp\Big(-\frac{2t}{C_{\mathsf{LSI}}}\Big) \mathsf{KL}(\pi_0 \parallel \pi) \,.$$

By *linearizing* the LSI, i.e., by taking $\rho = 1 + \varepsilon f$ for small $\varepsilon > 0$ and expanding both sides of the LSI in powers of ε, one can prove that the LSI implies a Poincaré inequality with constant $C_{\mathsf{PI}} \le C_{\mathsf{LSI}}$ (Exercise 1.8).

Example 1.2.26. For the Langevin diffusion, the LSI reads

$$\frac{2}{C_{\mathsf{LSI}}} \mathsf{KL}(\mu \parallel \pi) \le \mathbb{E}_\pi \Big\langle \nabla \frac{\mu}{\pi}, \nabla \ln \frac{\mu}{\pi} \Big\rangle = \mathbb{E}_\mu \Big[\Big\| \nabla \ln \frac{\mu}{\pi} \Big\|^2 \Big] = 4\, \mathbb{E}_\pi \Big[\Big\| \nabla \sqrt{\frac{\mu}{\pi}} \Big\|^2 \Big] \,.$$

The right-hand side of the above expression is important; it is known as the (relative) **Fisher information** $\mathsf{FI}(\mu \parallel \pi) := \mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2]$. In particular, the Fisher information plays a central role in the study of non-log-concave sampling in Chapter 11.

The LSI often appears in many equivalent forms. For example, another formulation is that for all functions $f : \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\mathrm{ent}_\pi(f^2) \le 2 C_{\mathsf{LSI}}\, \mathbb{E}_\pi[\|\nabla f\|^2] \,,$$

where for a function $g : \mathbb{R}^d \to \mathbb{R}_+$ we define $\mathrm{ent}_\pi(g) := \mathbb{E}_\pi(g \ln g) - \mathbb{E}_\pi g \ln \mathbb{E}_\pi g$. To verify the equivalence, consider $f = \sqrt{\mu/\pi}$.

Bakry–Émery condition. Although we have derived two criteria for convergence of the Markov process, namely, the Poincaré inequality and the log-Sobolev inequality, we have not yet addressed when these criteria hold. Introduce the following definition.

Definition 1.2.27. The **iterated carré du champ** is the operator Γ_2 defined via

$$\Gamma_2(f,g) := \frac{1}{2}\left\{\mathscr{L}\Gamma(f,g) - \Gamma(f,\mathscr{L}g) - \Gamma(g,\mathscr{L}f)\right\}.$$

Recalling that $\Gamma(f,g) = \frac{1}{2}\left\{\mathscr{L}(fg) - f\,\mathscr{L}g - g\,\mathscr{L}f\right\}$, we see that Γ_2 is defined analogously to Γ, except we replace the bilinear operation of multiplication, $(f,g) \mapsto fg$, by the carré du champ $(f,g) \mapsto \Gamma(f,g)$. Also, similarly to how the carré du champ appears when computing the time derivative of functionals such as the chi-squared divergence and the KL divergence, the iterated carré du champ appears when computing the *second* time derivative. After some calculations, one arrives at the following criterion.

Definition 1.2.28. The Markov semigroup is said to satisfy the **Bakry–Émery criterion** with constant $\alpha > 0$ if for all functions f,

$$\Gamma_2(f,f) \geq \alpha\,\Gamma(f,f).$$

This condition is also known as the **curvature-dimension condition** $\mathrm{CD}(\alpha,\infty)$.

We will prove the following theorem in Chapter 2.

Theorem 1.2.29 (Bakry–Émery)**.** *Consider a diffusion Markov semigroup. Assume that the curvature-dimension condition* $\mathrm{CD}(\alpha,\infty)$ *holds. Then, a log-Sobolev inequality holds with constant* $C_{\mathrm{LSI}} \leq 1/\alpha$.

We have not explained yet what a *diffusion* Markov semigroup is, but for now we can think of the Langevin diffusion as a fundamental example. The key point is that once the (iterated) carré du champ operators are known, the curvature-dimension condition amounts to an algebraic condition which can be easily checked, which in turn implies the log-Sobolev inequality (and hence the Poincaré inequality by Exercise 1.8). For the Langevin diffusion, this condition amounts to the following theorem.

Theorem 1.2.30. *For the Langevin diffusion* (1.E.1)*, the curvature-dimension condition* $\mathrm{CD}(\alpha,\infty)$ *holds if and only if the potential* V *is* α*-strongly convex.*

Although we have deferred the Markov semigroup proofs of these results to Chapter 2, we will shortly prove these results using the calculus of optimal transport.

Another point to address is the origin of the name "curvature-dimension condition". In fact this is part of a rich story in which Markov diffusions on Riemannian manifolds capture the geometric features of the ambient space, such as its curvature. A picture emerges in which curvature, concentration, and mixing of the diffusion all intertwine, and only in this context is it appreciated that the curvature-dimension condition is appropriately named. This is discussed more fully in Chapter 2.

1.3 The Geometry of Optimal Transport

In this section, we explain how the space of probability measures equipped with the 2-Wasserstein distance from optimal transport can be formally viewed as a Riemannian manifold. The textbook [Vil03] is a standard reference for optimal transport; see also [AGS08; Vil09; San15] for more detailed treatments of Wasserstein calculus. We remind readers that the "proofs" in this section are only sketched for intuition.

1.3.1 Introduction and Duality Theory

The optimal transport problem can be defined in great generality. Throughout this section, $\mathcal{P}(\mathcal{X})$ denotes the space of probability measures on a space $\mathcal{X}$.

Definition 1.3.1. Let $\mathcal{X}$ and $\mathcal{Y}$ be complete separable metric spaces, and consider a cost functional $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$. The **optimal transport cost** from $\mu \in \mathcal{P}(\mathcal{X})$ to $\nu \in \mathcal{P}(\mathcal{Y})$ with cost c is

$$\mathcal{T}_c(\mu, \nu) := \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int c(x, y) \, \mathrm{d}\gamma(x, y) \,, \tag{1.3.2}$$

where $\mathcal{C}(\mu, \nu)$ is the space of *couplings* of (μ, ν), i.e. the space of probability measures $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ whose marginals are μ and ν respectively.

A minimizer in this problem is known as an **optimal transport plan**.

An equivalent probabilistic formulation is that $\mathcal{T}_c(\mu, \nu)$ is the infimum of $\mathbb{E}\, c(X, Y)$ over all pairs of jointly defined random variables (X, Y) such that $X \sim \mu$ and $Y \sim \nu$.

Theorem 1.3.3. *If the cost c is lower semicontinuous, then an optimal transport plan always exists.*

Proof. One can show that the functional $\gamma \mapsto \int c \, \mathrm{d}\gamma$ is lower semicontinuous and that $\mathcal{C}(\mu, \nu)$ is compact, where we use the weak topology on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. It is a general fact that lower semicontinuous functions attain their minima on compact sets. □

Historically, optimal transport began with Monge who considered the Euclidean cost $c(x, y) := \|x - y\|$ on $\mathbb{R}^d \times \mathbb{R}^d$. Moreover, he considered a slightly different problem in which, rather than searching over all couplings in $\mathcal{C}(\mu, \nu)$, he restricted attention to couplings which are induced by a *mapping* $T : \mathbb{R}^d \to \mathbb{R}^d$ satisfying $T_\# \mu = \nu$; this is known as the **Monge problem**. In the probabilistic interpretation, this corresponds to a pair of random variables $(X, T(X))$ with $X \sim \mu$ and $T(X) \sim \nu$. The physical interpretation of this additional constraint is that no mass from μ be split up before it is transported, which may be reasonable from a modelling perspective but leads to an ill-posed mathematical problem. Indeed, there may not even exist any such mappings T, as is the case when $\mu = \delta_x$ places all of its mass on a single point and ν does not. Consequently, the solution to the Monge problem remained unknown for centuries.

The breakthrough arrived when Kantorovich formulated the relaxation of the Monge problem introduced in Definition 1.3.1, which is therefore known as the **Kantorovich problem**. As the product measure $\mu \otimes \nu$ always belongs to $\mathcal{C}(\mu, \nu)$, we at least know that the constraint set is non-empty, and Theorem 1.3.3 shows that the Kantorovich problem is well-behaved. Moreover, the Kantorovich problem is actually a *convex* problem on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$; indeed, the objective is linear and the constraint set $\mathcal{C}(\mu, \nu)$ is convex. Hence, one can bring to bear the power of convex duality to study the Kantorovich problem (historically, this study was actually the origin of linear programming).

Although a large part of optimal transport theory can be developed in a general framework as above, for the rest of the section we will focus on the case $c(x, y) := \|x - y\|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$ for the sake of simplicity.

Definition 1.3.4. The 2-**Wasserstein distance** between μ and ν, denoted $W_2(\mu, \nu)$, is defined via

$$W_2^2(\mu, \nu) := \inf_{\gamma \in \mathcal{C}(\mu, \nu)} \int \|x - y\|^2 \, \mathrm{d}\gamma(x, y) \, . \tag{1.3.5}$$

Write $\mathcal{P}_2(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \int \|\cdot\|^2 \, \mathrm{d}\mu < \infty\}$ for the space of probability measures on $\mathbb{R}^d$ with finite second moment.

Duality and optimality. For this section, it will actually be convenient to consider the cost $c(x, y) := \frac{1}{2} \|x - y\|^2$ instead, i.e. we consider $\frac{1}{2} W_2^2(\mu, \nu)$ instead of $W_2^2(\mu, \nu)$.

The key to solving the Kantorovich problem is duality. First observe that the constraint that the first marginal of γ is μ can be written as follows: for every function $f \in L^1(\mu)$, it holds that $\int f(x)\,\mathrm{d}\gamma(x,y) = \int f(x)\,\mathrm{d}\mu(x)$. Doing the same for the constraint on the second marginal of γ, we can write the Kantorovich problem as an *unconstrained* min-max problem

$$\frac{1}{2}W_2^2(\mu,\nu) = \inf_{\gamma\in\mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d)} \sup_{\substack{f\in L^1(\mu)\\ g\in L^1(\nu)}} \left\{ \int \frac{\|x-y\|^2}{2}\,\mathrm{d}\gamma(x,y) + \int f\,\mathrm{d}\mu - \int f(x)\,\mathrm{d}\gamma(x,y) \right.$$
$$\left. + \int g\,\mathrm{d}\nu - \int g(y)\,\mathrm{d}\gamma(x,y) \right\}.$$

Here, $\mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d)$ denotes the space of non-negative finite measures on $\mathbb{R}^d\times\mathbb{R}^d$. Next, if we switch the order of the infimum and the supremum, we arrive at the *dual* optimal transport problem:

$$\sup_{\substack{f\in L^1(\mu)\\ g\in L^1(\nu)}} \inf_{\gamma\in\mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d)} \left\{ \int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu + \int \left[\frac{\|x-y\|^2}{2} - f(x) - g(y)\right]\mathrm{d}\gamma(x,y) \right\}$$
$$= \sup_{(f,g)\in\mathcal{D}(\mu,\nu)} \left\{ \int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu \right\}$$

where $\mathcal{D}(\mu,\nu)$ is the set of *dual feasible potentials*

$$\mathcal{D}(\mu,\nu) := \left\{ (f,g)\in L^1(\mu)\times L^1(\nu) \;\middle|\; f(x)+g(y) \le \frac{\|x-y\|^2}{2} \text{ for all } x,y\in\mathbb{R}^d \right\}.$$

Definition 1.3.6. Let $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$. The **dual optimal transport problem** from μ to ν is the optimization problem

$$\sup_{(f,g)\in\mathcal{D}(\mu,\nu)} \left\{ \int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu \right\}. \tag{1.3.7}$$

Since $\inf\sup \ge \sup\inf$, the value of the dual problem is always at *most* $\frac{1}{2}W_2^2(\mu,\nu)$. On the other hand, if we find a transport plan $\gamma^\star$ and feasible dual potentials $f^\star$, $g^\star$ such that $\int \|x-y\|^2\,\mathrm{d}\gamma^\star(x,y) = \int f^\star\,\mathrm{d}\mu + \int g^\star\,\mathrm{d}\nu$, it implies that the primal and dual values coincide and that $\gamma^\star$, $f^\star$, and $g^\star$ are all optimal.

By carefully studying the dual problem, we will obtain a wealth of information about the optimal transport problem. Our main goal now is to sketch the following theorem.

Theorem 1.3.8 (fundamental theorem of optimal transport)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then, the following assertions hold.*

1. *(strong duality) The value of the dual optimal transport problem from μ to ν equals $\frac{1}{2} W_2^2(\mu, \nu)$.*
2. *(existence of optimal dual potentials) There exists an optimal pair $(f^\star, g^\star)$ for the dual optimal transport problem.*
3. *(characterization of optimality) The optimal dual potentials are of the form*

$$f^\star = \frac{\|\cdot\|^2}{2} - \varphi\,, \qquad g^\star = \frac{\|\cdot\|^2}{2} - \varphi^*\,, \tag{1.3.9}$$

   where $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a proper, convex, lower semicontinuous function and φ^* is its convex conjugate. If $\gamma^\star$ denotes the optimal transport plan, then for $\gamma^\star$-a.e. $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, it holds that $\varphi(x) + \varphi^*(y) = \langle x, y\rangle$, i.e., $\gamma^\star$ is supported on the subdifferential of φ.
4. *(**Brenier's theorem**) Suppose in addition that μ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Then, the optimal transport plan is unique, and moreover it is induced by an optimal transport map T. The mapping T is characterized as the (μ-almost surely) unique gradient of a proper convex lower semicontinuous function φ which pushes forward μ to ν: $T = \nabla\varphi$ and $(\nabla\varphi)_{\#}\mu = \nu$.*

Various parts of this theorem can be proven separately; for example, strong duality can be established by rigorously justifying the interchange of infimum and supremum via a high-powered minimax theorem. Instead, we will outline a proof of the theorem which simultaneously establishes all of the above facts.

Outline. In the proof, we abbreviate "proper convex lower semicontinuous function" to simply "closed convex function".

1. Optimal transport plans are cyclically monotone. Let $\gamma^\star$ be an optimal transport plan, and suppose that the pairs $(x_1, y_1), \ldots, (x_n, y_n)$ lie in the support of $\gamma^\star$. Then, it should be the case that we cannot "rematch" these points to lower the optimal transport cost, i.e. for every permutation σ of $[n]$ we should have

$$\sum_{i=1}^n \|x_i - y_i\|^2 \le \sum_{i=1}^n \|x_i - y_{\sigma(i)}\|^2\,.$$

Equivalently,

$$\sum_{i=1}^{n} \langle x_i, y_i \rangle \geq \sum_{i=1}^{n} \langle x_i, y_{\sigma(i)} \rangle \,. \tag{1.3.10}$$

Indeed, if this condition fails, then it is possible to construct a new transport plan from $\gamma^\star$ by slightly rearranging the mass which has strictly smaller transport cost, which is a contradiction; see [GM96, Theorem 2.3].

A subset $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is said to be **cyclically monotone** if for all $n \in \mathbb{N}^+$, all pairs $(x_1, y_1), \ldots, (x_n, y_n)$, and all permutations σ of $[n]$, the condition (1.3.10) holds. Thus, optimal transport plans are supported on cyclically monotone sets.

2. Characterization of cyclically monotone sets. Remarkably, a complete characterization of cyclically monotone sets is known. Suppose φ is convex and differentiable, let $x_1, \ldots, x_n \in \mathbb{R}^d$, and let σ be a permutation of $[n]$. Then, from convexity,

$$\varphi(x_{\sigma^{-1}(i)}) - \varphi(x_i) \geq \langle \nabla\varphi(x_i), x_{\sigma^{-1}(i)} - x_i \rangle \,. \tag{1.3.11}$$

Summing this over $i \in [n]$, we obtain

$$\sum_{i=1}^{n} \langle \nabla\varphi(x_i), x_i \rangle \geq \sum_{i=1}^{n} \langle \nabla\varphi(x_i), x_{\sigma^{-1}(i)} \rangle = \sum_{i=1}^{n} \langle \nabla\varphi(x_{\sigma(i)}), x_i \rangle \,.$$

More generally, if φ is not differentiable, then the *subdifferential* of φ at x_i is defined to be the set of vectors $y_i \in \mathbb{R}^d$ such that (1.3.11) holds with y_i replacing $\nabla\varphi(x_i)$. This reasoning shows that the set $\partial\varphi := \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mid y \in \partial\varphi(x)\}$ is a cyclically monotone subset of $\mathbb{R}^d \times \mathbb{R}^d$.

The converse is also true: if $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is cyclically monotone, then it is contained in the subdifferential of a closed convex function φ. To prove this, one can pick any $(x_0, y_0) \in S$ and consider

$$\varphi(x) := \sup\left\{ \sum_{i=0}^{n} \langle y_i, x_{i+1} - x_i \rangle \;\middle|\; n \in \mathbb{N}^+,\ (x_1, y_1), \ldots, (x_n, y_n) \in S,\ x_{n+1} = x \right\}.$$

This characterization is due to Rockafellar.

3. Characterization of dual optimality. Now that we see the connection between convexity and the primal problem, it is time to do the same for the dual problem.

Suppose (f, g) is a feasible dual pair; if we hold f fixed, can we improve g? The constraint on g says that for all $x, y \in \mathbb{R}^d$,

$$g(y) \le \frac{\|x-y\|^2}{2} - f(x)\,.$$

Hence, writing $\varphi := \|\cdot\|^2/2 - f$, the optimal choice is

$$g(y) = \inf_{x\in\mathbb{R}^d}\Big\{\frac{\|x-y\|^2}{2} - f(x)\Big\} = \frac{\|y\|^2}{2} - \sup_{x\in\mathbb{R}^d}\{\langle x, y\rangle - \varphi(x)\}\,.$$

The function φ^* defined by $\varphi^*(y) := \sup_{x\in\mathbb{R}^d}\{\langle x, y\rangle - \varphi(x)\}$ is known as the *convex conjugate* of φ. To summarize, we have shown that for fixed $f = \|\cdot\|^2/2 - \varphi$, the optimal choice for g is $\|\cdot\|^2/2 - \varphi^*$. Similarly, if we fix $g = \|\cdot\|^2/2 - \varphi^*$, the optimal choice for f is $\|\cdot\|^2/2 - \varphi^{**}$.

We have not yet established existence, but suppose for the moment that an optimal dual pair $(f^\star, g^\star)$ exists. The preceding reasoning shows that $f^\star = \|\cdot\|^2/2 - \varphi$ and $g^\star = \|\cdot\|^2/2 - \varphi^*$, where $\varphi^{**} = \varphi$; otherwise the dual pair could be improved. Next, it is known from convex analysis that $\varphi = \varphi^{**}$ if and only if φ is a closed convex function. Thus, optimal dual potentials have the representation (1.3.9).

4. Proof of strong duality. Now consider the optimal transport plan $\gamma^\star$ (which exists; see Theorem 1.3.3). We know that $\gamma^\star$ is supported on a cyclically monotone set, which in turn is contained in the subdifferential of a closed convex function φ. Define the functions $f^\star := \|\cdot\|^2/2 - \varphi$ and $g^\star := \|\cdot\|^2/2 - \varphi^*$; these are dual feasible potentials. Also, it is a standard fact of convex analysis that $(x, y) \in \partial\varphi$ if and only if $\varphi(x) + \varphi^*(y) = \langle x, y\rangle$. Since the support of $\gamma^\star$ is contained in $\partial\varphi$,

$$\begin{aligned}
\frac{1}{2}\int \|x-y\|^2\,\mathrm{d}\gamma^\star(x,y) &= \int \Big(\frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} - \langle x, y\rangle\Big)\,\mathrm{d}\gamma^\star(x,y)\\
&= \int \Big(\frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} - \varphi(x) - \varphi^*(y)\Big)\,\mathrm{d}\gamma^\star(x,y)\\
&= \int \Big(\frac{\|\cdot\|^2}{2} - \varphi\Big)\,\mathrm{d}\mu + \int \Big(\frac{\|\cdot\|^2}{2} - \varphi^*\Big)\,\mathrm{d}\nu\\
&= \int f^\star\,\mathrm{d}\mu + \int g^\star\,\mathrm{d}\nu\,.
\end{aligned}$$

This simultaneously proves that strong duality holds and that $(f^\star, g^\star)$ is a pair of optimal dual potentials.

5. For regular measures, optimal transport plans are induced by transport maps. Another fact from convex analysis is that convex functions enjoy some regularity: *a closed convex function φ is differentiable at Lebesgue-a.e. points of the interior of its domain.* Consequently, if μ is absolutely continuous w.r.t. Lebesgue measure, then φ is differentiable μ-a.e. This says that for μ-a.e. $x \in \mathbb{R}^d$, the gradient $\nabla\varphi(x)$ exists and $\partial\varphi(x) = \{\nabla\varphi(x)\}$. Therefore, we can write $\gamma^\star = (\mathrm{id}, \nabla\varphi)_\# \mu$. In particular, $(\nabla\varphi)_\# \mu = \nu$, and $\nabla\varphi$ is the optimal transport map from μ to ν.

   In our entire discussion so far, we started off with an arbitrary optimal transport plan $\gamma^\star$. Hence, we have shown that *every optimal transport plan is of the form $(\mathrm{id}, \nabla\varphi)_\# \mu$ for some closed convex function φ.*

6. Uniqueness of the optimal transport map. So far, we have not discussed uniqueness of the solution to the Kantorovich problem, and in general uniqueness does not hold. However, in the setting we are currently dealing with (the cost is the squared Euclidean distance and μ is absolutely continuous), we can use additional arguments to establish uniqueness. We will show that if $\bar{\gamma}^\star = (\mathrm{id}, \nabla\bar{\varphi})_\# \mu$ is another optimal transport plan where $\bar{\varphi}$ is a closed convex function, then $\nabla\varphi = \nabla\bar{\varphi}$ (μ-a.e.). Note that in particular, it implies that there is only one gradient of a closed convex function which pushes forward μ to ν.

   From our above arguments, we see that $(\|\cdot\|^2/2 - \bar{\varphi}, \|\cdot\|^2/2 - \bar{\varphi}^*)$ is a dual optimal pair. Therefore,

$$\begin{aligned}\int \{\bar{\varphi}(x) + \bar{\varphi}^*(y)\} \, \mathrm{d}\gamma^\star(x,y) &= \int \bar{\varphi} \, \mathrm{d}\mu + \int \bar{\varphi}^* \, \mathrm{d}\nu = \int \varphi \, \mathrm{d}\mu + \int \varphi^* \, \mathrm{d}\nu \\ &= \int \{\varphi(x) + \varphi^*(y)\} \, \mathrm{d}\gamma^\star(x,y) = \int \langle x, y\rangle \, \mathrm{d}\gamma^\star(x,y) \, .\end{aligned}$$

   Using $\gamma^\star = (\mathrm{id}, \nabla\varphi)_\# \mu$, it yields

$$\int \{\bar{\varphi}(x) + \bar{\varphi}^*(\nabla\varphi(x)) - \langle x, \nabla\varphi(x)\rangle\} \, \mathrm{d}\mu(x) = 0 \, .$$

   On the other hand, by the definition of $\bar{\varphi}^*$, we have $\bar{\varphi}(x) + \bar{\varphi}^*(y) \geq \langle x, y\rangle$ for all $x, y \in \mathbb{R}^d$, with equality if and only if $y \in \partial\bar{\varphi}(x)$. So, the integrand of the above expression is always non-negative but the integral is zero, which combined with the previous fact shows that $\nabla\varphi(x) \in \partial\bar{\varphi}(x)$ for μ-a.e. x. But for μ-a.e. x, we also know that $\partial\bar{\varphi}(x) = \{\nabla\bar{\varphi}(x)\}$, and we conclude that $\nabla\varphi = \nabla\bar{\varphi}$ (μ-a.e.). □

We refer to φ as a **Brenier potential**. From convex duality, $\nabla\varphi^* = (\nabla\varphi)^{-1}$. So, if ν is also absolutely continuous, then the optimal transport map from ν to μ is $\nabla\varphi^*$. We often write $T_{\mu\to\nu} = \nabla\varphi$ for the optimal transport map from μ to ν.

In light of this discussion, it is natural to focus on the following class of measures.

Definition 1.3.12. The space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is the set of measures in $\mathcal{P}_2(\mathbb{R}^d)$ which are absolutely continuous w.r.t. Lebesgue measure.

Remarks on other costs. Many of the arguments can be generalized to other costs c. For example, the supports of optimal transport plans can be characterized via *c-cyclical monotonicity* (generalizing cyclical monotonicity) and optimal dual potentials can be characterized via *c-concavity* (generalizing convexity). Arguing that the optimal transport plan is induced by a transport map requires additional information about the differentiability of c.

1.3.2 Riemannian Structure of the Wasserstein Space

Wasserstein space as a metric space. The following lemma is used to establish that the triangle inequality holds for W_2.

Lemma 1.3.13 (gluing lemma)**.** *If $\gamma_{1,2} \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ and $\gamma_{2,3} \in \mathcal{P}(\mathcal{X}_2 \times \mathcal{X}_3)$ have the same marginal distribution on $\mathcal{X}_2$, then there exists $\gamma \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3)$ such that its first two marginals are $\gamma_{1,2}$ and its last two marginals are $\gamma_{2,3}$.*

Proof. Let μ denote the common $\mathcal{X}_2$-marginal of $\gamma_{1,2}$ and $\gamma_{2,3}$. The idea is to first draw $X_2 \sim \mu$. Then, draw X_1 from its conditional distribution given X_2 (according to $\gamma_{1,2}$), and similarly draw X_3 from its conditional distribution given X_2 (according to $\gamma_{2,3}$). Then, take γ to be the law of the triple (X_1, X_2, X_3).

The way to formalize this argument is via disintegration of measure. □

Proposition 1.3.14. *The space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space.*

Proof. Clearly, W_2 is symmetric in its two arguments. It is also clear that $\mu = \nu$ implies $W_2(\mu, \nu) = 0$. Conversely, if $W_2(\mu, \nu) = 0$, then there exists a coupling (X, Y) of (μ, ν) such that $\|X - Y\|^2 = 0$ a.s., or equivalently $X = Y$ a.s., which gives $\mu = \nu$.

To verify the triangle inequality, we use the gluing lemma. Let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_2(\mathbb{R}^d)$, let $\gamma^\star_{1,2}$ be optimal for (μ_1, μ_2), and let $\gamma^\star_{2,3}$ be optimal for (μ_2, μ_3). Let γ be obtained by gluing

$\gamma^\star_{1,2}$ and $\gamma^\star_{2,3}$, and let $\gamma_{1,3} \in \mathcal{C}(\mu_1, \mu_3)$ denote the $(1,3)$-marginal of $\gamma^\star$. Then,

$$W_2(\mu_1, \mu_3) \le \sqrt{\int \|x_1 - x_3\|^2 \, \mathrm{d}\gamma_{1,3}(x_1, x_3)} \le \sqrt{\int \{\|x_1 - x_2\| + \|x_2 - x_3\|\}^2 \, \mathrm{d}\gamma(x_1, x_2, x_3)} \, .$$

Let $f(x_1, x_2, x_3) := \|x_1 - x_2\|$ and $g(x_1, x_2, x_3) := \|x_2 - x_3\|$. The above expression can be written as $\|f + g\|_{L^2(\gamma)}$. By applying the triangle inequality in $L^2(\gamma)$,

$$\begin{aligned} W_2(\mu_1, \mu_3) &\le \|f\|_{L^2(\gamma)} + \|g\|_{L^2(\gamma)} \\ &= \sqrt{\int \|x_1 - x_2\|^2 \, \mathrm{d}\gamma(x_1, x_2, x_3)} + \sqrt{\int \|x_2 - x_3\|^2 \, \mathrm{d}\gamma(x_1, x_2, x_3)} \\ &= \sqrt{\int \|x_1 - x_2\|^2 \, \mathrm{d}\gamma^\star_{1,2}(x_1, x_2)} + \sqrt{\int \|x_2 - x_3\|^2 \, \mathrm{d}\gamma^\star_{2,3}(x_2, x_3)} \\ &= W_2(\mu_1, \mu_2) + W_2(\mu_2, \mu_3) \, . \end{aligned}$$

□

Since the next result is technical, we omit the proof.

Proposition 1.3.15. *The metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is complete and separable. Also, we have $W_2(\mu_n, \mu) \to 0$ if and only if $\mu_n \to \mu$ weakly and $\int \|\cdot\|^2 \, \mathrm{d}\mu_n \to \int \|\cdot\|^2 \, \mathrm{d}\mu$.*

The continuity equation. Next, we are going to consider dynamics in the space of measures, i.e., curves of measures $t \mapsto \mu_t$. Throughout, we assume these curves are sufficiently nice, in the following sense.

Definition 1.3.16 (informal)**.** We say that a curve $t \mapsto \mu_t \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is **absolutely continuous** if for all t,

$$|\dot{\mu}|(t) := \lim_{s \to t} \frac{W_2(\mu_s, \mu_t)}{|s - t|} < \infty \, .$$

The quantity $|\dot{\mu}|$ is called the **metric derivative** of the curve.

More generally, the metric derivative can be defined on any metric space and represents the magnitude of the velocity of the curve, see [AGS08].

It is helpful to adopt a fluid dynamics analogy in which we think of μ_t as the mass density of a fluid at time t. There are two complementary perspectives on fluid flows:

the *Lagrangian* perspective which emphasizes particle trajectories, and the *Eulerian* perspective which tracks the evolution of the fluid density.

Suppose that $X_0 \sim \mu_0$ and that $t \mapsto X_t$ evolves according to the ODE $\dot{X}_t = v_t(X_t)$. Here, $(v_t)_{t\geq 0}$ is a family of vector fields, i.e. mappings $\mathbb{R}^d \to \mathbb{R}^d$. Since the ODE describes the evolution of the particle trajectory, it is the Lagrangian description of the dynamics. The corresponding Eulerian description is the continuity equation.

Theorem 1.3.17. *Let $t \mapsto v_t$ be a family of vector fields and suppose that the random variables $t \mapsto X_t$ evolve according to $\dot{X}_t = v_t(X_t)$. Then, the law μ_t of X_t evolves according to the* ***continuity equation***

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0\,. \tag{1.3.18}$$

Proof. Given a test function $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned}\int \varphi\, \partial_t \mu_t = \partial_t \int \varphi \mu_t = \partial_t\, \mathbb{E}\, \varphi(X_t) = \mathbb{E}\langle \nabla\varphi(X_t), \dot{X}_t\rangle = \mathbb{E}\langle \nabla\varphi(X_t), v_t(X_t)\rangle \\ = \int \langle \nabla\varphi, v_t\rangle\, \mu_t = -\int \varphi \operatorname{div}(\mu_t v_t)\,.\end{aligned}$$

Since this holds for every φ, we obtain $\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0$. □

The punchline is that *every* nice curve of measures $t \mapsto \mu_t$ can be interpreted as the fluid flow along a family of vector fields, i.e., we can find vector fields $t \mapsto v_t$ such that the continuity equation (1.3.18) holds. First, however, note that there is no uniqueness: if $\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0$ and for each t, w_t is a vector field satisfying $\operatorname{div}(\mu_t w_t) = 0$, then the continuity equation also holds with the new vector fields $\widetilde{v}_t := v_t + w_t$. We will show how to pick a distinguished choice of vector fields $t \mapsto v_t$ which can be described in two equivalent ways. First, among all vector fields making the continuity equation hold true, we can choose v_t to minimize $\int \|v_t\|^2\, \mathrm{d}\mu_t$, which has the physical interpretation of *minimizing kinetic energy*. Second, we can choose v_t to be the gradient of a function; we will see that this is natural in light of the characterization of optimal transport maps.

Theorem 1.3.19 (curves of measures as fluid flows)**.** *Let $t \mapsto \mu_t$ be an absolutely continuous curve of measures.*

1. *For any family of vector fields $t \mapsto \tilde{v}_t$ such that the continuity equation (1.3.18) holds, we have $|\dot{\mu}|(t) \leq \|\tilde{v}_t\|_{L^2(\mu_t)}$ for all t.*

2. *Conversely, there exists a unique choice of vector fields $t \mapsto v_t$ such that the continuity equation (1.3.18) holds and $\|v_t\|_{L^2(\mu_t)} \le |\dot{\mu}|(t)$ for all t. The choice of vector fields is also characterized by the following property: the continuity equation (1.3.18) holds and for each t, $v_t = \nabla\psi_t$ for a function $\psi_t : \mathbb{R}^d \to \mathbb{R}$.*

Moreover, the distinguished vector field v_t satisfies

$$v_t = \lim_{\delta \searrow 0} \frac{T_{\mu_t \to \mu_{t+\delta}} - \mathrm{id}}{\delta} \tag{1.3.20}$$

where $T_{\mu_t \to \mu_{t+\delta}}$ is the optimal transport map from μ_t to $\mu_{t+\delta}$.

Proof. 1. Proof of the first statement. Let $\delta > 0$ and consider the *flow map* $F_{t,t+\delta}$ defined as follows. Given any initial point $x_t \in \mathbb{R}^d$, consider the ODE $\dot{x}_t = \tilde{v}_t(x_t)$ started at x_t. Then, $F_{t,t+\delta}$ maps x_t to the solution $x_{t+\delta}$ of the ODE at time $t+\delta$.

If $X_t \sim \mu_t$, then the continuity equation implies $F_{t,t+\delta}(X_t) \sim \mu_{t+\delta}$, i.e., $F_{t,t+\delta}$ is a valid transport map from μ_t to $\mu_{t+\delta}$. Hence, we can estimate

$$\frac{W_2(\mu_t, \mu_{t+\delta})}{\delta} \le \sqrt{\int \frac{\|F_{t,t+\delta} - \mathrm{id}\|^2}{\delta^2} \, \mathrm{d}\mu_t} \, .$$

However, $F_{t,t+\delta} - \mathrm{id} = \delta\tilde{v}_t + o(\delta)$, so letting $\delta \searrow 0$ we obtain $|\dot{\mu}|(t) \le \|\tilde{v}_t\|_{L^2(\mu_t)}$. (Actually, to prove this statement we should also consider the limit $\delta \nearrow 0$ for negative δ, but it is clear that the same argument works.)

2. Uniqueness of the optimal vector field. Suppose we find $t \mapsto v_t$ satisfying the continuity equation and such that $\|v_t\|_{L^2(\mu_t)} \le |\dot{\mu}|(t)$. In light of the first statement, it implies that the zero vector field is the minimizer of $\|v_t + w_t\|_{L^2(\mu_t)}$ among all vector fields w_t such that $\mathrm{div}(\mu_t w_t) = 0$. This is a strictly convex problem so the minimizer is unique, meaning that the family $t \mapsto v_t$ is uniquely determined.

3. Gradient vector fields are optimal. Here, we show that if the continuity equation holds for the family of vector fields $t \mapsto v_t$ and that $v_t = \nabla\psi_t$ for all t, then the vector fields are optimal.

There are at least two ways of seeing why gradient vector fields should be optimal. First, the continuity equation is equivalent to requiring that for all test functions $\varphi : \mathbb{R}^d \to \mathbb{R}$, it holds that $\partial_t \int \varphi \, \mathrm{d}\mu_t = \int \langle \nabla\varphi, v_t \rangle \, \mathrm{d}\mu_t$. In this expression, the vector field v_t only enters through inner products with gradients. To put it another way, if we consider the space $S := \{\nabla\psi \mid \psi : \mathbb{R}^d \to \mathbb{R}\}$ of gradients, viewed as a subspace

of $L^2(\mu_t)$, then we can write $L^2(\mu_t) = S \oplus S^\perp$ (actually, to make this valid we should take the closure of S, but we will ignore this detail). If we decompose v_t according to this direct sum, then $v_t = \nabla\psi_t + w_t$ for some function ψ_t and some w_t which is orthogonal (in $L^2(\mu_t)$) to S. If we replace v_t by $\nabla\psi_t$, then the continuity equation continues to hold, but we have only made the norm $\|v_t\|_{L^2(\mu_t)}$ smaller, hence the optimal choice of v_t should be a gradient.

The second line of reasoning comes from the proof of the first statement: the reason why the metric derivative $|\dot\mu|(t)$ was upper bounded by $\|\tilde v_t\|_{L^2(\mu_t)}$ is because the flow map corresponding to $t \mapsto \tilde v_t$ furnishes a possibly suboptimal transport map. To fix this, the flow map for the optimal $t \mapsto v_t$ should be approximately equal to the *optimal* transport map, i.e. $v_t \approx (T_{\mu_t \to \mu_{t+\delta}} - \mathrm{id})/\delta$. From the fundamental theorem of optimal transport, however, $T_{\mu_t \to \mu_{t+\delta}}$ is the gradient of a convex function, so in the limit v_t should be as well.

Instead of using these arguments, we will instead provide a proof based on direct computation. If $v_t = \nabla\psi_t$, the continuity equation shows that

$$\begin{aligned}\frac{\int \psi_t \,\mathrm{d}\mu_{t+\delta} - \int \psi_t \,\mathrm{d}\mu_t}{\delta} &= \int \psi_t \,\partial_t \mu_t + o(1) = -\int \psi_t \,\mathrm{div}(\mu_t \nabla\psi_t) + o(1) \\ &= \int \|\nabla\psi_t\|^2 \,\mathrm{d}\mu_t + o(1)\,.\end{aligned}$$

On the other hand,

$$\begin{aligned}\frac{\int \psi_t \,\mathrm{d}\mu_{t+\delta} - \int \psi_t \,\mathrm{d}\mu_t}{\delta} &= \int \frac{\psi_t \circ T_{\mu_t \to \mu_{t+\delta}} - \psi_t}{\delta} \,\mathrm{d}\mu_t \\ &= \int \Big\langle \nabla\psi_t, \frac{T_{\mu_t \to \mu_{t+\delta}} - \mathrm{id}}{\delta} \Big\rangle \,\mathrm{d}\mu_t + o(1) \\ &\le \sqrt{\int \|\nabla\psi_t\|^2 \,\mathrm{d}\mu_t}\, \frac{W_2(\mu_t, \mu_{t+\delta})}{\delta} + o(1)\,.\end{aligned}$$

Taking $\delta \searrow 0$ yields $\|v_t\|_{L^2(\mu_t)} = \|\nabla\psi_t\|_{L^2(\mu_t)} \le |\dot\mu|(t)$.

4. Existence of optimal vector fields. Finally, one can show for instance that vector fields defined via limits of transport maps as in (1.3.20) indeed satisfy the continuity equation and are gradient vector fields, and are therefore optimal. However, the details are omitted. □

From the theorem, we learn that the optimal vector field v_t satisfies $\|v_t\|_{L^2(\mu_t)} = |\dot\mu|(t)$. On the other hand, the metric derivative is supposed to be the "magnitude of the velocity".

Our next goal is to interpret v_t as the velocity vector to the curve, and $\|v_t\|_{L^2(\mu_t)}$ as its norm, all through the lens of Riemannian geometry.

Background on Riemannian geometry. In the spirit of informality, we give a description of what a Riemannian manifold entails, rather than a precise definition. A **manifold** $\mathcal{M}$ is a space which is locally homeomorphic to a Euclidean space. At each point $p \in \mathcal{M}$, there is an associated vector space $T_p\mathcal{M}$, called the **tangent space** at p, which is the space of all possible velocities of curves passing through p. The whole structure should be smooth: the tangent spaces should vary smoothly in a suitable sense.

A **Riemannian metric** is a smoothly varying choice of inner products $p \mapsto \langle \cdot, \cdot \rangle_p$ on the tangent spaces. The metric allows us to, e.g., locally measure the angles between two intersecting curves. For our purposes, it is important to note that the metric allows us to define the steepest descent direction for an objective function, which in turn allows us to consider gradient flows.

The Riemannian metric induces a distance function (in the sense of metric spaces) via

$$\mathrm{d}(p, q) \coloneqq \inf\left\{ \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)} \,\mathrm{d}t \;\middle|\; \gamma : [0,1] \to \mathcal{M},\ \gamma(0) = p,\ \gamma(1) = q \right\}. \tag{1.3.21}$$

Here, $\dot\gamma(t)$ denotes the tangent vector to the curve at time t. Note that the norm of the tangent vector is measured w.r.t. the inner product on the tangent space $T_{\gamma(t)}\mathcal{M}$, hence we write $\|\dot\gamma(t)\|_{\gamma(t)}$. If the infimum is achieved by a curve γ, then γ is referred to as a **geodesic** (a shortest path); if $t \mapsto \|\dot\gamma(t)\|_{\gamma(t)}$ is constant, then it is called a **constant-speed geodesic**. From now on, we will only consider constant-speed geodesics, and the words "constant speed" will be dropped for brevity.

Given a functional $\mathcal{F} : \mathcal{M} \to \mathbb{R}$, the gradient of $\mathcal{F}$ at p is defined to be the unique element $\nabla\mathcal{F}(p) \in T_p\mathcal{M}$ such that for all curves $(p_t)_{t\in\mathbb{R}}$ passing through p at time 0 with velocity $v \in T_p\mathcal{M}$, it holds that $\partial_t|_{t=0}\mathcal{F}(p_t) = \langle \nabla\mathcal{F}(p), v\rangle_p$.

Wasserstein space as a Riemannian manifold. Based on our discussion thus far, it is natural to define the tangent space at $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ as

$$T_\mu\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \coloneqq \overline{\{\nabla\psi \mid \psi \in C_\mathrm{c}^\infty(\mathbb{R}^d)\}}^{L^2(\mu)},$$

where the notation denotes taking the $L^2(\mu)$ closure. Equivalently,

$$T_\mu\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) = \overline{\{\lambda\,(T - \mathrm{id}) \mid \lambda > 0,\ T \text{ is an optimal transport map}\}}^{L^2(\mu)}.$$

We equip the tangent space $T_\mu \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ with the $L^2(\mu)$ norm, which gives a Riemannian metric. This does not define a genuine Riemannian manifold (e.g., it is not locally homeomorphic to a Euclidean space or even a Hilbert space), but we will treat it as one for the purpose of developing calculation rules.

If the continuity equation $\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0$ holds and $v_t \in T_{\mu_t}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, then v_t is the tangent vector to the curve at time t. The condition $v_t \in T_{\mu_t}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is equivalent to saying that v_t is the optimal vector field considered in Theorem 1.3.19.

There are two questions to address. First, is this Riemannian structure compatible with the 2-Wasserstein distance? In other words, we know that a Riemannian metric induces a distance function; is the distance function induced by the Riemannian structure of $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ equal to W_2? Second, what are the geodesics? We answer both questions via the following theorem.

Theorem 1.3.22 (Wasserstein geodesics). *Let $\mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Then,*

$$W_2(\mu_0, \mu_1) = \inf\left\{ \int_0^1 \|v_t\|_{L^2(\mu_t)} \, \mathrm{d}t \;\middle|\; \partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0 \right\}. \tag{1.3.23}$$

The infimum is achieved as follows. Let $X_0 \sim \mu_0$ and $X_1 \sim \mu_1$ be optimally coupled random variables, let $X_t := (1-t)\,X_0 + t\,X_1$, and let $\mu_t := \operatorname{law}(X_t)$. Then, $t \mapsto \mu_t$ is the unique constant-speed geodesic joining μ_0 to μ_1.

Proof. Suppose that $\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0$. Then, $\int_0^1 \|v_t\|_{L^2(\mu_t)} \, \mathrm{d}t \geq \int_0^1 |\dot\mu|(t) \, \mathrm{d}t$. For a partition $0 \leq t_0 < t_1 < \cdots < t_k \leq 1$,

$$W_2(\mu_0, \mu_1) \leq \sum_{i=1}^k W_2(t_{i-1}, t_i) = \sum_{i=1}^k \frac{W_2(t_{i-1}, t_i)}{t_i - t_{i-1}} \, (t_i - t_{i-1}) \, .$$

As the size of the partition tends to zero, we obtain $W_2(\mu_0, \mu_1) \leq \int_0^1 |\dot\mu|(t) \, \mathrm{d}t$. This shows that $W_2(\mu_0, \mu_1)$ is at most the value of the infimum.

To show that equality holds, let X_t be defined as in the theorem statement and note that $\mathbb{E}[\|\dot X_t\|^2] = \|v_t\|^2_{L^2(\mu_t)}$ by the correspondence of the Lagrangian and Eulerian perspectives. (This can be verified by writing the vector field explicitly as $v_t = (T_1 - \mathrm{id}) \circ T_t^{-1}$, where $T_t := (1-t)\,\mathrm{id} + t\,T_{\mu_0 \to \mu_1}$—exercise!) Since $\mathbb{E}[\|\dot X_t\|^2] = \mathbb{E}[\|X_1 - X_0\|^2] = W_2^2(\mu_0, \mu_1)$ does not depend on time, the curve has constant speed, and $\int_0^1 \|v_t\|_{L^2(\mu_t)} \, \mathrm{d}t = W_2(\mu_0, \mu_1)$.

To show uniqueness, again work in the Lagrangian perspective: suppose we have an evolution $t \mapsto X_t$ of random variables such that $t \mapsto \mathbb{E}[\|\dot X_t\|^2]$ is constant, and $X_0 \sim \mu_0$,

$X_1 \sim \mu_1$. Then, we have

$$W_2^2(\mu_0, \mu_1) \le \mathbb{E}[\|X_1 - X_0\|^2] = \mathbb{E}\Big[\Big\|\int_0^1 \dot{X}_t \, \mathrm{d}t\Big\|^2\Big] \le \mathbb{E}\int_0^1 \|\dot{X}_t\|^2 \, \mathrm{d}t = \Big(\int_0^1 \|v_t\|_{L^2(\mu_t)} \, \mathrm{d}t\Big)^2$$

where the last equality follow from the constant-speed assumption. In order for the first inequality to be equality, (X_0, X_1) is an optimal coupling. In order for the second inequality to be equality, strict convexity of $\|\cdot\|^2$ implies that $\dot{X}_t$ is constant in time and equal to its average $\int_0^1 \dot{X}_t \, \mathrm{d}t = X_1 - X_0$. □

Definition 1.3.24. Let $\mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, and let $X_0 \sim \mu_0$, $X_1 \sim \mu_1$ be optimally coupled. Let $X_t := (1-t)\, X_0 + t\, X_1$, and let $\mu_t := \mathrm{law}(X_t)$. Then, the curve $t \mapsto \mu_t$ is called the **Wasserstein geodesic** joining μ_0 to μ_1. It is also called the **displacement interpolation** or **McCann's interpolation**.

Exponential map and logarithmic map. On a Riemannian manifold $\mathcal{M}$, the Riemannian **exponential map** $\exp_p : T_\mu\mathcal{M} \to \mathcal{M}$ takes a tangent vector $v \in T_\mu\mathcal{M}$ to the endpoint at time 1 of the constant-speed geodesic emanating from p with velocity v. The **logarithmic map** is then defined to be the inverse mapping $\log_p : \mathcal{M} \to T_\mu\mathcal{M}$. Actually, in general, the exponential map is only defined on a subset of the tangent space, because in many manifolds (e.g., the sphere), geodesics cannot continue indefinitely while remaining shortest paths between their endpoints. On Euclidean space $\mathbb{R}^d$, we have $\exp_p(v) = p + v$ and $\log_p(q) = q - p$.

We can identify these maps for $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$. If $(\mu_t)_{t\in[0,1]}$ is a Wasserstein geodesic and $\nabla\varphi_{\mu_0\to\mu_1}$ is the optimal transport map from μ_0 to μ_1, then the tangent vector to the geodesic at time 0 is $\nabla\varphi_{\mu_0\to\mu_1} - \mathrm{id}$. This implies that $\log_{\mu_0}(\mu_1) = \nabla\varphi_{\mu_0\to\mu_1} - \mathrm{id}$. The inverse mapping is then given as follows: if $\nabla\psi \in T_{\mu_0}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is such that $\mathrm{id} + \nabla\psi$ is the gradient of a convex function, then $\exp_{\mu_0}(\nabla\psi) = (\mathrm{id} + \nabla\psi)_{\#}\mu_0$.

Geodesically convex functionals. Over a Riemannian manifold $\mathcal{M}$, the correct way to define convexity is as follows.

Definition 1.3.25. Let $\mathcal{M}$ be a Riemannian manifold and let $\mathcal{F} : \mathcal{M} \to \mathbb{R} \cup \{\infty\}$ be smooth. For $\alpha \in \mathbb{R}$, we say that $\mathcal{F}$ is **α-geodesically convex** if one of the following equivalent conditions hold:

1. For all geodesics $(p_t)_{t\in[0,1]}$ and $t \in [0,1]$,

$$\mathcal{F}(p_t) \leq (1-t)\,\mathcal{F}(p_0) + t\,\mathcal{F}(p_1) - \frac{\alpha\, t\,(1-t)}{2}\,\mathrm{d}(p_0,p_1)^2\,,$$

   where d is the induced Riemannian distance (1.3.21).

2. For all $p, q \in \mathcal{M}$,

$$\mathcal{F}(q) \geq \mathcal{F}(p) + \langle \nabla\mathcal{F}(p), \log_p(q)\rangle_p + \frac{\alpha}{2}\,\mathrm{d}(p,q)^2\,.$$

   Here, ∇ denotes the Riemannian gradient.

3. For all constant-speed geodesics $(p_t)_{t\in[0,1]}$ with tangent vector $v_0 \in T_{p_0}\mathcal{M}$ at time 0, it holds that

$$\partial_t^2\big|_{t=0}\mathcal{F}(p_t) \geq \alpha\,\|v_0\|_{p_0}^2\,.$$

1.4 The Langevin SDE as a Wasserstein Gradient Flow

We are now ready to interpret the Langevin diffusion (1.E.1) as a gradient flow in the Wasserstein space $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$. Once we have done so, we will quickly deduce convergence results for the Langevin diffusion based on gradient flow computations.

1.4.1 Derivation of the Gradient Flow

Let $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}\cup\{\infty\}$ be a functional over the Wasserstein space. We now compute the Wasserstein gradient of $\mathcal{F}$ at μ, i.e., the element $\nabla_{W_2}\mathcal{F}(\mu) \in T_\mu\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ such that for every curve $t \mapsto \mu_t$ with $\mu_0 = \mu$, if v_0 is the tangent vector to the curve at time 0, then $\partial_t|_{t=0}\mathcal{F}(\mu_t) = \langle \nabla_{W_2}\mathcal{F}(\mu), v_0\rangle_\mu$, where $\langle\cdot,\cdot\rangle_\mu$ is the inner product on $T_\mu\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$.

We will give a formula in terms of the **first variation** of $\mathcal{F}$ at μ, denoted $\delta\mathcal{F}(\mu)$. The first variation is a function $\mathbb{R}^d \to \mathbb{R}$ which satisfies $\partial_t|_{t=0}\mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu)\,\partial_t|_{t=0}\mu_t$. This is almost the same as the $L^2(\mathfrak{m})$ gradient of $\mathcal{F}$, where $\mathfrak{m}$ is the Lebesgue measure on $\mathbb{R}^d$, except for a few differences: (1) there is no guarantee that $\delta\mathcal{F}(\mu) \in L^2(\mathfrak{m})$; (2) in order to consider the $L^2(\mathfrak{m})$ gradient, we would want $\mathcal{F}$ to be a functional defined over all of $L^2(\mathfrak{m})$, not just probability densities, and similarly we would have to consider all curves in $L^2(\mathfrak{m})$ rather than curves of probability densities.

As a consequence of looking only at probability densities, the first variation is only defined up to an additive constant. Indeed, $\partial_t|_{t=0}\mu_t$ always integrates to 0, so we can add

any constant to the first variation. This does not cause any ambiguity, as we now see.

Recall that v_t is the tangent vector to the curve of measures at time t if the continuity equation $\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0$ holds and $v_t \in T_{\mu_t} \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Using the continuity equation with a curve such that $\mu_0 = \mu$,

$$\partial_t\big|_{t=0} \mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu)\, \partial_t\big|_{t=0} \mu_t = -\int \delta\mathcal{F}(\mu) \operatorname{div}(v_0 \mu) = \int \langle \nabla \delta\mathcal{F}(\mu), v_0 \rangle \, \mathrm{d}\mu \,.$$

(Here, the ∇ is the Euclidean gradient.) Since $\nabla\delta\mathcal{F}(\mu)$ is the gradient of a function, from our characterization of the tangent space we know that $\nabla\delta\mathcal{F}(\mu) \in T_\mu \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Therefore, the equation above says that the Wasserstein gradient of $\mathcal{F}$ at μ is $\nabla\delta\mathcal{F}(\mu)$.

Theorem 1.4.1. *Let $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ be a functional. Then, its Wasserstein gradient at μ is*

$$\nabla_{W_2} \mathcal{F}(\mu) = \nabla \delta\mathcal{F}(\mu) \,,$$

where $\delta\mathcal{F}(\mu)$ is a first variation of $\mathcal{F}$ at μ.

Since we take the Euclidean gradient of the first variation, the fact that the first variation is only defined up to additive constant does not bother us.

The **Wasserstein gradient flow** of $\mathcal{F}$ is by definition a curve of measures $t \mapsto \mu_t$ such that its tangent vector v_t at time t is $v_t = -\nabla_{W_2} \mathcal{F}(\mu_t)$. Substituting this into the continuity equation (1.3.18), we obtain the gradient flow equation

$$\partial_t \mu_t = \operatorname{div}\big(\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)\big) = \operatorname{div}\big(\mu_t \nabla \delta\mathcal{F}(\mu_t)\big) \,.$$

Example 1.4.2. Consider $\mathcal{F} = \mathsf{KL}(\cdot \,\|\, \pi)$ where $\pi = \exp(-V)$. This functional can also be written as

$$\mathcal{F}(\mu) = \int \mu \ln \frac{\mu}{\pi} = \int V \, \mathrm{d}\mu + \int \mu \ln \mu \,.$$

These two terms have the interpretation of *energy* and (negative) *entropy*. From this, we can compute that

$$\delta\mathcal{F}(\mu) = V + \ln \mu + \text{constant}$$

and therefore

$$\nabla_{W_2}\mathcal{F}(\mu) = \nabla V + \nabla \ln \mu = \nabla \ln \frac{\mu}{\pi}.$$

The Wasserstein gradient flow of $\mathcal{F}$ satisfies

$$\partial_t \mu_t = \operatorname{div}\Big(\mu_t \nabla \ln \frac{\mu_t}{\pi}\Big).$$

Comparing with the Fokker–Planck equation $\partial_t \pi_t = \mathscr{L}^* \pi_t$ and the form of the adjoint generator $\mathscr{L}^*$ for the Langevin diffusion (see Example 1.2.8), we obtain a truly remarkable fact: *the law $t \mapsto \pi_t$ of the Langevin diffusion with potential V is the Wasserstein gradient flow of* $\mathsf{KL}(\cdot \,\|\, \pi)$.

The calculus of optimal transport was introduced by Otto in [Ott01], and it is often known as **Otto calculus**; the interpretation of the Langevin diffusion in this context was put forth in the seminal work [JKO98]. The paper [Ott01] also raises and answers a salient question: given that we can view dynamics as gradient flows in different ways (e.g. the Langevin diffusion can be either viewed as the gradient flow of the Dirichlet energy in $L^2(\pi)$, or the gradient flow of the KL divergence in $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$), what makes us prefer one gradient flow structure over another? Otto argues that the Wasserstein geometry is particularly natural because it cleanly separates out two aspects of the problem: the *geometry* of the ambient space, which is reflected in the metric on $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, and the *objective functional*. Moreover, the objective functional in the Wasserstein perspective is physically intuitive because it has an interpretation in *thermodynamics*. From a sampling standpoint, the Wasserstein geometry is undoubtedly more compelling and useful.

In our exposition, we focused on the calculation rules for Wasserstein gradient flows, but this is not how they are normally defined. Instead, one usually considers a sequence of discrete approximations to the gradient flow and proves that there is a limiting curve; this is called the minimizing movements scheme and it is developed in detail in [AGS08].

1.4.2 Convexity of the KL Divergence

The key to studying gradient flows is to understand the convexity properties of the objective functional. For the specific functional $\mathcal{F} \coloneqq \mathsf{KL}(\cdot \,\|\, \pi)$ with target $\pi = \exp(-V)$, our next goal is therefore to compute the Wasserstein Hessian of $\mathcal{F}$. When we computed Wasserstein gradients, we were free to differentiate $\mathcal{F}$ along any curve $t \mapsto \mu_t$ of measures, but we have to be more careful when computing the Hessian. If we take a function

$f : \mathbb{R}^d \to \mathbb{R}$ on Euclidean space and a curve $t \mapsto x_t$, then $\partial_t f(x_t) = \langle \nabla f(x_t), \dot{x}_t \rangle$ and

$$\partial_t^2 f(x_t) = \langle \nabla^2 f(x_t)\, \dot{x}_t, \dot{x}_t \rangle + \langle \nabla f(x_t), \ddot{x}_t \rangle \,.$$

Here, instead of just obtaining the Hessian, we have an additional term. However, if $t \mapsto x_t$ is a constant-speed *geodesic*, then it has no acceleration ($\ddot{x}_t = 0$), and the extra term vanishes.

In the same way, let $(\mu_t)_{t\in[0,1]}$ denote a Wasserstein geodesic. Explicitly, if T denotes the optimal transport map from μ_0 to μ_1, then $\mu_t = [(1-t)\,\mathrm{id} + t\,T]_{\#}\mu_0$. We will calculate $\partial_t^2|_{t=0}\mathcal{F}(\mu_t)$, as a function of the tangent vector $T - \mathrm{id} \in T_{\mu_0}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$; this is interpreted as $\langle \nabla^2_{W_2}\mathcal{F}(\mu_0)\,(T-\mathrm{id}), T-\mathrm{id}\rangle_{\mu_0}$. If we can lower bound this by $\alpha\,\|T-\mathrm{id}\|^2_{\mu_0}$, for all μ_0 and all optimal transport maps T, it means that $\mathcal{F}$ is α-strongly convex.

Write $\mathcal{E}(\mu) \coloneqq \int V\,\mathrm{d}\mu$ for the energy and $\mathcal{H}(\mu) \coloneqq \int \mu \ln \mu$ for the entropy. We deal with the two terms separately. First, for $X_t = (1-t)\,X_0 + t\,T(X_0)$ and $X_0 \sim \mu_0$,

$$\begin{aligned}
\partial_t \mathcal{E}(\mu_t) &= \partial_t\, \mathbb{E}\, V(X_t) = \mathbb{E}\langle \nabla V(X_t), \dot{X}_t\rangle = \mathbb{E}\langle \nabla V(X_t), T(X_0) - X_0\rangle \,, \\
\partial_t^2\big|_{t=0} \mathcal{E}(\mu_t) &= \mathbb{E}\langle \nabla^2 V(X_0)\,(T(X_0)-X_0), T(X_0)-X_0\rangle \,.
\end{aligned}$$

If V is α-strongly convex, then this is lower bounded by $\alpha\,\|T-\mathrm{id}\|^2_{\mu_0}$, which means that $\mathcal{E}$ is α-strongly convex.

The entropy is slightly trickier. Write $T_t \coloneqq (1-t)\,\mathrm{id} + t\,T$. Since $(T_t)_{\#}\mu_0 = \mu_t$, the change of variables formula shows that

$$\frac{\mu_0}{\mu_t \circ T_t} = \det \nabla T_t \,.$$

Therefore,

$$\begin{aligned}
\mathcal{H}(\mu_t) &= \int \mu_t \ln \mu_t = \int \mu_0 \ln(\mu_t \circ T_t) = \int \mu_0 \ln \frac{\mu_0}{\det \nabla T_t} \\
&= \mathcal{H}(\mu_0) - \int \mu_0 \ln \det\big((1-t)\,I_d + t\,\nabla T\big) \,.
\end{aligned}$$

Already from the fact that $-\ln\det$ is convex on the space of positive semidefinite matrices, we can see that $\partial_t^2 \mathcal{H}(\mu_t) \geq 0$. A more careful computation (Exercise 1.18) based on the derivatives of $-\ln\det$ shows that

$$\partial_t^2\big|_{t=0} \mathcal{H}(\mu_t) = \int \|\nabla T - I_d\|^2_{\mathrm{HS}}\,\mathrm{d}\mu_0 \geq 0 \,. \tag{1.4.3}$$

We have obtained the following result.

Theorem 1.4.4. *If $\pi \propto \exp(-V)$, where V is α-strongly convex, then $\mathsf{KL}(\cdot \,\|\, \pi)$ is also α-strongly convex along Wasserstein geodesics.*

Consequences of strong convexity. The strong convexity of the KL divergence implies the following statement (c.f. Definition 1.3.25).

Theorem 1.4.5. *If $\pi \propto \exp(-V)$, where V is α-strongly convex, then*

$$\mathsf{KL}(\nu \,\|\, \pi) \geq \mathsf{KL}(\mu \,\|\, \pi) + \left\langle \nabla \ln \frac{\mu}{\pi}, T_{\mu\to\nu} - \mathrm{id} \right\rangle_{\mu} + \frac{\alpha}{2} W_2^2(\mu, \nu)$$

for all $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$.

We now explore the implications of this fact for the gradient flow. If $t \mapsto \mu_t$ is the gradient flow for a functional $\mathcal{F}$ with $\inf \mathcal{F} = 0$, then

$$\partial_t \mathcal{F}(\mu_t) = \langle \nabla_{W_2} \mathcal{F}(\mu_t), \underbrace{-\nabla_{W_2} \mathcal{F}(\mu_t)}_{\text{tangent vector of the curve}} \rangle_{\mu_t} = -\|\nabla_{W_2} \mathcal{F}(\mu_t)\|_{\mu_t}^2 .$$

So, $t \mapsto \mathcal{F}(\mu_t)$ is always decreasing, and if the condition

$$\|\nabla_{W_2} \mathcal{F}(\mu)\|_{\mu}^2 \geq 2\alpha\, \mathcal{F}(\mu) \qquad \text{for all } \mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$$

holds, then the gradient flow converges exponentially fast, $\mathcal{F}(\mu_t) \leq \exp(-2\alpha t)\, \mathcal{F}(\mu_0)$, as a consequence of Grönwall's lemma (Lemma 1.1.21). This condition is known as a *gradient domination* condition, or a **Polyak–Łojasiewicz (PL) inequality**. It is implied by strong convexity: starting from

$$\mathcal{F}(\nu) \geq \mathcal{F}(\mu) + \langle \nabla_{W_2} \mathcal{F}(\mu), T_{\mu\to\nu} - \mathrm{id} \rangle_{\mu} + \frac{\alpha}{2} W_2^2(\mu, \nu) , \tag{1.4.6}$$

we take $\nu = \mu^\star := \arg\min \mathcal{F}$ so that $\mathcal{F}(\nu) = 0$, yielding

$$\mathcal{F}(\mu) \leq -\langle \nabla_{W_2} \mathcal{F}(\mu), T_{\mu\to\mu^\star} - \mathrm{id} \rangle_{\mu} - \frac{\alpha}{2} W_2^2(\mu, \mu^\star) \leq \frac{1}{2\alpha} \|\nabla_{W_2} \mathcal{F}(\mu)\|_{\mu}^2 ,$$

where the last line uses Young's inequality and $\|T_{\mu\to\mu^\star} - \mathrm{id}\|_{\mu} = W_2(\mu, \mu^\star)$. Therefore, strong convexity implies exponentially fast convergence of the gradient flow. If we apply this to the Langevin diffusion, we deduce that α-strong convexity of V implies

$$\mathsf{KL}(\mu \,\|\, \pi) \leq \frac{1}{2\alpha} \left\| \nabla \ln \frac{\mu}{\pi} \right\|_{\mu}^2 = \frac{1}{2\alpha} \mathsf{FI}(\mu \,\|\, \pi) \qquad \text{for all } \mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) . \tag{1.4.7}$$

This is precisely the log-Sobolev inequality (see Example 1.2.26); we have recovered the Bakry–Émery theorem (Theorem 1.2.29) that $\mathrm{CD}(\alpha, \infty)$ implies an LSI, as well as Theorem 1.2.25 which asserted that an LSI yields exponentially fast decay in the KL divergence.

Next, starting from the strong convexity inequality (1.4.6), we take $\mu = \mu^\star$ so that $\nabla_{W_2}\mathcal{F}(\mu^\star) = 0$, and hence

$$\mathcal{F}(\nu) \geq \frac{\alpha}{2} W_2^2(\nu, \mu^\star) \qquad \text{for all } \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) .$$

This is a *quadratic growth* inequality; as the name suggests, it asserts that $\mathcal{F}$ must grow quadratically away from its minimizer. For the Langevin diffusion, it says

$$\mathsf{KL}(\mu \,\|\, \pi) \geq \frac{\alpha}{2} W_2^2(\mu, \pi) \qquad \text{for all } \mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) . \tag{1.4.8}$$

This is known as **Talagrand's T_2 inequality** and it is an example of a *transportation inequality*. Such inequalities have been closely studied in relation to the concentration of measure phenomenon (see [Han16, Chapter 4]).

It is a general fact that the PL inequality implies the quadratic growth inequality. When applied to the Langevin diffusion, it says that the LSI implies the T_2 inequality, which is known as the **Otto–Villani theorem** [OV00]. See Exercise 1.16 for a proof.

Strong convexity also implies another fact: the gradient flow *contracts* exponentially fast. Namely, if we have two gradient flows $t \mapsto \mu_t$ and $t \mapsto \nu_t$ for a strongly convex functional $\mathcal{F}$, then

$$W_2^2(\mu_t, \nu_t) \leq \exp(-2\alpha t)\, W_2^2(\mu_0, \nu_0) . \tag{1.4.9}$$

In particular, if we take $\nu_t = \mu^\star$ for all t, then we obtain exponentially fast convergence to the minimizer in Wasserstein distance. For the Langevin diffusion, inequality (1.4.9) is implied by the following theorem (see Exercise 1.17).

Theorem 1.4.10. *Suppose that $\nabla^2 V \geq \alpha I_d$ for some $\alpha \in \mathbb{R}$. If $(Z_t)_{t\geq 0}$ and $(Z'_t)_{t\geq 0}$ denote two copies of the Langevin diffusion (1.E.1) with potential V and driven by the same Brownian motion, then*

$$\mathbb{E}[\|Z_t - Z'_t\|^2] \leq \exp(-2\alpha t)\, \mathbb{E}[\|Z_0 - Z'_0\|^2] .$$

Finally, in the case $\alpha = 0$, so that $\mathcal{F}$ is weakly convex, we can also obtain a convergence result by considering the Lyapunov functional $\mathcal{L}_t := t\,\mathcal{F}(\mu_t) + \frac{1}{2} W_2^2(\mu_t, \mu^\star)$. In order to differentiate this Lyapunov functional, we need the following theorem.

Theorem 1.4.11. *For $\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the Wasserstein gradient of $\mu \mapsto W_2^2(\mu,\nu)$ at μ is given by $-2\,(T_{\mu\to\nu} - \mathrm{id})$.*

Proof. See [Vil09, Theorem 23.9]. □

In general, on a Riemannian manifold, the gradient of $\mathrm{d}(\cdot,q)^2$ at p is $-2\log_p(q)$. Check that this formula makes sense on Euclidean space $\mathbb{R}^d$.

Differentiating in time and applying (1.4.6) and the lemma,

$$\partial_t \mathcal{L}_t = \mathcal{F}(\mu_t) - t\,\|\nabla_{W_2}\mathcal{F}(\mu_t)\|_{\mu_t}^2 + \underbrace{\langle \nabla_{W_2}\mathcal{F}(\mu_t), T_{\mu_t\to\mu^\star} - \mathrm{id}\rangle_{\mu_t}}_{\le -\mathcal{F}(\mu_t)} \le 0\,.$$

Hence, $\mathcal{L}_t \le \mathcal{L}_0$, which implies

$$\mathcal{F}(\mu_t) \le \frac{1}{2t}\, W_2^2(\mu_0,\mu^\star)\,. \tag{1.4.12}$$

1.5 Overview of the Convergence Results

1.5.1 Convergence Results

The main convergence results we have developed can be summarized as follows.

- $\mathsf{KL}(\cdot\,\|\,\pi)$ is α-strongly convex along W_2 geodesics if and only if V is strongly convex, if and only if: for all $\mu_0,\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, if $(\mu_t)_{t\ge0}$, $(\nu_t)_{t\ge0}$ are Langevin diffusions started at μ_0 and ν_0 respectively, then $W_2^2(\mu_t,\nu_t) \le \exp(-2\alpha t)\, W_2^2(\mu_0,\nu_0)$.
- The target π satisfies the log-Sobolev inequality (LSI) with constant $1/\alpha$ if and only if for all $\pi_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, along the Langevin dynamics $t \mapsto \pi_t$ started at π_0 it holds that $\mathsf{KL}(\pi_t\,\|\,\pi) \le \exp(-2\alpha t)\,\mathsf{KL}(\pi_0\,\|\,\pi)$. The LSI is a gradient domination condition in Wasserstein space.
- The target π satisfies the Poincaré inequality (PI) with constant $1/\alpha$ if and only if for all $\pi_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, along the Langevin dynamics $t \mapsto \pi_t$ started at π_0 it holds that $\chi^2(\pi_t\,\|\,\pi) \le \exp(-2\alpha t)\,\chi^2(\pi_0\,\|\,\pi)$. The Poincaré inequality is a spectral gap condition for the generator of the Langevin diffusion.

The conditions are listed from strongest to weakest: α-strong log-concavity implies α^{-1}-LSI, which implies α^{-1}-Poincaré. In addition:

- If the target π is log-concave, then along the Langevin dynamics $t \mapsto \pi_t$ it holds that $\mathsf{KL}(\pi_t \,\|\, \pi) \le \frac{1}{2t} W_2^2(\pi_0, \pi)$.

When we turn towards discretization analysis, there are two main ways in which the continuous-time result affects the analysis: the strength of the continuous-time result, and the metric in which we must perform the analysis.

Regarding the first point, the first two results are generally more useful because at initialization, we typically have $W_2^2(\pi_0, \pi),\ \mathsf{KL}(\pi_0 \,\|\, \pi) = O(d)$ (at least when π is strongly log-concave). Hence, exponential convergence in W_2^2 and KL both imply that the amount of time it takes for the Langevin diffusion to reach ε error is $O(\log(d/\varepsilon))$. In contrast, the chi-squared divergence is typically much larger at initialization: $\chi^2(\pi_0 \,\|\, \pi) = \exp(O(d))$. Therefore, the chi-squared result implies that the Langevin diffusion takes $O(d \vee \log(1/\varepsilon))$ time to reach ε error.

Regarding the second point, the W_2 contraction under strong log-concavity is the easiest to turn into a sampling guarantee for the discretized algorithm. This is because to bound the W_2 distance, we can use straightforward *coupling* techniques. On the other hand, a continuous-time result in KL or χ^2 often requires the discretization analysis to also be carried out in KL or χ^2, which is substantially trickier.

1.5.2 Appendix: Divergences between Probability Measures

As we have already seen, the analysis of Langevin introduces many different notions of divergences between probability measures. Therefore, it is important to develop a healthy understanding of the relationships between these divergences.

First of all, there is a distinction between the Wasserstein metric, which is a transport distance (measuring how far we must move the mass of one measure to the other), and information divergences which are defined directly in terms of the densities such as the KL divergence and the chi-squared divergence. Note that the latter two divergences are infinite unless the first argument is absolutely continuous w.r.t. the second, which is certainly not the case for the Wasserstein metric.

We introduce another important metric.

Definition 1.5.1. The **total variation** (**TV**) **distance** between probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is defined via

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| = \sup_{f : \mathcal{X} \to [0,1]} \left| \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu \right|$$

$$= \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \mathbb{1}\{x \neq y\} \, \mathrm{d}\gamma(x,y) = \frac{1}{2} \int \left| \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \right| \mathrm{d}\lambda \,,$$

where λ is a common dominating measure for μ and ν.

The TV metric is indeed a metric on the space $\mathcal{P}(\mathcal{X})$ (in fact, it can be extended to a norm on the space $\mathcal{M}(\mathcal{X})$ of signed measures). The TV distance can be thought of as both a transport metric (with cost $(x,y) \mapsto \mathbb{1}\{x \neq y\}$; in fact, the TV distance is a special case of the W_1 metric introduced in Exercise 1.11) and an information divergence.

The family of information divergences can be further expanded by introducing the following definition.

Definition 1.5.2. Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$, and let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $f(1) = 0$. Then, the f**-divergence** of μ from ν is

$$\mathcal{D}_f(\mu \,\|\, \nu) := \int f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu \,, \qquad \text{if } \mu \ll \nu \,.$$

In general, if $\mu \not\ll \nu$, we let p_μ, p_ν denote the respective densities of μ and ν w.r.t. a common dominating measure. Then,

$$\mathcal{D}_f(\mu \,\|\, \nu) := \int_{p_\nu > 0} f\left(\frac{p_\mu}{p_\nu}\right) \mathrm{d}\nu + f'(\infty) \, \mu\{p_\nu = 0\} \,.$$

For example, the TV distance corresponds to $f(x) = \frac{1}{2}|x-1|$, the KL divergence corresponds to $f(x) = x \ln x$, and the χ^2 divergence corresponds to $f(x) = (x-1)^2$. When f has superlinear growth, then $f'(\infty) = \infty$ and hence $\mathcal{D}_f(\mu \,\|\, \nu) = \infty$ unless $\mu \ll \nu$, but the second more general definition given above is necessary to recover the TV distance.

We always have $\chi^2 \geq \ln(1+\chi^2) \geq \mathsf{KL} \geq 2\|\cdot\|_{\mathrm{TV}}^2$ (the last inequality is **Pinsker's inequality**, see Exercise 2.13), and under a T_2 transport inequality with constant α^{-1} (which is implied by α^{-1}-LSI) we have $\mathsf{KL} \geq \frac{\alpha}{2} W_2^2$. This chain of inequalities helps to explain why, if the KL divergence is of order d, then the χ^2 divergence is of order $\exp d$.

In Section 2.2.4, we will also introduce the closely related family of divergences known as Rényi divergences.

We conclude by stating a few key facts (without complete proofs) about f-divergences. The first is the data-processing inequality.

Theorem 1.5.3 (data-processing inequality). *Suppose that $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and that P is any Markov kernel. Then, for any f-divergence, it holds that*

$$\mathcal{D}_f(\mu P \,\|\, \nu P) \le \mathcal{D}_f(\mu \,\|\, \nu)\,.$$

Equivalently, $\mathcal{D}_f(\cdot \,\|\, \cdot)$ is jointly convex in its two arguments.

Proof sketch. To simplify, we will abuse notation and identify all probability measures with densities. Then, by Jensen's inequality,

$$\begin{aligned}
\mathcal{D}_f(\mu \,\|\, \nu) &= \int f\Big(\frac{\mu(x)\,P(x,y)}{\nu(x)\,P(x,y)}\Big)\,\nu(\mathrm{d}x)\,P(x,\mathrm{d}y) \\
&= \int f\Big(\frac{\mu(x)\,P(x,y)}{\nu(x)\,P(x,y)}\Big)\,\frac{\nu(\mathrm{d}x)\,P(x,y)}{\nu P(y)}\,\nu P(\mathrm{d}y) \\
&\ge \int f\Big(\int \frac{\mu(x)\,P(x,y)}{\nu(x)\,P(x,y)}\,\frac{\nu(\mathrm{d}x)\,P(x,y)}{\nu P(y)}\Big)\,\nu P(\mathrm{d}y) = \int f\Big(\frac{\mu P(y)}{\nu P(y)}\Big)\,\nu P(\mathrm{d}y)\,. \qquad \square
\end{aligned}$$

The remaining facts are specific to the KL divergence. The Donsker–Varadhan theorem expresses the KL divergence via a variational principle.

Theorem 1.5.4 (Donsker–Varadhan variational principle). *Suppose that $\mu, \nu \in \mathcal{P}(\mathcal{X})$, where $\mathcal{X}$ is a Polish space. Then,*

$$\mathsf{KL}(\mu \,\|\, \nu) = \sup\big\{\mathbb{E}_\mu\, g - \ln \mathbb{E}_\nu \exp g \;\big|\; g : \mathcal{X} \to \mathbb{R} \text{ is bounded and measurable}\big\}\,.$$

The theorem asserts that the functionals $\mu \mapsto \mathsf{KL}(\mu \,\|\, \nu)$ and $g \mapsto \ln \mathbb{E}_\nu \exp g$ are convex conjugates of each other. See [DZ10, Lemma 6.2.13] or [RS15, Theorem 5.4] for careful proofs, or see the remark after Lemma 2.3.4.

Lastly, we have the chain rule for the KL divergence.

Lemma 1.5.5 (chain rule for KL divergence). *Let $\mathcal{X}_1$, $\mathcal{X}_2$ be Polish spaces and suppose we are given two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ with $\mu \ll \nu$. Let μ_1 be the $\mathcal{X}_1$ marginal of μ, and let $\mu_{2|1}(\cdot \mid \cdot)$ be the conditional distribution for μ on $\mathcal{X}_2$ conditioned*

on $\mathcal{X}_1$; likewise define ν_1 and $\nu_{2|1}$. Then, it holds that

$$\mathsf{KL}(\mu \,\|\, \nu) = \mathsf{KL}(\mu_1 \,\|\, \nu_1) + \int \mathsf{KL}\big(\mu_{2|1}(\cdot \mid x_1) \,\big\|\, \nu_{2|1}(\cdot \mid x_1)\big)\, \mu_1(\mathrm{d}x_1)\,.$$

We invite the reader to prove the chain rule in the discrete case ($\mathcal{X}_1$ and $\mathcal{X}_2$ are finite sets), free of measure-theoretic guilt.

Bibliographical Notes

Much of the material in this chapter is foundational, with entire textbooks giving comprehensive treatments of the topics. For stochastic calculus, there is of course a long list of textbooks, but as a starting place we suggest [Ste01; Le 16]. For Markov semigroup theory, see [BGL14; Han16]. For optimal transport, the core theory is developed in [Vil03; San15], and for a rigorous development of Otto calculus see [AGS08; Vil09].

The notion of solution used in Section 1.1.3 is more typically called a *strong solution* to the SDE, because given any Brownian motion B we can find a process X which is driven by B and which satisfies the SDE. There is also a notion of *weak solution*, in which we are allowed to construct the probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ and the Brownian motion B together with the solution X. We will not worry about the distinction in this book, since strong solutions suffice for our purposes.

See [San15, §1.6.3] for an elegant proof of strong duality for the optimal transport problem via convex duality.

The perspective of the Langevin diffusion as a Wasserstein gradient flow was introduced in [JKO98]; the application of Otto calculus to functional inequalities was given in [OV00]; and the calculation rules for Otto calculus were set out in [Ott01]. These three papers are seminal and are worth reading carefully. An alternative (but related) approach to functional inequalities via optimal transport is given in [Cor02]. The formal proof of the Otto–Villani theorem in Exercise 1.16 was made rigorous via entropic interpolations in [Gen+20]; see [BB18] for a generalization.

The Efron–Stein inequality in Exercise 1.2 is just one example of the use of martingales to derive concentration inequalities; see [BLM13; Han16] for more on this topic. We will also revisit the martingale method in the next chapter; see Exercise 2.15.

The upper bound (1.E.2) in Exercise 1.10 is surprisingly sharp: it holds that

$$\frac{1}{2}\,\|\Sigma_0^{1/2} - \Sigma_1^{1/2}\|_{\mathrm{HS}}^2 \leq W_2^2(\mu_0, \mu_1) - \|m_0 - m_1\|^2 \leq \|\Sigma_0^{1/2} - \Sigma_1^{1/2}\|_{\mathrm{HS}}^2\,,$$

see [CV21, Lemma 3.5].

The proof of the dynamical formulation of dual optimal transport in Exercise 1.12 is carried out rigorously in [Vil03, §8.1]. We mention that the Hamilton–Jacobi equation has a close connection with classical mechanics; in particular, the characteristics of the Hamilton–Jacobi equation are precisely Hamilton's equations of motion [Eva10, §3.3]. In the context of optimal transport, the Hamiltonian consists only of kinetic energy (no potential energy) and hence the characteristics are straight lines traversed at constant speed; this is of course consistent with the description of Wasserstein geodesics. The Hamilton–Jacobi equation, the Hopf–Lax semigroup, and their connection with optimal transport can also be generalized to other costs; see [Vil03, §5.4].

Exercises

A Primer on Stochastic Calculus

⊵ Exercise 1.1 (Doob's L^p maximal inequality)
Prove Doob's L^p maximal inequality (Corollary 1.1.31) for discrete-time submartingales.

Hint: Start with the inequality $\lambda \mathbb{P}(M_N^* \geq \lambda) \leq \mathbb{E}[M_N \mathbb{1}\{M_N^* \geq \lambda\}]$ from the proof of Theorem 1.1.30, where $M_N^* := \max_{k=0,1,\dots,N} M_k$. Compute the $L^p(\mathbb{P})$ norm of M_N^* by integrating the tails, and apply the above inequality together with Hölder's inequality.

⊵ Exercise 1.2 (orthogonality of martingale increments)
Let $(M_n)_{n\in\mathbb{N}}$ be a discrete-time martingale which is adapted to a filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$ and satisfies $\mathbb{E}[M_n^2] < \infty$ for all $n \in \mathbb{N}$. Let $\Delta_n := M_{n+1} - M_n$ denote the martingale increment.

1. Prove that for $m, n \in \mathbb{N}$ with $m \neq n$, $\mathbb{E}[\Delta_m \Delta_n] = 0$: the martingale increments are orthogonal. In particular, if $M_0 = 0$, then $\mathbb{E}[M_n^2] = \sum_{k=0}^{n-1} \mathbb{E}[\Delta_k^2]$.

2. Let $(X_i)_{i=1}^n$ be independent random variables taking values in some space $\mathcal{X}$, and suppose that the function $f : \mathcal{X}^n \to \mathbb{R}$ is bounded and measurable. Check that if $M_k := \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k]$, then the **Doob martingale** $(M_k)_{k=1}^n$ is indeed a martingale. Then, using the previous part, prove the following tensorization property of the variance:
$$\operatorname{var} f(X_1, \dots, X_n) \leq \mathbb{E} \sum_{k=1}^{n} \operatorname{var}\big(f(X_1, \dots, X_n) \,\big|\, X_{-k}\big),$$
where $X_{-k} := (X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)$.

3. Define the discrete derivative
$$D_k f(x) := \sup_{x_k' \in \mathcal{X}} f(x_1, \dots, x_{k-1}, x_k', x_{k+1}, \dots, x_n)$$

$$- \inf_{x_k' \in \mathcal{X}} f(x_1, \dots, x_{k-1}, x_k', x_{k+1}, \dots, x_n) \,.$$

Prove the inequality

$$\operatorname{var} f(X_1, \dots, X_n) \le \frac{1}{4} \mathbb{E} \sum_{k=1}^{n} \{D_k f(X_1, \dots, X_n)\}^2 \,.$$

This inequality, known as the **Efron–Stein inequality**, expresses the fact that a function f of independent random variables which is not too sensitive to any individual coordinate has controlled variance. This is a concentration inequality which has useful consequences in many probabilistic settings, see, e.g., [BLM13].

Hint: First prove that a random variable which takes values in $[a, b]$ has variance bounded by $\frac{1}{4}(b-a)^2$.

⊵ Exercise 1.3 (L^2 bounded martingale convergence theorem)
Let $(M_n)_{n\in\mathbb{N}}$ be a discrete-time martingale which is adapted to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Assume that the martingale is bounded in $L^2(\mathbb{P})$: $\sup_{n\in\mathbb{N}} \mathbb{E}[M_n^2] \le B^2 < \infty$. Prove that the martingale converges a.s. and in $L^2(\mathbb{P})$ to a limit M_∞ with $\mathbb{E}[M_\infty^2] \le B^2$.

Hint: For the a.s. convergence, it suffices to show that for all $\varepsilon > 0$, the event

$$\Big\{ \lim_{m\to\infty} \sup_{n\in\mathbb{N},\, n\ge m} |M_m - M_n| \ge \varepsilon \Big\}$$

has probability zero. To do so, use Doob's maximal inequality (Theorem 1.1.30) and orthogonality of martingale increments (Exercise 1.2).

⊵ Exercise 1.4 (explosion of ODEs)
Solve the following ODE on $\mathbb{R}$: $\dot{x}_t = b(x_t)$ with initial condition $x_0 \in \mathbb{R}$, where $b(x) = |x|^\alpha$, $\alpha > 0$. Show that

1. when $0 < \alpha < 1$, there are multiple solutions to the ODE (with initial condition $x_0 = 0$) so that uniqueness fails;
2. when $\alpha = 1$ (and hence b is globally Lipschitz), there is a unique solution to the ODE which is finite for all time;
3. when $\alpha > 1$, then the solution to the ODE blows up in finite time.

⊵ Exercise 1.5 (Ornstein–Uhlenbeck process)
One of the most important diffusions that we will encounter is the **Ornstein–Uhlenbeck (OU) process**, which solves the SDE

$$\mathrm{d}X_t = -X_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \,.$$

Give an explicit expression for X_t in terms of X_0 and an Itô integral involving $(B_t)_{t\geq 0}$. From this expression, can you read off the stationary distribution of this process?

Hint: Apply Itô's formula to $f(t, X_t) = X_t \exp t$. To find the stationary distribution, justify the following fact: if $(\eta_t)_{t\geq 0}$ is a *deterministic* function, then $\int_0^T \eta_t \, \mathrm{d}B_t$ is a Gaussian with mean zero and variance $\int_0^T \eta_t^2 \, \mathrm{d}t$.

Markov Semigroup Theory

▷ Exercise 1.6 (basic properties of the Markov semigroup)
Let $(P_t)_{t\geq 0}$ be a Markov semigroup with carré du champ Γ.

1. Prove that if $\phi : \mathbb{R} \to \mathbb{R}$ is convex, then $P_t\phi(f) \geq \phi(P_t f)$ and $\mathscr{L}\phi(f) \geq \phi'(f)\,\mathscr{L}f$ whenever the expressions are well-defined.

2. If $(X_t)_{t\geq 0}$ denotes the Markov process associated with the semigroup and f is smooth, then the process $t \mapsto f(X_t) - \int_0^t \mathscr{L}f(X_s)\,\mathrm{d}s$ is a continuous local martingale. In particular, $(f(X_t))_{t\geq 0}$ is a continuous local martingale if and only if $\mathscr{L}f = 0$.

3. Prove that for $f, g \in L^2(\pi)$ in the domain of the carré du champ, we have the Cauchy–Schwarz inequality $\Gamma(f, g) \leq \sqrt{\Gamma(f, f)\,\Gamma(g, g)}$.

   Hint: For $\lambda \in \mathbb{R}$, consider $0 \leq \Gamma(f + \lambda g, f + \lambda g)$ and use bilinearity.

▷ Exercise 1.7 (functional inequalities and exponential decay)
Prove the equivalence between the Poincaré inequality and exponential decay of variance (Theorem 1.2.20), and the equivalence between the log-Sobolev inequality and exponential decay of entropy (Theorem 1.2.21).

▷ Exercise 1.8 (log-Sobolev implies Poincaré)
Linearize the log-Sobolev inequality to obtain the Poincaré inequality.

Hint: Argue that if $f \in C_c^\infty(\mathbb{R}^d)$ satisfies $\int f \,\mathrm{d}\pi = 0$, then

$$\mathsf{KL}\big((1 + \varepsilon f)\,\pi \,\big\|\, \pi\big) = \frac{\varepsilon^2}{2} \int f^2 \,\mathrm{d}\pi + o(\varepsilon^2)\,. \tag{1.E.1}$$

▷ Exercise 1.9 (mixing of the Ornstein–Uhlenbeck process)
Consider the Ornstein–Uhlenbeck process $(X_t)_{t\geq 0}$ introduced in Exercise 1.5. Note that this is just an instance of the Langevin diffusion with potential $V(x) = \frac{\|x\|^2}{2}$.

1. Using the explicit solution of the OU process, show that the semigroup has the explicit expression

$$P_t f(x) = \mathbb{E}\, f\big(\exp(-t)\, x + \sqrt{1-\exp(-2t)}\, \xi\big)\,, \qquad \xi \sim \mathsf{normal}(0,1)\,.$$

   Using this expression for the semigroup, compute the generator by hand and check that it agrees with the general formula obtained in Example 1.2.4.

2. Show that for the OU process, $\nabla P_t f = \exp(-t)\, P_t \nabla f$. Next, by differentiating the Dirichlet energy $t \mapsto \mathscr{E}(P_t f, P_t f)$, show that $\mathscr{E}(P_t f, P_t f) \le \exp(-2t)\, \mathscr{E}(f,f)$. Explain why this implies a Poincaré inequality for the standard Gaussian distribution.

   Hint: For a general Markov semigroup, show that $\mathrm{var}_\pi f = 2\int_0^\infty \mathscr{E}(P_t f, P_t f)\, \mathrm{d}t$ by differentiating $t \mapsto \mathrm{var}_\pi(P_t f)$.

In Chapter 2, we will generalize these calculations to prove Theorem 1.2.29.

The Geometry of Optimal Transport

▷ Exercise 1.10 (optimal transport between Gaussians)
Let $\mu_0 := \mathsf{normal}(m_0, \Sigma_0)$ and $\mu_1 := \mathsf{normal}(m_1, \Sigma_1)$; assume that $\Sigma_0 \succ 0$. Compute the optimal transport map from μ_0 to μ_1, as well as the cost $W_2(\mu_0, \mu_1)$. [By Brenier's theorem, it suffices to find the gradient of a convex function which pushes forward μ_0 to μ_1.]

Also, exhibit a coupling to prove the upper bound

$$W_2^2(\mu_0,\mu_1) \le \|m_0 - m_1\|^2 + \|\Sigma_0^{1/2} - \Sigma_1^{1/2}\|_{\mathrm{HS}}^2\,. \tag{1.E.2}$$

Finally, suppose that ν_0, ν_1 are probability measures, and suppose that μ_0, μ_1 are Gaussians whose means and covariances match those of ν_0 and ν_1 respectively. Then, prove that $W_2(\nu_0,\nu_1) \ge W_2(\mu_0,\mu_1)$.

Hint: For the last statement, use the fact that the dual potentials for optimal transport between Gaussians are quadratic functions.

▷ Exercise 1.11 (optimal transport with other costs)
In this exercise, we consider optimal transport with a general cost function c as in (1.3.2).

1. By following the proof of Theorem 1.3.8, argue that the optimal dual potentials (f,g) are c-**conjugates**, i.e., $f = g^c$ and $g = f^c$ where

$$g^c(x) := \inf_{y\in\mathcal{Y}} \{c(x,y) - g(y)\}\,, \qquad f^c(y) := \sup_{x\in\mathcal{X}} \{c(x,y) - f(x)\}\,, \tag{1.E.3}$$

and that

$$\mathcal{T}_c(\mu,\nu) \geq \sup_{f\in L^1(\mu)} \left\{ \int f \,\mathrm{d}\mu + \int f^c \,\mathrm{d}\nu \right\}.$$

Under general conditions, equality holds; see [Vil09, Theorem 5.10].

Functions of the form (1.E.3) are called c-**concave**.

2. Let $\mathcal{X} = \mathcal{Y}$ be a metric space with metric d. For all $p \geq 1$, we can define the p-**Wasserstein distance**

$$W_p^p(\mu,\nu) = \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \int \mathrm{d}(x,y)^p \,\mathrm{d}\gamma(x,y)\,.$$

Let $\mathcal{P}_p(\mathcal{X})$ denote the space of probability measures μ over $\mathcal{X}$ such that for some $x_0 \in \mathcal{X}$, $\int \mathrm{d}(x_0,\cdot)^p \,\mathrm{d}\mu < \infty$. Show that $(\mathcal{P}_p(\mathcal{X}), W_p)$ is a metric space.

3. In the case $p = 1$, show that if (f,g) are d-conjugates, then $f = -g$ and f is 1-Lipschitz. Deduce the duality formula

$$W_1(\mu,\nu) = \sup\left\{ \int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu \;\middle|\; f : \mathcal{X} \to \mathbb{R} \text{ is 1-Lipschitz} \right\}. \tag{1.E.4}$$

▷ **Exercise 1.12** (dynamical formulations of optimal transport)
The formula (1.3.23) shows that the W_2 distance between μ_0 and μ_1 equals the smallest arc length of any curve joining μ_0 and μ_1. It is also true that the squared W_2 distance minimizes the *energy* or *action* of any curve joining μ_0 and μ_1, in the following sense:

$$W_2^2(\mu_0,\mu_1) = \inf\left\{ \int_0^1 \|v_t\|^2_{L^2(\mu_t)} \,\mathrm{d}t \;\middle|\; \partial_t\mu_t + \mathrm{div}(\mu_t v_t) = 0 \right\}. \tag{1.E.5}$$

Although the problems (1.3.23) and (1.E.5) both identify geodesics in the Wasserstein space, the latter problem has more favorable properties. Namely, the minimizing curves in (1.3.23) are geodesics, but they may not have constant speed (indeed, the arc length functional is invariant under time reparameterization of the curve); in contrast, minimizing curves in (1.E.5) necessarily have constant speed. Also, we can reparameterize problem (1.E.5) by introducing the *momentum density* $p_t := \mu_t v_t$ and write

$$W_2^2(\mu_0,\mu_1) = \inf\left\{ \int_0^1 \left(\int \frac{\|p_t\|^2}{\mu_t} \right) \mathrm{d}t \;\middle|\; \partial_t\mu_t + \mathrm{div}\, p_t = 0 \right\}, \tag{1.E.6}$$

which is now a strictly convex problem in the variables (μ, p). This convenient reformulation is known as the **Benamou–Brenier formula** [BB99].

Just as (1.E.5) describes the *dynamical* version of the *static* optimal transport problem (1.3.5), there is a dynamical formulation of the dual optimal transport problem (1.3.7), in which the dual potential evolves according to the **Hamilton–Jacobi equation**

$$\partial_t u_t + \frac{1}{2} \|\nabla u_t\|^2 = 0 \,. \tag{1.E.7}$$

Then, it holds that

$$\frac{1}{2} W_2^2(\mu_0, \mu_1) = \sup\Big\{ \int u_1 \,\mathrm{d}\mu_1 - \int u_0 \,\mathrm{d}\mu_0 \;\Big|\; \partial_t u_t + \frac{1}{2} \|\nabla u_t\|^2 = 0 \Big\} \,. \tag{1.E.8}$$

The goal of this exercise is to justify and understand these facts.

1. Show that the mapping $\mathbb{R}_{>0} \times \mathbb{R}^d \to \mathbb{R}$, $(\mu, p) \mapsto \|p\|^2/\mu$ is strictly convex. Also, compute the convex conjugate of this mapping. Deduce that the Benamou–Brenier reformulation (1.E.6) is a strictly convex problem.

2. Ignoring issues of regularity, show that the solution u_t of the Hamilton–Jacobi equation with initial condition $u_0 = f$ is described by the **Hopf–Lax semigroup**

$$u_t(x) = Q_t f(x) \coloneqq \inf_{y \in \mathbb{R}^d} \Big\{ f(y) + \frac{1}{2t} \|y - x\|^2 \Big\} \,.$$

3. Following the proof of Theorem 1.3.8, show that the dual optimal transport problem (1.3.7) can be written

$$\frac{1}{2} W_2^2(\mu_0, \mu_1) = \sup_{f \in L^1(\mu_0)} \Big\{ \int Q_1 f \,\mathrm{d}\mu_1 - \int f \,\mathrm{d}\mu_0 \Big\} \,,$$

   where Q_1 denotes the Hopf–Lax semigroup at time 1. From this, deduce that the formula (1.E.8) holds.

4. Although the previous part gives a proof of the dynamical formulation (1.E.8), it is unsatisfactory because it only involves an analysis of the static primal and dual problems. Here, we present a purely dynamical proof. The continuity constraint $\partial_t \mu_t + \operatorname{div} p_t = 0$ in (1.E.6) can be reformulated as follows: for any curve of functions $[0,1] \times \mathbb{R}^d \to \mathbb{R}$, $(t, x) \mapsto u_t(x)$,

$$\int u_1 \,\mathrm{d}\mu_1 - \int u_0 \,\mathrm{d}\mu_0 = \int_0^1 \Big(\partial_t \int u_t \,\mathrm{d}\mu_t \Big) \,\mathrm{d}t = \int_0^1 \Big(\int (\partial_t u_t \, \mu_t + u_t \, \partial_t \mu_t) \Big) \,\mathrm{d}t$$

$$= \int_0^1 \Big(\int \big(\partial_t u_t + \big\langle \nabla u_t, \frac{p_t}{\mu_t} \big\rangle \big) \, \mathrm{d}\mu_t \Big) \, \mathrm{d}t \, .$$

This can be incorporated as a Lagrange multiplier in (1.E.6):

$$\frac{1}{2} W_2^2(\mu_0, \mu_1) = \inf_{\substack{\mu : [0,1] \times \mathbb{R}^d \to \mathbb{R}_+ \\ p : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d}} \sup_{u : [0,1] \times \mathbb{R}^d \to \mathbb{R}} \Big\{ \int_0^1 \int \frac{\|p_t\|^2}{2\mu_t} \, \mathrm{d}t + \int u_1 \, \mathrm{d}\mu_1 - \int u_0 \, \mathrm{d}\mu_0 - \int_0^1 \Big(\int \big(\partial_t u_t + \big\langle \nabla u_t, \frac{p_t}{\mu_t} \big\rangle \big) \, \mathrm{d}\mu_t \Big) \, \mathrm{d}t \Big\}$$

Assume that the infimum and supremum can be interchanged (here, we are invoking an abstract minimax theorem, which crucially relies on the convexity of the problem established in the first part). Use this to prove that

$$\frac{1}{2} W_2^2(\mu_0, \mu_1) = \sup \Big\{ \int u_1 \, \mathrm{d}\mu_1 - \int u_0 \, \mathrm{d}\mu_0 \;\Big|\; \partial_t u_t + \frac{1}{2} \|\nabla u_t\|^2 \le 0 \Big\}$$

and that equality holds only if the Hamilton–Jacobi equation (1.E.7) holds, and if $\nabla u_t = v_t = p_t / \mu_t$. Note that this also establishes that the optimal vector fields $(v_t)_{t \in [0,1]}$ are gradients of functions.

5. Let $t \mapsto \mu_t$ be a Wasserstein geodesic and let $t \mapsto v_t$ be its associated curve of tangent vectors. Prove that $\partial_t v_t + \nabla v_t \, v_t = 0$. (This statement follows from the previous part by differentiating the Hamilton–Jacobi equation in space. Try to also give a more direct proof of this equation.)

   Hint: If $\dot{x}_t = v_t(x_t)$, then because particles travel with constant velocity along Wasserstein geodesics, $t \mapsto \dot{x}_t$ is constant.

▷ Exercise 1.13 (Wasserstein space has non-negative curvature)
Let $t \mapsto \mu_t$ denote a Wasserstein geodesic. By finding an appropriate coupling, prove that for all $t \in [0,1]$ and all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\mu_t, \nu) \ge (1-t) \, W_2^2(\mu_0, \nu) + t \, W_2^2(\mu_1, \nu) - t \, (1-t) \, W_2^2(\mu_0, \mu_1) \, . \tag{1.E.9}$$

Compare this to the following equality on $\mathbb{R}^d$: if $x_t = (1-t) \, x_0 + t \, x_1$, then

$$\|x_t - y\|^2 = (1-t) \, \|x_0 - y\|^2 + t \, \|x_1 - y\|^2 - t \, (1-t) \, \|x_0 - x_1\|^2 \, . \tag{1.E.10}$$

The equality (1.E.10) expresses the fact that $\mathbb{R}^d$ is *flat*, whereas the inequality (1.E.9) expresses the fact that $\mathcal{P}_2(\mathbb{R}^d)$ (equipped with the W_2 metric) is *non-negatively curved*, like a sphere. See Section 2.6.2.

The Langevin SDE as a Wasserstein Gradient Flow

▷ **Exercise 1.14** (reconciling the SDE and Wasserstein perspectives)
Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a smooth test function and let $\delta > 0$. First, consider the Langevin diffusion $\mathrm{d}Z_t = -\nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$ started at $Z_0 \sim \mu_0$, and compute $\mathbb{E}\,\varphi(Z_\delta)$ up to first order in δ.

Next, if $\partial_t \mu_t = \operatorname{div}(\mu_t \nabla \ln(\mu_t/\pi))$, with $\pi \propto \exp(-V)$, then we can interpret this as a fluid flow: let $X_0 \sim \mu_0$, and $\dot{X}_t = -\nabla \ln(\mu_t/\pi)(X_t)$, so that $X_t \sim \mu_t$. Compute the quantity $\mathbb{E}\,\varphi(X_\delta)$ up to first order in δ.

Check that the two expressions you computed match (up to first order in δ). Note that these calculations are implicit in §1.2 and §1.4, but it is illuminating to directly connect the Langevin diffusion to the Wasserstein gradient flow.

▷ **Exercise 1.15** (Wasserstein calculus for f-divergences)
Compute the Wasserstein gradient of the functional $\chi^2(\cdot \,\|\, \pi)$. Use the rules of Wasserstein calculus to compute $\partial_t \chi^2(\pi_t \,\|\, \pi)$, where $t \mapsto \pi_t$ is the law of the Langevin diffusion with stationary distribution π. Check that the result agrees with a calculation based on Markov semigroup theory.

More generally, let $f : \mathbb{R} \to \mathbb{R}_+$ and consider the f-divergence

$$D_f(\mu \,\|\, \pi) := \int f\Big(\frac{\mu}{\pi}\Big)\,\mathrm{d}\pi\,.$$

Compute the Wasserstein gradient of $D_f(\cdot \,\|\, \pi)$. For bonus points, calculate the Wasserstein Hessian as well.

▷ **Exercise 1.16** (Otto–Villani theorem)
Consider the gradient flow $t \mapsto \mu_t$ of a functional $\mathcal{F}$ with $\inf \mathcal{F} = 0$. Assume that the PL inequality $\|\nabla_{W_2}\mathcal{F}(\mu)\|_\mu^2 \geq 2\alpha\,\mathcal{F}(\mu)$ holds, and that the gradient flow converges to the minimizer of $\mathcal{F}$. Argue that $\partial_t W_2(\mu_t, \mu_0) \leq \|\nabla_{W_2}\mathcal{F}(\mu_t)\|_{\mu_t}$, and then show that

$$\partial_t\Big(\sqrt{\frac{\alpha}{2}}\,W_2(\mu_t, \mu_0) + \sqrt{\mathcal{F}(\mu_t)}\Big) \leq 0\,.$$

Conclude that a quadratic growth inequality holds.

▷ **Exercise 1.17** (contraction of the Langevin diffusion)
In this exercise, we explore different proofs of contraction.

1. Suppose that $V : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex. Let $t \mapsto x_t$, $t \mapsto y_t$ be two gradient flows for V. Show that $\|x_t - y_t\|^2 \leq \exp(-2\alpha t)\,\|x_0 - y_0\|^2$.

2. Next, prove Theorem 1.4.10. In fact, show that we have the almost sure contraction $\|Z_t - Z_t'\| \le \exp(-\alpha t)\,\|Z_0 - Z_0'\|$.

   Hint: Apply Itô's formula (Theorem 1.1.18) to $f(z, z') := \|z - z'\|^2$.

3. Let $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$ be an α-convex functional, and let $(\mu_t)_{t\ge 0}$, $(\nu_t)_{t\ge 0}$ be gradient flows for $\mathcal{F}$. Prove that

$$W_2^2(\mu_t, \nu_t) \le \exp(-2\alpha t)\, W_2^2(\mu_0, \nu_0)$$

   using the following steps. First, compute the derivative of $t \mapsto W_2^2(\mu_t, \nu_t)$ using Theorem 1.4.11. Next, apply the strong convexity inequality (1.4.6) to obtain two inequalities $\mathcal{F}(\mu_t) \ge \mathcal{F}(\nu_t) + \cdots$ and $\mathcal{F}(\nu_t) \ge \mathcal{F}(\mu_t) + \cdots$. Adding these two inequalities, deduce that $\partial_t W_2^2(\mu_t, \nu_t) \le -2\alpha\, W_2^2(\mu_t, \nu_t)$.

▷ Exercise 1.18 (smoothness along Wasserstein geodesics)
For a functional $\mathcal{F}$ over the Wasserstein space, let us say that it is β-smooth if, for all constant-speed geodesics $(\mu_t)_{t\in[0,1]}$ with initial tangent vector $v_0 = T - \mathrm{id}$, it holds that $\partial_t^2|_{t=0}\mathcal{F}(\mu_t) \le \beta\,\|T - \mathrm{id}\|_{\mu_0}^2$.

1. Show that the potential energy functional $\mathcal{E}$ corresponding to a potential V with $\nabla^2 V \preceq \beta I_d$ is β-smooth.

2. Establish the expression (1.4.3) for the entropy $\mathcal{H}$ and argue that $\mathcal{H}$ is non-smooth.

Overview of the Convergence Results

▷ Exercise 1.19 (divergences at initialization)
Let $\pi \propto \exp(-V)$ where $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Assume in addition that V is minimized at 0. Let $\kappa := \beta/\alpha$ denote the condition number. Show that if we initialize at the measure $\mu_0 = \mathrm{normal}(0, \beta^{-1} I_d)$, then $\ln \sup(\mu_0/\pi) \le \frac{d}{2} \ln \kappa$. What does this imply about the size of $\mathsf{KL}(\mu_0 \parallel \pi)$ and $\chi^2(\mu_0 \parallel \pi)$ at initialization?

What can you say about $W_2^2(\mu_0, \pi)$?

CHAPTER 2

Functional Inequalities

In this chapter, we explore the connection between functional inequalities, such as the Poincaré and log-Sobolev inequalities, and the concentration of measure phenomenon.

2.1 Overview of the Inequalities

2.1.1 Relationships between the Inequalities

The main inequalities that we study in this chapter are the following:

- the **Poincaré inequality** (**PI**), as specialized to the Langevin diffusion (see Example 1.2.22):

$$\mathrm{var}_\pi(f) \le C_{\mathsf{PI}}\, \mathbb{E}_\pi[\|\nabla f\|^2]\,, \qquad \text{for all smooth } f : \mathbb{R}^d \to \mathbb{R}\,,$$

- the **log-Sobolev inequality** (**LSI**), as specialized to the Langevin diffusion (see Example 1.2.26):

$$\mathrm{ent}_\pi(f^2) \le 2C_{\mathsf{LSI}}\, \mathbb{E}_\pi[\|\nabla f\|^2]\,, \qquad \text{for all smooth } f : \mathbb{R}^d \to \mathbb{R}\,,$$

- and **Talagrand's T_2 inequality**

$$\mathsf{KL}(\mu \,\|\, \pi) \ge \frac{1}{2C_{\mathsf{T}_2}}\, W_2^2(\mu, \pi)\,, \qquad \text{for all } \mu \in \mathcal{P}_2(\mathbb{R}^d)\,.$$

In addition, using the W_1 metric introduced in Exercise 1.11, we consider

- **Talagrand's T_1 inequality**

$$\mathsf{KL}(\mu \,\|\, \pi) \geq \frac{1}{2C_{\mathsf{T}_1}} W_1^2(\mu, \pi)\,, \qquad \text{for all } \mu \in \mathcal{P}_1(\mathbb{R}^d)\,.$$

In many cases, arguments involving Poincaré and log-Sobolev inequalities hold more generally in the context of reversible Markov processes, and when this is case we will try to use notation from Markov semigroup theory (e.g., writing $\mathbb{E}_\pi \Gamma(f,f)$ or $\mathscr{E}(f,f)$ instead of $\mathbb{E}_\pi[\|\nabla f\|^2]$) to indicate that this is the case. However, for clarity of exposition, we do not dwell on this point, and we urge readers to focus on the case in which the Markov process is the Langevin diffusion.

Although the Poincaré and log-Sobolev inequalities are stated above for smooth functions $f : \mathbb{R}^d \to \mathbb{R}$, once established they can be extended to a wider class of functions (e.g., locally Lipschitz functions) by arguing that smooth functions are dense w.r.t. appropriate norms. Throughout the chapter, we omit mention of such approximation arguments.

Write $\mathsf{PI}(C)$ to denote that the Poincaré inequality holds with constant C, and similarly for the other inequalities. We have the following relationships.

- The Bakry–Émery theorem (Theorem 1.2.29) shows that α-strong log-concavity of π implies that π satisfies $\mathsf{LSI}(\alpha^{-1})$.
- The Otto–Villani theorem (Exercise 1.16) shows that $\mathsf{LSI}(C)$ implies $\mathsf{T}_2(C)$.
- Since $W_1 \leq W_2$, then $\mathsf{T}_2(C)$ obviously implies $\mathsf{T}_1(C)$. On the other hand, we will show below that $\mathsf{T}_2(C)$ implies $\mathsf{PI}(C)$ as well. Combined with the previous point, this shows that $\mathsf{LSI}(C)$ implies $\mathsf{PI}(C)$, which was shown directly in Exercise 1.8.
- In general, PI and T_1 are incomparable (Exercise 2.14).

2.1.2 Linearization of Transport Inequalities

To prove that $\mathsf{T}_2(C)$ implies $\mathsf{PI}(C)$, we linearize the transport cost. It will be convenient for future purposes to prove a more general version of the linearization principle.

Proposition 2.1.1 (linearization of transport cost). *Let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a lower semicontinuous cost function. Assume that $c(x,x) = 0$ for all $x \in \mathbb{R}^d$, that there exists $\delta > 0$ for which $c(x,y) \geq \delta \|x-y\|^2$ for all $x, y \in \mathbb{R}^d$, and that there is a measurable*

mapping $x \mapsto H_x \succ 0$ such that for each compact $K \subseteq \mathbb{R}^d$,

$$\sup_{x \in K} \left| c(x+h, x) - \frac{1}{2} \langle h, H_x h \rangle \right| = o(\|h\|^2) \qquad \text{as } h \to 0 \,.$$

Then, for any $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $f \in C_c^\infty(\mathbb{R}^d)$ with $\int f \, \mathrm{d}\mu = 0$, it holds that

$$\liminf_{\varepsilon \searrow 0} \frac{1}{\varepsilon^2} \mathcal{T}_c\big(\mu, \ (1+\varepsilon f)\, \mu\big) \geq \frac{\left(\int f^2 \, \mathrm{d}\mu\right)^2}{2 \int \langle \nabla f(x), H_x^{-1} \nabla f(x) \rangle \, \mathrm{d}\mu(x)} \,.$$

Proof sketch. Fix $\lambda \in \mathbb{R}$. Using the dual formulation given in Exercise 1.11,

$$\mathcal{T}_c(\mu, \nu) \geq \int \lambda \varepsilon f \, \mathrm{d}\mu + \int (\lambda \varepsilon f)^c \, \mathrm{d}\nu = \int (\lambda \varepsilon f)^c \, \mathrm{d}\nu \,.$$

Here, $(\lambda \varepsilon f)^c(x) = \inf_{h \in \mathbb{R}^d} \{c(x+h, x) - \lambda \varepsilon f(x+h)\}$, and using the assumption on c together with the compact support of f, one can justify that the infimum is attained at a point h with $\|h\| = O(\varepsilon)$. Then,

$$\begin{aligned}(\lambda \varepsilon f)^c(x) &= \inf_{h \in \mathbb{R}^d} \Big\{ \frac{1}{2} \langle h, H_x h \rangle - \lambda \varepsilon f(x) - \lambda \varepsilon \langle \nabla f(x), h \rangle \Big\} + o(\varepsilon^2) \\ &\geq -\lambda \varepsilon f(x) - \frac{\lambda^2 \varepsilon^2}{2} \langle \nabla f(x), H_x^{-1} \nabla f(x) \rangle + o(\varepsilon^2) \,.\end{aligned}$$

Hence,

$$\mathcal{T}_c\big(\mu, \ (1+\varepsilon f)\, \mu\big) \geq -\lambda \varepsilon^2 \int f^2 \, \mathrm{d}\mu - \frac{\lambda^2 \varepsilon^2}{2} \int \langle \nabla f(x), H_x^{-1} \nabla f(x) \rangle \, \mathrm{d}\mu(x)$$

and the result follows by optimizing over λ. □

Corollary 2.1.2 (T_2 implies PI)**.** *If π satisfies $\mathsf{T}_2(C)$, then it satisfies $\mathsf{PI}(C)$.*

Proof. Let $f \in C_c^\infty(\mathbb{R}^d)$ and apply the linearization in the preceding proposition to the quadratic cost $c(x, y) = \frac{1}{2} \|x - y\|^2$ with $H_x = I_d$ for all $x \in \mathbb{R}^d$. Then, $\mathsf{T}_2(C)$ yields

$$2C \, \mathsf{KL}\big((1+\varepsilon f)\, \pi \,\big\|\, \pi\big) \geq W_2^2\big(\pi, \ (1+\varepsilon f)\, \pi\big) \geq \frac{\varepsilon^2 \left(\int f^2 \, \mathrm{d}\pi\right)^2}{\int \|\nabla f\|^2 \, \mathrm{d}\pi} + o(\varepsilon^2) \,.$$

On the other hand, the linearization (1.E.1) of the KL divergence in Exercise 1.8 yields

$$\mathsf{KL}\big((1+\varepsilon f)\,\pi \,\big\|\, \pi\big) = \frac{\varepsilon^2}{2}\int f^2 \,\mathrm{d}\pi + o(\varepsilon^2)\,.$$

Comparing terms proves the result. □

In Exercise 2.1, we explore a perhaps more intuitive approach to linearizing the 2-Wasserstein distance via the Monge–Ampère equation.

2.2 Proofs via Markov Semigroup Theory

2.2.1 Commutation and Curvature

In Section 1.2.3, we introduced the iterated carré du champ operator Γ_2, as well as the curvature-dimension condition $\Gamma_2(f,f) \geq \alpha\,\Gamma(f,f)$ (denoted $\mathrm{CD}(\alpha,\infty)$). Since this condition plays a key role in the subsequent calculations, our goal is to demystify this idea.

By definition, the iterated carré du champ is

$$\Gamma_2(f,f) = \frac{1}{2}\left\{\mathscr{L}\Gamma(f,f) - 2\,\Gamma(f,\mathscr{L}f)\right\}.$$

For the case of the Langevin diffusion with carré du champ $\Gamma(f,f) = \|\nabla f\|^2$,

$$\Gamma_2(f,f) = \frac{1}{2}\left\{\mathscr{L}(\|\nabla f\|^2) - 2\,\langle \nabla f, \nabla \mathscr{L} f\rangle\right\}. \tag{2.2.1}$$

Recall that $\mathscr{L}f = \Delta f - \langle \nabla V, \nabla f\rangle$, where V is the potential. Let us begin with the simple case in which $V = 0$, so $\mathscr{L}$ is the Laplacian Δ (the generator of $\sqrt{2}\,B$, where B is standard Brownian motion). In this case, the iterated carré du champ turns out to simply be the operator $\Gamma_2(f,f) = \|\nabla^2 f\|_{\mathrm{HS}}^2$, which is known as the **Bochner identity**:

$$\frac{1}{2}\,\Delta(\|\nabla f\|^2) = \langle \nabla \Delta f, \nabla f\rangle + \|\nabla^2 f\|_{\mathrm{HS}}^2\,. \tag{2.2.2}$$

Consequently, Δ satisfies $\mathrm{CD}(0,\infty)$.

It may seem strange at first sight to give such a fancy name to the seemingly innocuous identity (2.2.2), which is a simple exercise in calculus. However, the importance of the Bochner identity begins to reveal itself through the following fact: the identity continues to make sense on a Riemannian manifold, except that there is an extra term involving the *Ricci curvature* of the manifold.

$$\frac{1}{2}\,\Delta(\|\nabla f\|^2) = \langle \nabla \Delta f, \nabla f\rangle + \|\nabla^2 f\|_{\mathrm{HS}}^2 + \mathsf{Ric}(\nabla f, \nabla f)\,.$$

We will defer a fuller discussion of Riemannian geometry for later, but for now we can get a hint at the role of the curvature by observing that the Bochner identity (2.2.2) on $\mathbb{R}^d$ follows from the equation[1]

$$\nabla \Delta f - \Delta \nabla f = 0 \tag{2.2.3}$$

by taking the inner product with ∇f and applying the identity

$$\frac{1}{2}\,\Delta(\|\nabla f\|^2) = \operatorname{div}(\nabla^2 f\, \nabla f) = \langle \Delta \nabla f, \nabla f\rangle + \|\nabla^2 f\|_{\mathrm{HS}}^2\,.$$

In turn, the equation (2.2.3) shows that the Laplacian commutes with the gradient operator, which is true because partial derivatives commute on $\mathbb{R}^d$; this is a manifestation of the fact that $\mathbb{R}^d$ is *flat*. In contrast, the very definition of curvature on a Riemannian manifold is usually based upon the *lack* of commutativity of differential operators.[2]

Turning now to the Langevin generator $\mathscr{L}$, the identity (2.2.3) is replaced by

$$\nabla \mathscr{L} f - \mathscr{L} \nabla f = -\nabla^2 V\, \nabla f\,. \tag{2.2.4}$$

Hence, the commutator of ∇ and $\mathscr{L}$ brings out the curvature of the measure $\pi \propto \exp(-V)$, and the plan is to exploit this in order to prove functional inequalities. The identity (2.2.4) then yields the following formula for the iterated carré du champ:

$$\Gamma_2(f,f) = \|\nabla^2 f\|_{\mathrm{HS}}^2 + \langle \nabla f, \nabla^2 V\, \nabla f\rangle\,. \tag{2.2.5}$$

In particular, if $\nabla^2 V \geq \alpha I_d > 0$, then the curvature-dimension condition $\mathrm{CD}(\alpha,\infty)$ holds, which was asserted as Theorem 1.2.30.

2.2.2 The Brascamp–Lieb Inequality

As a first illustration of the use of curvature, we prove the Brascamp–Lieb inequality, which is a strong form of the Poincaré inequality. This inequality will also gain a natural interpretation via a diffusion process in Section 10.2.

The proof method in this section is known as **Hörmander's L^2 method**. The starting point is to write down a dual form of the Poincaré inequality.[3]

[1] Here, Δ acts on ∇u component by component.

[2] Loosely speaking, the idea of curvature is that travelling in direction u and then direction v is not exactly the same as travelling in direction v and direction u. Algebraically, this is captured by studying the difference between differentiating along vector field X and then vector field Y, or vice versa.

[3] The idea of dualizing the Poincaré inequality also appears in Exercise 2.1, in which the Poincaré inequality is deduced from an inequality on $(-\mathscr{L})^{-1}$.

Lemma 2.2.6. *Let $\pi \propto \exp(-V)$ be a probability measure on $\mathbb{R}^d$, where V is continuously differentiable; let $\mathscr{L}$ be the corresponding Langevin generator. Suppose that $A : \mathbb{R}^d \to \mathrm{PD}(d)$ is a matrix-valued function mapping into the space of symmetric positive definite matrices such that for all smooth $u : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}_\pi[(\mathscr{L}u)^2] \geq \mathbb{E}_\pi\langle \nabla u, A\,\nabla u\rangle\,. \tag{2.2.7}$$

Then, for all smooth $f : \mathbb{R}^d \to \mathbb{R}$ it holds that

$$\mathrm{var}_\pi f \leq \mathbb{E}_\pi\langle \nabla f, A^{-1}\,\nabla f\rangle\,.$$

Proof. Assume that $\mathbb{E}_\pi f = 0$. Recall that $\mathbb{E}_\pi \mathscr{L}u = 0$ for any u, so that $\mathbb{E}_\pi f = 0$ is a *necessary* condition for the solvability of the Poisson equation $-\mathscr{L}u = f$. For simplicity, we will assume that this condition is also sufficient.

If we express $\mathbb{E}_\pi[f^2]$ terms of u and apply integration by parts (Theorem 1.2.14) and Cauchy–Schwarz, we obtain

$$\begin{aligned}\mathbb{E}_\pi[f^2] &= 2\,\mathbb{E}_\pi[f\,(-\mathscr{L})\,u] - \mathbb{E}_\pi[(\mathscr{L}u)^2] \\ &\leq 2\,\mathbb{E}_\pi\langle \nabla f, \nabla u\rangle - \mathbb{E}\langle \nabla u, A\,\nabla u\rangle \\ &\leq 2\sqrt{\mathbb{E}_\pi\langle \nabla f, A^{-1}\,\nabla f\rangle\,\mathbb{E}_\pi\langle \nabla u, A\,\nabla u\rangle} - \mathbb{E}\langle \nabla u, A\,\nabla u\rangle \leq \mathbb{E}_\pi\langle \nabla f, A^{-1}\,\nabla f\rangle\,.\end{aligned}$$ □

The point is that the condition (2.2.7) can now be checked with the help of curvature. Suppose that $\pi \propto \exp(-V)$ where V is twice continuously differentiable and strictly convex. Then, using integration by parts (Theorem 1.2.14),

$$\mathbb{E}_\pi[(\mathscr{L}u)^2] = -\,\mathbb{E}_\pi\langle \nabla u, \nabla \mathscr{L}u\rangle = \underbrace{\mathbb{E}_\pi\Big[\Gamma_2(u,u) - \frac{1}{2}\,\mathscr{L}(\|\nabla u\|^2)\Big]}_{\text{by (2.2.1)}} = \underbrace{\mathbb{E}_\pi\,\Gamma_2(u,u)}_{\text{because } \mathbb{E}_\pi \mathscr{L}=0}$$
$$= \underbrace{\mathbb{E}_\pi[\|\nabla^2 u\|_{\mathrm{HS}}^2 + \langle \nabla u, \nabla^2 V\,\nabla u\rangle]}_{\text{by (2.2.5)}}\,.$$

Applying the lemma, we obtain the following result.

Theorem 2.2.8 (Brascamp–Lieb inequality)**.** *Let $\pi \propto \exp(-V)$, where V is strictly*

convex on $\mathbb{R}^d$ and twice continuously differentiable. Then, for all $f : \mathbb{R}^d \to \mathbb{R}$,

$$\operatorname{var}_\pi f \le \mathbb{E}_\pi \langle \nabla f, (\nabla^2 V)^{-1} \nabla f \rangle .$$

When $\nabla^2 V \succeq \alpha I_d \succ 0$, then this implies that a Poincaré inequality holds for π with constant $C_{\mathrm{PI}} \le 1/\alpha$. However, the Brascamp–Lieb inequality is much stronger, as it allows us to take advantage of non-uniform convexity.

In Exercise 2.3, we give another proof of Theorem 2.2.8 by linearizing a transport inequality. First, we introduce the transport cost.

Definition 2.2.9. The **Bregman transport cost** for the potential V, denoted $\mathcal{D}_V$, is the transport cost associated with the **Bregman divergence**

$$D_V(x, y) := V(x) - V(y) - \langle \nabla V(y), x - y \rangle ,$$

i.e., we set

$$\mathcal{D}_V(\mu, \nu) := \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int D_V(x, y) \, \mathrm{d}\gamma(x, y) .$$

The Bregman transport cost will also play a key role in Section 10.2, in which we will prove the following transport inequality.

Theorem 2.2.10 (Bregman transport inequality)**.** *Suppose that $V : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. Then, for $\pi \propto \exp(-V)$ and all $\mu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\mathcal{D}_V(\mu, \pi) \le \mathsf{KL}(\mu \,\|\, \pi) .$$

Actually, convexity of V is not necessary for the theorem to hold, although the Bregman transport cost $\mathcal{D}_V$ is only guaranteed to be non-negative when V is convex. Notice that when V is strongly convex, $\nabla^2 V \succeq \alpha I_d \succ 0$, then $D_V(x, y) \ge \frac{\alpha}{2} \|x - y\|^2$, so the Bregman transport inequality implies $\mathsf{T}_2(\alpha^{-1})$.

2.2.3 Proof of the Bakry–Émery Theorem

In this section, we generalize the calculation in Exercise 1.9 from the Ornstein–Uhlenbeck diffusion to the Langevin diffusion, and thereby prove the Bakry–Émery theorem (Theorem 1.2.29). Recall from that exercise that the Ornstein–Uhlenbeck semigroup satisfies the

identity $\nabla P_t f = \exp(-t)\, P_t \nabla f$. In the next result, we show that more generally, $\mathrm{CD}(\alpha, \infty)$ implies $\Gamma(P_t f, P_t f) \leq \exp(-2\alpha t)\, P_t \Gamma(f, f)$.

Theorem 2.2.11 (local Poincaré inequality). *Assume the Markov semigroup $(P_t)_{t\geq 0}$ is reversible and let $\alpha \in \mathbb{R}$. Then, the following are equivalent.*

1. *The curvature-dimension condition $\mathrm{CD}(\alpha, \infty)$ holds.*
2. *For all f and $t \geq 0$,*
$$\Gamma(P_t f, P_t f) \leq \exp(-2\alpha t)\, P_t \Gamma(f, f)\,.$$
3. *For all f and $t \geq 0$,*
$$P_t(f^2) - (P_t f)^2 \leq \frac{1 - \exp(-2\alpha t)}{\alpha}\, P_t \Gamma(f, f)\,.$$

Proof. (2) $\implies$ (3): Markov semigroup calculus yields

$$\partial_s [P_s((P_{t-s} f)^2)] = P_s\big(\mathscr{L}((P_{t-s} f)^2) - 2 P_{t-s} f\, \mathscr{L} P_{t-s} f\big)\,.$$

On the other hand, recall the definition of the carré du champ: $\mathscr{L}(f^2) - 2f\,\mathscr{L} f = 2\Gamma(f, f)$. Using this along with (2),

$$\partial_s [P_s((P_{t-s} f)^2)] = 2 P_s \Gamma(P_{t-s} f, P_{t-s} f) \leq 2 \exp\big(-2\alpha\,(t-s)\big)\, P_t \Gamma(f, f)\,.$$

Integrating this from $s = 0$ to $s = t$ yields (3).

(1) $\implies$ (2): Similarly, differentiating

$$\partial_s [P_s \Gamma(P_{t-s} f, P_{t-s} f)] = P_s\big(\mathscr{L}\Gamma(P_{t-s} f, P_{t-s} f) - 2\Gamma(P_{t-s} f, \mathscr{L} P_{t-s} f)\big)$$

and applying the definition of the iterated carré du champ yields

$$\partial_s [P_s \Gamma(P_{t-s} f, P_{t-s} f)] = 2 P_s \Gamma_2(P_{t-s} f, P_{t-s} f) \geq 2\alpha\, P_s \Gamma(P_{t-s} f, P_{t-s} f)\,.$$

Integrating this from $s = 0$ to $s = t$ yields $P_t \Gamma(f, f) \geq \exp(2\alpha t)\, \Gamma(P_t f, P_t f)$.

(3) $\implies$ (1): We leave this as Exercise 2.6. □

Observe that if we take expectations of both sides of the third statement above w.r.t. π and send $t \to \infty$, then we see that $\mathrm{CD}(\alpha, \infty)$ for $\alpha > 0$ implies $\mathrm{PI}(\alpha^{-1})$. However, the local decay asserted above is much stronger than $\mathrm{PI}(\alpha^{-1})$.

To proceed further, we introduce the notion of a diffusion semigroup.

Definition 2.2.12. The Markov semigroup $(P_t)_{t\geq 0}$ is a **diffusion semigroup** if for all functions $f, g \in L^2(\pi)$ in the domain of the carré du champ Γ and all $\phi : \mathbb{R} \to \mathbb{R}$, the chain rule holds:

$$\Gamma(\phi \circ f, g) = \phi'(f)\,\Gamma(f, g)\,.$$

More generally, for functions $f_1, \dots, f_k$ and $\Psi : \mathbb{R}^k \to \mathbb{R}$,

$$\Gamma\big(\Psi(f_1, \dots, f_k), g\big) = \sum_{i=1}^{k} (\partial_i \Psi)(f_1, \dots, f_k)\,\Gamma(f_i, g)\,.$$

The chain rule is satisfied for the Langevin diffusion whose carré du champ is given by $\Gamma(f, g) = \langle \nabla f, \nabla g \rangle$, and more generally this assumption encodes the fact that the Markov process is a diffusion. Since we are mainly interested in diffusion processes, this is not a restrictive assumption, but it indicates that the following proof will fail for Markov processes on discrete state spaces.

Proof of the Bakry–Émery theorem (Theorem 1.2.29). Given a smooth positive function f and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$, we differentiate $t \mapsto \int \phi(P_t f)\,\mathrm{d}\pi$. We are primarily interested in the case $\phi(x) := x \ln x$, but carrying out the calculation for a general ϕ clarifies the structure of the argument. Using the Markov semigroup calculus,

$$\partial_t \int \phi(P_t f)\,\mathrm{d}\pi = \int \phi'(P_t f)\,\mathscr{L} P_t f\,\mathrm{d}\pi = -\mathscr{E}(\phi' \circ P_t f, P_t f)\,.$$

This yields the representation

$$\begin{aligned} \mathrm{ent}_\pi^\phi f &:= \int \phi(f)\,\mathrm{d}\pi - \phi\Big(\int f\,\mathrm{d}\pi\Big) = -\int_0^\infty \Big(\partial_t \int \phi(P_t f)\,\mathrm{d}\pi\Big)\,\mathrm{d}t \\ &= \int_0^\infty \mathscr{E}(\phi' \circ P_t f, P_t f)\,\mathrm{d}t\,. \end{aligned}$$

We now specialize our calculations to the entropy function $\phi(x) = x \ln x$ and use reversibility of the semigroup.

$$\begin{aligned} \mathscr{E}(\phi' \circ P_t f, P_t f) &= \int (\ln P_t f)\,(-\mathscr{L}) P_t f\,\mathrm{d}\pi = \int (\ln P_t f)\,P_t(-\mathscr{L} f)\,\mathrm{d}\pi \\ &= \int P_t \ln P_t f\,(-\mathscr{L}) f\,\mathrm{d}\pi = \int \Gamma(P_t \ln P_t f, f)\,\mathrm{d}\pi \end{aligned}$$

$$\leq \sqrt{\int \frac{\Gamma(f,f)}{f}\,\mathrm{d}\pi \int f\,\Gamma(P_t \ln P_t f, P_t \ln P_t f)\,\mathrm{d}\pi}$$

where the last line uses the Cauchy–Schwarz inequality (Exercise 1.6). By the chain rule for the carré du champ, we have

$$\Gamma(\ln f, f) = \frac{\Gamma(f,f)}{f}\,.$$

By applying the local Poincaré inequality (Theorem 2.2.11) and the chain rule,

$$\begin{aligned}
\mathscr{E}(\ln P_t f, P_t f) &\leq \exp(-\alpha t)\sqrt{\int \Gamma(\ln f, f)\,\mathrm{d}\pi \int f\,P_t\Gamma(\ln P_t f, \ln P_t f)\,\mathrm{d}\pi} \\
&= \exp(-\alpha t)\sqrt{\int \Gamma(\ln f, f)\,\mathrm{d}\pi \int P_t f\,\Gamma(\ln P_t f, \ln P_t f)\,\mathrm{d}\pi} \\
&= \exp(-\alpha t)\sqrt{\int \Gamma(\ln f, f)\,\mathrm{d}\pi \int \Gamma(\ln P_t f, P_t f)\,\mathrm{d}\pi}
\end{aligned}$$

which is rearranged to yield

$$\mathscr{E}(\ln P_t f, f) \leq \exp(-2\alpha t)\,\mathscr{E}(\ln f, f)\,.$$

This shows that under $\mathsf{CD}(\alpha,\infty)$, the Fisher information (introduced in Example 1.2.26) decays exponentially fast.

Substituting this into the representation above,

$$\mathrm{ent}_\pi f = \int_0^\infty \mathscr{E}(\ln P_t f, P_t f)\,\mathrm{d}t \leq \mathscr{E}(\ln f, f)\int_0^\infty \exp(-2\alpha t)\,\mathrm{d}t \leq \frac{1}{2\alpha}\,\mathscr{E}(\ln f, f)\,,$$

which is the log-Sobolev inequality. □

2.2.4 Convergence in Rényi Divergence

One curiosity is that the log-Sobolev inequality directly implies a Poincaré inequality (Exercise 1.8), and yet the convergence guarantees implied by these inequalities for the Langevin diffusion are incomparable, because they apply to different metrics (χ^2 vs. KL). It turns out that these convergence guarantees can be placed in the same framework by introducing the family of *Rényi divergences*. Rényi divergences have also gained importance in recent research due to applications to differential privacy [Mir17].

Definition 2.2.13. For $q > 1$, the **Rényi divergence** of order q between μ and π is defined by

$$\mathcal{R}_q(\mu \,\|\, \pi) := \frac{1}{q-1} \ln \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^q \mathrm{d}\pi \tag{2.2.14}$$

if $\mu \ll \pi$, and $\mathcal{R}_q(\mu \,\|\, \pi) := +\infty$ otherwise.

Rényi divergences are monotonic in the order: if $1 < q \le q' < \infty$, then $\mathcal{R}_q \le \mathcal{R}_{q'}$ (this follows from Jensen's inequality). Some notable special cases include:

1. For $q \searrow 1$, we have $\mathcal{R}_q \to \mathsf{KL}$.
2. For $q = 2$, we have $\mathcal{R}_q = \ln(1 + \chi^2)$.
3. For $q \nearrow \infty$, we have $\mathcal{R}_q \nearrow \mathcal{R}_\infty$, where $\mathcal{R}_\infty(\mu \,\|\, \pi) := \ln \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|_{L^\infty(\pi)}$.

Remarkably, Vempala and Wibisono [VW19] show that a Poincaré inequality or a log-Sobolev inequality imply convergence of the Langevin diffusion in every Rényi divergence. We will prove the following theorem.

Theorem 2.2.15 ([VW19])**.** *Let $(P_t)_{t\ge 0}$ be a reversible diffusion Markov semigroup, and let $(\pi_t)_{t\ge 0}$ denote the law of the Markov process associated with the semigroup.*

1. *Suppose that a log-Sobolev inequality holds with constant C_{LSI}. Then, for all $q \ge 1$,*

$$\mathcal{R}_q(\pi_t \,\|\, \pi) \le \exp\left(-\frac{2t}{qC_{\mathsf{LSI}}}\right) \mathcal{R}_q(\pi_0 \,\|\, \pi)\,.$$

2. *Suppose that a Poincaré inequality holds with constant C_{PI}. Then, for all $q \ge 2$,*

$$\mathcal{R}_q(\pi_t \,\|\, \pi) \le \begin{cases} \mathcal{R}_q(\pi_0 \,\|\, \pi) - \dfrac{2t}{qC_{\mathsf{PI}}}\,, & \text{if } \mathcal{R}_q(\pi_t \,\|\, \pi) \ge 1\,, \\ \exp\left(-\dfrac{2t}{qC_{\mathsf{PI}}}\right) \mathcal{R}_q(\pi_0 \,\|\, \pi)\,, & \text{if } \mathcal{R}_q(\pi_0 \,\|\, \pi) \le 1\,. \end{cases}$$

Proof. We begin by differentiating the Rényi divergence in time. Let $\rho_t := \frac{\mathrm{d}\pi_t}{\mathrm{d}\pi} = P_t\rho_0$. Applying the chain rule for the carré du champ,

$$\partial_t \mathcal{R}_q(\pi_t \,\|\, \pi) = \frac{1}{q-1} \frac{\partial_t \int \rho_t^q \,\mathrm{d}\pi}{\int \rho_t^q \,\mathrm{d}\pi} = \frac{q}{q-1} \frac{\int \rho_t^{q-1} \mathscr{L}\rho_t \,\mathrm{d}\pi}{\int \rho_t^q \,\mathrm{d}\pi} = -\frac{q}{q-1} \frac{\int \Gamma(\rho_t^{q-1}, \rho_t) \,\mathrm{d}\pi}{\int \rho_t^q \,\mathrm{d}\pi}$$

$$= -\frac{4}{q} \frac{\mathscr{E}(\rho_t^{q/2}, \rho_t^{q/2})}{\int \rho_t^q \, \mathrm{d}\pi} \,.$$

Log-Sobolev case. The log-Sobolev inequality reads (due to the chain rule) as

$$2C_{\mathsf{LSI}} \, \mathscr{E}(f, f) \geq \mathrm{ent}_\pi(f^2) \,.$$

Applying this to $f = \rho^{q/2}$, we obtain

$$\begin{aligned} 2C_{\mathsf{LSI}} \, \mathscr{E}(\rho^{q/2}, \rho^{q/2}) &\geq q \int \rho^q \ln \rho \, \mathrm{d}\pi - \Big(\int \rho^q \, \mathrm{d}\pi \Big) \ln \Big(\int \rho^q \, \mathrm{d}\pi \Big) \\ &= q \, \partial_q \int \rho^q \, \mathrm{d}\pi - \Big(\int \rho^q \, \mathrm{d}\pi \Big) \ln \Big(\int \rho^q \, \mathrm{d}\pi \Big) \end{aligned}$$

and hence

$$\begin{aligned} \frac{4}{q} \frac{\mathscr{E}(\rho^{q/2}, \rho^{q/2})}{\int \rho^q \, \mathrm{d}\pi} &\geq \frac{2}{C_{\mathsf{LSI}}} \, \partial_q \ln \int \rho^q \, \mathrm{d}\pi - \frac{2}{q C_{\mathsf{LSI}}} \ln \int \rho^q \, \mathrm{d}\pi \\ &= \frac{2}{C_{\mathsf{LSI}}} \, \partial_q [(q-1) \, \mathcal{R}_q(\rho\pi \parallel \pi)] - \frac{2\,(q-1)}{q C_{\mathsf{LSI}}} \, \mathcal{R}_q(\rho\pi \parallel \pi) \\ &= \frac{2}{C_{\mathsf{LSI}}} \, \mathcal{R}_q(\rho\pi \parallel \pi) + \frac{2\,(q-1)}{C_{\mathsf{LSI}}} \underbrace{\partial_q \mathcal{R}_q(\rho\pi \parallel \pi)}_{\geq 0} - \frac{2\,(q-1)}{q C_{\mathsf{LSI}}} \, \mathcal{R}_q(\rho\pi \parallel \pi) \\ &\geq \frac{2}{q C_{\mathsf{LSI}}} \, \mathcal{R}_q(\rho\pi \parallel \pi) \end{aligned}$$

where we used the fact that the Rényi divergence is monotonic in the order.

Poincaré case. Next, applying a Poincaré inequality to $f = \rho^{q/2}$,

$$\begin{aligned} C_{\mathsf{PI}} \, \mathscr{E}(\rho^{q/2}, \rho^{q/2}) \geq \mathrm{var}_\pi(\rho^{q/2}) &= \int \rho^q \, \mathrm{d}\pi - \Big(\int \rho^{q/2} \, \mathrm{d}\pi \Big)^2 \\ &= \Big(\int \rho^q \, \mathrm{d}\pi \Big) \Big[1 - \frac{\exp((q-2) \, \mathcal{R}_{q/2}(\rho\pi \parallel \pi))}{\exp((q-1) \, \mathcal{R}_q(\rho\pi \parallel \pi))} \Big] \\ &\geq \Big(\int \rho^q \, \mathrm{d}\pi \Big) \big\{ 1 - \exp\big(-\mathcal{R}_q(\rho\pi \parallel \pi) \big) \big\} \end{aligned}$$

where we used the monotonicity of the Rényi divergence in the order. Hence,

$$\frac{4}{q} \frac{\mathscr{E}(\rho^{q/2}, \rho^{q/2})}{\int \rho^q \, \mathrm{d}\pi} \geq \frac{4}{q C_{\mathsf{PI}}} \big\{ 1 - \exp\big(-\mathcal{R}_q(\rho\pi \parallel \pi) \big) \big\}$$

$$\geq \frac{2}{qC_{\mathsf{PI}}} \begin{cases} 1\,, & \text{if } \mathcal{R}_q(\rho\pi \parallel \pi) \geq 1\,, \\ \mathcal{R}_q(\rho\pi \parallel \pi)\,, & \text{if } \mathcal{R}_q(\rho\pi \parallel \pi) \leq 1\,. \end{cases}$$ □

To interpret this theorem, the Poincaré result states that after an initial waiting period of time $O(qC_{\mathsf{PI}}\,\mathcal{R}_q(\pi_0 \parallel \pi))$, the Rényi divergence starts decaying exponentially fast. On the other hand, the log-Sobolev inequality implies exponentially fast convergence from the outset. In particular, for $q = 2$, we see that whereas a Poincaré inequality implies exponential decay of χ^2, a log-Sobolev inequality implies exponential decay of $\ln(1 + \chi^2)$, which is substantially stronger.

2.3 Operations Preserving Functional Inequalities

To further expand the class of distributions known to satisfy functional inequalities, we will show in this section that functional inequalities are stable under various common operations on probability measures.

We let $C_{\mathsf{PI}}(\pi)$ denote the Poincaré constant of a probability measure π, and similarly write $C_{\mathsf{LSI}}(\pi)$, $C_{\mathsf{T}_2}(\pi)$, etc.

2.3.1 Bounded Perturbation

Suppose that π satisfies a functional inequality, and that μ is another probability measure satisfying $0 < c \leq \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \leq C < \infty$. Then, it often follows that μ also satisfies the same functional inequality, with a worse constant. This furnishes a large class of examples of non-log-concave measures satisfying functional inequalities.

Proposition 2.3.1 (Holley–Stroock perturbation)**.** *Suppose that π satisfies either a Poincaré or log-Sobolev inequality. Then, if μ satisfies $0 < c \leq \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \leq C < \infty$, then μ also satisfies the corresponding functional inequality with constant*

$$C_{\mathsf{PI}}(\mu) \leq \frac{C}{c}\, C_{\mathsf{PI}}(\pi) \qquad \text{or} \qquad C_{\mathsf{LSI}}(\mu) \leq \frac{C}{c}\, C_{\mathsf{LSI}}(\pi)$$

respectively.

Proof. The key is to find a variational principle for the variance or for the entropy. For the variance, for any $\nu \in \mathcal{P}(\mathbb{R}^d)$ and $f : \mathbb{R}^d \to \mathbb{R}$,

$$\operatorname{var}_\nu f = \inf_{m \in \mathbb{R}} \mathbb{E}_\nu[|f - m|^2]\,.$$

From this,

$$\operatorname{var}_\mu f = \inf_{m\in\mathbb{R}} \mathbb{E}_\mu[|f-m|^2] \le C \inf_{m\in\mathbb{R}} \mathbb{E}_\pi[|f-m|^2] = C \operatorname{var}_\pi f$$
$$\le C\, C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\pi \Gamma(f,f) \le \frac{C}{c}\, C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\mu \Gamma(f,f)\,.$$

The proof for the log-Sobolev inequality is similar once we have the variational principle

$$\operatorname{ent}_\nu f = \inf_{t>0} \mathbb{E}_\nu \Big[\underbrace{f \ln \frac{f}{t} - f + t}_{\ge 0} \Big]$$

for any $f : \mathbb{R}^d \to \mathbb{R}_+$, which we leave as Exercise 2.9. □

One can also state a perturbation principle for the T_2 inequality, but it is more involved and the constants are less precise, see [BGL14, Proposition 9.6.3].

Proposition 2.3.2. *Suppose that π satisfies a T_2 inequality. If $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ satisfies $0 < C^{-1} \le \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \le C < \infty$, then μ also satisfies a T_2 inequality, where $C_{\mathsf{T}_2}(\mu)$ is bounded in terms of C and $C_{\mathsf{T}_2}(\pi)$ only.*

2.3.2 Contractive Mapping

Another simple but useful condition which enables us to transfer functional inequalities from π to μ is the existence of a Lipschitz mapping which pushes forward π to μ.

Proposition 2.3.3. *Suppose that $\pi \in \mathcal{P}(\mathbb{R}^d)$ satisfies either a Poincaré or a log-Sobolev inequality, and that there exists an L-Lipschitz map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $\mu = T_\# \pi$. Then, μ also satisfies the corresponding functional inequality with constant*

$$C_{\mathsf{PI}}(\mu) \le L^2\, C_{\mathsf{PI}}(\pi) \qquad \text{or} \qquad C_{\mathsf{LSI}}(\mu) \le L^2\, C_{\mathsf{LSI}}(\pi)$$

respectively.

Proof. Assume for simplicity that T is continuously differentiable, so that $\|\nabla T\|_{\mathrm{op}} \le L$. Then, for $f : \mathbb{R}^d \to \mathbb{R}$, by applying the Poincaré inequality for π,

$$\operatorname{var}_\mu f = \operatorname{var}_\pi (f \circ T) \le C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\pi[\|\nabla (f\circ T)\|^2] \le C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\pi[\|\nabla T\|_{\mathrm{op}}^2\, \|\nabla f \circ T\|^2]$$

$$\leq C_{\mathrm{PI}}(\pi)\, L^2\, \mathbb{E}_\pi[\|\nabla f \circ T\|^2] = C_{\mathrm{PI}}(\pi)\, L^2\, \mathbb{E}_\mu[\|\nabla f\|^2]\,.$$

The proof for the log-Sobolev inequality is similar. □

This result becomes particularly powerful when combined with **Caffarelli's contraction theorem**, which states that the optimal transport map from the standard Gaussian to an α-strongly log-concave measure is $\alpha^{-1/2}$-Lipschitz. As it is often easier to prove functional inequalities for the standard Gaussian, this principle then quickly implies Poincaré and log-Sobolev inequalities (as well as many other functional inequalities) for strongly log-concave measures. We will return to this in Section 3.5.

2.3.3 Convolution

Next, we show that if $\pi = \pi_1 * \pi_2$ is a convolution of two measures, where both π_1 and π_2 satisfy a functional inequality, then so does π. This is a consequence of the subadditivity of variance and entropy. We begin with a variational principle for the entropy.

Lemma 2.3.4 (variational principle for entropy). *For $f : \mathbb{R}^d \to \mathbb{R}_{>0}$,*

$$\mathrm{ent}_\pi f = \sup\{\mathbb{E}_\pi[fg] \mid g : \mathbb{R}^d \to \mathbb{R} \text{ such that } \mathbb{E}_\pi \exp g \leq 1\}\,.$$

Proof. We may assume that $\mathbb{E}_\pi \exp g = 1$, and define μ via $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} := \exp g$. Then,

$$\mathrm{ent}_\pi f = \mathbb{E}_\pi\Big[f \ln \frac{f}{\mathbb{E}_\pi f}\Big] = \underbrace{\mathbb{E}_\mu\Big[f \exp(-g) \ln \frac{f\exp(-g)}{\mathbb{E}_\mu[f\exp(-g)]}\Big]}_{=\mathrm{ent}_\mu(f\exp(-g))\geq 0} + \mathbb{E}_\pi[fg]\,.$$

Equality holds if $g = \ln(f/\mathbb{E}_\pi f)$. □

***Remark* 2.3.5.** The variational principle above is essentially a reformulation of the Donsker–Varadhan variational principle (Theorem 1.5.4).

Lemma 2.3.6 (subadditivity of variance and entropy). *If $X_1, \ldots, X_n$ are independent random variables, then*

$$\mathrm{var}\, f(X_1, \ldots, X_n) \leq \mathbb{E} \sum_{i=1}^n \mathrm{var}\big(f(X_1, \ldots, X_n) \,\big|\, X_{-i}\big)\,,$$

$$\operatorname{ent} f(X_1, \dots, X_n) \le \mathbb{E} \sum_{i=1}^{n} \operatorname{ent}\big(f(X_1, \dots, X_n) \,\big|\, X_{-i}\big) .$$

Here, $\operatorname{var}(\cdot \mid X_{-i})$ *and* $\operatorname{ent}(\cdot \mid X_{-i})$ *denote the conditional variance and entropy respectively when all variables except* X_i *are held fixed, i.e.,*

$$\operatorname{var}\big(f(X_1, \dots, X_n) \,\big|\, X_{-i} = x_{-i}\big) = \operatorname{var} f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n) ,$$
$$\operatorname{ent}\big(f(X_1, \dots, X_n) \,\big|\, X_{-i} = x_{-i}\big) = \operatorname{ent} f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n) .$$

Proof. The subadditivity of the variance was established in Exercise 1.2, so we turn towards the entropy. Let $Z := f(X_1, \dots, X_n)$ and

$$\Delta_i = \ln \mathbb{E}[Z \mid X_1, \dots, X_i] - \ln \mathbb{E}[Z \mid X_1, \dots, X_{i-1}] ,$$

so that

$$\operatorname{ent} Z = \mathbb{E}[Z \,(\ln Z - \ln \mathbb{E}\, Z)] = \mathbb{E}\Big[Z \sum_{i=1}^{n} \Delta_i\Big] .$$

Since

$$\mathbb{E}[\exp \Delta_i \mid X_{-i}] = \frac{\mathbb{E}[\mathbb{E}[Z \mid X_1, \dots, X_i] \mid X_{-i}]}{\mathbb{E}[Z \mid X_1, \dots, X_{i-1}]} = 1 ,$$

the variational principle yields

$$\mathbb{E}[Z\Delta_i] = \mathbb{E}\,\mathbb{E}[Z\Delta_i \mid X_{-i}] \le \mathbb{E} \operatorname{ent}(Z \mid X_{-i}) . \qquad \square$$

Proposition 2.3.7. *Suppose that* $\pi = \pi_1 * \pi_2 \in \mathcal{P}(\mathbb{R}^d)$*, where* π_1 *and* π_2 *both satisfy either a Poincaré or a log-Sobolev inequality. Then,* π *also satisfies the corresponding functional inequality with constant*

$$C_{\mathsf{PI}}(\pi) \le C_{\mathsf{PI}}(\pi_1) + C_{\mathsf{PI}}(\pi_2) \qquad \text{or} \qquad C_{\mathsf{LSI}}(\pi) \le C_{\mathsf{LSI}}(\pi_1) + C_{\mathsf{LSI}}(\pi_2)$$

respectively.

Proof. Let $X \sim \pi_1$ and $Y \sim \pi_2$ be independent, and let $f : \mathbb{R}^d \to \mathbb{R}$. Using the subadditivity of the variance,

$$\begin{aligned} \operatorname{var}_\pi f = \operatorname{var} f(X+Y) &\le \mathbb{E} \operatorname{var}\big(f(X+Y) \,\big|\, Y\big) + \mathbb{E} \operatorname{var}\big(f(X+Y) \,\big|\, X\big) \\ &\le \{C_{\mathsf{PI}}(\pi_1) + C_{\mathsf{PI}}(\pi_2)\}\, \mathbb{E}[\|\nabla f(X+Y)\|^2] , \end{aligned}$$

and a similar argument holds for the entropy. $\square$

2.3.4 Mixtures

Suppose that π is a mixture, $\pi = \mu P := \int P_x \, \mathrm{d}\mu(x)$, where $\mu \in \mathcal{P}(\mathcal{X})$ is the *mixing measure* and $(P_x)_{x\in\mathcal{X}}$ is a family of probability measures on $\mathbb{R}^d$ indexed by $\mathcal{X}$ (in other words, a Markov kernel). For example, when $\mathcal{X} = [k]$, then μP is a mixture of k distributions $P_1, \dots, P_k$ with mixing weights given by μ. When $\mathcal{X} = \mathbb{R}^d$ and P_x is the translation of a fixed probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$ by x, then $\mu P = \mu * \nu$ is the convolution of μ and ν.

Under general conditions on the mixture, it turns out that if each P_x satisfies a functional inequality, then so does the mixture μP. The simplest demonstration of this idea is for the Poincaré inequality. Although the arguments in this section apply more generally to mixtures μP on arbitrary state spaces, we focus on the $\mathbb{R}^d$ case for simplicity.

Proposition 2.3.8 (PI for mixtures, [Bar+18]). *Let μP be a mixture and assume that each P_x satisfies a Poincaré inequality with constant $C_{\mathsf{PI}}(P)$. Also, assume that*

$$C_{\chi^2} := \sup_{x,x' \in \mathrm{supp}(\mu)} \chi^2(P_x \,\|\, P_{x'}) < \infty \,. \tag{2.3.9}$$

Then, μP satisfies a Poincaré inequality with constant

$$C_{\mathsf{PI}}(\mu P) \le \big(1 + \frac{C_{\chi^2}}{2}\big)\, C_{\mathsf{PI}}(P) \,.$$

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}$, and let $X, X' \overset{\text{i.i.d.}}{\sim} \mu$. By the total law of variance,

$$\mathrm{var}_{\mu P} f = \mathbb{E}\, \mathrm{var}_{P_X} f + \mathrm{var}\, \mathbb{E}_{P_X} f \,.$$

The first term is easy to control, because we can apply the Poincaré inequality for P_X inside the expectation, so the main difficulty lies in the second term. Here,

$$\begin{aligned} \mathrm{var}\, \mathbb{E}_{P_X} f = \frac{1}{2}\, \mathbb{E}[|\mathbb{E}_{P_X} f - \mathbb{E}_{P_{X'}} f|^2] = \frac{1}{2}\, \mathbb{E}\Big[\Big|\int f \,\big(\frac{\mathrm{d}P_X}{\mathrm{d}P_{X'}} - 1\big)\, \mathrm{d}P_{X'}\Big|^2\Big] \\ \le \frac{1}{2}\, \mathbb{E}[(\mathrm{var}_{P_{X'}} f)\, \chi^2(P_X \,\|\, P_{X'})] \le \frac{C_{\chi^2}}{2}\, \mathbb{E}\, \mathrm{var}_{P_X} f \,. \end{aligned}$$

Hence,

$$\mathrm{var}_{\mu P} f \le \big(1 + \frac{C_{\chi^2}}{2}\big)\, \mathbb{E}\, \mathrm{var}_{P_X} f \le \big(1 + \frac{C_{\chi^2}}{2}\big)\, C_{\mathsf{PI}}(P)\, \underbrace{\mathbb{E}\, \mathbb{E}_{P_X}\, \Gamma(f,f)}_{= \mathbb{E}_{\mu P}\, \Gamma(f,f)} \,. \qquad \square$$

Our aim is to extend this idea to the log-Sobolev inequality, which will require a few preliminaries. Rather than aiming to directly prove a log-Sobolev inequality, we will instead prove a **defective log-Sobolev inequality**: for all $f : \mathbb{R}^d \to \mathbb{R}$,

$$\operatorname{ent}_\pi(f^2) \le 2C\,\mathbb{E}_\pi\,\Gamma(f,f) + D\,\mathbb{E}_\pi[f^2]\,. \tag{2.3.10}$$

Although the defective LSI involves an extra term on the right-hand side of the inequality, the extra term can be removed via a Poincaré inequality. The following two results show how this is achieved.

Lemma 2.3.11 (Rothaus lemma). *Let $f : \mathbb{R}^d \to \mathbb{R}$. For all $c \in \mathbb{R}$,*

$$\operatorname{ent}_\pi\big((f+c)^2\big) \le \operatorname{ent}_\pi(f^2) + 2\,\mathbb{E}_\pi[f^2]\,.$$

Proof. Omitted; see [BGL14, Lemma 5.1.4]. □

Lemma 2.3.12 (tightening a defective LSI). *Suppose that π satisfies the defective log-Sobolev inequality (2.3.10), together with a Poincaré inequality. Then, π satisfies an log-Sobolev inequality with constant*

$$C_{\mathrm{LSI}} \le C + C_{\mathrm{PI}}\left(1 + \frac{D}{2}\right).$$

Proof. Using the Rothaus lemma, the defective log-Sobolev inequality, and the Poincaré inequality, we obtain

$$\begin{aligned}\operatorname{ent}_\pi(f^2) \le \operatorname{ent}_\pi\big((f-\mathbb{E}_\pi f)^2\big) + 2\operatorname{var}_\pi(f) &\le 2C\,\mathbb{E}_\pi\,\Gamma(f,f) + (2+D)\operatorname{var}_\pi(f)\\ &\le \big(2C + C_{\mathrm{PI}}\,(2+D)\big)\,\mathbb{E}_\pi\,\Gamma(f,f)\,.\end{aligned}$$ □

We also need one change of measure lemma.

Lemma 2.3.13 ([CCN21]). *Suppose that μ, ν are probability measures and f is a positive function. Then,*

$$\mathbb{E}_\mu(f)\ln\frac{\mathbb{E}_\mu(f)}{\mathbb{E}_\nu(f)} \le \operatorname{ent}_\mu(f) + \mathbb{E}_\mu(f)\ln\big(1+\chi^2(\mu\,\|\,\nu)\big)\,.$$

Proof. By rescaling, we may assume $\mathbb{E}_\mu f = 1$. Recall the Donsker–Varadhan variational principle (Theorem 1.5.4), which states

$$\mathsf{KL}(\eta \,\|\, \eta') = \sup\left\{\mathbb{E}_\eta g - \ln \mathbb{E}_{\eta'} \exp g \;\middle|\; g : \mathcal{X} \to \mathbb{R} \text{ is bounded and measurable}\right\}.$$

If we take $\eta = f\mu$, $\eta' = \nu$, and $g = \ln(f/\mathbb{E}_\nu f)$, then

$$\mathbb{E}_\mu\Big[f \ln \frac{f}{\mathbb{E}_\nu f}\Big] = \mathbb{E}_\eta \ln \frac{f}{\mathbb{E}_\nu f} \le \mathsf{KL}(\eta \,\|\, \nu) + \underbrace{\ln \mathbb{E}_\nu \frac{f}{\mathbb{E}_\nu f}}_{=0} = \mathbb{E}_\mu\Big[f \ln\big(f \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\big)\Big].$$

By subtracting $\mathbb{E}_\mu(f \ln f)$ from both sides, we obtain

$$\ln \frac{1}{\mathbb{E}_\nu f} = \mathbb{E}_\mu\Big[f \ln \frac{1}{\mathbb{E}_\nu f}\Big] \le \mathbb{E}_\mu\Big[f \ln \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big].$$

Next, applying the Donsker–Varadhan principle a second time with $\eta = f\mu$, $\eta' = \mu$, and $g = \ln \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ yields

$$\mathbb{E}_\mu\Big[f \ln \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big] = \mathbb{E}_\eta \ln \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \le \mathsf{KL}(\eta \,\|\, \mu) + \ln \mathbb{E}_\mu \frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \mathrm{ent}_\mu f + \ln\big(1 + \chi^2(\mu \,\|\, \nu)\big),$$

which is what we wanted to show. □

We can now prove the log-Sobolev inequality for mixtures.

Proposition 2.3.14 (LSI for mixtures, [CCN21]). *Let μP be a mixture and assume that each P_x satisfies a log-Sobolev inequality with constant $C_{\mathsf{LSI}}(P)$. Also, assume that*

$$C_{\chi^2} := \sup_{x,x' \in \mathrm{supp}(\mu)} \chi^2(P_x \,\|\, P_{x'}) < \infty.$$

Then, μP satisfies a log-Sobolev inequality with constant

$$\begin{aligned} C_{\mathsf{LSI}}(\mu P) &\le C_{\mathsf{LSI}}(P) \left\{2 + \Big(1 + \frac{C_{\chi^2}}{2}\Big)\Big(1 + \frac{\ln(1 + C_{\chi^2})}{2}\Big)\right\} \\ &\le 6 C_{\mathsf{LSI}}(P)\,\big(1 \vee C_{\chi^2} \ln(1 + C_{\chi^2})\big). \end{aligned}$$

Proof. We begin, as in the proof of Proposition 2.3.8, with a decomposition of the entropy:

$$\mathrm{ent}_{\mu P}(f^2) = \mathbb{E}\,\mathrm{ent}_{P_X}(f^2) + \mathrm{ent}\,\mathbb{E}_{P_X}(f^2).$$

As before, it is the second term that is difficult to control.

Applying Lemma 2.3.13,

$$\begin{aligned}\operatorname{ent}\mathbb{E}_{P_X}(f^2) = \mathbb{E}\Big[\mathbb{E}_{P_X}(f^2)\ln\frac{\mathbb{E}_{P_X}(f^2)}{\mathbb{E}_{\mu P}(f^2)}\Big] &\le \mathbb{E}\Big[\operatorname{ent}_{P_X}(f^2) + \mathbb{E}_{P_X}(f^2)\ln\big(1+\chi^2(P_X\,\|\,\mu P)\big)\Big] \\ &\le \mathbb{E}\operatorname{ent}_{P_X}(f^2) + \ln(1+C_{\chi^2})\,\mathbb{E}_{\mu P}(f^2)\,.\end{aligned}$$

Hence,

$$\begin{aligned}\operatorname{ent}_{\mu P}(f^2) &\le 2\,\mathbb{E}\operatorname{ent}_{P_X}(f^2) + \ln(1+C_{\chi^2})\,\mathbb{E}_{\mu P}(f^2) \\ &\le 4C_{\mathsf{LSI}}(P)\,\mathbb{E}_{\mu P}\,\Gamma(f,f) + \ln(1+C_{\chi^2})\,\mathbb{E}_{\mu P}(f^2)\,,\end{aligned}$$

where we have applied the log-Sobolev inequality for P_X. This is a defective log-Sobolev inequality for μP; by applying the Poincaré inequality from Proposition 2.3.8 and tightening the inequality via Lemma 2.3.12, we conclude the proof. □

Example 2.3.15 (LSI for Gaussian mixtures)**.** Suppose that μ is supported on a ball $\mathsf{B}(0,R)$, and that for each $x\in\mathbb{R}^d$, $P_x = \mathsf{normal}(x,\sigma^2 I_d)$. Then, μP is the convolution $\mu * \mathsf{normal}(0,\sigma^2 I_d)$. Since $C_{\mathsf{LSI}}(P) = \sigma^2$ and

$$\chi^2(P_x\,\|\,P_{x'}) = \exp\frac{\|x-x'\|^2}{\sigma^2} - 1 \le \exp\frac{4R^2}{\sigma^2} - 1$$

for $x, x'\in\mathsf{B}(0,R)$, we deduce that μP satisfies a log-Sobolev inequality with constant

$$C_{\mathsf{LSI}}(\mu P) \lesssim \sigma^2 \vee \big(R^2\exp\frac{4R^2}{\sigma^2}\big)\,.$$

Hence, *Gaussian convolutions of measures with bounded support satisfy a log-Sobolev inequality.* The exponential dependence on R^2/σ^2 is unavoidable in general.

We extend the results of this section in Exercise 2.11.

2.3.5 Tensorization

A key feature of these functional inequalities which makes them crucial for the study of high-dimensional (or even infinite-dimensional phenomena) is that they often hold with dimension-free constants, as demonstrated in the next result.

Theorem 2.3.16 (tensorization). *Suppose that $\pi_1, \dots, \pi_N \in \mathcal{P}(\mathbb{R}^d)$ satisfy either a Poincaré inequality or a log-Sobolev inequality. Then, for any $N \in \mathbb{N}^+$, the product measure $\pi := \bigotimes_{i=1}^N \pi_i$ also satisfies the corresponding functional inequality with constant*

$$C_{\mathsf{PI}}(\pi) = \max_{i \in [N]} C_{\mathsf{PI}}(\pi_i) \qquad \text{or} \qquad C_{\mathsf{LSI}}(\pi) = \max_{i \in [N]} C_{\mathsf{LSI}}(\pi_i)$$

respectively.

Proof. The proof is a straightforward consequence of subadditivity (Lemma 2.3.6). Indeed, if $f : \mathbb{R}^{Nd} \to \mathbb{R}$ and if $X_i \sim \pi_i$ are independent for $i \in [N]$,

$$\begin{aligned} \operatorname{var}_\pi f = \operatorname{var} f(X_1, \dots, X_N) &\le \mathbb{E} \sum_{i=1}^N \operatorname{var}\big(f(X_1, \dots, X_N) \,\big|\, X_{-i}\big) \\ &\le \max_{i \in [N]} C_{\mathsf{PI}}(\pi_i)\, \mathbb{E} \sum_{i=1}^N \mathbb{E}[\|\nabla_i f(X_1, \dots, X_N)\|^2 \mid X_{-i}] \\ &= \max_{i \in [N]} C_{\mathsf{PI}}(\pi_i)\, \mathbb{E}_\pi[\|\nabla f\|^2]\,. \end{aligned}$$

The proof is the same for the log-Sobolev inequality. □

There is also a tensorization principle for transport inequalities, which however requires some additional work to prove. We formulate a general result which applies to many different transport inequalities (not just the T_1 and T_2 inequalities).

Theorem 2.3.17 (Marton's tensorization). *Let $\mathcal{X}_1, \dots, \mathcal{X}_N$ be Polish spaces equipped with probability measures $\pi_1, \dots, \pi_N$ respectively. Let $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ be equipped with the product measure $\pi := \pi_1 \otimes \cdots \otimes \pi_N$.*

Let $\varphi : [0, \infty) \to [0, \infty)$ be convex and for $i \in [N]$, let $c_i : \mathcal{X}_i \times \mathcal{X}_i \to [0, \infty)$ be a lower semicontinuous cost function. Suppose that

$$\inf_{\gamma_i \in \mathcal{C}(\pi_i, \nu_i)} \varphi\Big(\int c_i \, \mathrm{d}\gamma_i\Big) \le 2\sigma^2\, \mathsf{KL}(\nu_i \,\|\, \pi_i)\,, \qquad \forall \nu_i \in \mathcal{P}(\mathcal{X}_i)\,,\ \forall i \in [N]\,.$$

Then, it holds that

$$\inf_{\gamma\in\mathcal{C}(\pi,\nu)} \sum_{i=1}^{N} \varphi\Big(\int c_i(x_i,y_i)\,\mathrm{d}\gamma(x_{1:N},y_{1:N})\Big) \le 2\sigma^2\,\mathsf{KL}(\nu\parallel\pi)\,, \qquad \forall \nu\in\mathcal{P}(\mathcal{X})\,.$$

Proof. The proof goes by induction on N, with $N=1$ being trivial. So, assume that the result is true in dimension N, and let us prove it for dimension $N+1$.

Let $\nu\in\mathcal{P}(\mathcal{X})=\mathcal{P}(\mathcal{X}_1\times\cdots\times\mathcal{X}_{N+1})$, let $\nu_{1:N}$ denote its $\mathcal{X}_1\times\cdots\times\mathcal{X}_N$ marginal, and let $\nu_{N+1|1:N}$ denote the corresponding conditional kernel (and similarly for π). Let K denote the set of conditional kernels $y_{1:N}\mapsto\gamma_{N+1|1:N}(\cdot\mid y_{1:N})$ such that for $\nu_{1:N}$-a.e. $y_{1:N}\in\mathcal{X}_1\times\cdots\times\mathcal{X}_N$, it holds that $\gamma_{N+1|1:N}(\cdot\mid y_{1:N})\in\mathcal{C}(\pi_{N+1},\nu_{N+1|1:N}(\cdot\mid y_{1:N}))$. Instead of minimizing over all $\gamma\in\mathcal{C}(\nu,\pi)$, we can minimize over couplings γ such that for all bounded $f\in C(\mathcal{X}\times\mathcal{X})$,

$$\int f\,\mathrm{d}\gamma=\int\Big(\int f(x_{1:N+1},y_{1:N+1})\,\gamma_{N+1|1:N}(\mathrm{d}x_{N+1},\mathrm{d}y_{N+1}\mid y_{1:N})\Big)\,\gamma_{1:N}(\mathrm{d}x_{1:N},\mathrm{d}y_{1:N})\,,$$

for some $\gamma_{1:N}\in\mathcal{C}(\pi_{1:N},\nu_{1:N})$ and $\gamma_{N+1|1:N}\in\mathsf{K}$.[4] Thus,

$$\begin{aligned}
&\inf_{\gamma\in\mathcal{C}(\pi,\nu)} \sum_{i=1}^{N+1} \varphi\Big(\int c_i(x_i,y_i)\,\mathrm{d}\gamma(x_{1:N+1},y_{1:N+1})\Big)\\
&\quad\le \inf_{\gamma_{1:N}\in\mathcal{C}(\pi_{1:N},\nu_{1:N})}\Bigg\{\sum_{i=1}^{N}\varphi\Big(\int c_i(x_i,y_i)\,\mathrm{d}\gamma_{1:N}(x_{1:N},y_{1:N})\Big)\\
&\qquad\qquad + \inf_{\gamma_{N+1|1:N}\in\mathsf{K}}\varphi\Big(\int\Big(\int c_{N+1}\,\mathrm{d}\gamma_{N+1|1:N}(\cdot\mid y_{1:N})\Big)\,\mathrm{d}\gamma_{1:N}(x_{1:N},y_{1:N})\Big)\Bigg\}\\
&\quad\le \inf_{\gamma_{1:N}\in\mathcal{C}(\pi_{1:N},\nu_{1:N})}\Bigg\{\sum_{i=1}^{N}\varphi\Big(\int c_i(x_i,y_i)\,\mathrm{d}\gamma_{1:N}(x_{1:N},y_{1:N})\Big)\\
&\qquad\qquad + \inf_{\gamma_{N+1|1:N}\in\mathsf{K}}\int\varphi\Big(\int c_{N+1}\,\mathrm{d}\gamma_{N+1|1:N}(\cdot\mid y_{1:N})\Big)\,\mathrm{d}\gamma_{1:N}(x_{1:N},y_{1:N})\Bigg\}\,.
\end{aligned}$$

Then, after checking that the integrands are indeed measurable,

$$\inf_{\gamma_{N+1|1:N}\in\mathsf{K}}\int\varphi\Big(\int c_{N+1}\,\mathrm{d}\gamma_{N+1|1:N}(\cdot\mid y_{1:N})\Big)\,\mathrm{d}\gamma_{1:N}(x_{1:N},y_{1:N})$$

[4] Suppose $N=2$ and $(X_1,X_2)\sim\pi$ and $(Y_1,Y_2)\sim\nu$. Observe that a general coupling $p\in\mathcal{C}(\pi,\nu)$ factorizes as $p(x_1,x_2,y_1,y_2)=p_{X_1}(x_1)\,p_{X_2}(x_2)\,p_{Y_1,Y_2|X_1,X_2}(y_1,y_2\mid x_1,x_2)$. In contrast, we are restricting to couplings of the form $p(x_1,x_2,y_1,y_2)=p_{X_1}(x_1)\,p_{Y_1|X_1}(y_1\mid x_1)\,p_{X_2}(x_2)\,p_{Y_2|X_2,Y_1}(y_2\mid x_2,y_1)$.

$$
\begin{aligned}
&= \int \inf_{\gamma_{N+1|1:N} \in \mathcal{C}(\pi_{N+1}, \nu_{N+1|1:N}(\cdot|y_{1:N}))} \varphi\Big(\int c_{N+1} \, \mathrm{d}\gamma_{N+1|1:N}(\cdot \mid y_{1:N}) \Big) \, \mathrm{d}\gamma_{1:N}(x_{1:N}, y_{1:N}) \\
&\le 2\sigma^2 \int \mathsf{KL}\big(\nu_{N+1|1:N}(\cdot \mid y_{1:N}) \,\big\|\, \pi_{N+1}\big) \, \mathrm{d}\gamma_{1:N}(x_{1:N}, y_{1:N}) \\
&= 2\sigma^2 \int \mathsf{KL}\big(\nu_{N+1|1:N}(\cdot \mid y_{1:N}) \,\big\|\, \pi_{N+1}\big) \, \mathrm{d}\nu_{1:N}(y_{1:N}) \,,
\end{aligned}
$$

where we used the assumption. On the other hand, the inductive hypothesis is

$$
\inf_{\gamma_{1:N} \in \mathcal{C}(\pi_{1:N}, \nu_{1:N})} \sum_{i=1}^{N} \varphi\Big(\int c_i(x_i, y_i) \, \mathrm{d}\gamma_{1:N}(x_{1:N}, y_{1:N}) \Big) \le 2\sigma^2 \, \mathsf{KL}(\nu_{1:N} \,\|\, \pi_{1:N}) \,.
$$

The chain rule for the KL divergence (Lemma 1.5.5) yields

$$
\mathsf{KL}(\nu \,\|\, \pi) = \mathsf{KL}(\nu_{1:N} \,\|\, \pi_{1:N}) + \int \mathsf{KL}\big(\nu_{N+1|1:N}(\cdot \mid y_{1:N}) \,\big\|\, \pi_{N+1}\big) \, \mathrm{d}\nu_{1:N}(y_{1:N}) \,.
$$

Therefore, we have proven

$$
\inf_{\gamma \in \mathcal{C}(\pi, \nu)} \sum_{i=1}^{N+1} \varphi\Big(\int c_i(x_i, y_i) \, \mathrm{d}\gamma(x_{1:N+1}, y_{1:N+1}) \Big) \le 2\sigma^2 \, \mathsf{KL}(\nu \,\|\, \pi) \,. \qquad \square
$$

The preceding proof is supposed to be a straightforward proof by induction, but it is rather cumbersome to write out precisely.

As our first application of the tensorization principle, we will examine the tensorization properties of the T_1 inequality. Recall that on a general metric space, the Wasserstein distances are defined as in Exercise 1.11.

Example 2.3.18 (tensorization of T_1)**.** We will use the cost $c_i = \mathrm{d}_i$, where d_i is a lower semicontinuous metric on $\mathcal{X}_i$, and we take the convex function $\varphi(x) := x^2$. Suppose that for each $i \in [N]$, the measure $\pi_i \in \mathcal{P}(\mathcal{X})$ satisfies the T_1 inequality

$$
W_1^2(\nu_i, \pi_i) \le 2\sigma^2 \, \mathsf{KL}(\nu_i \,\|\, \pi_i), \qquad \forall \nu_i \in \mathcal{P}(\mathcal{X}_i) \,.
$$

Let $\pi := \pi_1 \otimes \cdots \otimes \pi_N$ be the product measure and let $\nu \in \mathcal{P}(\mathcal{X}_1 \times \cdots \times \mathcal{X}_N)$. Suppose also that $\alpha_1, \ldots, \alpha_N > 0$ are numbers with $\sum_{i=1}^N \alpha_i^2 = 1$. Then, Marton's tensorization (Theorem 2.3.17) yields

$$
2\sigma^2 \, \mathsf{KL}(\nu \,\|\, \pi) \ge \Big(\sum_{i=1}^{N} \alpha_i^2 \Big) \inf_{\gamma \in \mathcal{C}(\nu, \pi)} \sum_{i=1}^{N} \Big(\int \mathrm{d}_i(x_i, y_i) \, \mathrm{d}\gamma(x_{1:N}, y_{1:N}) \Big)^2
$$

$$\geq \inf_{\gamma \in \mathcal{C}(\nu,\pi)} \int \sum_{i=1}^{N} \alpha_i \mathrm{d}_i(x_i, y_i) \, \mathrm{d}\gamma(x_{1:N}, y_{1:N}) \,,$$

where we used the Cauchy–Schwarz inequality. This is a T_1 inequality for the weighted distance $\mathrm{d}_\alpha(x_{1:N}, y_{1:N}) := \sum_{i=1}^N \alpha_i \mathrm{d}_i(x_i, y_i)$.

Together with results from the next section, this tensorization result is already powerful enough to recover the bounded differences concentration inequality (see Exercise 2.15), but it is not fully satisfactory as it yields a transport inequality for a weighted metric.

A more satisfactory result is obtained by applying Marton's tensorization (Theorem 2.3.17) with the convex function $\varphi(x) := x$ and cost functions $c_i := \mathrm{d}_i^2$, with d_i a lower semicontinuous metric on $\mathcal{X}_i$. This immediately yields the following corollary.

Corollary 2.3.19 (tensorization of T_2)**.** *Suppose that for each $i \in [N]$, $\pi_i \in \mathcal{P}(\mathcal{X}_i)$ satisfies a T_2 inequality with parameter σ^2 with respect to the metric d_i. Then, the product measure $\pi_1 \otimes \cdots \otimes \pi_N$ satisfies a T_2 inequality with the same parameter σ^2 with respect to the metric $\mathrm{d}(x_{1:N}, y_{1:N})^2 := \sum_{i=1}^N \mathrm{d}(x_i, y_i)^2$ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_N$.*

2.4 Concentration of Measure

We now turn towards the close relationship between functional inequalities and the concentration of measure phenomenon, the latter of which is an indispensable tool in high-dimensional probability and statistics. Since many of the arguments hold on a general Polish space (that is, a complete separable metric space) $(\mathcal{X}, \mathrm{d})$ equipped with a probability measure π, we will work in this setting unless explicitly stated otherwise.

2.4.1 Blow-Up of Sets and Concentration of Lipschitz Functions

Loosely speaking, the concentration of measure phenomenon holds when a huge fraction of the mass of π is concentrated on a relatively small set. Another way of capturing this idea is to assert that whenever a set has a non-trivial amount of mass under π, then expanding the set slightly causes it to capture *almost all* of the mass of π. The following definitions formalize this idea.

Definition 2.4.1. For a Borel subset $A \subseteq \mathcal{X}$ and $\varepsilon > 0$, we let A^ε denote the **ε-blow-up**

of A, defined by

$$A^\varepsilon := \{x \in \mathcal{X} \mid \mathrm{d}(x, A) < \varepsilon\}\,.$$

The **concentration function** $\alpha_\pi : \mathbb{R}_+ \to [0,1]$ is defined via

$$\alpha_\pi(\varepsilon) = \sup\Big\{\pi\big((A^\varepsilon)^{\mathrm{c}}\big) \;\Big|\; A \subseteq \mathcal{X} \text{ is a Borel subset with } \pi(A) \geq \frac{1}{2}\Big\}\,.$$

Typically, we have $\alpha_\pi(\varepsilon) \leq C_0 \exp(-\varepsilon/C_1)$ or $\alpha_\pi(\varepsilon) \leq C_0 \exp(-\varepsilon^2/C_1)$ for some constants $C_0, C_1 > 0$; hence, as we increase ε, the blow-up A^ε captures a (often substantially) larger fraction of the mass of π. In the next sections, we will develop tools to upper bound concentration functions. For now, however, we wish to develop an equivalence between the formulation of concentration of measure via blow-up of sets, and with another involving concentration of Lipschitz functions. Although the former has a more striking geometric interpretation, the latter is often how concentration is used in applications.

Given a real-valued random variable X, we abuse notation and let $\operatorname{med} X$ denote any median of X, that is, any number m such that $\mathbb{P}\{X \leq m\} \wedge \mathbb{P}\{X \geq m\} \geq \frac{1}{2}$.

Theorem 2.4.2 (blow-up and Lipschitz functions)**.** *Suppose that $(\mathcal{X}, \mathrm{d}, \pi)$ has concentration function α_π. Then, for any 1-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ and $\varepsilon \geq 0$,*

$$\pi\{f \geq \operatorname{med} f + \varepsilon\} \leq \alpha_\pi(\varepsilon)\,.$$

Conversely, suppose that for all 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$, it holds that

$$\pi\{f \geq \operatorname{med} f + \varepsilon\} \leq \beta(\varepsilon)\,.$$

Then, the concentration function α_π of $(\mathcal{X}, \mathrm{d}, \pi)$ satisfies $\alpha_\pi \leq \beta$.

Proof. ($\Longrightarrow$) Consider the set $A := \{f \leq \operatorname{med} f\}$. We claim that $A^\varepsilon \subseteq \{f - \operatorname{med} f < \varepsilon\}$. To prove this, let $x \in A^\varepsilon$. By definition, there exists $y \in A$ such that $\mathrm{d}(x,y) < \varepsilon$, so $f(x) - \operatorname{med} f = f(y) - \operatorname{med} f + f(x) - f(y) \leq \mathrm{d}(x,y) < \varepsilon$. Hence,

$$\pi\{f - \operatorname{med} f < t\} \geq \pi(A^\varepsilon) \geq 1 - \alpha(\varepsilon)\,.$$

($\Longleftarrow$) The function $f := \mathrm{d}(\cdot, A)$ is 1-Lipschitz, and if $\pi(A) \geq \frac{1}{2}$ then 0 is a median of f. Thus, it holds that

$$\pi(A^\varepsilon) = \pi\{f - \operatorname{med} f < \varepsilon\} \geq 1 - \beta(\varepsilon)\,. \qquad \square$$

More broadly, it is a general principle that many statements about sets have an equivalent reformulation in terms of functions. We will see more instances of this idea throughout the book.

Some statements regarding concentration, such as the theorem above, are more easily phrased in terms of concentration around the median rather than around the mean. The following result shows that, up to numerical constants, the mean and the median are equivalent. To state the result in generality, we introduce the idea of an Orlicz norm.

Definition 2.4.3 (Orlicz norm)**.** If $\psi : [0,\infty) \to [0,\infty)$ is a convex strictly increasing function with $\psi(0) = 0$ and $\psi(x) \to \infty$ as $x \to \infty$, then it is an **Orlicz function**.

For a real-valued random variable X, its **Orlicz norm** is defined to be

$$\|X\|_\psi := \inf\big\{t > 0 \;\big|\; \mathbb{E}\,\psi\big(\frac{|X|}{t}\big) \le 1\big\}.$$

Examples of Orlicz functions include $\psi(x) = x^p$ for $p \ge 1$, for which the corresponding Orlicz norm is the $L^p(\mathbb{P})$ norm, and $\psi_2(x) := \exp(x^2) - 1$ for which the Orlicz norm $\|X\|_{\psi_2}$ captures the sub-Gaussianity of X.

Lemma 2.4.4 (mean and median)**.** *Let ψ be an Orlicz function and let X be a real-valued random variable. Then,*

$$\frac{1}{2}\,\|X - \mathbb{E}\,X\|_\psi \le \|X - \operatorname{med} X\|_\psi \le 3\,\|X - \mathbb{E}\,X\|_\psi\,.$$

Proof. We can assume that X is not constant; from the properties of Orlicz functions, $\psi^{-1}(t)$ is well-defined for any $t > 0$. Then,

$$\begin{aligned}\|X - \mathbb{E}\,X\|_\psi &\le \|X - \operatorname{med} X\|_\psi + \|\operatorname{med} X - \mathbb{E}\,X\|_\psi \\ &= \|X - \operatorname{med} X\|_\psi + |\operatorname{med} X - \mathbb{E}\,X|\;\|1\|_\psi \\ &\le \|X - \operatorname{med} X\|_\psi + \mathbb{E}|X - \operatorname{med} X|\;\|1\|_\psi\,.\end{aligned}$$

Since

$$\mathbb{E}\,\psi\Big(\frac{|X - \operatorname{med} X|}{\mathbb{E}|X - \operatorname{med} X|\;\|1\|_\psi}\Big) \ge \psi\Big(\frac{\mathbb{E}|X - \operatorname{med} X|}{\mathbb{E}|X - \operatorname{med} X|\;\|1\|_\psi}\Big) = \psi\big(\frac{1}{\|1\|_\psi}\big) = 1\,,$$

it implies $\mathbb{E}|X - \operatorname{med} X|\;\|1\|_\psi \le \|X - \operatorname{med} X\|_\psi$.

Next, assume that $\operatorname{med} X \ge \mathbb{E}\,X$ (or else replace X by $-X$). Then,

$$\frac{1}{2} \le \mathbb{P}\{X \ge \operatorname{med} X\} \le \mathbb{P}\{|X - \mathbb{E}\,X| \ge \operatorname{med} X - \mathbb{E}\,X\}$$

$$\leq \frac{1}{\psi((\operatorname{med} X - \mathbb{E}\, X)/\|X - \mathbb{E}\, X\|_\psi)}\,,$$

so that

$$|\operatorname{med} X - \mathbb{E}\, X| \leq \psi^{-1}(2)\, \|X - \mathbb{E}\, X\|_\psi\,.$$

Therefore,

$$\|X - \operatorname{med} X\|_\psi \leq \|X - \mathbb{E}\, X\|_\psi + \|\mathbb{E}\, X - \operatorname{med} X\|_\psi \leq \big(1 + \|1\|_\psi\, \psi^{-1}(2)\big)\, \|X - \mathbb{E}\, X\|_\psi\,.$$

Note, however, that $\|1\|_\psi = 1/\psi^{-1}(1)$. Since $\psi(\psi^{-1}(2)/2) \leq 1$ by convexity (and the property $\psi(0) = 0$), it implies $\psi^{-1}(2) \leq 2\psi^{-1}(1)$, and we obtain the result. □

2.4.2 The Herbst Argument

In this section, we specialize to the case where $(\mathcal{X}, \mathrm{d})$ is the Euclidean space $\mathbb{R}^d$.

To put it succinctly, the idea of the Herbst argument is to apply functional inequalities, such as the Poincaré inequality or the log-Sobolev inequality, to the moment-generating function of a 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ in order to deduce a concentration inequality for f. We illustrate this with the log-Sobolev inequality, which implies, for any $\lambda \in \mathbb{R}$,

$$\begin{aligned} \operatorname{ent}_\pi \exp(\lambda f) &\leq 2C_{\mathrm{LSI}}\, \mathbb{E}_\pi\Big[\Big\|\frac{\lambda \exp(\lambda f/2)}{2}\, \nabla f\Big\|^2\Big] = \frac{C_{\mathrm{LSI}}\, \lambda^2}{2}\, \mathbb{E}_\pi[\exp(\lambda f)\, \|\nabla f\|^2] \\ &\leq \frac{C_{\mathrm{LSI}}\, \lambda^2}{2}\, \mathbb{E}_\pi \exp(\lambda f)\,. \end{aligned} \tag{2.4.5}$$

The next lemma shows how to apply this inequality.

Lemma 2.4.6 (Herbst argument)**.** *Suppose that a random variable X satisfies*

$$\operatorname{ent} \exp(\lambda X) \leq \frac{\lambda^2 \sigma^2}{2}\, \mathbb{E} \exp(\lambda X) \qquad \textit{for all } \lambda \geq 0\,.$$

Then, it holds that

$$\mathbb{E} \exp\{\lambda\, (X - \mathbb{E}\, X)\} \leq \exp \frac{\lambda^2 \sigma^2}{2} \qquad \textit{for all } \lambda \geq 0\,.$$

In particular, via a standard Chernoff inequality,

$$\mathbb{P}\{X \geq \mathbb{E}X + t\} \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \qquad \text{for all } t \geq 0\,.$$

Proof. Let $\tau(\lambda) := \lambda^{-1} \ln \mathbb{E} \exp\{\lambda\,(X - \mathbb{E}X)\}$. We leave it to the reader to check the calculus identity

$$\tau'(\lambda) = \frac{1}{\lambda^2} \frac{\operatorname{ent} \exp(\lambda X)}{\mathbb{E} \exp(\lambda X)}\,. \tag{2.4.7}$$

Since $\tau(\lambda) \to 0$ as $\lambda \searrow 0$, the assumption of the lemma yields $\tau(\lambda) \leq \lambda\sigma^2/2$. □

The calculation in (2.4.5) shows that the assumption of the Herbst argument is satisfied for *all* 1-Lipschitz functions f, with $\sigma^2 = C_{\mathsf{LSI}}$. Hence, we deduce a concentration inequality for Lipschitz functions, which we formally state in the next theorem together with the corresponding result under a Poincaré inequality. The Poincaré case is left as Exercise 2.12.

Theorem 2.4.8. *Let $\pi \in \mathcal{P}(\mathbb{R}^d)$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be a 1-Lipschitz function.*

1. *If π satisfies a Poincaré inequality with constant C_{PI}, then for all $t \geq 0$,*

$$\pi\{f - \mathbb{E}_\pi f \geq t\} \leq 3 \exp\left(-\frac{t}{\sqrt{C_{\mathsf{PI}}}}\right).$$

2. *If π satisfies a log-Sobolev inequality with constant C_{LSI}, then for all $t \geq 0$,*

$$\pi\{f - \mathbb{E}_\pi f \geq t\} \leq \exp\left(-\frac{t^2}{2C_{\mathsf{LSI}}}\right).$$

Example 2.4.9. Suppose that γ is the standard Gaussian measure on $\mathbb{R}^d$. From the Bakry–Émery theorem (Theorem 1.2.29), γ satisfies the log-Sobolev inequality with $C_{\mathsf{LSI}} = 1$. For $Z \sim \gamma$, since $\mathbb{E}[\|Z\|^2] = d$, the Poincaré inequality applied to the norm $\|\cdot\|$ shows that $\operatorname{var} \|Z\| \leq 1$, i.e., $\sqrt{d-1} \leq \mathbb{E}\|Z\| \leq \sqrt{d}$.

The concentration result above now shows that the standard Gaussian "lives" on a thin spherical shell of radius $\sqrt{d}$ and width $O(1)$.

2.4.3 Transport Inequalities and Concentration

Next, we will show that a T_1 transport inequality is equivalent to sub-Gaussian concentration of Lipschitz functions, which was proven by Bobkov and Götze. The proof shows that in a sense, the two statements are dual to each other.

Theorem 2.4.10 (Bobkov–Götze). *Let $\pi \in \mathcal{P}_1(\mathcal{X})$. The following are equivalent.*

1. *The function f is σ^2-sub-Gaussian with respect to π for every 1-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ which is mean-zero under π.*
2. *The measure π satisfies $\mathsf{T}_1(\sigma^2)$.*

Proof. Let $\mathrm{Lip}_1(\mathcal{X})$ denote the space of 1-Lipschitz and mean-zero functions on $\mathcal{X}$. Lipschitz concentration can be stated as

$$\sup_{\lambda \in \mathbb{R}} \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \left\{ \ln \int \exp(\lambda f) \, \mathrm{d}\pi - \frac{\lambda^2 \sigma^2}{2} \right\} \le 0 \,.$$

By Donsker–Varadhan duality (Theorem 1.5.4), this is equivalent to

$$\sup_{\lambda \in \mathbb{R}} \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \sup_{\nu \in \mathcal{P}(\mathcal{X})} \left\{ \lambda \left(\int f \, \mathrm{d}\nu - \int f \, \mathrm{d}\pi \right) - \mathsf{KL}(\nu \,\|\, \pi) - \frac{\lambda^2 \sigma^2}{2} \right\} \le 0 \,,$$

where we recall that $\int f \, \mathrm{d}\pi = 0$ for $f \in \mathrm{Lip}_1(\mathcal{X})$. If we first evaluate the supremum over $\lambda \in \mathbb{R}$, then we obtain the statement

$$\sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \sup_{\nu \in \mathcal{P}(\mathcal{X})} \left\{ \frac{1}{2\sigma^2} \left(\int f \, \mathrm{d}\nu - \int f \, \mathrm{d}\pi \right)^2 - \mathsf{KL}(\nu \,\|\, \pi) \right\} \le 0 \,,$$

If we next evaluate the supremum over functions $f \in \mathrm{Lip}_1(\mathcal{X})$ using the Kantorovich duality formula (1.E.4) for W_1 from Exercise 1.11, we obtain

$$\sup_{\nu \in \mathcal{P}(\mathcal{X})} \left\{ \frac{W_1^2(\nu, \pi)}{2\sigma^2} - \mathsf{KL}(\nu \,\|\, \pi) \right\} \le 0 \,,$$

which is the T_1 inequality. □

Using the fact that the W_1 distance for the trivial metric $\mathrm{d}(x, y) = \mathbb{1}\{x \neq y\}$ coincides with the TV distance[5], the Bobkov–Götze theorem implies that two classical inequalities in

[5] One has to be slightly careful since for the trivial metric, $(\mathcal{X}, \mathrm{d})$ is usually not separable.

probability theory, Hoeffding's inequality and Pinsker's inequality, are in fact equivalent to each other (see Exercise 2.13).

Although the T_1 inequality implies sub-Gaussian concentration for *all* Lipschitz functions, it is in fact equivalent to sub-Gaussian concentration of a single function, the distance function $\mathsf{d}(\cdot, x_0)$ for some $x_0 \in \mathcal{X}$. The next theorem is not used often because the quantitative dependence of the equivalence can be crude, but it is worth knowing.

Theorem 2.4.11. *Let $\pi \in \mathcal{P}_1(\mathcal{X})$ and $x_0 \in \mathcal{X}$. The following are equivalent:*

1. *π satisfies a T_1 inequality.*
2. *There exists $c > 0$ such that $\mathbb{E}_\pi \exp(c\, \mathsf{d}(\cdot, x_0)^2) < \infty$.*

Transport inequalities offer a flexible and powerful method for characterizing and proving concentration inequalities, as we will see in the next section. Before doing so, however, we wish to also demonstrate how concentration of measure, formulated via blow-up of sets, can be deduced directly from a T_1 inequality.

Suppose that $\mathsf{T}_1(\sigma^2)$ holds, i.e.,

$$W_1^2(\mu, \pi) \le 2\sigma^2\, \mathsf{KL}(\mu \,\|\, \pi) \qquad \text{for all } \mu \in \mathcal{P}_1(\mathcal{X}),\ \mu \ll \pi\,.$$

For any disjoint sets A, B, with $\pi(A)\,\pi(B) > 0$, if we let $\pi(\cdot \mid A)$ (resp. $\pi(\cdot \mid B)$) denote the distribution π conditioned on A (resp. B), then

$$\begin{aligned}\mathsf{d}(A,B) \le W_1\big(\pi(\cdot \mid A), \pi(\cdot \mid B)\big) &\le W_1\big(\pi(\cdot \mid A), \pi\big) + W_1\big(\pi(\cdot \mid B), \pi\big) \\ &\le \sqrt{2\sigma^2\, \mathsf{KL}\big(\pi(\cdot \mid A) \,\big\|\, \pi\big)} + \sqrt{2\sigma^2\, \mathsf{KL}\big(\pi(\cdot \mid B) \,\big\|\, \pi\big)}\,.\end{aligned}$$

However,

$$\mathsf{KL}\big(\pi(\cdot \mid A) \,\big\|\, \pi\big) = \int_A \frac{\pi(\mathrm{d}x)}{\pi(A)} \ln \frac{1}{\pi(A)} = \ln \frac{1}{\pi(A)}\,,$$

so that

$$\mathsf{d}(A,B) \le \sqrt{2\sigma^2 \ln \frac{1}{\pi(A)}} + \sqrt{2\sigma^2 \ln \frac{1}{\pi(B)}}\,.$$

In particular, if we take $B = (A^\varepsilon)^{\mathrm{c}}$ where $\pi(A) \ge \frac{1}{2}$, then $\mathsf{d}(A,B) \ge \varepsilon$. Hence, for all $\varepsilon \ge 2\sqrt{2\sigma^2 \ln 2}$, it holds that $\frac{\varepsilon}{2} \le \sqrt{2\sigma^2 \ln \frac{1}{\pi(B)}}$, or

$$\pi\big((A^\varepsilon)^{\mathrm{c}}\big) \le \exp\big(-\frac{\varepsilon^2}{8\sigma^2}\big) \qquad \text{for all } \varepsilon \ge \sqrt{8 \ln 2}\,\sigma\,. \tag{2.4.12}$$

2.4.4 Tensorization and Gozlan's Theorem

Our goal is now to investigate the relationship between concentration and tensorization. Although results like the Bobkov–Götze theorem (Theorem 2.4.10) provide us with powerful tools to establish concentration results, so far there is nothing inherently *high-dimensional* about these phenomena.

Indeed, to discuss dimensionality, we should move to the product space $\mathcal{X}^N$ and ask when concentration results can hold *independently* of N. If such a statement holds, then the concentration inequality typically becomes stronger[6] as N becomes larger.

For instance, when $\mathcal{X} = \mathbb{R}$, then we know from Theorem 2.3.16 that the Poincaré and log-Sobolev inequalities both tensorize: if they hold for $\pi \in \mathcal{P}(\mathbb{R})$ with a constant C, then they also hold for $\pi^{\otimes N} \in \mathcal{P}(\mathbb{R}^N)$ with the *same* constant C. Since these inequalities imply powerful concentration results (Theorem 2.4.8), they yield examples of genuinely high-dimensional concentration.

For transport inequalities, the tensorization for the T_1 inequality is unsatisfactory in the sense that once we equip $\mathcal{X}^N$ with the product metric $\mathrm{d}(x_{1:N}, x'_{1:N})^2 := \sum_{i=1}^N \mathrm{d}(x_i, x'_i)^2$, the validity of $\mathsf{T}_1(C)$ for $\pi \in \mathcal{P}(\mathcal{X})$ does *not* imply the validity of $\mathsf{T}_1(C)$ for $\pi^{\otimes N} \in \mathcal{P}(\mathcal{X}^N)$ with the same constant C. In fact, from Example 2.3.18, we expect that the T_1 constant for $\pi^{\otimes N}$ can grow as $\sqrt{N}$. On the other hand, from Corollary 2.3.19, we know that the T_2 inequality tensorizes. Since the T_2 inequality on $\mathcal{X}^N$ implies the T_1 inequality on $\mathcal{X}^N$ (trivially), it in turn implies high-dimensional concentration via the Bobkov–Götze equivalence (Theorem 2.4.10).

In this section, we will prove the surprising fact that high-dimensional concentration is actually *equivalent* to the T_2 inequality, in a sense that we shall make precise shortly.

First, we need a few preliminary results, which we shall not prove. The first one is a straightforward technical lemma (see Exercise 2.16).

Lemma 2.4.13. *Let $\pi \in \mathcal{P}_2(\mathcal{X})$.*

1. *The mapping $(x_1, \ldots, x_N) \mapsto W_2(N^{-1} \sum_{i=1}^N \delta_{x_i}, \pi)$ is $N^{-1/2}$-Lipschitz.*
2. *(Wasserstein law of large numbers) Suppose that $(X_i)_{i=1}^\infty \overset{\text{i.i.d.}}{\sim} \pi$, and that for some*

[6] Here, the word "stronger" is not precisely defined but it means something akin to "more useful" or "produces more surprising consequences".

$x_0 \in \mathcal{X}$ and some $\varepsilon > 0$, it holds that $\mathbb{E}[\mathrm{d}(x_0, X_1)^{2+\varepsilon}] < \infty$. Then,

$$\mathbb{E}\, W_2\Big(\frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}, \pi\Big) \to 0 \qquad \text{as } N \to \infty\,.$$

The second result, Sanov's theorem, is a foundational theorem from large deviations. Although Sanov's theorem is of fundamental importance in its own right, it would take us too far afield to develop large deviations theory here, so we invoke it as a black box.

Theorem 2.4.14 (Sanov's theorem). *Let $(X_i)_{i=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \pi$ and let $\pi_N := N^{-1}\sum_{i=1}^{N} \delta_{X_i}$ denote the empirical measure. Then, for any Borel set $A \subseteq \mathcal{P}(\mathcal{X})$, it holds that*

$$\begin{aligned} -\inf_{\operatorname{int} A} \mathsf{KL}(\cdot \,\|\, \pi) &\le \liminf_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\{\pi_N \in A\} \\ &\le \limsup_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\{\pi_N \in A\} \le -\inf_{\overline{A}} \mathsf{KL}(\cdot \,\|\, \pi)\,. \end{aligned}$$

We are now ready to establish the equivalence.

Theorem 2.4.15 (Gozlan). *The measure $\pi \in \mathcal{P}_2(\mathcal{X})$ satisfies $\mathsf{T}_2(\sigma^2)$ if and only if for all $N \in \mathbb{N}^+$ and all 1-Lipschitz $f : \mathcal{X}^N \to \mathbb{R}$, the centered function $f - \mathbb{E}_{\pi^{\otimes N}} f$ is σ^2-sub-Gaussian under $\pi^{\otimes N}$.*

Proof. It remains to prove the converse implication. Fix $t > 0$ and apply the assumption statement to the $N^{-1/2}$-Lipschitz function $(x_1, \ldots, x_N) \mapsto W_2(N^{-1}\sum_{i=1}^{N} \delta_{x_i}, \pi)$. It implies

$$\mathbb{P}\{W_2(\pi_N, \pi) > t\} \le \exp\Big(-\frac{N\,\{t - \mathbb{E}\, W_2(\pi_N, \pi)\}^2}{2\sigma^2}\Big),$$

where $\pi_N := N^{-1}\sum_{i=1}^{N} \delta_{X_i}$, with $(X_i)_{i\in\mathbb{N}^+} \overset{\text{i.i.d.}}{\sim} \pi$. On the other hand, the lower semicontinuity of W_2 implies that $\{\nu \in \mathcal{P}(\mathcal{X}) \mid W_2(\mu, \nu) > t\}$ is open. By Sanov's theorem (Theorem 2.4.14), we obtain

$$\begin{aligned} -\inf\{\mathsf{KL}(\nu \,\|\, \pi) \mid W_2(\nu, \pi) > t\} &\le \liminf_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\{W_2(\pi_N, \pi) > t\} \\ &\le -\limsup_{N\to\infty} \frac{\{t - \mathbb{E}\, W_2(\pi_N, \pi)\}^2}{2\sigma^2} = -\frac{t^2}{2\sigma^2}\,, \end{aligned}$$

where the last inequality comes from the Wasserstein law of large numbers (our assumption implies that π has sub-Gaussian tails, which in particular means $\mathbb{E}[\mathsf{d}(x, X_1)^p] < \infty$ for any $x \in \mathcal{X}$ and any $p \geq 1$).

We have proven that $W_2(\nu, \pi) > t$ implies $\mathsf{KL}(\nu \,\|\, \pi) \geq t^2/(2\sigma^2)$, which is seen to be equivalent to the T_2 inequality. □

Observe in particular that this theorem implies the Otto–Villani theorem (Exercise 1.16): due to tensorization (Theorem 2.3.16) and the Herbst argument (Lemma 2.4.6), a log-Sobolev inequality implies high-dimensional sub-Gaussian concentration of Lipschitz functions, which by Gözlan's theorem is equivalent to a T_2 inequality.

2.5 Isoperimetric Inequalities

In Section 2.4.1, we introduced the concentration function α_π of a measure π. Thus far, we have provided tools to upper bound the concentration function; for example, in (2.4.12), we showed that if π satisfies $\mathsf{T}_1(\sigma^2)$, then

$$\alpha_\pi(\varepsilon) \leq \exp\Big(-\frac{\varepsilon^2}{8\sigma^2}\Big) \qquad \text{for all } \varepsilon \geq \sqrt{8 \ln 2}\,\sigma\,.$$

Observe that this provides no information when ε is small, whereas by definition we know that $\alpha_\pi(0) = \frac{1}{2}$. In this section, we will study finer questions about the concentration function. More generally, for $p \in (0, 1)$ and $\varepsilon > 0$, we introduce the quantity

$$\omega_\pi(p, \varepsilon) := \inf\{\pi(A^\varepsilon) \mid A \text{ is a Borel set with } \pi(A) = p\}\,,$$

and we can ask about the deviation of $\omega_\pi(p, \varepsilon)$ from p when ε is small. In some special cases, we can even determine the function ω_π exactly. The study of this question will bring us to the classical geometric problem of isoperimetry. In its simplest guise, it asks: among all plane curves which enclose an area of a prescribed area, which ones have the least perimeter? Unsurprisingly, among regular curves, it is well-known that circles provide the answer to this question. As we shall see, the isoperimetric question, once generalized to abstract spaces, contains a wealth of information about concentration phenomena.

2.5.1 Classical Isoperimetry Results

The connection between concentration and isoperimetry began with the work of Lévy, who found the isoperimetric inequality on the sphere.

Theorem 2.5.1 (spherical isoperimetry). *Let σ_d denote the uniform measure on the d-dimensional unit sphere $\mathbb{S}^d$, and let $A \subseteq \mathbb{S}^d$ be a Borel subset with $\mu(A) \notin \{0, 1\}$. Let C be a spherical cap with the same measure as A. Then, for all $\varepsilon > 0$,*

$$\sigma_d(A^\varepsilon) \geq \sigma_d(C^\varepsilon)\,.$$

We give a few reminders about spherical geometry. We equip $\mathbb{S}^d$ with its geodesic metric d, so that the distance between two points $x, y \in \mathbb{S}^d$ is equal to the angle between x and y. A spherical cap is a geodesic ball, that is, it is a set of the form $\mathrm{B}(x_0, r)$ for some $x_0 \in \mathbb{S}^d$ and $r > 0$, where the balls are defined w.r.t. d.

Note that Theorem 2.5.1 identifies the exact function ω_{σ_d}.

To see how an isoperimetric result naturally leads to a concentration result, suppose that A has measure $\frac{1}{2}$. Then, the corresponding spherical cap C can be taken to be half of the sphere, $C = \mathrm{B}(x_0, \frac{\pi}{2})$, and so

$$\sigma_d(A^\varepsilon) \geq \sigma_d(C^\varepsilon) = \sigma_d\Big(\mathrm{B}\big(x_0,\ \frac{\pi}{2} + \varepsilon\big)\Big)\,.$$

To obtain an upper bound on the concentration function α_{σ_d}, it therefore suffices to lower bound the volume of the spherical cap. It leads to the following result, which we leave as an exercise (Exercise 2.17).

Theorem 2.5.2 (concentration on the sphere). *Let σ_d be the uniform measure on the unit sphere $\mathbb{S}^d$ in dimension $d \geq 2$, equipped with the geodesic distance* d. *Then,*

$$\alpha_{\sigma_d}(\varepsilon) \leq \exp\Big(-\frac{(d-1)\,\varepsilon^2}{2}\Big)\,, \qquad \text{for all } \varepsilon > 0\,.$$

There is also an isoperimetric result for the standard Gaussian measure γ_d on $\mathbb{R}^d$. In this case, the optimal sets are given by half-spaces, i.e., sets of the form

$$H_{x_0,t} := \{x \in \mathbb{R}^d \mid \langle x, x_0\rangle \leq t\}\,.$$

Theorem 2.5.3 (Gaussian isoperimetry). *Let γ_d denote the standard Gaussian measure on $\mathbb{R}^d$, and let $A \subseteq \mathbb{R}^d$ be a Borel subset with $\gamma_d(A) \neq \{0, 1\}$. Let $H_{x_0,t}$ be a half-space*

with the same measure as A. Then, for all $\varepsilon > 0$,

$$\gamma_d(A^\varepsilon) \geq \gamma_d(H^\varepsilon_{x_0,t}) .$$

We can write this result more explicitly as follows. By rotational invariance of the Gaussian, we can take x_0 to be any unit vector e, in which case the measure of $H_{e,t}$ is $\gamma_d(H_{e,t}) = \Phi(t)$, where Φ is the Gaussian CDF. Since $H^\varepsilon_{e,t} = H_{e,t+\varepsilon}$, then

$$\gamma_d(A^\varepsilon) \geq \Phi\big(\Phi^{-1}\big(\gamma_d(A)\big) + \varepsilon\big) . \tag{2.5.4}$$

In particular, if $\gamma_d(A) = \frac{1}{2}$, then $\Phi^{-1}(\frac{1}{2}) = 0$, so

$$\alpha_{\gamma_d}(\varepsilon) \leq \Phi(-\varepsilon) \leq \frac{1}{2}\exp\big(-\frac{\varepsilon^2}{2}\big) .$$

We now pause to give a remark on proofs. Since these isoperimetric inequalities require a detailed understanding of the measure (including the optimal sets in the inequality), they are considerably more difficult to prove than the other results we have seen so far (e.g., a log-Sobolev inequality). In particular, usually they can only established for measures which are simple in some regard, e.g., they enjoy many symmetries. Hence, we will not prove them here.

It is often convenient to pass to a differential form of the isoperimetric inequality, which is obtained by sending $\varepsilon \searrow 0$. This is formalized as follows.

Definition 2.5.5 (Minkowski content)**.** Given a non-empty Borel set A and a measure π on a Polish space $(\mathcal{X}, \mathrm{d})$, the **Minkowski content** of A under π is

$$\pi^+(A) := \liminf_{\varepsilon \searrow 0} \frac{\pi(A^\varepsilon) - \pi(A)}{\varepsilon} .$$

Definition 2.5.6 (isoperimetric profile)**.** For a measure π on a Polish space $(\mathcal{X}, \mathrm{d})$, the **isoperimetric profile** of π, denoted $\mathcal{I}_\pi : [0,1] \to \mathbb{R}_+$, is the function

$$\mathcal{I}_\pi(p) := \inf\{\pi^+(A) \mid A \text{ is measurable with } \pi(A) = p\} .$$

For the standard Gaussian, a Taylor expansion of (2.5.4) yields

$$\gamma_d(A^\varepsilon) \geq \gamma_d(A) + \phi\big(\Phi^{-1}\big(\gamma_d(A)\big)\big)\,\varepsilon + o(\varepsilon) ,$$

where $\phi = \Phi'$ is the Gaussian density. Hence, we can identify the isoperimetric profile of the standard Gaussian as

$$\mathcal{I}_{\gamma_d}(p) = \phi\big(\Phi^{-1}(p)\big)\,. \tag{2.5.7}$$

Actually, the result (2.5.7) is *equivalent* to the Gaussian isoperimetric inequality (2.5.4). Here, (2.5.4) is called the *integral* form of the inequality, whereas (2.5.7) is called the *differential* form. The following theorem shows how to convert between the two forms.

Theorem 2.5.8 ([BH97])**.** *Let $\mathcal{I} : (0,1) \to \mathbb{R}_{>0}$. Define the increasing function F such that $F(0) = \frac{1}{2}$ and $f \circ F^{-1} = \mathcal{I}$, where $f = F'$; equivalently, we can take $F^{-1}(p) = \int_{1/2}^{p} \mathcal{I}(t)^{-1}\,\mathrm{d}t$. Then, the following statements are equivalent.*

1. *For all $\varepsilon > 0$ and all Borel A with $\pi(A) \notin \{0,1\}$,*

$$\pi(A^\varepsilon) \geq F\big(F^{-1}\big(\pi(A)\big) + \varepsilon\big)\,.$$

2. *For all Borel A with $\pi(A) \notin \{0,1\}$,*

$$\pi^+(A) \geq \mathcal{I}\big(\pi(A)\big)\,.$$

Proof sketch. Let $\bar{\omega}(p,\varepsilon) := F(F^{-1}(p) + \varepsilon)$. Then, $\bar{\omega}$ satisfies the semigroup property $\bar{\omega}(\bar{\omega}(p,\varepsilon),\varepsilon') = \bar{\omega}(p,\varepsilon+\varepsilon')$, and using this one can show that to prove $\pi(A^\varepsilon) \geq \omega(\pi(A))$ it suffices to consider $\varepsilon \searrow 0$. A Taylor expansion yields

$$\begin{aligned}\pi(A^\varepsilon) &\geq \pi(A) + \pi^+(A)\,\varepsilon + o(\varepsilon)\,,\\ \bar{\omega}(\pi(A),\varepsilon) &= \pi(A) + \mathcal{I}\big(\pi(A)\big)\,\varepsilon + o(\varepsilon)\,,\end{aligned}$$

from which we deduce that $\pi^+(A) \geq \mathcal{I}(\pi(A))$ for all A if and only if $\pi(A^\varepsilon) \geq \bar{\omega}(\pi(A),\varepsilon)$ for all A and all $\varepsilon > 0$. □

2.5.2 Cheeger Isoperimetry

We now consider a class of probability measures which is characterized by a lower bound on the isoperimetric profile.

Definition 2.5.9. A probability measure π satisfies a **Cheeger isoperimetric in-**

equality with constant $\mathsf{Ch} > 0$ if for all Borel sets $A \subseteq \mathcal{X}$,

$$\pi^+(A) \geq \frac{1}{\mathsf{Ch}}\, \pi(A)\, \pi(A^c)\,. \tag{2.5.10}$$

For the two-sided exponential density $x \mapsto \mu(x) := \frac{1}{2}\exp(-|x|)$, the isoperimetric profile is known to be $\mathcal{I}_\mu(p) = \min(p, 1-p)$. Hence, the Cheeger isoperimetric inequality roughly asserts that the isoperimetric properties of π are at least as good as those of μ.

The inequality in (2.5.10) is the differential form of the inequality. By applying Theorem 2.5.8, one shows that the inequality (2.5.10) implies, for any $\varepsilon \in [0, \mathsf{Ch}]$,

$$\pi(A^\varepsilon) - \pi(A) \geq \frac{\varepsilon}{2\,\mathsf{Ch}}\, \pi(A)\, \pi(A^c)\,. \tag{2.5.11}$$

For all $\varepsilon > 0$, Theorem 2.5.8 also implies that

$$\alpha_\pi(\varepsilon) \leq \exp\Big(-\frac{\varepsilon}{\mathsf{Ch}}\Big)\,,$$

so π enjoys at least subexponential concentration.

Such isoperimetric inequalities will play a key role when we study Metropolis-adjusted sampling algorithms in Chapter 7. For now, however, our goal is to establish an equivalence between the Cheeger isoperimetric inequality and a functional version of it.

To pass from a functional inequality to an inequality involving sets, we can usually apply the functional inequality to the indicator of a set. To go the other way around, we need to represent a function via its level sets, which is achieved via the coarea inequality.

Theorem 2.5.12 (coarea inequality). *Let $f : \mathcal{X} \to \mathbb{R}$ be Lipschitz. Then,*

$$\int \|\nabla f\|\, \mathrm{d}\pi \geq \int_{-\infty}^{\infty} \pi^+\{f > t\}\, \mathrm{d}t\,.$$

***Remark* 2.5.13.** On a general metric space $(\mathcal{X}, \mathrm{d})$, we define

$$\|\nabla f\|(x) := \limsup_{y \in \mathcal{X},\, \mathrm{d}(x,y) \searrow 0} \frac{|f(x) - f(y)|}{\mathrm{d}(x,y)}\,.$$

In "nice" spaces, the coarea inequality is actually an equality, but we will not need this.

Proof. By an approximation argument we may assume that f is bounded, and by adding a constant to f we may suppose $f \geq 0$. Let $f_\varepsilon(x) := \sup_{\mathrm{d}(x,\cdot)<\varepsilon} f$ and $A_t := \{f > t\}$. We

can check that $A_t^\varepsilon = \{f_\varepsilon > t\}$, and for $g \geq 0$ we have the formula $\int g \, \mathrm{d}\pi = \int_0^\infty \pi\{g > t\} \, \mathrm{d}t$. By applying this to $g = f$ and $g = f_\varepsilon$,

$$\int \frac{f_\varepsilon - f}{\varepsilon} \, \mathrm{d}\pi = \int_0^\infty \frac{\pi(A_t^\varepsilon) - \pi(A_t)}{\varepsilon} \, \mathrm{d}t \, .$$

Now let $\varepsilon \searrow 0$ using Fatou's lemma and dominated convergence. □

Theorem 2.5.14. *Let $\pi \in \mathcal{P}_1(\mathcal{X})$ and let $\mathsf{Ch} > 0$. The following are equivalent.*

1. *π satisfies a Cheeger isoperimetric inequality with constant Ch.*
2. *For all Lipschitz $f : \mathcal{X} \to \mathbb{R}$, it holds that*

$$\mathbb{E}_\pi |f - \mathbb{E}_\pi f| \leq 2 \, \mathsf{Ch} \, \mathbb{E}_\pi \|\nabla f\| \, . \tag{2.5.15}$$

Proof sketch. (2) $\implies$ (1): Apply (2.5.15) to an approximation f of the indicator function $\mathbb{1}_A$, so that $\mathbb{E}_\pi |f - \mathbb{E}_\pi f| \approx 2\,\pi(A)\,(1 - \pi(A))$ and $\mathbb{E}_\pi \|\nabla f\| \approx \pi^+(A)$.

(1) $\implies$ (2): Let $A_t := \{f > t\}$. Applying the coarea inequality and the Cheeger isoperimetric inequality,

$$\begin{aligned}
2 \, \mathsf{Ch} \, \mathbb{E}_\pi \|\nabla f\| &\geq 2 \, \mathsf{Ch} \int_{-\infty}^\infty \pi^+\{f > t\} \, \mathrm{d}t \\
&\geq 2 \int_{-\infty}^\infty \pi(A_t) \, \pi(A_t^\mathsf{c}) \, \mathrm{d}t = \int_{-\infty}^\infty \mathbb{E}_\pi |\mathbb{1}_{A_t} - \pi(A_t)| \, \mathrm{d}t \\
&\geq \sup_{\|g\|_{L^\infty(\pi)} \leq 1} \int_{-\infty}^\infty \Big(\int g \, \{\mathbb{1}_{A_t} - \pi(A_t)\} \, \mathrm{d}\pi \Big) \, \mathrm{d}t \\
&= \sup_{\|g\|_{L^\infty(\pi)} \leq 1} \int_{-\infty}^\infty \Big(\int \{g - \mathbb{E}_\pi g\} \, \mathbb{1}_{A_t} \, \mathrm{d}\pi \Big) \, \mathrm{d}t = \sup_{\|g\|_{L^\infty(\pi)} \leq 1} \int \{g - \mathbb{E}_\pi g\} \, f \, \mathrm{d}\pi \\
&= \mathbb{E}_\pi |f - \mathbb{E}_\pi f| \, .
\end{aligned}$$

□

2.5.3 L^p–L^q Poincaré Inequalities

In this section, we work on Euclidean space for simplicity.

The inequality (2.5.15) can be considered an "L^1 variant" of the Poincaré inequality. More generally, we can define the following family of inequalities.

Definition 2.5.16 (L^p–L^q Poincaré inequality)**.** For $p, q \in [1, \infty]$ with $q \geq p$, the L^p–L^q **Poincaré inequality** asserts that for all smooth $f : \mathbb{R}^d \to \mathbb{R}$,

$$\|f - \mathbb{E}_\pi f\|_{L^p(\pi)} \leq C_{p,q} \left\| \|\nabla f\| \right\|_{L^q(\pi)}.$$

In this new notation, the usual Poincaré inequality is an L^2–L^2 Poincaré inequality with $C_{2,2} = \sqrt{C_{\mathrm{PI}}}$, whereas the inequality (2.5.15) is an L^1–L^1 Poincaré inequality.

These inequalities form a hierarchy via Hölder's inequality.

Proposition 2.5.17 ([Mil09])**.** *Suppose $p, q, \hat{p}, \hat{q} \in [1, \infty]$ are such that $p \leq \hat{p}$ and $q \leq \hat{q}$, and $p^{-1} - q^{-1} = \hat{p}^{-1} - \hat{q}^{-1}$. Then,*

$$C_{\hat{p},\hat{q}} \lesssim \frac{\hat{p}}{p} C_{p,q}.$$

Proof. Let f satisfy $\mathrm{med}_\pi f = 0$, which we can arrange by adding a constant. Define the function $g := (\operatorname{sgn} f)\, |f|^{\hat{p}/p}$, which still satisfies $\mathrm{med}_\pi g = 0$. By using the equivalence between the mean and the median (Lemma 2.4.4) and applying the L^p–L^q Poincaré inequality to g together with Hölder's inequality,

$$\begin{aligned}
\|f - \mathrm{med}_\pi f\|_{L^{\hat{p}}(\pi)}^{\hat{p}/p} = \|g - \mathrm{med}_\pi g\|_{L^p(\pi)} &\lesssim \|g - \mathbb{E}_\pi g\|_{L^p(\pi)} \leq C_{p,q} \left\| \|\nabla g\| \right\|_{L^q(\pi)} \\
&= \frac{\hat{p}}{p} C_{p,q} \left\| |f|^{\hat{p}/p - 1} \|\nabla f\| \right\|_{L^q(\pi)} \\
&= \frac{\hat{p}}{p} C_{p,q} \|f - \mathrm{med}_\pi f\|_{L^{\hat{p}}(\pi)}^{\hat{p}/p - 1} \left\| \|\nabla f\| \right\|_{L^{\hat{q}}(\pi)},
\end{aligned}$$

where we leave it to the reader to check that the exponents work out correctly. If we rearrange this inequality and apply Lemma 2.4.4 again, then

$$\|f - \mathbb{E}_\pi f\|_{L^{\hat{p}}(\pi)} \lesssim \|f - \mathrm{med}_\pi f\|_{L^{\hat{p}}(\pi)} \lesssim \frac{\hat{p}}{p} C_{p,q} \left\| \|\nabla f\| \right\|_{L^{\hat{q}}(\pi)}. \qquad \square$$

Thus, we have the following implications: for any $p \in (2, \infty)$,

$$(L^1\text{–}L^1) \implies (L^2\text{–}L^2) \implies \cdots \implies (L^p\text{–}L^p).$$

In particular, the first implication together with the equivalence in Theorem 2.5.14 shows that the Cheeger isoperimetric inequality implies the Poincaré inequality.

Also, given any L^p–L^q Poincaré inequality, by Jensen's inequality we can trivially make it weaker by decreasing p or increasing q; hence, every L^p–L^q Poincaré inequality implies an L^1–L^∞ Poincaré inequality. On the other hand, for any $1 \le p \le q < \infty$, an L^1–L^1 Poincaré implies an L^p–L^p Poincaré, which trivially implies an L^p–L^q Poincaré inequality. We conclude that *among these inequalities, the L^1–L^1 inequality is the strongest and the L^1–L^∞ inequality is the weakest.*

We now wish to sketch the proof of a deep result by E. Milman, which states that for log-concave measures, the hierarchy can be reversed. The formal statement is as follows.

Theorem 2.5.18 (reversing the hierarchy)**.** *Let $\pi \in \mathcal{P}(\mathbb{R}^d)$ be log-concave. Then,*

$$C_{1,1} \lesssim C_{1,\infty}\,.$$

As a consequence, suppose that π is α-strongly log-concave. By the Bakry–Émery theorem (Theorem 1.2.29), π satisfies a Poincaré inequality with $C_{2,2}^2 = C_{\mathsf{PI}} \le 1/\alpha$. By reversing the hierarchy, we see that this implies a Cheeger isoperimetric inequality.

Corollary 2.5.19. *If $\pi \in \mathcal{P}(\mathbb{R}^d)$ is α-strongly log-concave, then π satisfies a Cheeger isoperimetric inequality with constant* $\mathsf{Ch} \lesssim 1/\sqrt{\alpha}$.

The proof of Milman's theorem will require some preparations. The first fact that we need is a deep result in its own right. Typically it is proven with geometric measure theory by studying the isoperimetric problem, and we omit the proof.

Theorem 2.5.20. *If π is log-concave, then its isoperimetric profile $\mathcal{I}_\pi$ is concave.*

The isoperimetric profile satisfies $\mathcal{I}_\pi(0) = 0$, and it is symmetric around $\frac{1}{2}$, so it suffices to consider $p \in [0, \frac{1}{2}]$. By concavity,

$$\mathcal{I}_\pi(p) \ge \mathcal{I}_\pi\big(\frac{1}{2}\big)\, p\,. \tag{2.5.21}$$

Hence, in order to prove Cheeger's isoperimetric inequality, we need only find a suitable lower bound for $\mathcal{I}_\pi(\frac{1}{2})$.

The next idea is that instead of applying the L^1–L^∞ directly to an indicator function $\mathbb{1}_A$, we will first regularize $\mathbb{1}_A$ using the Langevin semigroup $(P_t)_{t\ge 0}$ with stationary distribution π. We start with a semigroup calculation.

Proposition 2.5.22. *Assume that the Markov semigroup $(P_t)_{t\geq 0}$ is reversible and satisfies the curvature-dimension condition $\mathsf{CD}(\alpha,\infty)$ for some $\alpha \in \mathbb{R}$.*

1. *For all f and $t \geq 0$,*

$$P_t(f^2) - (P_t f)^2 \geq \frac{\exp(2\alpha t) - 1}{\alpha}\, \Gamma(P_t f, P_t f)\,, \tag{2.5.23}$$

where we interpret $\frac{\exp(2\alpha t)-1}{\alpha} = 2t$ when $\alpha = 0$.

2. *If $\alpha = 0$, then for all $t > 0$ and $p \in [2,\infty]$,*

$$\left\| \sqrt{\Gamma(P_t f, P_t f)} \right\|_{L^p(\pi)} \leq \frac{1}{\sqrt{2t}} \|f\|_{L^p(\pi)} \tag{2.5.24}$$

and

$$\|f - P_t f\|_{L^1(\pi)} \leq \sqrt{2t} \left\| \sqrt{\Gamma(f,f)} \right\|_{L^1(\pi)}. \tag{2.5.25}$$

Proof. We recall the calculation that we performed for the local Poincaré inequality (Theorem 2.2.11):

$$\partial_s [P_s((P_{t-s}f)^2)] = 2P_s\Gamma(P_{t-s}f, P_{t-s}f)\,.$$

From the second statement of Theorem 2.2.11, we have

$$\Gamma(P_s g, P_s g) \leq \exp(-2\alpha s)\, P_s\Gamma(g,g)\,.$$

Taking $g = P_{t-s}f$, we deduce that

$$P_s\Gamma(P_{t-s}f, P_{t-s}f) \geq \exp(2\alpha s)\, \Gamma(P_t f, P_t f)\,.$$

Integrating this from $s = 0$ to $s = t$,

$$P_t(f^2) - (P_t f)^2 \geq 2\Gamma(P_t f, P_t f) \int_0^t \exp(2\alpha s)\, \mathrm{d}s = \frac{\exp(2\alpha t) - 1}{\alpha}\, \Gamma(P_t f, P_t f)\,.$$

This establishes the first inequality. In particular, for $\alpha = 0$,

$$\sqrt{\Gamma(P_t f, P_t f)} \leq \frac{1}{\sqrt{2t}} \sqrt{P_t(f^2)}$$

so that

$$\left\| \sqrt{\Gamma(P_t f, P_t f)} \right\|_{L^p(\pi)} \leq \frac{1}{\sqrt{2t}} \left\{ \mathbb{E}_\pi P_t(|f|^p) \right\}^{1/p} \leq \frac{1}{\sqrt{2t}} \|f\|_{L^p(\pi)} .$$

The last inequality is the dual of the $p = \infty$ case. Indeed, for g with $\|g\|_{L^\infty(\pi)} \leq 1$,

$$\partial_t \int (f - P_t f)\, g \,\mathrm{d}\pi = - \int P_t \mathscr{L} f \, g \,\mathrm{d}\pi = \int \Gamma(f, P_t g) \,\mathrm{d}\pi$$

and hence, by the Cauchy–Schwarz inequality for the carré du champ (Exercise 1.6),

$$\begin{aligned} \int (f - P_t f)\, g \,\mathrm{d}\pi &= \int_0^t \left(\int \Gamma(f, P_s g) \,\mathrm{d}\pi \right) \mathrm{d}s \leq \int_0^t \left(\int \sqrt{\Gamma(f,f)\, \Gamma(P_s g, P_s g)} \,\mathrm{d}\pi \right) \mathrm{d}s \\ &\leq \left\| \sqrt{\Gamma(f,f)} \right\|_{L^1(\pi)} \int_0^t \left\| \sqrt{\Gamma(P_s g, P_s g)} \right\|_{L^\infty(\pi)} \mathrm{d}s \\ &\leq \left\| \sqrt{\Gamma(f,f)} \right\|_{L^1(\pi)} \|g\|_{L^\infty(\pi)} \int_0^t \frac{1}{\sqrt{2s}} \,\mathrm{d}s \leq \sqrt{2t} \left\| \sqrt{\Gamma(f,f)} \right\|_{L^1(\pi)} . \end{aligned}$$ □

***Remark* 2.5.26.** The inequality (2.5.23) can be regarded as a "reverse Poincaré" inequality because it upper bounds the size of the gradient via a variance term. Similarly to Theorem 2.2.11, the inequality (2.5.23) can also be shown to be equivalent to $\mathrm{CD}(\alpha, \infty)$.

We are now ready to prove Milman's theorem.

Proof of Milman's theorem, Theorem 2.5.18. By approximating the indicator function $\mathbb{1}_A$ with a smooth function and applying the inequality (2.5.25), we can justify the bound

$$\sqrt{2t}\, \pi^+(A) \geq \|\mathbb{1}_A - P_t \mathbb{1}_A\|_{L^1(\pi)} .$$

Next, a calculation shows that

$$\begin{aligned} \mathbb{E}_\pi |\mathbb{1}_A - P_t \mathbb{1}_A| &= 2 \left\{ \pi(A)\, \pi(A^{\mathrm{c}}) - \mathbb{E}\left[\left(\mathbb{1}_A - \pi(A) \right) \left(P_t \mathbb{1}_A - \pi(A) \right) \right] \right\} \\ &\geq 2 \left\{ \pi(A)\, \pi(A^{\mathrm{c}}) - \underbrace{\|\mathbb{1}_A - \pi(A)\|_{L^\infty(\pi)}}_{\leq 1} \|P_t \mathbb{1}_A - \pi(A)\|_{L^1(\pi)} \right\} . \end{aligned}$$

From the L^1–L^∞ Poincaré inequality and (2.5.24),

$$\|P_t \mathbb{1}_A - \pi(A)\|_{L^1(\pi)} \leq C_{1,\infty} \left\| \|\nabla P_t \mathbb{1}_A\| \right\|_{L^\infty(\pi)} \leq \frac{C_{1,\infty}}{\sqrt{2t}} \|\mathbb{1}_A\|_{L^\infty(\pi)} \leq \frac{C_{1,\infty}}{\sqrt{2t}} .$$

Hence, we have

$$\sqrt{2t}\,\pi^+(A) \geq 2\left\{\pi(A)\,\pi(A^c) - \frac{C_{1,\infty}}{\sqrt{2t}}\right\}.$$

Now choose $\sqrt{2t} = 2C_{1,\infty}/(\pi(A)\,\pi(A^c))$ to obtain

$$\pi^+(A) \geq \frac{1}{2C_{1,\infty}}\,\pi(A)^2\,\pi(A^c)^2\,.$$

This inequality is not fully satisfactory, but if we take $\pi(A) = \frac{1}{2}$ then we deduce from this that $\mathcal{I}_\pi(\frac{1}{2}) \geq 1/(32C_{1,\infty})$, and from (2.5.21) we conclude. □

2.5.4 Gaussian Isoperimetry

The Cheeger isoperimetric inequality asserts that for small p, $\mathcal{I}_\pi(p) \gtrsim p$. On the other hand, one can check that the Gaussian isoperimetric profile $\mathcal{I}_{\gamma_d}(p) = \phi(\Phi^{-1}(p))$ has the asymptotics $\mathcal{I}_{\gamma_d}(p) \sim p\sqrt{2\ln(1/p)}$ as $p \searrow 0$.

As with the Cheeger isoperimetric inequality, isoperimetry of Gaussian type can also be captured via a functional inequality. The following result is due to Bobkov.

Theorem 2.5.27 (Gaussian isoperimetry, functional form)**.** *Suppose that $\pi \in \mathcal{P}(\mathbb{R}^d)$ is α-strongly log-concave for some $\alpha > 0$. Then, for all $f : \mathbb{R}^d \to [0,1]$,*

$$\sqrt{\alpha}\,\mathcal{I}_{\gamma_d}(\mathbb{E}_\pi f) \leq \mathbb{E}_\pi \sqrt{\alpha\,\mathcal{I}_{\gamma_d}(f)^2 + \Gamma(f,f)}\,. \tag{2.5.28}$$

As this formulation suggests, Theorem 2.5.27 has a proof via Markov semigroup theory, for which we refer readers to [BGL14, §8.5.2]. By converting the functional inequality back into an isoperimetric statement, one can deduce the following comparison theorem.

Theorem 2.5.29 (Gaussian isoperimetry comparison theorem)**.** *Suppose that the measure $\pi \in \mathcal{P}(\mathbb{R}^d)$ is α-strongly log-concave for some $\alpha > 0$. Then,*

$$\mathcal{I}_\pi \geq \sqrt{\alpha}\,\mathcal{I}_{\gamma_d}\,.$$

We explore these results further in the exercises.

2.6 Metric Measure Spaces

In this section, we revisit the curvature-dimension condition. As we hinted at in Section 2.2.1, the commutation relation which underlies the curvature-dimension condition captures the underlying curvature of both the ambient space and the measure. This observation leads not only to an extension of the ideas we have been considering thus far to weighted Riemannian manifolds, but in fact provides an avenue towards developing geometric analysis on non-smooth spaces which a priori have no differential structure. The key to this program is that, whereas the Ricci curvature tensor cannot be defined on such spaces, the convexity of the KL divergence w.r.t. an appropriate Wasserstein space continues to make sense as a "synthetic" notion of a Ricci curvature lower bound.

To aid the reader who is unfamiliar with Riemannian geometry, we will provide a brief review of the main concepts. In this section, we shall omit many of the proofs, as the goal is simply to acquaint the reader with the general picture in a geometric context without delving into the details.

2.6.1 Riemannian Geometry

Basic concepts. We recall some of the definitions from Section 1.3.2. A **Riemannian manifold** $\mathcal{M}$ is a space which is locally homeomorphic to a Euclidean space, such that at every point $p \in \mathcal{M}$ there is an associated vector space $T_p\mathcal{M}$, called the **tangent space** to $\mathcal{M}$ at p, equipped with an inner product $\langle \cdot, \cdot \rangle_p$. The tangent space $T_p\mathcal{M}$ represents the velocities of all curves passing through p. We can collect together the different tangent spaces into a single object called the **tangent bundle**,

$$T\mathcal{M} := \bigcup_{p \in \mathcal{M}} (\{p\} \times T_p\mathcal{M}) .$$

The Riemannian metric $p \mapsto \langle \cdot, \cdot \rangle_p$ is required to be smooth in a suitable sense.

A smooth function $f : \mathcal{M} \to \mathbb{R}$ has a **differential** $df : T\mathcal{M} \to \mathbb{R}$, defined as follows. Given a point $p \in \mathcal{M}$ and a tangent vector $v \in T_p\mathcal{M}$, let $(p_t)_{t \in \mathbb{R}}$ be a curve on $\mathcal{M}$ with $p_0 = p$ and with velocity v at time 0. Then, $(df)_p v := \partial_t|_{t=0} f(p_t)$. One can check that this definition does not depend on the choice of curve $(p_t)_{t \in \mathbb{R}}$ and that $(df)_p$ is a linear function on $T_p\mathcal{M}$. Note that the differential can be defined on any manifold, even if it does not have a Riemannian structure, but $(df)_p$ is not an element of $T_p\mathcal{M}$; it is an element of the dual space $T_p^*\mathcal{M}$, called the **cotangent space**. The Riemannian metric allows us to identify $(df)_p$ with an element of $T_p\mathcal{M}$: there is a unique vector $\nabla f(p) \in T_p\mathcal{M}$ such that for all $v \in T_p\mathcal{M}$, it holds that $(df)_p v = \langle \nabla f(p), v \rangle_p$. The vector $\nabla f(p)$ is called the **gradient** of f at p. The gradient depends on the choice of the metric, and we can then

define the **gradient flow** of f to be a curve $(p_t)_{t\geq 0}$ such that the velocity $\dot{p}_t$ of the curve equals $-\nabla f(p_t)$ for all $t \geq 0$.

We pause to give a simple example. Suppose that $\mathcal{M} = \mathbb{R}^d$, and we pick a smooth mapping $p \mapsto A_p$ where A_p is a positive definite $d \times d$ matrix for each $p \in \mathbb{R}^d$. This induces a Riemannian metric via $\langle u, v\rangle_p := \langle u, A_p\, v\rangle$, where $\langle \cdot, \cdot \rangle$ (without a subscript) denotes the usual Euclidean inner product. If $(p_t)_{t\in\mathbb{R}}$ is a smooth curve in $\mathbb{R}^d$ and its usual time derivative is $(\dot{p}_t)_{t\in\mathbb{R}}$, then we know that $\partial_t f(p_t) = \langle \nabla f(p_t), \dot{p}_t\rangle = \langle A_{p_t}^{-1}\, \nabla f(p_t), \dot{p}_t\rangle_{p_t}$. Hence, the manifold gradient $\nabla_{\mathcal{M}} f$ is given by $\nabla_{\mathcal{M}} f(p) = A_p^{-1}\, \nabla f$, where ∇f is the Euclidean gradient. When $A_p = \nabla^2\phi(p)$ is obtained as the Hessian of a mapping ϕ, then $\mathcal{M}$ is called a **Hessian manifold**.

A **vector field** on $\mathcal{M}$ is a mapping $X : \mathcal{M} \to T\mathcal{M}$ such that $X(p) \in T_p\mathcal{M}$ for all $p \in \mathcal{M}$.[7] A single vector $v \in T_p\mathcal{M}$ can be thought of as a differential operator; for $f \in C^\infty(\mathcal{M})$, we can define the action of v on f via $v(f) := (df)_p v$. Similarly, a vector field X acts on f and produces a function $Xf : \mathcal{M} \to \mathbb{R}$, defined by $Xf(p) = X(p)f$ for all $p \in \mathcal{M}$. For example, on $\mathbb{R}^d$, a vector field can be identified with a mapping $\mathbb{R}^d \to \mathbb{R}^d$, and it differentiates functions via $Xf(p) = \langle \nabla f(p), X(p)\rangle$.

We would also like to differentiate vector fields along other vector fields, and there are two main ways of doing so. The first is called the Lie derivative, and it can be defined on any smooth manifold without the need for a Riemannian metric, and is consequently less important for our discussion. The second is the **Levi–Civita connection**, which given vector fields X and Y, outputs another vector field $\nabla_X Y$. This connection is characterized by various properties, including compatibility with the Riemannian metric: for all vector fields X, Y, and Z, we have the chain rule

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle\,. \tag{2.6.1}$$

Here, the vector field Z is differentiating the scalar function $p \mapsto \langle X(p), Y(p)\rangle_p$. Since we do not aim to perform many Riemannian calculations here, we omit most of the other properties for simplicity. However, we mention one key fact, which is that for any smooth function f, it holds that $\nabla_{fX} Y = f\, \nabla_X Y$, where fX is the vector field $(fX)(p) = f(p)\, X(p)$. This property implies that the mapping $(X, Y) \mapsto \nabla_X Y$ is *tensorial* in its first argument, that is, $(\nabla_X Y)(p)$ only depends on the value $X(p)$ of X at p.

The tensorial property of the Levi–Civita connection allows us to compute the derivative of a vector field Y along a curve $c : \mathbb{R} \to \mathcal{M}$. Namely, for $t \in \mathbb{R}$, we can define $D_c Y(t) := (\nabla_{\dot{c}(t)} Y)(c(t))$, which makes sense because we can extend $\dot{c}$ to a vector field X on $\mathcal{M}$ and deduce that $(\nabla_X Y)(c(t))$ only depends on $X(c(t)) = \dot{c}(t)$ (and not on the choice of extension X). Then, $D_c Y$ is called the **covariant derivative** of Y along the curve

[7] Geometers would say that X is a *section* of the tangent bundle.

c. From there, we can define the **parallel transport** of a vector $v_0 \in T_{c(0)}\mathcal{M}$ along the curve c to be the unique vector field $(v(t))_{t\in\mathbb{R}}$ defined along the curve c with $v(0) = v_0$ such that the covariant derivative vanishes: $D_c v = 0$. The parallel transport is a canonical way of identifying two different tangent spaces on $\mathcal{M}$. Due to compatibility with the metric, it has the property that if $c(0) = p$, $c(1) = q$, and $P_c v \in T_q\mathcal{M}$ denotes the parallel transport of $v \in T_p\mathcal{M}$ along c for time 1, then $P_c : T_p\mathcal{M} \to T_q\mathcal{M}$ is an *isometry*.

We already have seen the idea of a length-minimizing curve, or a **geodesic**. Recall that the Riemannian metric induces a distance on $\mathcal{M}$ via

$$\mathrm{d}(p,q) = \inf\left\{ \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)}\,\mathrm{d}t \;\middle|\; \gamma(0) = p,\ \gamma(1) = q \right\}.$$

If there is a minimizing constant-speed curve γ in this variational problem, we say that γ is a geodesic joining p and q. By taking the first variation of this problem, one shows that a necessary condition for γ to be a geodesic is for the covariant derivative of its velocity to vanish: $D_\gamma\dot\gamma = 0$. We will write this, however, with the more familiar notation $\ddot\gamma = 0$, which in Euclidean space means that there is zero acceleration (and hence Euclidean geodesics are straight lines). The converse is not true; if $\ddot\gamma = 0$, it does not imply that γ must be a shortest path between its endpoints (but it means that γ is locally a shortest path).

If $p \in \mathcal{M}$ and $v \in T_p\mathcal{M}$, then $\exp_p(v)$ is defined to be the endpoint (at time 1) of a constant-speed geodesic emanating from p with velocity v, if such a geodesic exists. In general, the exponential map may only be defined in a neighborhood of 0 on $T_p\mathcal{M}$. The logarithmic map is the inverse of the exponential map: given $q \in \mathcal{M}$, $\log_p(q)$ is the unique vector $v \in T_p\mathcal{M}$, if this is well-defined, such that $\exp_p(v) = q$.

Given a vector field X on $\mathcal{M}$, the **divergence** of X is the function $\operatorname{div} X : \mathcal{M} \to \mathbb{R}$ defined as follows: $(\operatorname{div} X)(p)$ is the trace[8] of the linear mapping $v \mapsto (\nabla_v X)(p)$ on $T_p\mathcal{M}$. Also, $f \in C^\infty(\mathcal{M})$, we define the **Hessian** of f at p to be the bilinear mapping $\nabla^2 f(p) : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ given by

$$\nabla^2 f(p)[v,w] = \langle \nabla_v \nabla f(p), w\rangle_p .$$

Even though we have not given all of the definitions precisely, we will now work through one example to give the reader the flavor of the computations. Suppose that $(p_t)_{t\in\mathbb{R}}$ is a curve on $\mathcal{M}$; then we know that $\partial_t f(p_t) = \langle \nabla f(p_t), \dot p_t\rangle_{p_t}$. If $g(p) := \langle \nabla f(p), \dot p\rangle_p$, then by definition we have $\partial_t^2 f(p_t) = \dot p_t(g)(p_t)$. By compatibility of the Levi–Civita connection with the metric (2.6.1), this equals $\langle \nabla_{\dot p_t}\nabla f(p_t), \dot p_t\rangle_{p_t} + \langle \nabla f(p_t), \nabla_{\dot p_t}\dot p_t\rangle_{p_t}$. The first term is

[8] The trace is defined as usual, namely if A is a linear mapping on $T_p\mathcal{M}$, then after choosing an arbitrary orthonormal basis $e_1,\dots,e_d$ of $T_p\mathcal{M}$ (w.r.t. the Riemannian metric), we have $\operatorname{tr} A = \sum_{i=1}^d \langle e_i, A\,e_i\rangle_p$.

$\nabla^2 f(p_t)[\dot{p}_t, \dot{p}_t]$. For the second term, if p is a geodesic, then the term $\nabla_{\dot{p}_t}\dot{p}_t$ vanishes. This shows that it is convenient to pick geodesic curves when computing Hessians.[9]

For $f \in C^\infty(\mathcal{M})$, the **Laplacian** of f is the function $\Delta f : \mathcal{M} \to \mathbb{R}$ defined by $\Delta f := \operatorname{tr} \nabla^2 f$. In the Riemannian setting, Δ is usually called the **Laplace–Beltrami operator**. The Riemannian metric induces a **volume measure**, which we always denote via $\mathfrak{m}$. Throughout, when we abuse notation to refer to the density of an absolutely continuous measure $\mu \in \mathcal{P}(\mathcal{M})$, we always refer to the density w.r.t. the volume measure, i.e., $\frac{\mathrm{d}\mu}{\mathrm{d}\mathfrak{m}}$. We have the integration by parts formula

$$\int \Delta f \, g \, \mathrm{d}\mathfrak{m} = \int f \, \Delta g \, \mathrm{d}\mathfrak{m} = -\int \langle \nabla f, \nabla g \rangle \, \mathrm{d}\mathfrak{m} \,,$$

provided that there are no boundary terms.

Curvature. For a two-dimensional surface, it is easier to define the notion of curvature: one has the **Gaussian curvature**, which associates to each point $p \in \mathcal{M}$ a single number $K(p) \in \mathbb{R}$. It is the product of the two principal curvatures at p. The celebrated *Theorema Egregium* ("remarkable theorem") of Gauss asserts that the Gaussian curvature is unchanged under local isometries, i.e., the Gaussian curvature is intrinsic to the surface. (In contrast, there are other *extrinsic* notions of curvature, such as the mean curvature, which rely on the embedding of the manifold in Euclidean space.)

In higher dimensions, we are not so fortunate and it requires much more geometric information to fully capture the idea of curvature. In fact, at each point $p \in \mathcal{M}$, we associate to it a 4-tensor, called the **Riemann curvature tensor**. It is defined as follows: given vector fields W, X, Y, and Z,

$$\mathsf{Riem}(W, X, Y, Z) := \langle \nabla_X \nabla_W Y - \nabla_W \nabla_X Y + \nabla_{[W,X]} Y, Z \rangle \,.$$

Here, $[W, X]$ is the **Lie bracket** of W and X, which is the vector field U defined as the commutator: $Uf := WXf - XWf$. This tensor is obviously an unwieldy object, and it is unclear whether anyone fully understands its complexities. Nevertheless, we may begin to get a handle on it by observing that at its core, it measures the lack of commutativity of certain differential operators, which we stated was the basis for curvature in Section 2.2.1.

[9] When computing first-order derivatives, it is only important that the first-order behavior of the curve is correct (i.e., the curve has the correct tangent vector). When computing second-order derivatives, it should come at no surprise that the second-order behavior of the curve begins to matter.

Incidentally, if $\nabla f(p_t) = 0$, i.e., we are at a *stationary point*, then the second term vanishes regardless of the curve p. Hence, the Hessian of f can be defined on any smooth manifold without the need for a Riemannian metric, provided that we restrict ourselves to stationary points of f. This observation is used heavily in Morse theory.

On Euclidean space, it vanishes: $\mathsf{Riem} = 0$. Also, the Riemann curvature tensor is fully determined by the **sectional curvatures** of $\mathcal{M}$: given a two-dimensional subspace S of $T_p\mathcal{M}$, the sectional curvature of S can be defined as the Gaussian curvature of the two-dimensional surface obtained by following geodesics with directions in S. Thus, we can view the Riemann curvature tensor as collecting together all of the curvature information from two-dimensional slices.

Luckily, the Riemann curvature tensor contains information that is too detailed for our purposes. With an eye towards probabilistic applications, we focus mainly on properties such as the distortion of volumes of balls along geodesics, which only requires looking at certain averages of the Riemann curvature. More specifically, for $u, v \in T_p\mathcal{M}$, let

$$\mathsf{Ric}_p(u, v) := \operatorname{tr} \mathsf{Riem}(u, \cdot, v, \cdot) .$$

The tensor Ric is called the **Ricci curvature tensor**. It is a powerful fact that many useful geometric and probabilistic consequences, such as diameter bounds and functional inequalities, are consequences of lower bounds on the Ricci curvature.

We also mention that one can further take the trace of the Ricci curvature tensor to arrive at a single scalar function, known as the **scalar curvature**, but we shall not use it in this book.

Diffusions on manifolds. Recall that on $\mathbb{R}^d$, the generator of the standard Brownian motion is $\frac{1}{2}\Delta$, where Δ is the Laplacian operator. On a manifold $\mathcal{M}$, we define standard Brownian motion $(B_t)_{t\geq 0}$ to be the unique $\mathcal{M}$-valued stochastic process with generator $\frac{1}{2}\Delta$, where Δ is now the Laplace–Beltrami operator. This means that for all smooth functions $f : \mathcal{M} \to \mathbb{R}$, we require $t \mapsto f(B_t) - f(B_0) - \int_0^t \frac{1}{2}\Delta f(B_s)\,\mathrm{d}s$ to be a local martingale.

More generally, a stochastic process $(Z_t)_{t\geq 0}$ has generator $\mathscr{L}$ if for all smooth functions $f : \mathcal{M} \to \mathbb{R}$, the process $t \mapsto f(Z_t) - f(Z_0) - \int_0^t \mathscr{L}f(Z_s)\,\mathrm{d}s$ is a local martingale. When the generator is $\mathscr{L}f = \Delta f - \langle \nabla V, \nabla f\rangle$ for a smooth function $V : \mathcal{M} \to \mathbb{R}$, this corresponds to a Langevin diffusion on the manifold. We informally write $\mathrm{d}Z_t = -\nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$, although the "+" symbol has to be interpreted carefully. Under some assumptions, the stationary distribution π of the Langevin diffusion has density $\pi \propto \exp(-V)$ w.r.t. the volume measure $\mathfrak{m}$.

Under appropriate assumptions on ∇V, the existence and uniqueness of the diffusion process on the manifold can be proven, e.g., via embedding the manifold in Euclidean space and using similar arguments as in Section 1.1.3.

Optimal transport on Riemannian manifolds. We conclude this section by discussing how the optimal transport problem can be generalized to Riemannian manifolds.

Recall from Exercise 1.11 that the optimal transport problem can be posed with other costs; in particular, we take the cost to be $c(x,y) = \mathrm{d}(x,y)^2$, where d is the distance induced by the Riemannian metric. Suppose, for simplicity, that $\mathcal{M}$ is compact and that μ is absolutely continuous (w.r.t. the volume measure). Then, there is a unique optimal transport map T from μ to ν, which is of the form $T(x) = \exp_x(\nabla\psi(x))$, where $-\psi$ is $\mathrm{d}^2/2$-concave.

Moreover, there is a formal Riemannian structure on $\mathcal{P}_{2,\mathrm{ac}}(\mathcal{M})$. We can formally define the tangent space at μ to be

$$T_\mu \mathcal{P}_{2,\mathrm{ac}}(\mathcal{M}) := \overline{\{\nabla\psi \mid \psi \in C^\infty(\mathcal{M})\}}^{L^2(\mu)},$$

equipped with the norm $\|\nabla\psi\|_\mu := \sqrt{\int \|\nabla\psi\|^2 \,\mathrm{d}\mu}$. Also, curves of measures are again characterized by the continuity equation

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0\,,$$

where the equation is to be interpreted in a weak sense: for any test function $\varphi \in C^\infty(\mathcal{M})$, for a.e. t, it holds that

$$\partial_t \int \varphi \,\mathrm{d}\mu_t = \int \langle \nabla\varphi, v_t\rangle \,\mathrm{d}\mu_t\,.$$

In short, aside from new technicalities introduced in the Riemannian setting (such as the presence of a cut locus[10]), most of the facts familiar to us from the Euclidean setting continue to hold when generalized appropriately. We refer to [Vil09] for more details.

2.6.2 Metric Geometry

We now depart from the setting of smooth manifolds and consider metric spaces $(\mathcal{X}, \mathrm{d})$.

Definition 2.6.2 (length)**.** Given a continuous curve $\gamma : [0,1] \to \mathcal{X}$, we define the **length** of γ to be

$$\operatorname{len}\gamma := \sup\left\{\sum_{i=1}^{n} \mathrm{d}\big(\gamma(t_i), \gamma(t_{i-1})\big) \;\middle|\; 0 \le t_0 < t_1 < \cdots < t_n \le 1\right\}.$$

We can check that this definition agrees with the usual notion of length on $\mathbb{R}^d$. By the triangle inequality, if $\gamma(0) = p$ and $\gamma(1) = q$, then $\mathrm{d}(p,q) \le \operatorname{len}\gamma$.

[10] Loosely speaking, the presence of a cut locus means that there are multiple minimizing geodesics connecting two points. Think for instance of the two poles of a sphere.

Definition 2.6.3. We say that $(\mathcal{X}, \mathrm{d})$ is a **geodesic space** if for all $p, q \in \mathcal{X}$, there is a constant-speed curve $\gamma : [0,1] \to \mathcal{X}$ such that $\gamma(0) = p$, $\gamma(1) = q$, and $\mathrm{d}(p,q) = \operatorname{len}\gamma$. Here, "constant speed" implies that for all $s, t \in [0,1]$,

$$\mathrm{d}(\gamma(s), \gamma(t)) = |s - t|\, \mathrm{d}(p,q)\,.$$

The curve γ is called the **geodesic** joining p to q.

Geodesic spaces are a broader class of spaces than Riemannian manifolds. In particular, they do not have to have a smooth structure, and they can have "kinks". For example, the Wasserstein space $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$ is not truly a Riemannian manifold, as it is infinite-dimensional (along with other issues, e.g., it is not locally homeomorphic to a Hilbert space), but from the considerations in Section 1.3.2 it follows that the Wasserstein space is a geodesic space. The study of geodesic spaces is called **metric geometry**, and a comprehensive treatment of this subject can be found in [BBI01].

There is a way to generalize the idea of a uniform bound on the sectional curvature to the setting of geodesic spaces. It is based on comparing the sizes of triangles in $\mathcal{X}$ with the corresponding sizes in a model space.

Definition 2.6.4 (model space)**.** Let $\kappa \in \mathbb{R}$. The **model space** $\mathbb{M}^2_\kappa$ of curvature κ is the standard two-dimensional Riemannian manifold with constant sectional curvature equal to κ, that is:

1. the hyperbolic plane $\mathbb{H}^2$ of curvature κ (that is, the usual hyperbolic plane but with metric rescaled by $1/\sqrt{-\kappa}$) if $\kappa < 0$;
2. the Euclidean plane $\mathbb{R}^2$ if $\kappa = 0$;
3. the rescaled sphere $\mathbb{S}^2/\sqrt{\kappa}$ if $\kappa > 0$.

Definition 2.6.5 (Alexandrov curvature)**.** Let $(\mathcal{X}, \mathrm{d})$ be a geodesic space and let $\kappa \in \mathbb{R}$. We say that $(\mathcal{X}, \mathrm{d})$ has **Alexandrov curvature bounded from below by** κ (resp. **from above by** κ) if the following holds. For any triple of points $a, b, c \in \mathcal{X}$, and any corresponding triple of points $\bar{a}, \bar{b}, \bar{c}$ in the model space $\mathbb{M}^2_\kappa$ such that

$$\mathrm{d}(a,b) = \mathrm{d}(\bar{a},\bar{b})\,, \qquad \mathrm{d}(a,c) = \mathrm{d}(\bar{a},\bar{c})\,, \qquad \mathrm{d}(b,c) = \mathrm{d}(\bar{b},\bar{c})\,,$$

for any $p \in \mathcal{X}$ in the geodesic joining a to c, and any $\bar{p} \in \mathbb{M}^2_\kappa$ in the geodesic joining $\bar{a}$ to $\bar{c}$ with $\mathrm{d}(a,p) = \mathrm{d}(\bar{a},\bar{p})$, it holds that $\mathrm{d}(b,p) \geq \mathrm{d}(\bar{b},\bar{p})$ (resp. $\mathrm{d}(b,p) \leq \mathrm{d}(\bar{b},\bar{p})$).

If such a curvature bound holds, then $(\mathcal{X}, \mathrm{d})$ is called an **Alexandrov space**.

Thus, triangles in $\mathcal{X}$ are thicker (resp. thinner) than their counterparts in $\mathbb{M}^2_\kappa$. The advantage of this definition is that it can be stated using only the metric (and geodesic) structure of $\mathcal{X}$. For the case when $\kappa = 0$, there is another useful reformulation.

Proposition 2.6.6. *Let $(\mathcal{X}, \mathrm{d})$ be a geodesic space. Then, $(\mathcal{X}, \mathrm{d})$ has Alexandrov curvature bounded below by 0 (resp. bounded above by 0) if and only if the following holds. For any constant-speed geodesic $(p_t)_{t\in[0,1]}$ in $\mathcal{X}$, any $q \in \mathcal{X}$, and any $t \in [0,1]$,*

$$\mathrm{d}(p_t, q)^2 \quad \geq \text{ (resp. } \leq) \quad (1-t)\,\mathrm{d}(p_0,q)^2 + t\,\mathrm{d}(p_1,q)^2 - t\,(1-t)\,\mathrm{d}(p_0,p_1)^2\,.$$

We saw in Exercise 1.13 that $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$ has non-negative Alexandrov curvature. One can show that a Riemannian manifold has section curvature bounded by κ if and only if the corresponding Alexandrov curvature bound holds.

Alexandrov curvature bounds enforce enough regularity that a satisfactory infinitesimal theory can be developed for Alexandrov spaces. For instance, one can define the notion of a *tangent cone*[11], and in the case of the Wasserstein space, its tangent cone coincides with the definition of the tangent space that we gave in Section 1.3.2; see [AGS08, §12.4] for details.

2.6.3 Geometry of Markov Semigroups

We now indicate how Markov semigroup proofs can be extended to the setting of a weighted Riemannian manifold $\mathcal{M}$ with a reference measure π which admits a density $\pi \propto \exp(-V)$ w.r.t. the volume measure $\mathfrak{m}$.

Consider the Langevin diffusion on $\mathcal{M}$ with generator $\mathscr{L}$ given by

$$\mathscr{L} f := \Delta f - \langle \nabla V, \nabla f \rangle\,.$$

As before, we can compute the carré du champ to be

$$\Gamma(f,f) = \|\nabla f\|^2\,.$$

For the iterated carré du champ,

$$\Gamma_2(f,f) = \frac{1}{2}\left\{\mathscr{L}(\|\nabla f\|^2) - 2\,\langle \nabla f, \nabla \mathscr{L} f\rangle\right\}$$

[11] In general, this is only a cone and not a vector space, because of the possibility of kinks.

$$= \frac{1}{2}\left\{\Delta(\|f\|^2) - 2\langle \nabla f, \nabla \Delta f\rangle\right\} + \langle \nabla f, \nabla^2 V\, \nabla f\rangle\,.$$

Unlike in Section 2.2.1, however, we now have to apply the Bochner identity

$$\frac{1}{2}\Delta(\|\nabla f\|^2) = \langle \nabla f, \nabla \Delta f\rangle + \|\nabla^2 f\|_{\mathrm{HS}}^2 + \mathrm{Ric}(\nabla f, \nabla f) \tag{2.6.7}$$

which shows that

$$\Gamma_2(f,f) = \|\nabla^2 f\|_{\mathrm{HS}}^2 + \langle \nabla f, (\mathrm{Ric} + \nabla^2 V)\,\nabla f\rangle\,.$$

Observe that in this formula, the curvature of the ambient space and the curvature of the measure are placed on an equal footing through the tensor $\mathrm{Ric} + \nabla^2 V$. If $\mathrm{Ric} + \nabla^2 V \geq \alpha$, in the sense that $\mathrm{Ric}(X,X) + \langle X, \nabla^2 V\, X\rangle \geq \alpha\,\|X\|^2$ for any vector field X on $\mathcal{M}$, then the curvature-dimension condition $\Gamma_2 \geq \alpha\Gamma$ holds. Since the proof of the Bakry–Émery theorem (Theorem 1.2.29) only relied on the $\mathrm{CD}(\alpha,\infty)$ condition (together with calculus rules for the Markov semigroup, such as the chain rule), the theorem continues to hold in the setting of weighted Riemannian manifolds.

Actually, we can refine the condition further as follows. If $\dim \mathcal{M} = d$, then

$$\|\nabla^2 f\|_{\mathrm{HS}}^2 \geq \frac{1}{d}\left(\mathrm{tr}\,\nabla^2 f\right)^2 = \frac{1}{d}\left(\Delta f\right)^2.$$

This observation motivates the following definition.

Definition 2.6.8. A Markov semigroup is said to satisfy the **curvature-dimension condition** with curvature lower bound α and dimension bound d, denoted $\mathrm{CD}(\alpha, d)$, if for all functions f,

$$\Gamma_2(f,f) \geq \alpha\,\Gamma(f,f) + \frac{1}{d}\left(\mathscr{L} f\right)^2. \tag{2.6.9}$$

As the name suggests, the following theorem holds.

Theorem 2.6.10. *Let $\mathcal{M}$ be a complete Riemannian manifold with volume measure $\mathfrak{m}$, and let $\alpha > 0$, $d \geq 1$. Consider the Markov semigroup associated with standard Brownian motion on $\mathcal{M}$. Then, the following two statements are equivalent.*

1. $\mathrm{CD}(\alpha, d)$ *holds.*
2. $\mathrm{Ric} \geq \alpha$ *and* $\dim \mathcal{M} \leq d$.

As an example, one can show that the unit sphere $\mathbb{S}^d$ satisfies $\mathrm{Ric} = d - 1$, so that the $\mathrm{CD}(d-1, d)$ condition holds. Then, using the Bakry–Émery theorem (Theorem 1.2.29), or by using Markov semigroup calculus to prove that the curvature-dimension condition implies Bobkov's functional form of the Gaussian isoperimetric inequality (Theorem 2.5.27; see [BGL14, Corollary 8.5.4]), one can now deduce results such as concentration on the sphere (Theorem 2.5.2).

Besides providing an abstract framework for deriving functional inequalities, it is worth noting that the condition (2.6.9) no longer makes any mention of the ambient space except through the Markov semigroup $(P_t)_{t\geq 0}$ and its associated operators $\mathscr{L}$, Γ, and Γ_2. This has led to a line of research investigating to what extent we can study the intrinsic geometry intrinsic associated with a Markov semigroup. Although we do not intend to survey the literature here, we show one illustrative example to give the flavor of the results. First, one shows that the $\mathrm{CD}(\alpha, d)$ condition implies a Sobolev inequality.

Theorem 2.6.11. *Consider a diffusion Markov semigroup satisfying the* $\mathrm{CD}(\alpha, d)$ *condition for some* $\alpha > 0$ *and* $d > 2$. *Then, for all* $p \in [1, \frac{2d}{d-2}]$ *and all functions* f,

$$\frac{1}{p-2}\left\{\left(\int |f|^p \,\mathrm{d}\pi\right)^{2/p} - \int f^2 \,\mathrm{d}\pi\right\} \leq \frac{d-1}{\alpha d}\int \Gamma(f,f)\,\mathrm{d}\pi\,. \tag{2.6.12}$$

From this Sobolev inequality, one can then deduce a diameter bound for the Markov semigroup. Here, the diameter is defined as follows:

$$\mathrm{diam}\big((P_t)_{t\geq 0}\big) := \sup\Big\{\pi\text{-}\operatorname*{ess\,sup}_{x,y\in\mathcal{X}} |f(x) - f(y)| \;\Big|\; \|\Gamma(f,f)\|_{L^\infty(\pi)} \leq 1\Big\}\,.$$

Theorem 2.6.13. *Suppose that the Markov semigroup* $(P_t)_{t\geq 0}$ *satisfies the Sobolev inequality* (2.6.12). *Then,*

$$\mathrm{diam}\big((P_t)_{t\geq 0}\big) \leq \pi\sqrt{\frac{d-1}{\alpha}}\,.$$

The diameter bound is sharp, as it is attained by the sphere, and together with Theorem 2.6.11 it recovers the classical **Bonnet–Myers diameter bound** from Riemannian geometry. Other geometric results obtained in this fashion include volume growth comparison results and heat kernel bounds.

2.6.4 The Lott–Sturm–Villani Theory of Synthetic Ricci Curvature

The other perspective with which we can encode geometry is the optimal transport perspective. Namely, in Section 1.4, we informally argued that in the Euclidean context, the α-strong convexity of the KL divergence $\mathsf{KL}(\cdot \,\|\, \pi)$ on $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is equivalent to the α-strong convexity of the potential V. At this stage, it is a perhaps expected, although still remarkable, fact that on a general weighted Riemannian manifold $\mathcal{M}$, the α-strong convexity of $\mathsf{KL}(\cdot \,\|\, \pi)$ on $(\mathcal{P}_2(\mathcal{M}), W_2)$ is equivalent to the $\mathsf{CD}(\alpha, \infty)$ condition, which in turn is equivalent to $\mathrm{Ric} + \nabla^2 V \geq \alpha$.

There are also ways to formulate the general $\mathsf{CD}(\alpha, d)$ condition via displacement convexity, but they are considerably more complicated, and we omit them for simplicity.

If $(\mathcal{X}, \mathsf{d})$ is a geodesic space, then $(\mathcal{P}_2(\mathcal{X}), W_2)$ is also a geodesic space, which is sufficient to define displacement convexity. Hence, we can work in the setting of Section 2.6.2, together with the additional data of a reference measure $\pi \in \mathcal{P}(\mathcal{X})$. In general, technical issues arise when geodesics on $\mathcal{X}$ can "branch" off into multiple geodesics, and so we ought to impose a mild non-branching assumption; however, we will ignore this technicality. We can then formulate the following definition.

Definition 2.6.14. Let $(\mathcal{X}, \mathsf{d}, \pi)$ be a metric measure space, where $(\mathcal{X}, \mathsf{d})$ is a geodesic space. Then, we say that $(\mathcal{X}, \mathsf{d}, \pi)$ satisfies the $\mathsf{CD}(\alpha, \infty)$ condition if for all measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathcal{X})$, there exists a constant-speed geodesic $(\mu_t)_{t \in [0,1]}$ joining μ_0 to μ_1 with

$$\mathsf{KL}(\mu_t \,\|\, \pi) \leq (1-t)\,\mathsf{KL}(\mu_0 \,\|\, \pi) + t\,\mathsf{KL}(\mu_1 \,\|\, \pi) - \frac{\alpha\, t\,(1-t)}{2}\, W_2^2(\mu_0, \mu_1)\,,$$

for all $t \in [0, 1]$.

We now pause to discuss the motivation behind the introduction of this definition. Unlike the statement $\mathrm{Ric} \geq \alpha$, which only makes sense on Riemannian manifolds (and hence requires a smooth structure), the above definition makes sense on a wider class of spaces, including non-smooth spaces. The question of to what extent the concept of curvature makes sense on non-smooth spaces is perhaps an interesting question in its own right, but it also arises even when one is solely interested in smooth Riemannian manifolds. Suppose, for instance, that we have a sequence of Riemannian manifolds $(\mathcal{M}_k)_{k \in \mathbb{N}}$ that is converging in some sense to a limit space $\mathcal{M}$; what properties of the sequence are preserved in the limit?

If we want to pass to the limit in the condition $\mathrm{Ric}^{\mathcal{M}_k} \geq \alpha$, then typically we would need the Ricci curvature tensors $\mathrm{Ric}^{\mathcal{M}_k}$ to be converging in the limit. Since curvature involves two derivatives of the metric, this holds if the sequence converges in a C^2 sense.

However, for some applications, this notion of convergence is too strong. Instead, it is common to work with **Gromov–Hausdorff convergence**, which is based on a notion of distance between metric spaces. More specifically, it metrizes the space[12] of compact metric spaces. Moreover, this notion of convergence is weak enough that it admits a useful compactness theorems.

As a consequence of the compactness theorem, a sequence of Riemannian manifolds $(\mathcal{M}_k)_{k\in\mathbb{N}}$ with a uniform upper bound on the diameter and a uniform lower bound on the Ricci curvature converges to a limit space $\mathcal{M}$ in the Gromov–Hausdorff topology. However, in this topology, the space of Riemannian manifolds with diameter $\leq D$ and with $\mathrm{Ric} \geq \alpha$ is not closed; the limit space $\mathcal{M}$ is not necessarily a Riemannian manifold. So what then is $\mathcal{M}$? It is a geodesic space, but understanding whether it can be said to satisfy "$\mathrm{Ric}^{\mathcal{M}} \geq \alpha$" requires developing a theory of Ricci curvature lower bounds that makes sense on such spaces.

An analogy is in order. For a function $f : \mathbb{R}^d \to \mathbb{R}$, convexity can be described via the Hessian, $\nabla^2 f \geq 0$, or via the property

$$f\big((1-t)\,x + t\,y\big) \leq (1-t)\,f(x) + t\,f(y)\,, \qquad \text{for all } x, y \in \mathbb{R}^d,\ t \in [0,1]\,.$$

The former definition only makes sense for C^2 functions, whereas the latter definition makes sense for any function. The former is called the *analytic* definition, whereas the definition is called *synthetic* definition. Although the analytic definition is often more intuitive, the synthetic definition is more general and more useful for technical arguments. For example, from the synthetic definition is apparent that convexity is preserved under pointwise convergence, whereas from the analytic definition one needs the stronger notion of C^2 convergence.

From this perspective, the definition of Alexandrov curvature bounds in Section 2.6.2 is the synthetic counterpart to *sectional curvature bounds* from Riemannian geometry. However, as we have already seen, sectional curvature bounds are often too strong for geometric purposes, as we can obtain a wide array of geometric consequences (spectral gap estimates, log-Sobolev and Sobolev inequalities, diameter bounds, volume growth estimates, heat kernel bounds, etc.) from *Ricci curvature lower bounds*. Here, the curvature-dimension condition provides us with **synthetic Ricci curvature lower bounds**.

By deducing geometric facts from the $\mathsf{CD}(\alpha,\infty)$ condition, one shows that spaces satisfying the $\mathsf{CD}(\alpha,\infty)$ condition, despite the lack of smoothness, enjoy many of the good properties shared by Riemannian manifolds satisfying $\mathrm{Ric} \geq \alpha$. To complete the program described in this section, we should ask whether synthetic Ricci curvature lower bounds are preserved under a weak notion of convergence. The correct notion to consider

[12] The space of all compact metric spaces is too large to be a set (it is a proper class). However, if we choose one representative from each isometry class of metric spaces, then this is a bona fide set.

is an extension of Gromov–Hausdorff convergence to take into account the reference measure, called measured Gromov–Hausdorff convergence.

Definition 2.6.15. Let $(\mathcal{X}_k, \mathrm{d}_k, \pi_k)_{k\in\mathbb{N}}$ be a sequence of compact metric measure spaces. We say that the sequence converges to $(\mathcal{X}, \mathrm{d}, \pi)$ in the **measured Gromov–Hausdorff topology** if there is a sequence $(f_k)_{k\in\mathbb{N}}$ of maps $f_k : \mathcal{X}_k \to \mathcal{X}$ with:

1. $\sup_{x_k, x_k' \in \mathcal{X}_k} |\mathrm{d}(f_k(x_k), f_k(x_k')) - \mathrm{d}_k(x_k, x_k')| = o(1)$;
2. $\sup_{x\in\mathcal{X}} \inf_{x_k\in\mathcal{X}_k} |\mathrm{d}(f_k(x_k), x)| = o(1)$;
3. $(f_k)_\# \pi_k \to \pi$ weakly.

The following stability result is a key achievement of the theory of synthetic Ricci curvature, arrived at simultaneously by Lott and Villani [LV09] and Sturm [Stu06a; Stu06b].

Theorem 2.6.16 (stability of synthetic Ricci curvature bounds). *Let $(\mathcal{X}_k, \mathrm{d}_k, \pi_k)_{k\in\mathbb{N}} \to (\mathcal{X}, \mathrm{d}, \pi)$ in the measured Gromov–Hausdorff topology. Let $\alpha \in \mathbb{R}$ and $d \geq 1$. If each $(\mathcal{X}_k, \mathrm{d}_k, \pi_k)$ satisfies $\mathrm{CD}(\alpha, d)$, then so does $(\mathcal{X}, \mathrm{d}, \pi)$.*

Note that we have not defined the $\mathrm{CD}(\alpha, d)$ condition for $d < \infty$ in this context; we refer readers to the original sources for the full treatment.

2.6.5 Discussion

A remark on the settings of the results. Throughout this chapter, we have not been careful to state in what generality the various results hold. Certainly the results hold on the Euclidean space $\mathbb{R}^d$, and with appropriate modifications they continue to hold on weighted Riemannian manifolds.

The results based on optimal transport (e.g., results on transport inequalities) typically hold on general Polish spaces. The theory of synthetic Ricci curvature makes sense on geodesic spaces (with mild regularity conditions).

The results based on Markov semigroup theory only require an abstract space $\mathcal{X}$ on which there is a Markov semigroup $(P_t)_{t\geq 0}$ satisfying various properties (e.g., a chain rule for the carré du champ). Although this usually arises from a diffusion on a Riemannian manifold, one can also start with a Dirichlet energy functional on a metric space and develop a theory of non-smooth analysis. See [AGS15] for further discussion on how the two approaches may be reconciled in a quite general setting.

Comparison between the two approaches. The discussion thus far has been rather abstract, and it may be difficult to grasp how the two main approaches (Bakry–Émery theory and optimal transport) can capture geometric information such as the curvature. Here, we will briefly provide some intuition for this connection following [Vil09, §14].

Starting with the optimal transport perspective, fix $x_0 \in \mathcal{M}$ and a mapping $\nabla\psi$. For $t \geq 0$, let $x_t := \exp(t\,\nabla\psi(x_0))$, and let $\delta > 0$. If $e_1, \dots, e_d$ be an orthonormal basis of $T_{x_0}\mathcal{M}$, in an abuse of notation let $x_0 + \delta e_i$ denote a point obtained by travelling along a curve emanating from x_0 with velocity e_i for time δ. The points $(x_0 + \delta e_i)_{i=1}^d$ form the vertices of a parallelepiped A_0^δ. On the other hand, for $t > 0$, we can consider pushing the point $x_0 + \delta e_i$ along the exponential map to obtain a new point $\exp_{x_0+\delta e_i}(t\,\nabla\psi(x_0 + \delta e_i))$. These points form the vertices of a new parallelepiped A_t^δ.

In terms of measures, let μ_0^δ denote the uniform measure on A_0^δ, and $\mu_t^\delta = \exp(t\,\nabla\psi)_\# \mu_0^\delta$, so that μ_t^δ is approximately the uniform measure on A_t^δ. Then, the displacement convexity of entropy states that

$$\ln \frac{1}{\mathfrak{m}(A_t^\delta)} \leq (1-t) \ln \frac{1}{\mathfrak{m}(A_0^\delta)} + t \ln \frac{1}{\mathfrak{m}(A_1^\delta)} + o(1)$$

as $\delta \searrow 0$. On the other hand, the infinitesimal change in volume is governed by the Jacobian determinant

$$\frac{\mathfrak{m}(A_t^\delta)}{\mathfrak{m}(A_0^\delta)} \to \mathcal{J}(t,x) := \det J(t,x)\,,$$

where $J_i(t,x) := \partial_\delta|_{\delta=0} \exp_{x_0+\delta e_i}(t\,\nabla\psi(x_0+\delta e_i))$. Hence, the displacement convexity yields

$$\ln \mathcal{J}(t,x) \geq (1-t) \ln \mathcal{J}(0,x) + t \ln \mathcal{J}(1,x)\,. \tag{2.6.17}$$

In Euclidean space, we have the formula $\mathcal{J}(t,x) = |\det(I_d + t\,\nabla^2\psi(x))|$, but the situation is more complicated on a Riemannian manifold because there is also a change of volume due to curvature. To account for this, one can derive an equation for J, known as the **Jacobi equation**:

$$\ddot{J}(t,x) + R(t,x)\,J(t,x) = 0\,,$$

where $R(t,x) := \mathrm{Riem}_{x_t}(\dot{x}_t, \cdot, \dot{x}_t, \cdot)$. By taking the trace and performing some computations, we arrive at

$$\partial_t^2 \mathcal{J}(t,x) = -\|J^{-1}(t,x)\,\dot{J}(t,x)\|_{\mathrm{HS}}^2 - \mathrm{Ric}_{x_t}(\dot{x}_t, \dot{x}_t)\,. \tag{2.6.18}$$

By comparing (2.6.17) and (2.6.18), we now obtain a hint as to how optimal transport captures curvature: displacement convexity of the entropy is related to concavity of the Jacobian determinant, which in turn is tied to Ricci curvature lower bounds.

The calculations above are performed with the Lagrangian description of fluid flows, as they follow a single trajectory $t \mapsto x_t$. If we switch to the Eulerian perspective, then we are led to define the vector field $\nabla\psi_t$ as follows: $\nabla\psi_t(x)$ is the velocity $\dot{x}_t$ of the curve $t \mapsto \exp_x(t\,\nabla\psi(x))$ at time t. By reformulating the Jacobi equation in the Eulerian perspective, we arrive precisely at the Bochner identity (2.6.7) for ψ which, as we saw in Section 2.6.3, underlies the curvature-dimension condition from the Bakry–Émery perspective. In this sense, the two approaches to curvature are dual.

2.7 Discrete Space and Time

Up until this point, we have been focusing on continuous-time Markov processes on a continuous state space. In this section, we give a few pointers on what may break down in discrete space or discrete time. Our treatment here is far from comprehensive.

Discrete space. For Markov processes on a discrete space space, we can still define the Markov semigroup, generator, carré du champ, and Dirichlet form. The main difference is that the carré du champ is now a finite difference operator, rather than a differential operator, and consequently it fails to satisfy a chain rule.

Crucially, this difference manifests itself for the log-Sobolev inequality, which we have written in this chapter as

$$\mathrm{ent}_\pi(f^2) \le 2C\,\mathscr{E}(f,f) \qquad \text{for all } f\,. \tag{2.7.1}$$

On the other hand, recall from Theorem 1.2.21 (which still holds for discrete state spaces) that the exponential decay of the KL divergence is equivalent to the inequality

$$\mathrm{ent}_\pi(f) \le \frac{C}{2}\,\mathscr{E}(f,\ln f) \qquad \text{for all } f \ge 0\,. \tag{2.7.2}$$

When the carré du champ satisfies a chain rule, then (2.7.1) and (2.7.2) are equivalent, but in general the first inequality (2.7.1) is strictly stronger.

Lemma 2.7.3. *The inequality* (2.7.1) *implies inequality* (2.7.2).

See Exercise 2.20. The first inequality (2.7.1) is often simply called the *log Sobolev inequality*, whereas the second inequality (2.7.2) is called a **modified log-Sobolev inequality** (**MLSI**). In many cases, the log-Sobolev inequality is too strong in that it does

not hold with a good constant C; hence, the modified log-Sobolev inequality is often the more appropriate inequality for the discrete setting.

We have already seen a concentration inequality for discrete spaces in Exercise 2.15. In general, concentration of measure on discrete spaces is a rich subject, with many applications to computer science and probability, and at the same time subtle, involving new ideas such as asymmetric transport inequalities or careful use of hypercontractivity. See, e.g., [BLM13; Han16] for more detailed treatments.

Discrete time. Similarly, for discrete-time Markov chains we can no longer use semigroup calculus, although the basic principles of Poincaré inequalities (spectral gap inequalities) and modified log-Sobolev inequalities can be adapted to this setting. In addition, there are new techniques based on the notion of *conductance*. As we shall need to study discrete-time Markov chains in detail for sampling algorithms, we defer a fuller discussion of this theory to Chapter 7.

Discrete curvature. Inspired by the geometric connections in Section 2.6, many researchers have attempted to define notions of curvature on discrete spaces. We do not attempt to survey this literature here, but we give a few pointers to the literature.

Ollivier [Oll07; Oll09] introduced the following notion of curvature.

Definition 2.7.4. A metric space $(\mathcal{X}, \mathsf{d})$ equipped with a Markov kernel P is said to have **coarse Ricci curvature** bounded below by $\kappa \in [0, 1]$ if for all $x, y \in \mathcal{X}$,

$$W_1\big(P(x,\cdot), P(y,\cdot)\big) \leq (1-\kappa)\,\mathsf{d}(x,y)\,.$$

In other words, the Markov chain with kernel P is a W_1 contraction. The definition is motivated by the following observation: on a d-dimensional Riemannian manifold with $\mathrm{Ric} \geq \alpha$, let $P(x,\cdot)$ be the uniform measure on $\mathsf{B}(x,\varepsilon)$. Then, provided that $\mathsf{d}(x,y) = O(\varepsilon)$, it holds that

$$W_1\big(P(x,\cdot), P(y,\cdot)\big) \leq \Big(1 - \frac{\alpha\varepsilon^2}{2\,(d+2)} + O(\varepsilon^3)\Big)\,\mathsf{d}(x,y)\,.$$

A lower bound on the coarse Ricci curvature is often too strong of an assumption for the purpose of studying mixing times of Markov chains, although there are refinements in [Oll07; Oll09]. However, when a lower bound on the coarse Ricci curvature holds, then it implies a number of useful consequences, such as concentration estimates and functional inequalities. We mention the following result in particular.

Theorem 2.7.5 ([Oll09]). *Suppose that P is a Markov kernel on a metric space $(\mathcal{X}, \mathrm{d})$, and that P has coarse Ricci curvature bounded below by κ. Then, P satisfies a Poincaré inequality with constant at most $1/\kappa$.*

Refer to Chapter 7 for a precise definition of the Poincaré inequality used here.

Other approaches for studying the curvature of discrete Markov processes include: studying the displacement convexity of entropy (using different interpolating curves rather than W_2 geodesics) [OV12; Goz+14; Léo17]; using ideas from Bakry–Émery theory [Kla+16; FS18]; and defining a modified W_2 distance for which the Markov process becomes a gradient flow of the KL divergence [Maa11; EM12; Mie13].

We emphasize that although we only described the coarse Ricci curvature approach in any detail, there is not a single approach which supersedes the others in the discrete setting. Each approach has its own merits and shortcomings.

Bibliographical Notes

The monographs [BGL14; Han16] are excellent sources to learn more about Markov semigroup theory.

The Monge–Ampère equation introduced in Exercise 2.1, being a fully non-linear PDE, is fairly difficult to study. See [Vil03, §4] for an overview of rigorous results on the Monge–Ampère equation, including the celebrated regularity theory of Caffarelli. The proofs of Proposition 2.1.1 and Exercise 2.3 are taken from [Cor17]. One might wonder whether a "log-Sobolev" version of the Brascamp–Lieb inequality holds, but the answer is unfortunately negative [BL00].

In the proof of Lemma 2.2.6, we assumed the solvability of the Poisson equation; this can be avoided via a density argument, see [CFM04; BC13]. The proof of the dimensional Brascamp–Lieb inequality in Exercise 2.4 is taken from the paper [BGG18], and Exercise 2.5 is from [HS94]. The bound on $\mathrm{var}_\pi V$ obtained in the exercise was used in [Che21b] to show that the entropic barrier is an optimal self-concordant barrier. Finally, we caution the reader that the Brascamp–Lieb inequality in Theorem 2.2.8 should not be confused with another family of inequalities, which are unfortunately also known as Brascamp–Lieb inequalities, described in, e.g., [Vil03, §6.3].

Although the device in the proof of Theorem 2.2.11 of differentiating $s \mapsto P_s((P_{t-s}f)^2)$ may seem mysterious at first glance, it forms the basis for a great number of useful inequalities. The key is that the chain rule for the carré du champ also implies a chain rule for the generator: $\mathscr{L}(\phi \circ f) = \phi'(f)\,\mathscr{L}f + \phi''(f)\,\Gamma(f,f)$. Using this, one can differentiate

$s \mapsto P_s\phi(P_{t-s}f)$ for a general function $\phi : \mathbb{R} \to \mathbb{R}$, and obtain the nice identity

$$\partial_s[P_s\phi(P_{t-s}f)] = P_s\big(\mathscr{L}\phi(P_{t-s}f) - \phi'(P_{t-s}f)\,\mathscr{L}P_{t-s}f\big) = P_s\big(\phi''(P_{t-s}f)\,\Gamma(P_{t-s}f, P_{t-s}f)\big)\,.$$

The book [BGL14] is a treasure trove of applications of this principle.

The convergence in Rényi divergence of the Langevin diffusion was obtained earlier, under stronger assumptions in [CLL19]. A natural question to ask is whether there are functional inequalities that interpolate between the Poincaré and log-Sobolev inequalities, which imply intermediate rates of convergence for the Langevin diffusion. One answer is given by the family of **Latała–Oleszkiewicz inequalities** (**LOI**) [LO00]. The convergence of the Langevin diffusion under an LO inequality is given in [Che+21a]. One can also consider variants of Sobolev inequalities [Cha04].

In [KLS95], Kannan, Lovász, and Simonovits conjectured that any log-concave measure π on $\mathbb{R}^d$ which is isotropic (i.e., if $X \sim \pi$ then $\operatorname{cov} X = I_d$) satisfies a Poincaré inequality with a dimension-free constant $C_{\mathsf{PI}} \lesssim 1$. This is known as the **Kannan–Lovász–Simonovits** (**KLS**) **conjecture**. By considering linear test functions of the form $x \mapsto \langle a, x\rangle$, one has $C_{\mathsf{PI}} \geq 1$, so the conjecture asserts that linear functions nearly saturate the spectral gap inequality for log-concave measures. The KLS conjecture has inspired a considerable amount of research (including Theorem 2.5.18), see [GM11; Eld13; LV17; Che21a; KL22], culminating in the current state-of-the-art result of [Kla23] which asserts that $C_{\mathsf{PI}} \lesssim \log d$.

The Prékopa–Leindler inequality given in Exercise 2.8 can be used to deduce a number of other functional inequalities, such as the log-Sobolev inequality and the Bregman transport inequality; see [BL00].

Exercise 2.11 essentially contains the main results of [CCN21] (actually the paper assumes a slightly weaker condition than (2.E.3), namely that the p-th moment of the chi-squared divergence is bounded for some $p > 1$, but this is handled with the same arguments as in Exercise 2.11).

There are many treatments on concentration of measure, e.g., [Led01; BLM13; BGL14; Han16; Ver18]. The proof of Lemma 2.4.4 is from [Mil09]. A proof of the characterization of the T_1 inequality in Theorem 2.4.11 can be found in, e.g., [BV05].

The proof of Sanov's theorem can be found in many textbooks on large deviations, e.g., [DZ10; RS15].

The monographs [BH97; Led01; BGL14] are excellent sources to learn about isoperimetry. The exposition of the functional form of Cheeger's inequality (Theorem 2.5.14) as well as Milman's theorem (Theorem 2.5.18) were inspired by the treatment in [AB15]. It would be hard to survey the various developments on this subject here, but we would like to mention a few nice additions to the story. First, as we saw in Theorem 2.5.14 and Proposition 2.5.17, isoperimetric inequalities are typically stronger than their functional inequality counterparts, and often strictly so. In order to obtain inequalities involving sets which

are *equivalent* to, say, the Poincaré and log-Sobolev inequalities, one should turn towards *measure capacity inequalities*, for which we refer the reader to [BGL14, §8]. Also, more refined "two-level" isoperimetric inequalities have been pioneered by Talagrand in [Tal91], which has applications in its own right.

The Gaussian isoperimetric inequality is due to Sudakov and Tsirelson [SC74] and Borell [Bor75]. It has since been extended and refined in various ways, e.g., in the context of noise stability [Bor85; IM12; Eld15; MN15; KKO18].

Section 2.6 draws upon many resources on geometry, which we list here: [Car92] for Riemannian geometry; [Hsu02] for diffusions on manifolds; [Vil09] for optimal transport on manifolds, including synthetic Ricci curvature bounds and the discussion in Section 2.6.5; [BBI01; Gro07] for metric geometry; and [Led00; BGL14] for the geometry of Markov semigroups. The curvature-dimension condition and its equivalences have been explored in a vast number of works, e.g., [Stu06a; Stu06b; LV09; Wan11].

Although the bounded differences inequality from Exercise 2.15 is already quite powerful, there are situations in which it does not give the correct answer, in which case we must turn towards more powerful tools. Among these, we mention Talagrand's convex distance inequality [Tal96], which can be established via the tensorization argument of Theorem 2.3.17 (see [Mar96]).

Exercises

Overview of the Inequalities

▷ **Exercise 2.1** (linearization of the Monge–Ampère equation)
In general, when $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ have smooth densities and $\nabla\varphi$ denotes the optimal transport map from μ to ν, then from the change of variables formula we expect

$$\frac{\mu}{\nu \circ \nabla\varphi} = \det \nabla^2 \varphi \,.$$

This is known as the **Monge–Ampère equation**. It is a non-linear PDE in the variable φ, which is a convex function (by Brenier's theorem, see Theorem 1.3.8). In this exercise, we linearize the Monge–Ampère equation to gain insight into the infinitesimal behavior of the optimal transport problem.

Let μ be a probability measure on $\mathbb{R}^d$ with a smooth density, and let $f \in C_c^\infty(\mathbb{R}^d)$ satisfy $\int f \,\mathrm{d}\mu = 0$. Let $\mu_\varepsilon := (1 + \varepsilon f)\,\mu$, and let $\nabla\varphi_\varepsilon$ denote the optimal transport map from μ to μ_ε. Assuming that $\varphi_\varepsilon(x) = \frac{\|x\|^2}{2} + \varepsilon u(x) + o(\varepsilon)$ for some function $u : \mathbb{R}^d \to \mathbb{R}$, perform an expansion of the Monge–Ampère equation in ε and argue that u satisfies the

following linear PDE, known as the **Poisson equation**:

$$-\mathscr{L}u = f\,, \qquad \text{where} \qquad \mathscr{L}u \coloneqq \Delta u - \Big\langle \nabla \ln \frac{1}{\mu}\,, \nabla u \Big\rangle\,.$$

Note that $\mathscr{L}$ is the generator of the Langevin diffusion with stationary distribution μ (see Example 1.2.4). Use this to formally argue that

$$\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon^2}\, W_2^2\big(\mu,\ (1+\varepsilon f)\,\mu\big) = \int \|\nabla u\|^2 \,\mathrm{d}\mu = \int u\,(-\mathscr{L})\,u \,\mathrm{d}\mu = \int f\,(-\mathscr{L})^{-1} f \,\mathrm{d}\mu\,.$$

Here, $\int \|\nabla u\|^2 \,\mathrm{d}\mu = \int u\,(-\mathscr{L})\,u\,\mathrm{d}\mu$ is the squared **Sobolev norm** $\|u\|^2_{\dot{H}^1(\mu)}$, where the dot is used to distinguish this from the usual Sobolev norm $\|u\|^2_{H^1(\mu)} = \|u\|^2_{L^2(\mu)} + \|u\|^2_{\dot{H}^1(\mu)}$. Similarly, $\int f\,(-\mathscr{L})^{-1} f\,\mathrm{d}\mu$ is the squared **inverse Sobolev norm** $\|f\|^2_{\dot{H}^{-1}(\mu)}$. Therefore, the linearization result shows that $W_2^2(\mu,\nu) \sim \|\mu-\nu\|^2_{\dot{H}^{-1}(\mu)}$ as $\nu \to \mu$.

Using the linearization (1.E.1) of the KL divergence from Exercise 1.8, deduce that the $\mathsf{T}_2(C)$ inequality implies

$$C\int f^2 \,\mathrm{d}\mu \geq \int f\,(-\mathscr{L})^{-1} f \,\mathrm{d}\mu\,.$$

In light of the spectral gap interpretation of the Poincaré inequality, why does the above inequality suggest that $\mathsf{T}_2(C)$ implies $\mathsf{PI}(C)$?

The astute reader should also work out how the Poisson equation can be obtained starting with the continuity equation (1.3.18).

Proofs via Markov Semigroup Theory

▷ Exercise 2.2 (curvature-dimension condition)
Verify the commutation identity (2.2.4) and deduce the formula (2.2.5) for the iterated carré du champ operator.

▷ Exercise 2.3 (Bregman transport inequality)
Let $\nabla\varphi$ denote the optimal transport map from π to μ, so that the Monge–Ampère equation holds (see Exercise 2.1):

$$\frac{\pi}{\mu \circ \nabla\varphi} = \det \nabla^2 \varphi\,.$$

Take logarithms of both sides of this equation and integrate w.r.t. π to prove the Bregman transport inequality (Theorem 2.2.10). Then, by applying Proposition 2.1.1, give another proof of the Brascamp–Lieb inequality (Theorem 2.2.8).

▷ Exercise 2.4 (dimensional improvement of the Brascamp–Lieb inequality)
In finite-dimensional space, one can improve upon the Brascamp–Lieb inequality (Theorem 2.2.8) by subtracting a non-negative term from the right-hand side. There are different ways to do this, but in this exercise we explore an approach which utilizes the extra term $\|\nabla^2 u\|_{\mathrm{HS}}^2$ in the iterated carré du champ operator.

1. Let $\pi \propto \exp(-V)$, where as before we assume that V is twice continuously differentiable and strictly convex. Let f satisfy $\mathbb{E}_\pi f = 0$, and consider another function u (not necessarily the solution to $-\mathscr{L}u = f$). Show that

$$\mathbb{E}_\pi[f^2] \leq \mathbb{E}_\pi[(f + \mathscr{L}u)^2] + \mathbb{E}_\pi\langle \nabla f, (\nabla^2 V)^{-1} \nabla f\rangle - \mathbb{E}_\pi[\|\nabla^2 u\|_{\mathrm{HS}}^2] \,.$$

2. Prove that $\mathbb{E}_\pi[\|\nabla^2 u\|_{\mathrm{HS}}^2] \geq d^{-1} (\mathbb{E}_\pi \Delta u)^2$, and that

$$\mathbb{E}_\pi \Delta u = \mathrm{cov}_\pi(f, V) - \mathbb{E}_\pi[(f + \mathscr{L}u)\, V] \,.$$

3. Choose u to solve $-\mathscr{L}u = f + \lambda (V - \mathbb{E}_\pi V)$ for some $\lambda \geq 0$ and substitute this into the previous parts. Optimize over λ and prove that

$$\mathrm{var}_\pi f \leq \mathbb{E}_\pi\langle \nabla f, (\nabla^2 V)^{-1} \nabla f\rangle - \frac{\mathrm{cov}_\pi(f, V)^2}{d - \mathrm{var}_\pi V} \,.$$

4. In particular, deduce that

$$\mathrm{var}_\pi V \leq \frac{d\, \mathbb{E}_\pi\langle \nabla V, (\nabla^2 V)^{-1} \nabla V\rangle}{d + \mathbb{E}_\pi\langle \nabla V, (\nabla^2 V)^{-1} \nabla V\rangle} \leq d \,.$$

▷ Exercise 2.5 (Helffer–Sjöstrand identity)
Let $\mathscr{L}$ denote the generator corresponding to $\pi \propto \exp(-V)$, $\nabla^2 V > 0$. Heuristically derive the following *identity*: for all $f, g : \mathbb{R}^d \to \mathbb{R}$,

$$\mathrm{cov}_\pi(f, g) = \mathbb{E}_\pi\langle \nabla f, (-\mathscr{L} + \nabla^2 V)^{-1} \nabla g\rangle \,.$$

Note that since $-\mathscr{L} \geq 0$, this implies the Brascamp–Lieb inequality (Theorem 2.2.8)!

Hint: Let $(\cdot)^*$ denote the adjoint in $L^2(\pi)$. It suffices to prove that $\nabla^* (-\mathscr{L} + \nabla^2 V)^{-1} \nabla$ is the orthogonal projection onto $1^\perp$. Suppose that $L^2(\pi)$ admits a basis of eigenfunctions for $-\mathscr{L}$, and let $u : \mathbb{R}^d \to \mathbb{R}$ be such an eigenfunction with $-\mathscr{L}u = \lambda u$, $\lambda > 0$. Study the effect of the operator on u, recalling (1.2.15) and (2.2.4).

▷ **Exercise 2.6** (local Poincaré inequality)
Prove the implication (3) $\implies$ (1) in Theorem 2.2.11.

Hint: Perform a Taylor expansion of both sides of (3) up to order $o(t^2)$.

▷ **Exercise 2.7** (hypercontractivity)
Let $(P_t)_{t\geq 0}$ be a reversible Markov semigroup with stationary distribution π, and let $\alpha \geq 0$. Show that the log-Sobolev inequality in the form

$$\operatorname{ent}_\pi(f^2) \leq 2C_{\mathsf{LSI}}\,\mathscr{E}(f,f)$$

for all f is equivalent to the following **hypercontractivity** statement: for all functions f, $t \geq 0$, and $p \geq 1$, if we set $p(t) := 1 - (p-1)\exp(2t/C_{\mathsf{LSI}})$, then

$$\|P_t f\|_{L^{p(t)}(\pi)} \leq \|f\|_{L^p(\pi)}\,.$$

This is a strengthening of the fact that the semigroup is a contraction on any $L^p(\pi)$ space and shows that in fact the semigroup maps $L^p(\pi)$ into $L^{p'}(\pi)$ for some $p' > p$.

Hint: Differentiate $t \mapsto \ln \|P_t f\|_{L^{p(t)}(\pi)}$.

▷ **Exercise 2.8** (Prékopa–Leindler inequality)
In this exercise, we introduce another important functional inequality, known as the **Prékopa–Leindler inequality**.

1. Let π be α-strongly log-concave, let $t \in [0,1]$, and let $f, g, h : \mathbb{R}^d \to \mathbb{R}_{>0}$ be three functions such that for all $x, y \in \mathbb{R}^d$,

$$\ln h\big((1-t)\,x + t\,y\big) \geq (1-t)\ln f(x) + t\ln g(y) - \frac{\alpha}{2}\,t\,(1-t)\,\|x-y\|^2\,.$$

Prove that

$$\ln \int h\,\mathrm{d}\pi \geq (1-t)\ln \int f\,\mathrm{d}\pi + t\ln \int g\,\mathrm{d}\pi\,. \tag{2.E.1}$$

Hint: Let μ_0, μ_1 achieve equality in the Donsker–Varadhan variational principle (Theorem 1.5.4), so that

$$\ln \mathbb{E}_\pi f = \mathbb{E}_{\mu_0} \ln f - \mathsf{KL}(\mu_0 \,\|\, \pi)\,,$$
$$\ln \mathbb{E}_\pi g = \mathbb{E}_{\mu_1} \ln g - \mathsf{KL}(\mu_1 \,\|\, \pi)\,.$$

Let μ_t be along the Wasserstein geodesic from μ_0 to μ_1. Apply the Donsker–Varadhan principle again, together with the assumption on f, g, h as well as strong convexity of the KL divergence, in order to lower bound $\ln \mathbb{E}_\pi h$.

2. The inequality (2.E.1) continues to hold when π is replaced by Lebesgue measure, if we set $\alpha = 0$ in the assumption.[13] Use this to prove that if π is a log-concave measure over $\mathbb{R}^{d_1+d_2}$, then the marginal π^1 of π on $\mathbb{R}^{d_1}$ is also log-concave.

   Hint: Partition elements of $\mathbb{R}^{d_1+d_2}$ as (x, y). Apply the Prékopa–Leindler inequality on $\mathbb{R}^{d_2}$ with $f := \ln \pi(x_0, \cdot)$, $g := \ln \pi(x_1, \cdot)$, and $h := \ln \pi((1-t)\, x_0 + t\, x_1, \cdot)$.

3. We now aim to generalize the fact in the previous part. Suppose that π is a density on $\mathbb{R}^{d_1+d_2}$ such that $(x, y) \mapsto \ln \pi(x, y) + \frac{1}{2} \langle (x, y), \Sigma^{-1} (x, y) \rangle$ is concave, where

$$\Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix} .$$

   Prove that for the marginal π^1 of π on $\mathbb{R}^{d_1+d_2}$, $x \mapsto \ln \pi^1(x) + \frac{1}{2} \langle x, (\Sigma_{1,1})^{-1} x \rangle$ is concave. (Use the result in the previous part.)

4. Show that if π is α-strongly log-concave, then the convolution $\pi * \mathsf{normal}(0, tI_d)$ is $\alpha/(1+\alpha t)$-strongly log-concave.

5. Show that the Prékopa–Leindler inequality for the Lebesgue measure is equivalent to the **Brunn–Minkowski inequality**: for compact sets $A, B \subseteq \mathbb{R}^d$,

$$\mathrm{vol}\big((1-t)\, A + t\, B\big) \geq \mathrm{vol}(A)^{1-t}\, \mathrm{vol}(B)^t . \tag{2.E.2}$$

   By scaling A and B and choosing t, show that (2.E.2) can be upgraded to

$$\mathrm{vol}(A+B)^{1/d} \geq \mathrm{vol}(A)^{1/d} + \mathrm{vol}(B)^{1/d} .$$

Operations Preserving Functional Inequalities

▷ Exercise 2.9 (variational principle for entropies)
Let $\phi : \mathbb{R}_{>0} \to \mathbb{R}$ be a convex function and let D_ϕ denote the associated Bregman divergence (c.f. Definition 2.2.9). For any positive random variable X with $\mathbb{E}|\phi(X)| < \infty$ and any $t > 0$, prove that that $\mathbb{E}\, D_\phi(X, t) - \mathbb{E}\, D_\phi(X, \mathbb{E}\, X) = D_\phi(\mathbb{E}\, X, t)$, and deduce that $\mathbb{E}\, \phi(X) - \phi(\mathbb{E}\, X) = \inf_{t>0} \mathbb{E}\, D_\phi(X, t)$. Use this to prove the variational principle for the entropy used in the proof of Holley–Stroock perturbation (Proposition 2.3.1).

▷ Exercise 2.10 (transport inequality in one dimension)
Let π be the standard Gaussian on $\mathbb{R}$, and let $\mu \ll \pi$. In one dimension, the optimal transport map T from π to μ is the monotone rearrangement that satisfies, for each $x \in \mathbb{R}$, $\mu((-\infty, T(x)]) = \pi((-\infty, x])$.

[13] For example, one could first consider the case when f, g, h are compactly supported, take π to be the uniform distribution over a ball $\mathsf{B}(0, R)$, and take $R \to \infty$.

1. Differentiate this relation to obtain a formula for $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(T(x))$.

2. Substitute this into the KL divergence $\mathsf{KL}(\mu \,\|\, \pi) = \int \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(T(x))\, \mathrm{d}\pi(x)$ and use the inequality $t - 1 - \ln t \geq 0$ for all $t > 0$ in order to prove the Gaussian T_2 inequality in one dimension. Deduce the Gaussian T_2 inequality in general dimension via a tensorization argument.

3. Can you generalize this calculation to a density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, where V is smooth and α-strongly convex for some $\alpha > 0$?

▷ Exercise 2.11 (generalizing the LSI for mixtures)
In this exercise, we generalize the log-Sobolev inequality for mixtures (Proposition 2.3.14).

1. First, show that Example 2.3.15 is sharp up to universal constants as follows. Consider the case when $\mu = \frac{1}{2}(\delta_{-R} + \delta_{+R})$ on $\mathbb{R}$, so that μP is a mixture of two Gaussians. Construct a test function $f : \mathbb{R} \to \mathbb{R}$ for the Poincaré inequality which shows that $C_{\mathsf{PI}}(\mu P) \gtrsim R^2 \exp(\Omega(R^2/\sigma^2))$ if $R/\sigma \gtrsim 1$.

2. Next, consider the setting of Proposition 2.3.8 except that we replace the assumption (2.3.9) with the weaker condition

$$C_{\chi^2,2} := \sqrt{\mathbb{E}[\chi^2(P_X \,\|\, P_{X'})^2]} < \infty\,, \tag{2.E.3}$$

where $X, X' \overset{\text{i.i.d.}}{\sim} \mu$. Now, rather than writing $\operatorname{var} \mathbb{E}_{P_X} f = \frac{1}{2}\mathbb{E}[|\mathbb{E}_{P_X} f - \mathbb{E}_{P_{X'}} f|^2]$, instead write $\operatorname{var} \mathbb{E}_{P_X} f = \mathbb{E}[|\mathbb{E}_{P_X} f - \mathbb{E}_{\mu P} f|^2]$. By bounding this quantity in two different ways deduce that

$$\begin{aligned}\operatorname{var} \mathbb{E}_{P_X} f &\leq \mathbb{E} \min\{(\operatorname{var}_{P_X} f)\, \chi^2(\mu P \,\|\, P_X), (\operatorname{var}_{\mu P} f)\, \chi^2(P_X \,\|\, \mu P)\} \\ &\leq \mathbb{E} \sqrt{(\operatorname{var}_{P_X} f)\, \chi^2(\mu P \,\|\, P_X)\, (\operatorname{var}_{\mu P} f)\, \chi^2(P_X \,\|\, \mu P)}\,.\end{aligned}$$

Use this to prove that a Poincaré inequality holds for μP, and give an upper bound on $C_{\mathsf{PI}}(\mu P)$.

3. Now consider the setting of Proposition 2.3.14 except that we again assume the weaker condition (2.E.3). Previously, we bounded

$$\mathbb{E}\big[\mathbb{E}_{P_X}(f^2) \ln\big(1 + \chi^2(P_X \,\|\, P_{X'})\big)\big] \leq \mathbb{E}_{\mu P}(f^2) \ln(1 + C_{\chi^2})\,,$$

which relies on L^1–L^∞ duality. This time, we want to use duality between "$L \log L$" and "$\exp L$". Namely, use the variational principle for the entropy (Lemma 2.3.4) to prove that for a suitable constant $C > 0$ (depending on $C_{\chi^2,2}$),

$$2\,\mathbb{E}\big[\mathbb{E}_{P_X}(f^2)\,\big\{\ln\big(1+\chi^2(P_X \,\|\, P_{X'})\big) - C\big\}\big] \le \operatorname{ent}\mathbb{E}_{P_X}(f^2)\,.$$

Use this to prove that a log-Sobolev inequality holds for μP, and give an upper bound on $C_{\mathsf{LSI}}(\mu P)$.

4. Consider Example 2.3.15 again, except instead of assuming that μ is supported on $\mathrm{B}(0,R)$, we assume that μ has sub-Gaussian tails:

$$\iint \exp\frac{\|x-x'\|^2}{\sigma_{\mathsf{sG}}^2}\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \le C_{\mathsf{sG}}\,.$$

Prove that if $\sigma \gtrsim \sigma_{\mathsf{sG}}$ for a sufficiently large implied constant, then the Gaussian mixture μP satisfies a log-Sobolev inequality, and give an upper bound on $C_{\mathsf{LSI}}(\mu P)$. Also, show how this can recover the result of Example 2.3.15.

Concentration of Measure

⊵ Exercise 2.12 (Herbst argument)
Consider the Herbst argument from Section 2.4.2.

1. Verify the calculus identity (2.4.7) in the Herbst argument.

2. Suppose that X is a real-valued random variable satisfying the following condition: for all $\lambda \ge 0$, it holds that

$$\operatorname{var}\exp\frac{\lambda X}{2} \le \frac{\lambda^2\sigma^2}{4}\,\mathbb{E}_\pi \exp(\lambda X)\,.$$

Let $\eta(\lambda) \coloneqq \mathbb{E}\exp(\lambda X)$ and deduce an inequality for $\eta(\lambda)$ in terms of $\eta(\lambda/2)$. Solve this recursion to prove that for $\lambda < 2/\sigma$,

$$\mathbb{E}\exp\{\lambda\,(X - \mathbb{E}\,X)\} \le \frac{2+\lambda\sigma}{2-\lambda\sigma}\,.$$

3. Prove the Poincaré case of Theorem 2.4.8.

⊵ Exercise 2.13 (Hoeffding's lemma and Pinsker's inequality)
This exercise establishes the equivalence of Pinsker's inequality with a statement about sub-Gaussian concentration.

1. **Hoeffding's lemma** states that for any mean-zero random variable X with values in $[a, b]$ a.s., it holds that X is $(b-a)^2/4$-sub-Gaussian. Prove this lemma as follows. For $\lambda \in \mathbb{R}$, let $\psi(\lambda) := \ln \mathbb{E} \exp(\lambda X)$. Differentiate ψ twice and show that $\psi''(\lambda)$ can be interpreted as the variance of a random variable under a change of measure and hence $\psi''(\lambda) \le (b-a)^2/4$.

2. **Pinsker's inequality** states that for any two probability measures μ and ν on the same space, $\|\mu - \nu\|_{\mathrm{TV}}^2 \le \frac{1}{2} \mathsf{KL}(\mu \,\|\, \nu)$. Prove this inequality as follows. First, by the data-processing inequality (Theorem 1.5.3), for any event A,

$$\mathsf{KL}(\mu \,\|\, \nu) \ge \mathsf{KL}\big((\mathbb{1}_A)_\# \mu \,\big\|\, (\mathbb{1}_A)_\# \nu\big) = \mathsf{KL}\big(\mathsf{Bernoulli}(\mu(A)) \,\big\|\, \mathsf{Bernoulli}(\mu(A))\big)\,.$$

   Next, for any $q \in (0,1)$, differentiate $p \mapsto k_q(p) := \mathsf{KL}(\mathsf{Bernoulli}(p) \,\|\, \mathsf{Bernoulli}(q))$ twice to show that k_q is 4-strongly convex, and deduce that $k_q(p) \ge 2\,|p-q|^2$. Finally, take the supremum over events A.

3. Apply the Bobkov–Götze theorem (Theorem 2.4.10) to show that Hoeffding's lemma and Pinsker's inequality are equivalent to each other.

▷ Exercise 2.14 (inequivalence between PI and T_1)

In this exercise, we show that the Poincaré inequality and the T_1 inequality are incomparable, i.e., one does not necessarily imply the other.

1. Use Theorem 2.4.11 to provide an example of a measure $\pi \in \mathcal{P}_1(\mathbb{R}^d)$ which satisfies a T_1 inequality but which does not satisfy a Poincaré inequality.

   Hint: Explain why a Poincaré inequality necessarily requires the support of the measure to be connected.

2. For the converse direction, let μ be the exponential distribution on $\mathbb{R}$, so that the density is $\mu(x) = \exp(-x)\,\mathbb{1}\{x > 0\}$. Let $f : \mathbb{R}_+ \to \mathbb{R}$; we may assume that $f(0) = 0$. Now apply the identity $f(x)^2 = 2\int_0^x f(s)\,f'(s)\,\mathrm{d}s$ to the integral $\int f^2\,\mathrm{d}\mu$ and prove that μ satisfies $\mathsf{PI}(4)$. Explain why μ cannot satisfy a T_1 inequality.

▷ Exercise 2.15 (bounded differences inequality)

This exercise establishes a broadly useful concentration inequality.

1. Prove the **Azuma–Hoeffding inequality**: let $(\mathscr{F}_i)_{i=0}^n$ be a filtration, let $(\Delta_i)_{i=1}^n$ be a martingale difference sequence (that is, Δ_i is $\mathscr{F}_i$-measurable and $\mathbb{E}[\Delta_i \mid \mathscr{F}_{i-1}] = 0$), and assume that for each i there exist $\mathscr{F}_{i-1}$-measurable random variables A_i and B_i such that $A_i \le \Delta_i \le B_i$ a.s. Then, $\sum_{i=1}^n \Delta_i$ is $\sum_{i=1}^n \|B_i - A_i\|_{L^\infty(\mathbb{P})}^2/4$-sub-Gaussian.

   Hint: Apply Hoeffding's lemma from Exercise 2.13 conditionally.

2. Use this to prove the **bounded differences inequality**: if $X_1, \dots, X_n$ are independent, then $f(X_1, \dots, X_n) - \mathbb{E} f(X_1, \dots, X_n)$ is $\sum_{i=1}^n \|D_i f\|_{\sup}^2/4$-sub-Gaussian.

   Hint: Recall the proof of the Efron–Stein inequality from Exercise 1.2.

3. Next, apply Marton's tensorization (Theorem 2.3.17) to Pinsker's inequality from Exercise 2.13 (see Example 2.3.18) to obtain a transport inequality for the product space $\mathcal{X}^N$. Using the Bobkov–Götze equivalence (Theorem 2.4.10), give a second proof of the bounded differences inequality.

▷ **Exercise 2.16** (a loose end in Gozlan's theorem)
Prove the first statement of Lemma 2.4.13.

Isoperimetric Inequalities

▷ **Exercise 2.17** (isoperimetry on the sphere)
Prove Theorem 2.5.2 from the spherical isoperimetric inequality (Theorem 2.5.1). To do so, use the fact that the measure of $\mathsf{B}(x_0, r)$ is

$$\sigma_d\big(\mathsf{B}(x_0, r)\big) = \frac{\int_0^r (\sin\theta)^{d-1}\,\mathrm{d}\theta}{\int_0^\pi (\sin\theta)^{d-1}\,\mathrm{d}\theta},$$

or prove this fact yourself. It is also acceptable to establish a weaker bound of the form $\alpha_{\sigma_d}(\varepsilon) \le C\exp(-cd\varepsilon^2)$ for universal constants $c, C > 0$.

▷ **Exercise 2.18** (Gaussian isoperimetry)
Consider the Gaussian isoperimetric inequality in Theorem 2.5.3.

1. In the spirit of Theorem 2.5.14, show that the functional inequality (2.5.28) is equivalent to an isoperimetric statement. Consequently, deduce the comparison theorem (Theorem 2.5.29) from Theorem 2.5.27.

2. Show that the functional form of the Gaussian isoperimetric inequality in (2.5.28) is preserved (up to constants) under Lipschitz mappings (in other words, prove the analogue of Proposition 2.3.3 for (2.5.28)).

Metric Measure Spaces

▷ **Exercise 2.19** (Lichnerowicz inequality)
Under the $\mathsf{CD}(\alpha, d)$ condition (2.6.9), the spectral gap estimate for $-\mathscr{L}$ can be sharpened to $\lambda_{\min}(-\mathscr{L}) \ge \alpha d/(d-1)$, an estimate that is attributed to Lichnerowicz. Prove this as

follows: assume that f is such that $-\mathscr{L}f = \lambda f$. Show that $\lambda \int \Gamma(f,f)\,\mathrm{d}\pi = \int \Gamma_2(f,f)\,\mathrm{d}\pi$. Apply $\mathrm{CD}(\alpha, d)$ and deduce that $\lambda \geq \alpha d/(d-1)$.

Discrete Space and Time

⊵ **Exercise 2.20** (LSI implies MLSI)
Prove Lemma 2.7.3.

Hint: Prove that $4\,(\sqrt{a} - \sqrt{b})^2 = (\int_a^b t^{-1/2}\,\mathrm{d}t)^2 \leq (\ln a - \ln b)\,(a-b)$ for all $a, b > 0$.

CHAPTER 3

Additional Topics in Stochastic Analysis

In this chapter, we further expand our toolbox of stochastic analysis. Namely, we introduce Girsanov's theorem, which furnishes a formula for the Radon–Nikodym derivative of the laws of two SDEs w.r.t. to each other, and we discuss the time reversal of an SDE. In order to highlight the flexibility and power of these ideas, we then study some interesting applications, not all of which are directly relevant to log-concave sampling but nevertheless fit within the broader themes of this book.

3.1 Quadratic Variation

We now take a more general view of the ideas that led to the construction of the Itô integral as well as Itô's formula (Theorem 1.1.18).

Finite variation vs. quadratic variation. As a first step towards understanding the difficulties we faced when constructing the Itô integral, we recall that the classical condition under which it is possible to integrate a continuous process $(\eta_t)_{t\in[0,T]}$ against another continuous process $(A_t)_{t\in[0,T]}$, i.e., when we can consider the integral $\int_{[0,T]} \eta_t \, \mathrm{d}A_t$, is when the process A is of **finite variation**. This means that for any partition $0 = t_0 < t_1 < \cdots < t_n = T$ of $[0, T]$, if we define the **mesh** of the partition to be

$$\mathrm{mesh}(t_i : i = 0, 1, \ldots, n) \coloneqq \max_{i\in[n]} |t_i - t_{i-1}| ,$$

then it holds that

$$\lim_{\mathrm{mesh}(t_i:i=0,1,\dots,n)\searrow 0} \sum_{i=1}^{n} |A_{t_i} - A_{t_{i-1}}| < \infty\,.$$

The above limit is called the **total variation** of A on $[0,T]$. Under this condition, there is a signed measure μ_A such that for all $t \in [0,T]$, we have $\mu_A([0,t]) = A_t - A_0$. Moreover, we can define a norm $\|\cdot\|_{\mathrm{TV}}$ on the space of signed measures, called the **total variation norm**, for which $\|\mu_A\|_{\mathrm{TV}}$ equals the total variation of A as defined above.[1] In this case, we can simply define the integral $\int_{[0,T]} \eta_t \,\mathrm{d}A_t := \int_{[0,T]} \eta_t \,\mathrm{d}\mu_A(t)$.

Note that if $t \mapsto A_t$ is differentiable, then the total variation of A equals $\int_{[0,T]} |\dot{A}_t| \,\mathrm{d}t$, and the integral becomes $\int_{[0,T]} \eta_t \,\mathrm{d}A_t = \int_{[0,T]} \eta_t \dot{A}_t \,\mathrm{d}t$.

Hence, the condition that A is of finite variation is enough to develop a satisfactory theory of integration. The drawback, however, is that Brownian motion is *not* of finite variation. To see this, take $t_i := iT/n$ for $i = 0, 1, \dots, n$, so that the mesh of the partition is T/n. Since $B_{t_i} - B_{t_{i-1}} \sim \mathrm{normal}(0, T/n)$, we expect (heuristically) that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \underbrace{|B_{t_i} - B_{t_{i-1}}|}_{\asymp\sqrt{T/n}} \gtrsim \lim_{n\to\infty} n \cdot \sqrt{\frac{T}{n}} = \infty\,.$$

On the other hand, if we change the definition slightly, then we expect (heuristically) that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \underbrace{|B_{t_i} - B_{t_{i-1}}|^2}_{\asymp T/n} \lesssim \lim_{n\to\infty} n \cdot \frac{T}{n} < \infty\,.$$

We say that Brownian motion has finite **quadratic variation**. We will show in fact that the above limit is well-defined in the sense of convergence in probability.

More generally, for a process of the form

$$X_t = X_0 + \int_0^t b_t \,\mathrm{d}t + \int_0^t \sigma_t \,\mathrm{d}B_t\,, \qquad t \in [0,T]\,,$$

the second term is a process of finite variation (provided that $\int_{[0,T]} |b_t| \,\mathrm{d}t < \infty$ almost surely), whereas the third term requires consideration of quadratic variation.

[1] Indeed, the notation $\|\mu - \nu\|_{\mathrm{TV}}$ for the total variation distance between μ and ν is in accordance with this more general notion of a norm on the space of signed measures, up to a factor of 2 in the conventions.

Definition of the quadratic variation. More formally, we have the following theorem.

Theorem 3.1.1 (quadratic variation). *Let $(M_t)_{t\in[0,T]}$ be a continuous local martingale; then, there is an a.s. unique increasing process $t \mapsto [M,M]_t$ such that $t \mapsto M_t^2 - [M,M]_t$ is a continuous local martingale. Also, suppose that for each $n \in \mathbb{N}^+$, $(t_i : i = 0,1,\dots,n)$ is a partition of $[0,t]$, with mesh tending to zero as $n \to \infty$. Then,*

$$[M,M]_t = \lim_{n\to\infty} \sum_{i=1}^{n} (M_{t_i} - M_{t_{i-1}})^2 \qquad \textit{in probability}.$$

Definition 3.1.2 (quadratic variation). The process $[M,M]$ of Theorem 3.1.1 is called the **quadratic variation** of M.

We will not prove Theorem 3.1.1 in full generality. However, we will verify that the quadratic variation of one-dimensional Brownian motion $(B_t)_{t\in[0,T]}$ is $[B,B]_T = T$, which gives an idea of the general result. By independence of the Brownian increments,

$$\begin{aligned}
\mathbb{E}\Big[\Big|\sum_{i=1}^{n}\{(B_{t_i} - B_{t_{i-1}})^2 - (t_i - t_{i-1})\}\Big|^2\Big] &= \sum_{i=1}^{n} \mathbb{E}[|(B_{t_i} - B_{t_{i-1}})^2 - (t_i - t_{i-1})|^2] \\
&\le \sum_{i=1}^{n} \mathbb{E}[(B_{t_i} - B_{t_{i-1}})^4] = 3\sum_{i=1}^{n} (t_i - t_{i-1})^2 \\
&\le 3\,\mathrm{mesh}(t_i : i = 0,1,\dots,n) \underbrace{\sum_{i=1}^{n}(t_i - t_{i-1})}_{=T} \to 0\,.
\end{aligned}$$

Hence, $\sum_{i=1}^{n} (B_{t_i} - B_{t_{i-1}})^2 \xrightarrow{\mathbb{P}} T$ as $n \to \infty$. We also know that $t \mapsto B_t^2 - t$ is a martingale (see, e.g., Exercise 1.6).

Semimartingales. We often consider solutions to SDEs with a non-zero drift coefficient, which means that the resulting process are not continuous local martingales. To accommodate this addition, we consider the following definition.

Definition 3.1.3 (semimartingale). A process $(X_t)_{t\in[0,T]}$ is a **continuous semimartingale** if we can write $X = A + M$, where A is a process of finite variation with

$A_0 = 0$ and M is a continuous local martingale.

The decomposition $X = A + M$ is then unique. Indeed, suppose that $X = \widetilde{A} + \widetilde{M}$ for another finite variation process $\widetilde{A}$ (with $\widetilde{A}_0 = 0$) and a continuous local martingale $\widetilde{M}$. Then, from $\Delta := M - \widetilde{M} = \widetilde{A} - A$, we deduce that Δ is both a continuous local martingale and a process of finite variation. Since Δ is of finite variation,

$$\sum_{i=1}^{n} (\Delta_{t_i} - \Delta_{t_{i-1}})^2 \leq \operatorname{mesh}(t_i : i = 0, 1, \ldots, n) \underbrace{\sum_{i=1}^{n} |\Delta_{t_i} - \Delta_{t_{i-1}}|}_{\text{bounded as } n \to \infty} \to 0$$

as the mesh size tends to zero. This shows that $[\Delta, \Delta] = 0$, and thus Δ^2 is a continuous local martingale. If we knew that Δ^2 were a genuine martingale, then together with $\Delta_0 = 0$ it would imply that $\Delta = 0$, establishing uniqueness of the semimartingale decomposition. We omit the localization argument required to finish the proof.

We can also define the quadratic variation of the semimartingale X as $[X, X] := [M, M]$. To see why this makes sense, observe that

$$\begin{aligned}
&\left|\sum_{i=1}^{n} (X_{t_i} - X_{t_{i-1}})^2 - \sum_{i=1}^{n} (M_{t_i} - M_{t_{i-1}})^2\right| \\
&\qquad = \left|\sum_{i=1}^{n} (A_{t_i} - A_{t_{i-1}})^2 + 2\sum_{i=1}^{n} (A_{t_i} - A_{t_{i-1}})\,(M_{t_i} - M_{t_{i-1}})\right| \\
&\qquad \leq \sum_{i=1}^{n} (A_{t_i} - A_{t_{i-1}})^2 + 2\sqrt{\left(\sum_{i=1}^{n} (A_{t_i} - A_{t_{i-1}})^2\right)\left(\sum_{i=1}^{n} (M_{t_i} - M_{t_{i-1}})^2\right)},
\end{aligned}$$

which tends to zero using the same argument as in the uniqueness of the semimartingale decomposition: finite variation processes have zero quadratic variation.

The bracket of two semimartingales. Given two semimartingales X and Y, we define their bracket via polarization:

$$[X, Y] := \frac{1}{2}\left([X + Y, X + Y] - [X, X] - [Y, Y]\right).$$

Equivalently, if $X = A_X + M_X$ and $Y = A_Y + M_Y$ are the respective decompositions, then $[X, Y] = [M_X, M_Y]$. The following theorem gives a concrete way of computing the bracket for processes driven by Brownian motion.

Theorem 3.1.4 (bracket of processes driven by Brownian motion). *Suppose that X and Y are $\mathbb{R}^d$-valued processes with*

$$\begin{aligned} \mathrm{d}X_t &= b_t^X \,\mathrm{d}t + \sigma_t^X \,\mathrm{d}B_t\,, \\ \mathrm{d}Y_t &= b_t^Y \,\mathrm{d}t + \sigma_t^Y \,\mathrm{d}B_t\,, \end{aligned}$$

where we assume $\int_{[0,T]} \|b_t^X\| \,\mathrm{d}t$, $\int_{[0,T]} \|b_t^Y\| \,\mathrm{d}t$, $\int_{[0,T]} \|\sigma_t^X\|_{\mathrm{HS}}^2 \,\mathrm{d}t$, and $\int_{[0,T]} \|\sigma_t^Y\|_{\mathrm{HS}}^2 \,\mathrm{d}t$ are all finite almost surely. Then, X and Y are continuous semimartingales, and

$$[X,Y]_t = \int_0^t \langle \sigma_s^X, \sigma_s^Y \rangle \,\mathrm{d}s\,, \qquad \text{for } t \in [0,T]\,.$$

Itô's formula revisited. Finally, we conclude this section by revisiting Itô's formula (Theorem 1.1.18) using our new calculus.

Theorem 3.1.5 (Itô's formula revisited). *Let X be an $\mathbb{R}^d$-valued semimartingale, and write $X = (X^1, \dots, X^d)$. Let $f \in C^2(\mathbb{R}^d)$. Then, $f(X)$ is also a semimartingale, and*

$$f(X_t) = f(X_0) + \sum_{i=1}^d \int_0^t \partial_i f(X_s) \,\mathrm{d}X_s^i + \frac{1}{2} \sum_{i,j=1}^d \int_0^t \partial_{i,j} f(X_s) \,\mathrm{d}[X^i, X^j]_s\,.$$

If we interpret $[X,X]$ as the matrix whose (i,j)-entry is $[X^i, X^j]$, then this can be written in matrix notation as

$$f(X_t) = f(X_0) + \int_0^t \langle \nabla f(X_s), \mathrm{d}X_s \rangle + \frac{1}{2} \int_0^t \left\langle \nabla^2 f(X_s), \mathrm{d}[X,X]_s \right\rangle. \tag{3.1.6}$$

For d-dimensional standard Brownian motion $(B_t)_{t\in[0,T]}$, we have $[B,B]_t = tI_d$, so that $\mathrm{d}[B,B]_t = I_d \,\mathrm{d}t$ and we recover the original statement of Itô's formula in Theorem 1.1.18. The point is that the quadratic variation is a convenient way of streamlining Itô calculations, as it formalizes the idea that only the Brownian motion part of a process contributes in the second-order term in Itô's formula.

3.2 Change of Measure in Path Space

In this section, we begin to investigate measures on path space, for which it is convenient to adopt the following canonical setup. Let $(B_t)_{t\in[0,T]}$ be standard Brownian motion, and

let $\mathbf{W}$ denote its law, the **Wiener measure**. Recall that in probability theory, all of our random variables are defined on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Since we can take this probability space to be whatever we wish, as long as it is sufficiently rich, we may as well take $\Omega = C([0,T])$ to be the space of continuous paths with $\mathcal{F}$ being the Borel σ-algebra, equipped with the Wiener measure $\mathbb{P} = \mathbf{W}$. Then, for $\omega \in \Omega$, the Brownian motion at time t simply becomes the evaluation functional, $B_t(\omega) = \omega_t$.

3.2.1 The Cameron–Martin Theorem

Our goal now is to understand when two measures $\mathbf{P}$, $\mathbf{Q}$ on the path space $C([0,T])$ are absolutely continuous w.r.t. each other, and if so, to write down a formula for the Radon–Nikodym derivative $\frac{\mathrm{dP}}{\mathrm{dQ}}$. The final result, known as *Girsanov's theorem*, will be used for the analysis of sampling algorithms later in this book, and more broadly it is an indispensable tool for stochastic analysis.

Before reaching this goal, however, it may be helpful to provide some mathematical context. We begin with the following question. For a curve $h \in C([0,T])$, the translation operator $T_h : C([0,T]) \to C([0,T])$ is defined simply by the mapping $\omega \mapsto \omega + h$. What happens to the Wiener measure under translations?

More generally, we can define the translation operator T_h on any Banach space $\mathcal{B}$, with $h \in \mathcal{B}$. If $\mathcal{B} = \mathbb{R}^d$ and μ is the Lebesgue measure, then we know that μ is invariant under translations, in the sense that $(T_h)_{\#}\mu = \mu$ for all $h \in \mathbb{R}^d$, and that the Lebesgue measure is the unique measure with this property up to rescaling. However, as soon as we move to infinite dimensions, a classical result of analysis states that there is *no* non-trivial measure μ which is invariant under translations, i.e., infinite-dimensional Lebesgue measure does not exist. This makes it difficult to decide upon a "canonical" reference measure for infinite-dimensional analysis.

Although invariance is impossible, we can at least ask for *quasi-invariance*: does there exist μ such that $(T_h)_{\#}\mu \ll \mu$ for all $h \in \mathcal{B}$? For example, the standard Gaussian measure is quasi-invariant on $\mathbb{R}^d$. In infinite dimensions, the answer is still *no*; in particular, there is no infinite-dimensional standard Gaussian. To understand this point more concretely, suppose that $\mathcal{B} = \mathcal{H}$ is actually a Hilbert space, and let $(e_k)_{k\in\mathbb{N}}$ be an orthonormal basis. An obvious attempt to build "the standard Gaussian measure on $\mathcal{H}$" is to take an i.i.d. sequence $(\xi_k)_{k\in\mathbb{N}}$ of standard Gaussians on $\mathbb{R}$, and to take μ to be the law of $\sum_{k\in\mathbb{N}} \xi_k e_k$. However, since $\|\sum_{k=0}^N \xi_k e_k\|^2 = \sum_{k=0}^N \xi_k^2$, standard probability theory tells us that almost surely, this is not a convergent sum in $\mathcal{H}$.

Despite this obstruction, we will now see that the Wiener measure behaves in some sense like a standard Gaussian measure on a Hilbert space! The theorem below begins by precisely characterize the set of h for which $(T_h)_{\#}\mathbf{W} \ll \mathbf{W}$.

Theorem 3.2.1 (Cameron–Martin). *Let* $\mathbf{W}$ *be the Wiener measure on* $C([0,T])$. *Then,* $(T_h)_\# \mathbf{W} \ll \mathbf{W}$ *if and only if*

$$h \in \mathcal{H} := \left\{ h \in C([0,T]) \;\middle|\; h(0) = 0,\ \int_0^T \|\dot{h}(t)\|^2 \,\mathrm{d}t < \infty \right\}.$$

If this holds, then

$$\frac{\mathrm{d}(T_h)_\# \mathbf{W}}{\mathrm{d}\mathbf{W}}(\omega) = \exp\left(\int_0^T \langle \dot{h}(t), \mathrm{d}\omega_t \rangle - \frac{1}{2} \int_0^T \|\dot{h}(t)\|^2 \,\mathrm{d}t \right). \tag{3.2.2}$$

Here, $\mathcal{H}$ *is called the* ***Cameron–Martin space*** *associated with Brownian motion.*

The Cameron–Martin theorem will be subsumed by Girsanov's theorem, so we will not prove it here. Instead, we will focus on its interpretation.

Interpretation of the Cameron–Martin theorem. To interpret (3.2.2), let γ denote the standard Gaussian measure on $\mathbb{R}^d$ and let $h \in \mathbb{R}^d$. Then, on $\mathbb{R}^d$, we have the formula

$$\frac{\mathrm{d}(T_h)_\# \gamma}{\mathrm{d}\gamma}(x) = \exp\left(\langle h, x \rangle - \frac{1}{2} \|h\|^2 \right).$$

This bears a striking resemblance to (3.2.2). Namely, for $h_0, h_1 \in \mathcal{H}$, let us define the inner product $\langle h_0, h_1 \rangle_{\mathcal{H}} := \int_0^T \langle \dot{h}_0(t), \dot{h}_1(t) \rangle \,\mathrm{d}t$. If we interpret the stochastic integral $\int_0^T \langle \dot{h}(t), \mathrm{d}\omega_t \rangle$ as $\langle h, \omega \rangle_{\mathcal{H}}$, then the density ratio in (3.2.2) behaves as if

$$\mathrm{d}\mathbf{W}(\omega) \text{ “}\propto\text{” } \exp\left(-\frac{1}{2} \|\omega\|_{\mathcal{H}}^2 \right) \mathrm{d}\omega\,. \tag{3.2.3}$$

The charming part about (3.2.3) is that not a single aspect of it makes any sense. We know that $\mathbf{W}$-a.s. $\omega \in \Omega$ does not even belong to $\mathcal{H}$, since Brownian paths are non-differentiable (in fact, they are Hölder continuous of any exponent less than $1/2$, but no better, whereas we would require Lipschitz continuity to have a.e. differentiability). Also, (3.2.3) tries to express the density of $\mathbf{W}$, but with respect to what measure? We have just stated that there is no “Lebesgue measure” on $\mathcal{H}$.

Despite these objections, Theorem 3.2.1 is a perfectly rigorous manifestation of the intuition that $\mathbf{W}$ is a standard Gaussian measure on $\mathcal{H}$. To reconcile this, it will turn out that a “standard $\mathcal{H}$-Gaussian measure” *can exist*, but the catch is that it no longer “fits” in $\mathcal{H}$ (indeed, $\mathbf{W}$ is supported on $C([0,T]) \supsetneq \mathcal{H}$). Indeed, the fact that $\mathbf{W}$ is usually defined on $C([0,T])$ is somewhat of a red herring, and many of the deeper properties of Brownian motion (e.g., Schilder's theorem in large deviations) are best understood via $\mathcal{H}$.

Abstract Wiener space. More generally, let $\mathcal{H}$ be an infinite-dimensional Hilbert space and let us try to construct the standard Gaussian measure on $\mathcal{H}$. We tried earlier to use the sum $\sum_{k\in\mathbb{N}} \xi_k e_k$, but this does not converge in the norm of $\mathcal{H}$. To proceed forward, the idea is rather simple: we can just use another norm. Namely, if we can find a Banach space $\mathcal{B} \supseteq \mathcal{H}$ with corresponding norm $\|\cdot\|_{\mathcal{B}}$, such that the sum $\sum_{k\in\mathbb{N}} \xi_k e_k$ converges in $\|\cdot\|_{\mathcal{B}}$, then $\sum_{k\in\mathbb{N}} \xi_k e_k$ makes sense as a random element of $\mathcal{B}$, and we can take μ to be its law. Note that μ is supported on $\mathcal{B}$ rather than on $\mathcal{H}$. It turns out that a suitable orthonormal basis $(e_k)_{k\in\mathbb{N}}$ and a norm $\|\cdot\|_{\mathcal{B}}$ can always be found to make this procedure work.

For Brownian motion, we take $\|\cdot\|_{\mathcal{B}}$ to be the supremum norm (i.e., the norm of $C([0,T])$) and the orthonormal basis can be chosen as a certain (integrated) wavelet basis. In fact, the abstract construction described above is implicit in Lévy's usual construction of Brownian motion.

It is also possible to flip this process around. Namely, suppose that μ is a Gaussian measure on a Banach space $\mathcal{B}$, which means that for any linear functionals $\ell_1, \dots, \ell_n \in \mathcal{B}^*$ and $X \sim \mu$, the vector $(\ell_1(X), \dots, \ell_n(X))$ is jointly Gaussian. Then, one can find a Hilbert space $\mathcal{H}$ associated to $(\mathcal{B}, \mu)$, which is called the Cameron–Martin space, and an appropriate analogue of the Cameron–Martin theorem (Theorem 3.2.1) holds. In fact, one has $\|x\|_{\mathcal{H}} := \sup\{|\ell(\omega)| \mid \ell \in \mathcal{B}^*,\ \|\ell\|_{L^2(\mu)} \le 1\}$ and $\mathcal{H} := \{x \in \mathcal{B} \mid \|x\|_{\mathcal{H}} < \infty\}$. The triple $(\mathcal{B}, \mathcal{H}, \mu)$ is known as an **abstract Wiener space**.

3.2.2 Girsanov's Theorem

The Cameron–Martin theorem (Theorem 3.2.1) provides a formula for the density ratio of the laws of two diffusions that differ by a deterministic drift. Girsanov's theorem generalizes this to diffusions which differ by a random drift.

Why do we only consider a change of drift? The answer is that two diffusions

$$\begin{aligned} \mathrm{d}X_t &= b_t^X \,\mathrm{d}t + \sigma_t^X \,\mathrm{d}B_t\,, \\ \mathrm{d}Y_t &= b_t^Y \,\mathrm{d}t + \sigma_t^Y \,\mathrm{d}B_t\,, \end{aligned}$$

with $\sigma^X (\sigma^X)^\mathsf{T} \neq \sigma^Y (\sigma^Y)^\mathsf{T}$, have mutually *singular* laws, as a consequence of the existence of the quadratic variation (Theorem 3.1.1). Indeed, the laws of X and Y are concentrated on the disjoint events that the quadratic variation equals $\int_0^\cdot \sigma^X (\sigma^X)^\mathsf{T}$ or $\int_0^\cdot \sigma^Y (\sigma^Y)^\mathsf{T}$ respectively. Nevertheless, Girsanov's theorem will show that a change of drift is enough to obtain any other path measure which is absolutely continuous w.r.t. the original one.

To understand the intuition behind Girsanov's theorem, consider the diffusion

$$\mathrm{d}X_t = b_t \,\mathrm{d}t + \mathrm{d}B_t \iff X_t = \int_0^t b_s \,\mathrm{d}s + B_t\,, \tag{3.2.4}$$

where $(b_t)_{t\geq 0}$ is an adapted process. If $(b_t)_{t\geq 0}$ is in fact deterministic, then recall that the law of X_t is a centered Gaussian with covariance $\int_0^t (b_s \otimes b_s)\,\mathrm{d}s$. In general, however, the law of X_t is not easily describable. Nevertheless, it is possible to understand the *joint law* of $(X_t)_{t\in[0,T]}$, which is a measure $\mathbf{P}$ on path space.

To see why this is the case, consider a discrete-time analogue of (3.2.4):

$$X_{k+1} := F(X_k) + \xi_k\,, \qquad k = 0, 1, 2, \ldots\,,$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is a deterministic map and $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. Gaussian variables. Again, due to the non-linear mapping F, the law of X_k is hard to describe exactly, yet once we *condition* on X_k, the law of X_{k+1} is an explicit Gaussian. This observation makes it straightforward to write down an explicit and simple expression for the joint law of $(X_0, X_1, \ldots, X_N)$.

Returning to (3.2.4), we can think of it in the same way: namely, conditioned on the past, the conditional law of the diffusion in the next instant is a Gaussian with some mean and covariance. Moreover, by using the formula for the density ratio of two Gaussians, one can guess the formula

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{W}} = \exp\Big(\int_0^T \langle b_s, \mathrm{d}B_s\rangle - \frac{1}{2}\int_0^T \|b_s\|^2\,\mathrm{d}s\Big)\,. \tag{3.2.5}$$

Note that this exactly mirrors Theorem 3.2.1, except that now we allow for adapted processes $(b_t)_{t\geq 0}$.

Let us now discuss how we would establish (3.2.5) carefully. Actually, we will proceed in the opposite order from our informal discussion: we will first *define* $\mathbf{P}$ via the formula (3.2.5), and then investigate the effect of the change of measure from $\mathbf{W}$ to $\mathbf{P}$ on our stochastic processes. This requires a change of perspective: instead of considering two processes $(B_t)_{t\in[0,T]}$ and $(X_t)_{t\in[0,T]}$, we instead think of a single process $(B_t)_{t\in[0,T]}$ defined on our canonical filtered space $(\Omega = C([0,T]), \mathscr{F}, (\mathscr{F}_t)_{t\in[0,T]})$. Recall that $(B_t)_{t\in[0,T]}$ is just the coordinate process $B_t(\omega) = \omega_t$. When we endow our space with the Wiener measure $\mathbf{W}$, then $(B_t)_{t\in[0,T]}$ becomes a standard Brownian motion. On the other hand, if we instead endow our space with the measure $\mathbf{P}$, we will show that $(B_t)_{t\in[0,T]}$ is, in some sense, a Brownian motion *with drift*.

To carry out our plan, the first step is to show that (3.2.5) defines a valid probability measure $\mathbf{P}$. In other words, we need the $\mathbf{W}$-expectation of the right-hand side of (3.2.5) to equal 1. Actually, for any $t \in [0,T]$, let us write $\mathbf{W}_t$ to be the restriction of $\mathbf{W}$ to $\mathscr{F}_t$ (and similarly write $\mathbf{P}_t$). In order for our putative $\mathbf{P}_t$ to be a probability measure for each $t \in [0,T]$, we would require

$$t \mapsto \frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{W}_t} \overset{?}{=} \exp\Big(\int_0^t \langle b_s, \mathrm{d}B_s\rangle - \frac{1}{2}\int_0^t \|b_s\|^2\,\mathrm{d}s\Big)$$

to have constant $\mathbf{W}$-expectation, equal to 1 for all $t \in [0,T]$. This would follow if we knew that this defined a $\mathbf{W}$-martingale.

Assume that $\mathbb{E}^{\mathbf{W}} \int_0^T \|b_s\|^2 \, \mathrm{d}s < \infty$. Then, the process $t \mapsto \int_0^t \langle b_s, \mathrm{d}B_s \rangle$ is a $\mathbf{W}$-martingale, and $t \mapsto \int_0^t \|b_s\|^2 \, \mathrm{d}s$ is its quadratic variation. We will simply write $M := \int_0^\cdot \langle b, \mathrm{d}B \rangle$ for the martingale and $[M,M]$ for its quadratic variation. Then,

$$\mathcal{E}(M) := \exp\Big(M - \frac{1}{2}[M,M]\Big)$$

is called the **exponential martingale** associated with M. Is it actually a martingale? Applying Itô's formula in the form (3.1.6),

$$\mathrm{d}\mathcal{E}(M)_t = \mathcal{E}(M)_t \left(\mathrm{d}M_t - \frac{1}{2}\,\mathrm{d}[M,M]_t + \frac{1}{2}\,\mathrm{d}[M,M]_t\right) = \mathcal{E}(M)_t \,\mathrm{d}M_t\,,$$

so $\mathcal{E}(M)$ is a stochastic integral. From Proposition 1.1.15, this tells us that $\mathcal{E}(M)$ is a continuous *local* $\mathbf{W}$-martingale. In other words, it is possible for $\mathcal{E}(M)$ to *fail* to be a martingale if some integrability conditions are violated. For now, we will *assume* that $\mathcal{E}(M)$ is an honest[2] martingale, treating this point as a technical issue, although later we will see that there is a clear understanding of what happens when this assumption fails.

Under this assumption, the measure $\mathbf{P}$ defined via (3.2.5) is a probability measure on path space. We now claim that under $\mathbf{P}$, *the process* $t \mapsto \tilde{B}_t := B_t - [B,M]_t = B_t - \int_0^t b_s \, \mathrm{d}s$ *is a standard Brownian motion.* Actually, this is not too hard to check using (3.2.5) and characteristic functions; we leave it as Exercise 3.1.

We have arrived at the following theorem.

Theorem 3.2.6 (Girsanov). *Let $(B_t)_{t\in[0,T]}$ be a standard Brownian motion under the Wiener measure $\mathbf{W}$ and let $(b_t)_{t\in[0,T]}$ be a progressive process with $\mathbb{E}^{\mathbf{W}} \int_0^T \|b_s\|^2 \, \mathrm{d}s < \infty$. Let $M_t := \int_0^t \langle b_s, \mathrm{d}B_s \rangle$ for $t \in [0,T]$ and let $[M,M]_t := \int_0^t \|b_s\|^2 \, \mathrm{d}s$ denote the quadratic variation. Define the exponential martingale*

$$\mathcal{E}(M) := \exp\Big(M - \frac{1}{2}[M,M]\Big)\,.$$

Assume that $\mathcal{E}(M)$ is a $\mathbf{W}$-martingale and define the measure $\mathbf{P}$ on path space via

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{W}} = \mathcal{E}(M)_T\,.$$

[2] We borrow the terminology from [Ste01].

Then, under $\mathbf{P}$*,*

$$t \mapsto \tilde{B}_t := B_t - [B, M]_t = B_t - \int_0^t b_s \, \mathrm{d}s \qquad \textit{is a standard Brownian motion}\,.$$

At present, Girsanov's theorem may seem rather abstract, and perhaps the best way to learn its meaning is to see it in action. We will put it to work in Section 4.4.

When is the exponential martingale an honest martingale? Since $\mathcal{E}(M)$ is a non-negative local martingale, then it is a *supermartingale*, i.e., we always have $\mathbb{E}^{\mathbf{W}} \mathcal{E}(M)_T \leq 1$. The only situation in which we encounter difficulties is when $\mathbb{E}^{\mathbf{W}} \mathcal{E}(M)_T < 1$, which would lead $\mathbf{P}$ defined via (3.2.5) to be a *sub*-probability measure. One might suspect that this is related to some probability mass "running off to ∞", and indeed one can show that $1 - \mathbb{E}^{\mathbf{W}} \mathcal{E}(M)_T$ is precisely the probability that the diffusion has exploded by time T, in the sense discussed in Section 1.1.3. Therefore, the following criteria for $\mathcal{E}(M)$ to be a martingale are really criteria for non-explosion.

The standard sufficient condition for $\mathcal{E}(M)$ to be a martingale is **Novikov's condition**, $\mathbb{E}^{\mathbf{W}} \exp(\frac{1}{2}[M, M]_T) < \infty$. An even weaker condition, known as **Kazamaki's condition**, requires only that $\sup_{t \in [0,T]} \mathbb{E}^{\mathbf{W}} \exp(\frac{1}{2} M_t) < \infty$. In principle, one of these conditions should be checked before applying Girsanov's theorem. However, if one is only interested in bounding a quantity such as the KL divergence or a Rényi divergence, one could use the technique of localization, mentioned in Section 1.1.1, together with the lower semicontinuity of the KL divergence, to avoid these conditions altogether.

3.3 Doob's Transform

In this section, we will introduce a more sophisticated use of change of measure on path space, known as *Doob's transform*. Some applications include obtaining SDEs for processes conditioned on an endpoint and deriving the Föllmer process in the next section.

Suppose that $\mathbf{Q}$ is a reference measure on path space, describing the law of the SDE

$$\mathrm{d}X_t = b_t(X_t)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}B_t\,, \qquad X_0 \sim \pi_0\,. \tag{3.3.1}$$

Let $P_{s,t}$ denote the transition operator from time s to time t (note that our reference process is time-inhomogeneous). The question we address in this section is the following: suppose that $\mathbf{P}$ is another probability measure on path space such that its Radon–Nikodym derivative w.r.t. $\mathbf{Q}$ only depends on X_T, the path at time T:

$$\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{Q}_T} = h_T(X_T)\,.$$

Since $\mathbf{P}_T \ll \mathbf{Q}_T$, we know from the previous section that the process X under $\mathbf{P}$ corresponds to the original process under $\mathbf{Q}$ with a change of drift. Can we solve for this drift?

The process $t \mapsto \frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{Q}_t}$ must be a martingale, and we make the ansatz that at time t, it only depends on the path at time t: $\frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{Q}_t} = h_t(X_t)$. Applying Itô's formula to $(h_t(X_t))_{t\in[0,T]}$,

$$\mathrm{d}h_t(X_t) = (\partial_t h_t + \mathscr{L}_t h_t)(X_t) + \langle \nabla h_t(X_t), \sigma_t(X_t)\,\mathrm{d}B_t\rangle\,, \tag{3.3.2}$$

and we deduce that this is a martingale if and only if

$$\partial_t h_t + \mathscr{L}_t h_t = 0\,,$$

where $\mathscr{L}_t f \coloneqq \frac{1}{2}\langle \sigma_t\sigma_t^\mathsf{T}, \nabla^2 f\rangle + \langle b_t, \nabla f\rangle$ is the generator at time t. This is the *backward* heat equation (indeed, if X is a Brownian motion, then the equation reads $\partial_t h_t + \frac{1}{2}\Delta h_t = 0$), which makes sense since we have a terminal time condition for h_T. In the case when the process is time-homogeneous (i.e., the coefficients do not depend on t), setting $h_t^{\leftarrow} \coloneqq h_{T-t}$, we see that $h^{\leftarrow}$ satisfies the forward heat equation $\partial_t h_t^{\leftarrow} = \mathscr{L} h_t^{\leftarrow}$, which has the solution $h_t^{\leftarrow} = P_{0,t} h_0^{\leftarrow}$. Switching back to h, we deduce that $h_t = P_{0,T-t} h_T$.

Now that we have a formula for $\frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{Q}_t}$, let us solve for the change of drift. On one hand, we know that if $\tilde{B} = B - [B, M]$ is a standard Brownian motion under $\mathbf{P}$, where $M \coloneqq \int_0^\cdot \langle \Delta, \mathrm{d}B\rangle$, then Girsanov's theorem (Theorem 3.2.6) yields

$$\mathrm{d}\Big(\ln \frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{Q}_t}\Big) = \langle \Delta_t, \mathrm{d}B_t\rangle - \frac{1}{2}\,\|\Delta_t\|^2\,\mathrm{d}t\,.$$

On the other hand, Itô's formula and (3.3.2) yield

$$\begin{aligned}\mathrm{d}\big(\ln h_t(X_t)\big) &= \frac{1}{h_t(X_t)}\,\mathrm{d}h_t(X_t) - \frac{1}{2\,h_t(X_t)^2}\,\mathrm{d}[h_\cdot(X), h_\cdot(X)]_t \\ &= \frac{1}{h_t(X_t)}\,\langle \nabla h_t(X_t), \sigma_t(X_t)\,\mathrm{d}B_t\rangle - \frac{1}{2\,h_t(X_t)^2}\,\|\sigma_t(X_t)^\mathsf{T}\,\nabla h_t(X_t)\|^2\,\mathrm{d}t\end{aligned}$$

and we quickly deduce that

$$\Delta_t = \sigma_t(X_t)^\mathsf{T}\,\nabla \ln h_t(X_t)\,.$$

Finally, we have obtained

$$\begin{aligned}\mathrm{d}X_t &= \big(b_t(X_t) + \sigma_t(X_t)\,\Delta_t\big)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}\tilde{B}_t \\ &= \big(b_t(X_t) + \sigma_t(X_t)\,\sigma_t(X_t)^\mathsf{T}\,\nabla \ln h_t(X_t)\big)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}\tilde{B}_t\,.\end{aligned}$$

This is our expression for the SDE under $\mathbf{P}_T$. We recall that in the time-homogeneous case, we actually have $h_t = P_{0,T-t} h_T$.

3.3.1 Conditioning on an Endpoint

As a first application, we will show how the Doob transform allows us to condition on $X_T = x_T$, for some fixed $x_T \in \mathbb{R}^d$. To see what this means, let $t < T$ and let η_t be a bounded $\mathscr{F}_t$-measurable random variable; let us try to compute $\mathbb{E}^{\mathbf{Q}}[\eta_t \mid X_T]$. By the definition of the conditional expectation, we wish to compute $\mathbb{E}^{\mathbf{Q}}[\eta_t\, f(X_T)]$ for any bounded measurable $f : \mathbb{R}^d \to \mathbb{R}$. Using the transition operator $P_{t,T}$ (which, in an abuse of notation, we identify with the transition kernel itself), and writing $\pi_t := \mathrm{law}^{\mathbf{Q}}(X_t)$,

$$\begin{aligned}\mathbb{E}^{\mathbf{Q}}[\eta_t\, f(X_T)] &= \mathbb{E}^{\mathbf{Q}}\big[\eta_t\, \mathbb{E}^{\mathbf{Q}}[f(X_T) \mid X_t]\big] = \mathbb{E}^{\mathbf{Q}}\Big[\eta_t \int f(x_T)\, P_{t,T}(X_t, \mathrm{d}x_T)\Big] \\ &= \mathbb{E}^{\mathbf{Q}}\Big[\eta_t \int f(x_T)\, \frac{\mathrm{d}P_{t,T}(X_t,\cdot)}{\mathrm{d}\pi_T}(x_T)\, \pi_T(\mathrm{d}x_T)\Big] \\ &= \int f(x_T)\, \mathbb{E}^{\mathbf{Q}}\Big[\eta_t\, \frac{\mathrm{d}P_{t,T}(X_t,\cdot)}{\mathrm{d}\pi_T}(x_T)\Big]\, \pi_T(\mathrm{d}x_T)\,.\end{aligned}$$

We conclude that if $h_t^{x_T}(x) := \frac{\mathrm{d}P_{t,T}(x,\cdot)}{\mathrm{d}\pi_T}(x_T)$, then

$$\mathbb{E}^{\mathbf{Q}}[\eta_t \mid X_T = x_T] = \mathbb{E}^{\mathbf{Q}}[\eta_t\, h_t^{x_T}(X_t)]\,.$$

Since this holds for every η_t, it says that if $\mathbf{P}$ denotes the measure $\mathbf{Q}$ conditioned on $X_T = x_T$, then

$$\frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{Q}_t} = h_t^{x_T}(X_t) \qquad \text{for all } 0 \le t < T\,.$$

We are now in the setting of Doob's transform. In particular, under $\mathbf{P}$,

$$\mathrm{d}X_t = \big(b_t(X_t) + \sigma_t(X_t)\, \sigma_t(X_t)^{\mathsf{T}}\, \nabla \ln h_t^{x_T}(X_t)\big)\, \mathrm{d}t + \sigma_t(X_t)\, \mathrm{d}\tilde{B}_t\,, \qquad 0 \le t < T\,. \tag{3.3.3}$$

Example 3.3.4 (Brownian bridge)**.** Suppose that $B = X$ is a standard Brownian motion under $\mathbf{Q}$. Then, $P_{s,t} = \mathrm{normal}(0, (t-s)\, I_d)$ and $\pi_t = P_{0,t}$. If we condition B to hit x_T at time T, then this process is known as a **Brownian bridge**.

We can calculate

$$h_t^{x_T}(x) \propto \exp\Big(-\frac{\|x - x_T\|^2}{2\,(T-t)}\Big) \qquad \Longrightarrow \qquad \nabla \ln h_t^{x_T}(x) = -\frac{x - x_T}{T-t}\,.$$

Hence, we arrive at the SDE representation for Brownian bridge,

$$\mathrm{d}X_t = \frac{x_T - X_t}{T-t}\, \mathrm{d}t + \mathrm{d}\tilde{B}_t\,.$$

Note that the drift is singular as $t \nearrow T$. Of course, it must be, in order to drive the process to hit a single point x_T at time T.

3.3.2 Reversing the SDE

Next, we describe how to construct the *time reversal* of the SDE $(X_t)_{t\in[0,T]}$. Namely, suppose that $\pi_t := \mathrm{law}(X_t)$ is the marginal law of X at time t. We will construct another SDE $X^{\leftarrow}$ such that $\mathrm{law}(X_t^{\leftarrow}) = \pi_{T-t}$ for all $t \in [0,T]$. This construction will play an important role in the study of the proximal sampler in Chapter 8.

Perhaps the most straightforward approach is to start with the Fokker–Planck equation for X. Let X denote the general SDE (3.3.1), but for the sake of simplifying calculations we shall assume that the diffusion matrix σ_t does not depend on the spatial variable, for all $t \in [0,T]$. Then, we have the Fokker–Planck equation

$$\partial_t \pi_t = \frac{1}{2} \langle \sigma_t \sigma_t^{\mathsf{T}}, \nabla^2 \pi_t \rangle - \mathrm{div}(\pi_t b_t) \,.$$

Therefore, the time reversal $\pi_t^{\leftarrow} := \pi_{T-t}$ satisfies

$$\partial_t \pi_t^{\leftarrow} = -\frac{1}{2} \langle \sigma_{T-t} \sigma_{T-t}^{\mathsf{T}}, \nabla^2 \pi_t^{\leftarrow} \rangle + \mathrm{div}(\pi_t^{\leftarrow} b_{T-t}) \,.$$

Next, we note that

$$\langle \sigma_{T-t} \sigma_{T-t}^{\mathsf{T}}, \nabla^2 \pi_t^{\leftarrow} \rangle = \mathrm{div}(\sigma_{T-t} \sigma_{T-t}^{\mathsf{T}} \nabla \pi_t^{\leftarrow}) = \mathrm{div}\big(\pi_t^{\leftarrow} (\sigma_{T-t} \sigma_{T-t}^{\mathsf{T}} \nabla \ln \pi_t^{\leftarrow})\big)$$

hence we can write

$$\partial_t \pi_t^{\leftarrow} = \frac{1}{2} \langle \sigma_{T-t} \sigma_{T-t}^{\mathsf{T}}, \nabla^2 \pi_t^{\leftarrow} \rangle + \mathrm{div}\big(\pi_t^{\leftarrow} (b_{T-t} - \sigma_{T-t} \sigma_{T-t}^{\mathsf{T}} \nabla \ln \pi_t^{\leftarrow})\big) \,. \tag{3.3.5}$$

We can therefore read off the SDE

$$\mathrm{d}X_t^{\leftarrow} = \{-b_{T-t}(X_t^{\leftarrow}) + \sigma_{T-t} \sigma_{T-t}^{\mathsf{T}} \nabla \ln \pi_t^{\leftarrow}(X_t^{\leftarrow})\} \,\mathrm{d}t + \sigma_{T-t} \,\mathrm{d}B_t \,. \tag{3.3.6}$$

If we initialize this SDE with $X_0^{\leftarrow} \sim \pi_T$, then $X_t^{\leftarrow} \sim \pi_t^{\leftarrow} = \pi_{T-t}$ for all $t \in [0,T]$.

If we initialize this SDE at $X_0^{\leftarrow} = x_T$, does it then follow that $X_T^{\leftarrow} \sim \pi_{0|T}(\cdot \mid x_T)$, where $\pi_{0|T}$ denotes the conditional distribution of X_0 given X_T? This would follow if we knew that $(X_0^{\leftarrow}, X_T^{\leftarrow})$ has the same joint distribution as (X_T, X_0), but this is not clear from the above derivation, which produced the process $X^{\leftarrow}$ by matching only the *marginal* laws. To see that this statement indeed holds, we will instead apply the conditioning argument from the previous section.

In the previous section, recalling that $\mathbf{P}$ is the measure $\mathbf{Q}$ conditioned on $X_T = x_T$, we know that $\mathrm{law}^{\mathbf{P}}(X_t) = \pi_{t|T}(\cdot \mid x_T)$. Therefore, from (3.3.3) and writing $\pi_{t|T} \coloneqq \pi_{t|T}(\cdot \mid x_T)$ to lighten the notation, we deduce that

$$\partial_t \pi_{t|T} = \frac{1}{2} \langle \sigma_t \sigma_t^\mathsf{T}, \nabla^2 \pi_{t|T} \rangle - \mathrm{div}\Big(\pi_{t|T} \left(b_t + \sigma_t \sigma_t^\mathsf{T} \nabla \ln \frac{\mathrm{d}P_{t,T}}{\mathrm{d}\pi_T} \right) \Big) .$$

The time reversal $\pi^{\leftarrow}_{t|T} \coloneqq \pi_{T-t|T}$ satisfies

$$\partial_t \pi^{\leftarrow}_{t|T} = -\frac{1}{2} \langle \sigma_{T-t} \sigma_{T-t}^\mathsf{T}, \nabla^2 \pi^{\leftarrow}_{t|T} \rangle + \mathrm{div}\Big(\pi^{\leftarrow}_{t|T} \left(b_{T-t} + \sigma_{T-t} \sigma_{T-t}^\mathsf{T} \nabla \ln \frac{\mathrm{d}P_{T-t,T}}{\mathrm{d}\pi_T} \right) \Big) . \tag{3.3.7}$$

As an application of the Bayes rule,

$$\frac{\mathrm{d}P_{T-t,T}(x, \cdot)}{\mathrm{d}\pi_T}(x_T) = \frac{\pi_{T|T-t}(x_T \mid x)}{\pi_T(x_T)} = \frac{\pi_{T-t|T}(x \mid x_T)}{\pi_{T-t}(x)}$$

and

$$\nabla \ln \frac{\pi_{T-t|T}(\cdot \mid x_T)}{\pi_{T-t}} = \nabla \ln \pi^{\leftarrow}_{t|T} - \nabla \ln \pi^{\leftarrow}_t .$$

Substituting this into (3.3.7) and applying the logarithmic derivative trick,

$$\partial_t \pi^{\leftarrow}_{t|T} = \frac{1}{2} \langle \sigma_{T-t} \sigma_{T-t}^\mathsf{T}, \nabla^2 \pi^{\leftarrow}_{t|T} \rangle + \mathrm{div}\big(\pi^{\leftarrow}_{t|T} \left(b_{T-t} - \sigma_{T-t} \sigma_{T-t}^\mathsf{T} \nabla \ln \pi^{\leftarrow}_t \right) \big) . \tag{3.3.8}$$

Observe that this is the same Fokker–Planck equation as (3.3.5), except that it is now satisfied by $t \mapsto \pi^{\leftarrow}_{t|T}$ rather than $t \mapsto \pi^{\leftarrow}_t$. Therefore, the SDE corresponding to (3.3.8) is the same SDE (3.3.6) above. Since $\pi^{\leftarrow}_{0|T} = \delta_{x_T}$, this confirms that initializing (3.3.6) with $X^{\leftarrow}_0 = x_T$ yields $X^{\leftarrow}_t \sim \pi^{\leftarrow}_{t|T}$ for all $t \in [0, T]$.

Note that *the time reversal of the SDE depends on the initial distribution.*

3.4 Föllmer Drift

As an application of the Doob transform, we now introduce a process attributed to [Föl85]. Under $\mathbf{Q}$, let $(X_t)_{t \in [0,1]}$ be a standard Brownian motion started at 0. Then, the law of X_1 is standard Gaussian, $\gamma \coloneqq \mathrm{normal}(0, I_d)$. Next, we take $\mathbf{P}$ to be a path measure under which $X_1 \sim \mu$, for some other probability measure μ with $\mu \ll \gamma$.

To achieve this, we can use the construction in Section 3.3, namely, we take $\frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\mathbf{Q}_1} = h_1(X_1)$ where $h_1 = \frac{\mathrm{d}\mu}{\mathrm{d}\gamma}$. This yields $h_t = P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}$, and

$$\mathrm{d}X_t = \nabla \ln P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\,\mathrm{d}t + \mathrm{d}\tilde{B}_t\,, \qquad X_0 = 0\,,$$

where $\tilde{B}$ is the $\mathbf{P}$-Brownian motion. This process is known as the **Föllmer process**, and the added drift term is known as the **Föllmer drift**. The fundamental property enjoyed by this process (or more specifically, by $\mathbf{P}$) is that

$$\mathsf{KL}(\mathbf{P} \,\|\, \mathbf{Q}) = \mathbb{E}^{\mathbf{P}} \ln \frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_1) = \mathbb{E}_\mu \ln \frac{\mathrm{d}\mu}{\mathrm{d}\gamma} = \mathsf{KL}(\mu \,\|\, \gamma)\,. \tag{3.4.1}$$

On the other hand, if $\hat{\mathbf{P}}$ is any other path measure under which $X_1 \sim \mu$, then by the data-processing inequality (Theorem 1.5.3) we have $\mathsf{KL}(\hat{\mathbf{P}} \,\|\, \mathbf{Q}) \geq \mathsf{KL}(\mu \,\|\, \gamma)$. This reflects a certain *entropy optimality* property for the Föllmer process. Moreover, if $\hat{B}$ is a $\hat{\mathbf{P}}$-Brownian motion and under $\hat{\mathbf{P}}$,

$$\mathrm{d}X_t = \hat{b}_t(X_t)\,\mathrm{d}t + \mathrm{d}\hat{B}_t\,,$$

then by Girsanov's theorem (Theorem 3.2.6), this entropy optimality property becomes

$$\mathsf{KL}(\mathbf{P} \,\|\, \mathbf{Q}) = \frac{1}{2}\,\mathbb{E}^{\mathbf{P}} \int_0^1 \left\|\nabla \ln P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\right\|^2 \mathrm{d}t \leq \frac{1}{2}\,\mathbb{E}^{\hat{\mathbf{P}}} \int_0^1 \|\hat{b}_t(X_t)\|^2\,\mathrm{d}t = \mathsf{KL}(\hat{\mathbf{P}} \,\|\, \mathbf{Q})\,. \tag{3.4.2}$$

We say that among all drifts that drive the process to satisfy $X_1 \sim \mu$, the Föllmer drift has minimal "energy".

Moreover, by the chain rule for the KL divergence,

$$\mathsf{KL}(\mathbf{P} \,\|\, \mathbf{Q}) = \mathsf{KL}(\mu \,\|\, \gamma) + \int \mathsf{KL}(\mathbf{P}^{|X_1=x} \,\|\, \mathbf{Q}^{|X_1=x})\,\mu(\mathrm{d}x)\,,$$

where $\mathbf{P}^{|X_1=x} = \mathrm{law}_{\mathbf{P}}((X_t)_{0<t<1} \mid X_1 = x)$ and similarly for $\mathbf{Q}^{|X_1=x}$. The optimality property (3.4.1) for the Föllmer process entails that the second term above vanishes, which means that under $\mathbf{P}$, the conditional law of the path given $X_1 = x$ is the same as the corresponding conditional law under $\mathbf{Q}$. The latter corresponds to the Brownian bridge (see Example 3.3.4).

3.4.1 Application to Functional Inequalities

The use of the Föllmer drift as a potent tool for establishing functional inequalities was perhaps pioneered by Lehec [Leh13], although he attributes the idea earlier, e.g., to Borell [Bor00]. Here, we demonstrate its power to establish the Gaussian log-Sobolev and T_2 inequalities, and refer to [Leh13] and subsequent literature for further applications.

Transport inequality. Under $\mathbf{P}$, we know that $X_1 \sim \mu$ and $\tilde{B}_1 \sim \gamma$. Hence,

$$W_2^2(\mu,\gamma) \le \mathbb{E}^{\mathbf{P}}[\|X_1 - \tilde{B}_1\|^2] = \mathbb{E}^{\mathbf{P}}\Big[\Big\|\int_0^1 \nabla \ln P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\,\mathrm{d}t\Big\|^2\Big]$$
$$\le \mathbb{E}^{\mathbf{P}}\int_0^1 \Big\|\nabla \ln P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\Big\|^2\,\mathrm{d}t = 2\,\mathsf{KL}(\mu \parallel \gamma)\,,$$

where we used (3.4.1) and (3.4.2) in the last line.

Log-Sobolev inequality. Let $h_t := P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}$ and recall from the construction of the Doob transform that $(h_t(X_t))_{t\in[0,1]}$ is a $\mathbf{Q}$-martingale. We claim that $(\nabla \ln h_t(X_t))_{t\in[0,1]}$ is a $\mathbf{P}$-martingale. To prove this, let $s \le t$ and let $A_s \in \mathscr{F}_s$. Then,

$$\mathbb{E}^{\mathbf{P}_t}[\nabla \ln h_t(X_t)\,\mathbb{1}_{A_s}] = \mathbb{E}^{\mathbf{P}_t}\Big[\frac{\nabla h_t(X_t)}{h_t(X_t)}\,\mathbb{1}_{A_s}\Big] = \mathbb{E}^{\mathbf{Q}_t}[\nabla h_t(X_t)\,\mathbb{1}_{A_s}] = \mathbb{E}^{\mathbf{Q}_t}[\nabla P_{1-t}h_1(X_t)\,\mathbb{1}_{A_s}]$$
$$= \mathbb{E}^{\mathbf{Q}_t}[P_{1-t}\nabla h_1(X_t)\,\mathbb{1}_{A_s}] = \mathbb{E}^{\mathbf{Q}_s}\big[\mathbb{E}[P_{1-t}\nabla h_1(X_t) \mid \mathscr{F}_s]\,\mathbb{1}_{A_s}\big]$$
$$= \mathbb{E}^{\mathbf{Q}_s}[P_{1-s}\nabla h_1(X_s)\,\mathbb{1}_{A_s}]\,.$$

Applying this equality for $s = t$ yields $\mathbb{E}^{\mathbf{P}_t}[\nabla \ln h_t(X_t)\,\mathbb{1}_{A_s}] = \mathbb{E}^{\mathbf{P}_s}[\nabla \ln h_s(X_s)\,\mathbb{1}_{A_s}]$, or

$$\mathbb{E}^{\mathbf{P}}[\nabla \ln h_t(X_t) \mid \mathscr{F}_s] = \nabla \ln h_s(X_s)\,.$$

In particular, $t \mapsto \|\nabla \ln h_t(X_t)\|^2$ is a $\mathbf{P}$-submartingale.

Now, using the equality for the KL divergence and the submartingale property,

$$\mathsf{KL}(\mu \parallel \gamma) = \frac{1}{2}\,\mathbb{E}^{\mathbf{P}}\int_0^1 \Big\|\nabla \ln P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\Big\|^2\,\mathrm{d}t \le \frac{1}{2}\,\mathbb{E}^{\mathbf{P}}\Big[\Big\|\nabla \ln \frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_1)\Big\|^2\Big] = \frac{1}{2}\,\mathsf{FI}(\mu \parallel \gamma)\,.$$

3.4.2 Connection to Stochastic Localization

In this section, we relate the Föllmer process to Eldan's **stochastic localization** scheme (introduced in [Eld13]), which by now has solidified its status as a core tool in high-dimensional probability. Although we do not have space in this book to describe its

many applications, we present some basic ideas here to help the reader understand the connections with the extant literature; see [KP21] for more details.

Stochastic localization is a method of understanding a probability measure μ by decomposing it into simpler parts, with the goal of, e.g., establishing functional inequalities or other useful properties for μ. It was inspired by earlier work on deterministic localization schemes [KLS95], but instead seeks to produce a *random* measure-valued process $(p_t)_{t\geq 0}$. This process is such that $p_0 = \mu$, $p_\infty = \delta_X$ for some random variable X, and $(p_t)_{t\geq 0}$ is a martingale. The last property implies that $X \sim \mu$, and indeed, $\mathbb{E}\, p_t = \mu$ for all $t \geq 0$. This is the decomposition of μ into "simpler parts" as alluded to earlier.

How might we build such a process? We motivate the process via a Bayesian interpretation. Consider the process $\theta_t := tX + B_t$, where as usual $(B_t)_{t\geq 0}$ is a Brownian motion, independent of X. At time $t \approx 0$, the Brownian motion dominates, so θ_t contains almost no information about X. At time $t \to \infty$, the linear term tX dominates and θ_t contains nearly perfect information about X. Therefore, if we set p_t to be the conditional law of X given the observation θ_t, then we expect $p_0 = \mu$ and $p_\infty = \delta_X$. One can check that $(p_t)_{t\geq 0}$ is indeed a martingale.

We can relate this process to the usual heat flow via rescaling: let $\breve{\theta}_t := \theta_t/t$, so that $\breve{\theta}_t = X + B_t/t$. The time inversion property of Brownian motion implies that $(\breve{B}_t)_{t\geq 0}$ is also a Brownian motion, where $\breve{B}_{1/t} := B_t/t$. Hence, $\breve{\theta}_t = X + \breve{B}_{1/t}$ is the output of the heat flow, started at X, after time $1/t$ (note the time inversion). If we let $\rho_{0,s}$ denote the joint distribution of the heat flow, started at μ, at times 0 and s, and denote conditional distributions accordingly, we can write $p_t = \mathrm{law}(X \mid \breve{\theta}_t) = \rho_{0|t^{-1}}(\cdot \mid \theta_t/t)$. Thus,

$$p_t(x) \propto \mu(x) \exp\Big(-\frac{t\,\|x - \theta_t/t\|^2}{2}\Big).$$

We now want to take logarithms in this expression and apply Itô's formula to derive a stochastic evolution equation. Before doing so, note that the equation $\theta_t = tX + B_t$, which seems to give $\mathrm{d}\theta_t = X\,\mathrm{d}t + \mathrm{d}B_t$, does not express θ as a Markov process (since X is not adapted to the filtration at time $t < \infty$). However, one can replace this equation by the equivalent Markov evolution $\mathrm{d}\theta_t = \mathbb{E}[X \mid \mathscr{F}_t]\,\mathrm{d}t + \mathrm{d}B_t = a_t\,\mathrm{d}t + \mathrm{d}B_t$, where we set $a_t := \int x\, p_t(\mathrm{d}x)$; see [KP21]. Then, a calculation starting from

$$\mathrm{d}\ln p_t(x) = -\mathrm{d}\Big(\frac{t\,\|x - \theta_t/t\|^2}{2}\Big) - \mathrm{d}\ln \int \exp\Big(-\frac{t\,\|y - \theta_t/t\|^2}{2}\Big)\,\mu(\mathrm{d}y)$$

eventually yields (Exercise 3.2)

$$\mathrm{d}p_t(x) = p_t(x)\,\langle x - a_t, \mathrm{d}B_t\rangle\,, \tag{3.4.3}$$

which was the form in which stochastic localization was originally introduced.

To see the connection with the Föllmer process $(F_t)_{t\in[0,1]}$ with $F_1 \sim \mu$, recall that given F_1, the law of $(F_t)_{0<t<1}$ is a Brownian bridge. In other words, $F_t = tF_1 + \mathsf{BB}_t$, where $(\mathsf{BB}_t)_{t\in[0,1]}$ is the Brownian bridge process starting and ending at 0. We can identify $F_1 = X$, in which case this expression for F nearly resembles the expression for the tilt process θ in stochastic localization. To make this precise, we claim that the Brownian bridge can be constructed as $\mathsf{BB}_t = (1-t)\, B_{t/(1-t)}$. With this identification, then $F_t = (1-t)\, \theta_{t/(1-t)}$, i.e., the Föllmer process is a rescaled time compression of the tilt process θ from $\mathbb{R}_+$ to the interval $[0,1]$; see Exercise 3.3 for details.

This discussion also shows that these concepts are related to the idea of running the heat flow *backward* in time (e.g., we consider the conditional distribution $\rho_{0|t^{-1}}$ above). These ideas will reappear in Chapter 8 as the proximal sampler.

3.5 Schrödinger Bridge

In this section, we consider a generalization of the Föllmer process. The setup arises from a hot gas *Gedankenexperiment* due to Schrödinger. Let μ and ν be two probability measures over $\mathbb{R}^d$, representing the observed distribution of a cloud of particles at times 0 and 1 respectively. In the absence of the observation of ν, we may have modelled the evolution of the gas particles as a scaled Brownian motion: $X_t = X_0 + \sqrt{\varepsilon}\, \mathrm{d}B_t$ for $t \in [0,1]$, where $X_0 \sim \mu$ and $(B_t)_{t\in[0,1]}$ are independent, and $\varepsilon > 0$ represents the noise level of the process. However, if the observed distribution ν differs from the law $\mu * \mathsf{normal}(0, \varepsilon I_d)$ of X_1 in our model, what then is our best guess for the law of the trajectory $(X_t)_{t\in[0,1]}$? The law of the trajectory is said to *bridge* the distributions μ_0 and μ_1.

Schrödinger formulated this as a KL minimization problem:

$$\underset{\mathbf{P}\in\mathcal{P}(C([0,1]))}{\text{minimize}} \quad \mathsf{KL}(\mathbf{P} \parallel \mathbf{W}^{\mu,\varepsilon}) \qquad \text{such that} \qquad (X_0)_{\#}\mathbf{P} = \mu\,,\ (X_1)_{\#}\mathbf{P} = \nu\,.$$

Here, the minimization takes place over the set of path measures (probability measures over $C([0,1])$), and $\mathbf{W}^{\mu,\varepsilon}$ denotes the path measure corresponding to Brownian motion, rescaled by $\sqrt{\varepsilon}$ and started at μ. The choice of KL divergence as our criterion can be motivated by large deviations theory.

We can solve this problem as follows. If we condition on the endpoints X_0 and X_1 and apply the KL chain rule, we end up with

$$\mathsf{KL}(\mathbf{P} \parallel \mathbf{W}^{\mu,\varepsilon}) = \mathsf{KL}\big(\mathrm{law}_{\mathbf{P}}(X_0, X_1) \,\big\|\, \mathrm{law}_{\mathbf{W}^{\mu,\varepsilon}}(X_0, X_1)\big) + \mathbb{E}^{\mathbf{P}}\, \mathsf{KL}(\mathbf{P}^{|X_0,X_1} \parallel \mathbf{W}^{\mu,\varepsilon|X_0,X_1})\,,$$

where $\mathbf{P}^{|X_0,X_1}$, $\mathbf{W}^{\mu,\varepsilon|X_0,X_1}$ denote the path measures $\mathbf{P}$, $\mathbf{W}^{\mu,\varepsilon}$ conditioned on (X_0, X_1) respectively. Also, $\mathbf{W}^{\mu,\varepsilon|X_0,X_1}$ is a Brownian bridge (rescaled by $\sqrt{\varepsilon}$), and the second term above can be made zero by setting $\mathbf{P}^{|X_0,X_1} = \mathbf{W}^{\mu,\varepsilon|X_0=X_1}$.

Thus far, the development has closely mirrored our discussion of the Föllmer process, which is the special case of the Schrödinger bridge when $\mu = \delta_0$. The interesting new features of this more general setting arise, however, when we consider minimizing the first term in the KL chain rule, which is a minimization over joint distributions for (X_0, X_1) with $X_0 \sim \mu$ and $X_1 \sim \nu$, i.e., a *coupling* of μ and ν. In the Föllmer case, this minimization problem was trivial, essentially because the space of couplings of a Dirac measure δ_0 and any other measure ν is also trivial (consisting solely of $\delta_0 \otimes \nu$). Our goal now is to understand this minimization problem when the space of couplings is non-trivial.

3.5.1 Entropically Regularized Optimal Transport

Our first step is to note that for $\eta := \mathrm{law}_{\mathbf{W}^{\mu,\varepsilon}}(X_0, X_1)$,

$$\eta(\mathrm{d}x, \mathrm{d}y) \propto \mu(\mathrm{d}x) \exp\Big(-\frac{\|y-x\|^2}{2\varepsilon}\Big)\,\mathrm{d}y\,.$$

Therefore, we can explicitly write, for $\gamma := \mathrm{law}_{\mathbf{P}}(X_0, X_1)$,

$$\begin{aligned}\mathsf{KL}(\gamma \parallel \eta) &= \frac{1}{2\varepsilon}\int \|x-y\|^2\,\gamma(\mathrm{d}x,\mathrm{d}y) + \int \ln\frac{\gamma(x,y)}{\mu(x)}\,\gamma(\mathrm{d}x,\mathrm{d}y) + \text{const.}\\ &= \frac{1}{2\varepsilon}\int \|x-y\|^2\,\gamma(\mathrm{d}x,\mathrm{d}y) + \mathsf{KL}(\gamma \parallel \mu\otimes\nu) + \text{const.}\end{aligned}$$

where we used the fact that $\int \ln \nu(y)\,\gamma(\mathrm{d}x,\mathrm{d}y) = \int \ln \nu\,\mathrm{d}\nu$ does not depend on γ, allowing us to absorb it into the constant term. Hence, the problem of finding the optimal coupling γ between μ and ν for the Schrödinger bridge problem is an entropically regularized variant of the optimal transport problem from Section 1.3. More generally, we have:

Definition 3.5.1. Let $\mathcal{X}$ and $\mathcal{Y}$ be complete separable metric spaces, let $\varepsilon > 0$, and let $c : \mathcal{X}\times\mathcal{Y} \to [0,\infty]$ be a cost function. The **entropically regularized optimal transport cost** from $\mu \in \mathcal{P}(\mathcal{X})$ to $\nu \in \mathcal{P}(\mathcal{Y})$ with cost c is

$$\mathcal{T}_{c,\varepsilon}(\mu,\nu) := \inf_{\gamma\in\mathcal{C}(\mu,\nu)}\Big[\int c(x,y)\,\gamma(\mathrm{d}x,\mathrm{d}y) + \varepsilon\,\mathsf{KL}(\gamma \parallel \mu\otimes\nu)\Big]\,.$$

Entropic optimal transport was introduced to speed up computation of optimal transport costs in [Cut13]; see the bibliographical notes for further discussion. One can show that if c is lower semicontinuous, then there always exists a unique entropic optimal transport plan. Note that unlike Theorem 1.3.8, which required μ to have a density in

order for uniqueness of the optimal transport plan with quadratic cost, here uniqueness always holds as a consequence of the strict convexity of the KL divergence.

To summarize our observations thus far, we have argued that the solution to the Schrödinger bridge problem is to first draw (X_0, X_1) from the entropic optimal transport plan between μ and ν with cost $c(x,y) = \frac{1}{2}\|x-y\|^2$, and then to join X_0 and X_1 by a Brownian bridge (rescaled by $\sqrt{\varepsilon}$). Although in principle this completes the description of the Schrödinger bridge, we can go further by characterizing the entropic optimal transport plan via a duality principle.

Duality works here similarly as Kantorovich duality did for unregularized optimal transport, and we simply quote the main theorem here.

Theorem 3.5.2 (duality for entropic optimal transport)**.** *There exist maximizers f_ε, g_ε to the dual problem*

$$\sup_{(f,g)\in L^1(\mu)\times L^1(\nu)} \left\{ \int f \,\mathrm{d}\mu + \int g \,\mathrm{d}\nu - \varepsilon \iint \exp\Big(\frac{f\oplus g - c}{\varepsilon}\Big) \,\mathrm{d}(\mu\otimes\nu) + \varepsilon \right\}$$

which are unique up to adding a constant to f_ε and subtracting that same constant from g_ε. The optimal value of the dual problem equals the entropic optimal transport cost from μ to γ, and the entropic optimal transport plan γ_ε is of the form

$$\gamma_\varepsilon(\mathrm{d}x,\mathrm{d}y) = \exp\Big(\frac{f_\varepsilon(x) + g_\varepsilon(y) - c(x,y)}{\varepsilon}\Big)\,\mu(\mathrm{d}x)\,\nu(\mathrm{d}y)\,. \tag{3.5.3}$$

The expression (3.5.3) characterizes the optimal solution in the following sense. If γ_ε is any coupling of μ and ν of the form (3.5.3) for some f_ε, g_ε, then γ_ε is the entropic optimal transport plan, and $(f_\varepsilon, g_\varepsilon)$ is a pair of optimal dual potentials.

Note that compared to the dual problem for Kantorovich duality, we have replaced the "hard" constraint of $f\oplus g \le c$ with the "soft" constraint of adding the penalty term $\iint \exp((f\oplus g - c)/\varepsilon)\,\mathrm{d}(\mu\otimes\nu)$ into the objective.

Let us now specialize to the case of quadratic cost, $c(x,y) = \frac{1}{2}\|x-y\|^2$. In this case, it is natural to work with $\varphi_\varepsilon := \frac{1}{2}\|\cdot\|^2 - f_\varepsilon$ and $\psi_\varepsilon := \frac{1}{2}\|\cdot\|^2 - g_\varepsilon$, since then (provided that we fix a normalization for the potentials, e.g., $\int \varphi_\varepsilon \,\mathrm{d}\mu = \int \psi_\varepsilon \,\mathrm{d}\nu$) we have the convergence $\varphi_\varepsilon \to \varepsilon$ and $\psi_\varepsilon \to \psi$ of the entropic potentials to their unregularized counterparts as $\varepsilon \searrow 0$

(see [NW22]). The condition that γ_ε has marginals μ and ν yields the coupled equations

$$\begin{aligned} \varphi_\varepsilon(x) &= \varepsilon \ln \int \exp\Big(\frac{\langle x, y\rangle - \psi_\varepsilon(y)}{\varepsilon}\Big)\, \nu(\mathrm{d}y)\,, \\ \psi_\varepsilon(y) &= \varepsilon \ln \int \exp\Big(\frac{\langle x, y\rangle - \varphi_\varepsilon(x)}{\varepsilon}\Big)\, \mu(\mathrm{d}x)\,. \end{aligned} \tag{3.5.4}$$

From these expressions, one can prove the following lemma (see Exercise 3.4).

Lemma 3.5.5. *The following relations hold:*

$$\nabla\varphi_\varepsilon(x) = \mathbb{E}_{\gamma_\varepsilon}[Y \mid X = x]\,, \qquad \nabla\psi_\varepsilon(y) = \mathbb{E}_{\gamma_\varepsilon}[X \mid Y = y]\,.$$

Also,

$$\nabla^2\varphi_\varepsilon(x) = \frac{1}{\varepsilon}\operatorname{cov}_{\gamma_\varepsilon}(Y \mid X = x)\,, \qquad \nabla^2\psi_\varepsilon(y) = \frac{1}{\varepsilon}\operatorname{cov}_{\gamma_\varepsilon}(X \mid Y = y)\,.$$

Since covariance matrices are always positive semidefinite, these expressions witness the convexity of φ_ε and ψ_ε, and provide another explanation for Brenier's theorem (Theorem 1.3.8).

3.5.2 Caffarelli's Contraction Theorem

We now provide an application of the theory of entropic optimal transport to functional inequalities. We will establish bounds on the Hessian of entropic potential which, as $\varepsilon \searrow 0$, furnish bounds on the Brenier potential for the unregularized optimal transport problem. This will yield a proof of Caffarelli's contraction theorem.

Theorem 3.5.6 ([CP23])**.** *Suppose that μ is β-log-smooth and that ν is α-strongly log-concave, i.e., $\nabla^2 \log(1/\mu) \preceq \beta I_d$ and $\nabla^2 \log(1/\nu) \succeq \alpha I_d$. Then, the entropic Brenier potential φ_ε from μ to ν satisfies*

$$\nabla^2\varphi_\varepsilon \preceq \frac{1}{2}\left(\sqrt{4\beta/\alpha + \varepsilon^2\beta^2} - \varepsilon\beta\right) I\,.$$

Letting $\varepsilon \searrow 0$, one readily obtains (see [CP23]):

Corollary 3.5.7 (Caffarelli's contraction theorem). *Suppose that μ is β-log-smooth and that ν is α-strongly log-concave, i.e., $\nabla^2 \log(1/\mu) \preceq \beta I_d$ and $\nabla^2 \log(1/\nu) \succeq \alpha I_d$. Then, the Brenier map $\nabla\varphi$ from μ to ν is $\sqrt{\beta/\alpha}$-Lipschitz.*

The proof of Theorem 3.5.6 will exploit the representation of the Hessians of the entropic Brenier potentials as covariance matrices (Lemma 3.5.5), together with a pair of covariance inequalities.

Theorem 3.5.8 (Cramér–Rao inequality). *Let $\pi \propto \exp(-V)$ be a probability measure over $\mathbb{R}^d$. For any well-behaved function $f : \mathbb{R}^d \to \mathbb{R}$, it holds that*

$$\operatorname{var}_\pi f \geq \langle \mathbb{E}_\pi f, (\mathbb{E}_\pi \nabla^2 V)^{-1} \mathbb{E}_\pi f \rangle .$$

Proof. Integration by parts and $\mathbb{E}_\pi \nabla V = 0$ yield

$$\mathbb{E}_\pi \nabla f = \int \nabla f \, \mathrm{d}\pi = \int \left(f \nabla \ln \frac{1}{\pi}\right) \mathrm{d}\pi = \mathbb{E}_\pi [(f - \mathbb{E}_\pi f) \nabla V] .$$

Therefore,

$$\begin{aligned} \langle \mathbb{E}_\pi \nabla f, (\mathbb{E}_\pi \nabla^2 V)^{-1} \mathbb{E}_\pi \nabla f \rangle &= \mathbb{E}_\pi \left[(f - \mathbb{E}_\pi f) \langle \nabla V, (\mathbb{E}_\pi \nabla^2 V)^{-1} \mathbb{E}_\pi \nabla f \rangle \right] \\ &\leq \sqrt{(\operatorname{var}_\pi f) \, \mathbb{E}_\pi \left\langle \mathbb{E}_\pi \nabla f, (\mathbb{E}_\pi \nabla^2 V)^{-1} (\nabla V)^{\otimes 2} (\mathbb{E}_\pi \nabla^2 V)^{-1} \mathbb{E}_\pi \nabla f \right\rangle} . \end{aligned}$$

Another integration by parts shows that $\mathbb{E}_\pi [(\nabla V)^{\otimes 2}] = \mathbb{E}_\pi \nabla^2 V$, so the result follows by rearranging the above expression. □

Corollary 3.5.9 (covariance bounds). *Let $\pi \propto \exp(-V)$ be a probability measure over $\mathbb{R}^d$ and let cov_π denote its covariance matrix. Then,*

$$(\mathbb{E}_\pi \nabla^2 V)^{-1} \preceq \operatorname{cov}_\pi \preceq \mathbb{E}_\pi [(\nabla^2 V)^{-1}] .$$

Proof. The lower and upper bounds follow respectively from the Cramér–Rao inequality (Theorem 3.5.8) and the Brascamp–Lieb inequality (Theorem 2.2.8) by taking test functions $f = \langle e, \cdot \rangle$ for unit vectors $e \in \mathbb{R}^d$. □

Proof of Theorem 3.5.6. We write $\mu \propto \exp(-V)$ and $\nu \propto \exp(-W)$. For any $x \in \mathbb{R}^d$, from Lemma 3.5.5 and the upper bound in Corollary 3.5.9, we obtain

$$\nabla^2 \varphi_\varepsilon(x) = \varepsilon^{-1} \operatorname{cov}_{\gamma_\varepsilon^{Y|X=x}} \preceq \varepsilon^{-1} \mathbb{E}_{\gamma_\varepsilon^{Y|X=x}} [(\varepsilon^{-1} \nabla^2 \psi_\varepsilon + \nabla^2 W)^{-1}] \preceq \mathbb{E}_{\gamma_\varepsilon^{Y|X=x}} [(\nabla^2 \psi_\varepsilon + \varepsilon\alpha I)^{-1}] .$$

For any $y \in \mathbb{R}^d$, from Lemma 3.5.5 and the lower bound in Corollary 3.5.9,

$$\nabla^2 \psi_\varepsilon(y) = \varepsilon^{-1} \operatorname{cov}_{\gamma_\varepsilon^{X|Y=y}} \geq \varepsilon^{-1} \left(\mathbb{E}_{\gamma_\varepsilon^{X|Y=y}} [\varepsilon^{-1} \nabla^2 \varphi_\varepsilon + \nabla^2 V] \right)^{-1} \geq \left(\mathbb{E}_{\gamma_\varepsilon^{X|Y=y}} [\nabla^2 \varphi_\varepsilon + \varepsilon \beta I] \right)^{-1}.$$

Let $L_\varepsilon := \sup_{x \in \mathbb{R}^d} \lambda_{\max}(\nabla^2 \varphi_\varepsilon(x))$. From the two inequalities above, we can conclude that

$$\lambda_{\max}\left(\nabla^2 \varphi_\varepsilon(x)\right) \leq \left((L_\varepsilon + \varepsilon\beta)^{-1} + \varepsilon\alpha \right)^{-1}$$

and hence

$$L_\varepsilon \leq \left((L_\varepsilon + \varepsilon\beta)^{-1} + \varepsilon\alpha \right)^{-1}.$$

Solving the quadratic inequality yields the upper bound on L_ε in the theorem. □

As discussed in Section 2.3, if we apply Caffarelli's contraction theorem taking μ as the standard Gaussian measure (so $\beta = 1$), we deduce that the optimal transport map from the standard Gaussian to any α-strongly log-concave measure is $\alpha^{-1/2}$-Lipschitz. Together with the preservation of functional inequalities under Lipschitz mappings (Proposition 2.3.3), it allows us to transfer functional inequalities satisfied by the standard Gaussian measure to all strongly log-concave measures. In this way, Caffarelli's contraction theorem is a "universal blueprint" for proving such inequalities. This point was already made in Caffarelli's original paper [Caf00].

For example, one can use it to transfer the functional inequalities established for the standard Gaussian in Section 3.4 to strongly log-concave measures. As another example, one can prove the Gaussian isoperimetric inequality in Theorem 2.5.27 by first establishing it for Gaussians and appealing to Caffarelli contraction.

Bibliographical Notes

The discussion of quadratic variation and Girsanov's theorem is heavily inspired by the treatment in [Le 16].

The study of functions of bounded variation and the relationship with absolute continuity and total variation can be found in any standard graduate text on real analysis, e.g., [Fol99]. See [Ste01, Proposition 8.6] for a simple case of Theorem 3.1.5.

The discussion of abstract Wiener spaces and the Cameron–Martin theorem is the starting point of calculus on path space, which is usually called the **Malliavin calculus**. To illustrate, suppose we have a functional $F(B)$ of the Brownian path B, and we ask how $F(B)$ changes under infinitesimal perturbations $B \mapsto B + h$. The key point is that one should restrict to directions h which belong to the Cameron–Martin space, which leads to the concept of the Malliavin derivative.

Girsanov's theorem is also used heavily in mathematical finance, and for this purpose the book [Ste01] is warmly recommended.

See [CGP21] for an introduction to the Schrödinger bridge problem. [TODO: Literature on entropic optimal transport.]

Exercises

Quadratic Variation

Change of Measure in Path Space

⊵ **Exercise 3.1** (proof of Girsanov's theorem)
Let $0 = t_0 < t_1 < \cdots < t_n \leq T$ and $\theta_1, \dots, \theta_n \in \mathbb{R}^d$. Let b, B, and $\tilde{B}$ be as in Theorem 3.2.6. Using the formula for the Radon–Nikodym derivative of $\mathbf{P}$ w.r.t. $\mathbf{W}$, compute $\mathbb{E}^{\mathbf{P}} \exp(\mathrm{i} \sum_{i=1}^n \langle \theta_i, \tilde{B}_{t_i} - \tilde{B}_{t_{i-1}} \rangle)$ and deduce that $\tilde{B}$ is a $\mathbf{P}$-Brownian motion.

Doob's Transform

Föllmer Drift

⊵ **Exercise 3.2** (derivation of stochastic localization)
Derive the evolution equation (3.4.3) for stochastic localization.

⊵ **Exercise 3.3** (Föllmer and stochastic localization)
Let $\mathrm{BB}_t := (1-t)\, B_{t/(1-t)}$. By solving the SDE that we derived for Brownian bridge in Example 3.3.4 (with $x_T = 0$), show that BB is indeed a Brownian bridge. Then, verify that $F_t = (1-t)\, \theta_{t/(1-t)}$.

Schrödinger Bridge

⊵ **Exercise 3.4** (entropic potentials)
Derive (3.5.4) and use it to prove Lemma 3.5.5 for the entropic potentials.

Part II

Complexity of Sampling

CHAPTER 4

Analysis of Langevin Monte Carlo

In this chapter, we will provide several analyses of the **Langevin Monte Carlo** (**LMC**) algorithm, i.e., the iteration

$$X_{(k+1)h} \coloneqq X_{kh} - h\,\nabla V(X_{kh}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,. \tag{LMC}$$

This is known as the **Euler–Maruyama discretization** of the Langevin diffusion.

Although LMC does not achieve state-of-the-art complexity bounds, it is one of the most fundamental sampling algorithms. Through the quantitative convergence analysis of LMC, we will develop techniques for discretization analysis that are broadly useful for studying more complex algorithms.

To emphasize the kinship of optimization and sampling as the core theme of this book, we include "optimization boxes" which provide background and context for the corresponding results in optimization.

4.1 Proof via Wasserstein Coupling

Perhaps the most straightforward analysis of LMC is based on coupling together the discrete-time algorithm with the continuous-time diffusion, and using this coupling to bound the discretization error in Wasserstein distance. The underlying continuous-time result we use here is the fact that strong log-concavity implies contraction in the Wasserstein metric for the Langevin diffusion. On one hand, this proof is robust and can

be applied to more complicated processes; on the other hand, its reliance on contractivity means it is not applicable under weaker assumptions such as an LSI.

Before proceeding, we review the corresponding result for gradient descent.

Optimization Box 4.1.1. Let $V : \mathbb{R}^d \to \mathbb{R}$ be strongly convex and smooth, i.e., $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. The **gradient descent** (GD) algorithm with fixed step size $h > 0$ is the iteration $x_{k+1} = x_k - h\nabla V(x_k)$. Using strong convexity, we can show that GD converges exponentially fast to the minimizer $x_\star$ of V. First, note that for any $y \in \mathbb{R}^d$, by expanding the square and applying strong convexity,

$$\begin{aligned}\|x_{k+1} - y\|^2 &= \|x_k - h\nabla V(x_k) - y\|^2 \\ &= \|x_k - y\|^2 - 2h\langle \nabla V(x_k), x_k - y\rangle + h^2 \|\nabla V(x_k)\|^2 \\ &\le (1 - \alpha h)\|x_k - y\|^2 - 2h\{V(x_k) - V(y)\} + h^2\|\nabla V(x_k)\|^2 .\end{aligned}$$

Now take $y = x_\star$. Using the smoothness of V,

$$V(x_{k+1}) - V(x_k) \le \langle \nabla V(x_k), x_{k+1} - x_k\rangle + \frac{\beta}{2}\|x_{k+1} - x_k\|^2 = -h\left(1 - \frac{\beta h}{2}\right)\|\nabla V(x_k)\|^2 .$$

For any $h \le 1/\beta$, it yields $\|\nabla V(x_k)\|^2 \le \frac{2}{h}\{V(x_k) - V(x_{k+1})\} \le \frac{2}{h}\{V(x_k) - V(x_\star)\}$. Substituting this above, it yields $\|x_{k+1} - x_\star\|^2 \le (1-\alpha h)\|x_k - x_\star\|^2$. Choosing $h = 1/\beta$, one obtains $\|x_N - x_\star\| \le \varepsilon$ in $N \le O(\kappa \log(\|x_0 - x_\star\|/\varepsilon))$ iterations.

We now consider the corresponding result for LMC.

Theorem 4.1.2. *For $k \in \mathbb{N}$, let μ_{kh} denote the law of the k-th iterate of LMC with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies $\nabla V(0) = 0$ and $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, provided $h \le \frac{1}{3\beta}$, for all $N \in \mathbb{N}$,*

$$W_2^2(\mu_{Nh}, \pi) \le \exp(-\alpha N h)\, W_2^2(\mu_0, \pi) + O\left(\frac{\beta^4 d h^2}{\alpha^3} + \frac{\beta^2 d h}{\alpha^2}\right). \tag{4.1.3}$$

In particular, if we initialize at $\mu_0 = \delta_0$ and take $h \asymp \frac{\alpha\varepsilon^2}{\beta^2 d}$, then for any $\varepsilon \in [0, \sqrt{d}]$ we obtain the guarantee $\sqrt{\alpha}\, W_2(\mu_{Nh}, \pi) \le \varepsilon$ after

$$N = O\left(\frac{\kappa^2 d}{\varepsilon^2}\log\frac{d}{\varepsilon}\right) \qquad \textit{iterations}.$$

***Remark* 4.1.4.** We pause to make a few comments about the assumptions and result.

1. Typically we assume $\nabla V(0) = 0$ without loss of generality, i.e., that the potential V is minimized at 0. This is because the complexity of finding the minimizer of V via convex optimization is typically less than the complexity of sampling.

2. It is convenient to use the metric $\sqrt{\alpha}\, W_2$ instead of W_2 because it is scale-invariant. Namely, for $\lambda > 0$, if we define the scaling map $s_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ via $x \mapsto \lambda x$, then information divergences such as KL satisfy $\mathsf{KL}((s_\lambda)_\# \mu \,\|\, (s_\lambda)_\# \pi) = \mathsf{KL}(\mu \,\|\, \pi)$. On the other hand, W_2 is not invariant, $W_2((s_\lambda)_\#\mu, (s_\lambda)_\#\pi) = \lambda\, W_2(\mu, \pi)$, but $\sqrt{\alpha}\, W_2$ is (because the distribution $(s_\lambda)_\#\pi$ is α/λ^2-strongly convex).

   Recall also that the T_2 transport inequality, implied by α-strong log-concavity, asserts that $\sqrt{\alpha}\, W_2(\cdot, \pi) \le \sqrt{2\, \mathsf{KL}(\cdot \,\|\, \pi)}$. Therefore, $\sqrt{\alpha}\, W_2$ is a more natural metric.

3. The result in (4.1.3) is not sharp; in Section 4.3, via a more sophisticated analysis and averaging, we will improve the iteration complexity to $\widetilde{O}(d\kappa/\varepsilon^2)$.

4. The inequality (4.1.3) has the following interpretation: for fixed $h > 0$, the first term tends to zero exponentially fast, which reflects the fact that LMC converges to its stationary distribution μ_∞. However, the stationary distribution is *biased*, $\mu_\infty \neq \pi$, and the second term provides an upper bound on the bias $W_2(\mu_\infty, \pi)$. Note the contrast with Optimization Box 4.1.1, in which there is no bias; this will be discussed further in Section 4.3.

First, let us see why the first statement of Theorem 4.1.2 implies the second.

Lemma 4.1.5. *Let $\pi \propto \exp(-V)$ satisfy $\nabla^2 V \succeq \alpha I_d$ and $\nabla V(0) = 0$. Then, we have the second moment bound $\int \|\cdot\|^2 \,\mathrm{d}\pi \le d/\alpha$.*

Moreover, if $h \le \frac{1}{\beta}$, then the LMC iterates initialized at $\mu_0 = \delta_0$ with step size $h > 0$ have uniformly bounded second moment: $\sup_{k \in \mathbb{N}} \mathbb{E}[\|X_{kh}\|^2] \lesssim d/\alpha$.

Proof. See Exercise 4.2. □

By taking $h \asymp \frac{\alpha\varepsilon^2}{\beta^2 d}$, we can make the second term in (4.1.3) at most $\frac{\varepsilon^2}{2\alpha}$, and then for all $N \gtrsim \frac{1}{\alpha h} \log(\alpha\, W_2^2(\delta_0, \pi)/\varepsilon^2) \asymp \frac{d\kappa^2}{\varepsilon^2} \log(d/\varepsilon)$ the first term is also at most $\frac{\varepsilon^2}{2\alpha}$. This proves the second statement of Theorem 4.1.2.

We now prove the first statement.

Proof of Theorem 4.1.2. **1. One-step discretization bound.** Suppose that the continuous-time Langevin diffusion and the LMC algorithm are both initialized at the same measure μ_0. We will first bound the discretization error $W_2^2(\mu_h, \pi_h)$ in one step.

We couple the two processes by taking $X_0 = Z_0$ and using the same Brownian motion:

$$X_h = Z_0 - h\,\nabla V(Z_0) + \sqrt{2}\,B_h\,,$$
$$Z_h = Z_0 - \int_0^h \nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,B_h\,.$$

Then,

$$W_2^2(\mu_h, \pi_h) \le \mathbb{E}[\|X_h - Z_h\|^2] \le \mathbb{E}\Big[\Big\|\int_0^h \nabla V(Z_t)\,\mathrm{d}t - h\,\nabla V(Z_0)\Big\|^2\Big]$$
$$\le h\int_0^h \mathbb{E}[\|\nabla V(Z_t) - \nabla V(Z_0)\|^2]\,\mathrm{d}t\,.$$

Therefore, we just have to bound the movement $\|Z_t - Z_0\| = \|-\int_0^t \nabla V(Z_s)\,\mathrm{d}s + \sqrt{2}\,B_t\|$ of the Langevin diffusion in time t. Roughly, we expect $\|\int_0^t \nabla V(Z_s)\,\mathrm{d}s\| = O(\sqrt{d}\,t)$ if the size of the gradient is $O(\sqrt{d})$, and $\|B_t\| = O(\sqrt{dt})$. For small t, it is the Brownian motion term which is dominant, which is a common intuition for discretization proofs.

To rigorously bound this term, we appeal to stochastic calculus, see Lemma 4.1.9. For $t \le \frac{1}{3\beta}$, it yields the bound

$$\mathbb{E}[\|Z_t - Z_0\|^2] \le 8\beta^2 t^2\,\mathbb{E}[\|Z_0\|^2] + 8dt$$

and hence

$$W_2^2(\mu_h, \pi_h) \le \beta^2 h\int_0^h \mathbb{E}[\|Z_t - Z_0\|^2]\,\mathrm{d}t \le 3\beta^4 h^4\,\mathbb{E}[\|Z_0\|^2] + 4\beta^2 dh^3\,.$$

2. Multi-step discretization bound. We produce a coupling of $\mu_{(k+1)h}$ and π as follows. First, let $X_{kh} \sim \mu_{kh}$ and $Z_{kh} \sim \pi$ be optimally coupled. Using the same Brownian motion, we set

$$X_{(k+1)h} := X_{kh} - h\,\nabla V(X_{kh}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,,$$
$$Z_t := Z_{kh} - \int_{kh}^t \nabla V(Z_s)\,\mathrm{d}s + \sqrt{2}\,(B_t - B_{kh})\,, \qquad \text{for } t \in [kh, (k+1)h]\,.$$

Clearly $X_{(k+1)h} \sim \mu_{(k+1)h}$; also, since π is stationary for the Langevin diffusion, then $Z_{(k+1)h} \sim \pi$. We also introduce an auxiliary process: let

$$\bar{X}_t := X_{kh} - \int_{kh}^{t} \nabla V(\bar{X}_s)\, \mathrm{d}s + \sqrt{2}\,(B_t - B_{kh}) \qquad \text{for } t \in [kh, (k+1)h]$$

denote the Langevin diffusion started at X_{kh}. We bound

$$\begin{aligned} W_2(\mu_{(k+1)h}, \pi) &\le \sqrt{\mathbb{E}[\|X_{(k+1)h} - Z_{(k+1)h}\|^2]} \\ &\le \sqrt{\mathbb{E}[\|\bar{X}_{(k+1)h} - Z_{(k+1)h}\|^2]} + \sqrt{\mathbb{E}[\|X_{(k+1)h} - \bar{X}_{(k+1)h}\|^2]}\,. \end{aligned}$$

Now we examine the two terms. In the first term, both $\bar{X}$ and Z evolve via the Langevin diffusion for an α-strongly convex potential, so we have the following contraction (which is established by a direct coupling argument, see Theorem 1.4.10):

$$\mathbb{E}[\|\bar{X}_{(k+1)h} - Z_{(k+1)h}\|^2] \le \exp(-2\alpha h)\, \mathbb{E}[\|X_{kh} - Z_{kh}\|^2] = \exp(-2\alpha h)\, W_2^2(\mu_{kh}, \pi)\,.$$

For the second term, X is the LMC algorithm and $\bar{X}$ is the continuous-time Langevin diffusion, both initialized at the same distribution μ_{kh}. Hence, we can apply our one-step discretization bound from before and deduce that

$$\mathbb{E}[\|X_{(k+1)h} - \bar{X}_{(k+1)h}\|^2] \le 3\beta^4 h^4\, \mathbb{E}[\|X_{kh}\|^2] + 4\beta^2 d h^3 \lesssim \frac{\beta^4 d h^4}{\alpha} + \beta^2 d h^3\,,$$

where we used the bound from Lemma 4.1.5 on the second moment.

Combining everything and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for $x, y \ge 0$,

$$W_2(\mu_{(k+1)h}, \pi) \le \exp(-\alpha h)\, W_2(\mu_{kh}, \pi) + O\Big(\frac{\beta^2 d^{1/2} h^2}{\alpha^{1/2}} + \beta d^{1/2} h^{3/2}\Big)\,.$$

After iterating this recursion, it implies

$$W_2(\mu_{Nh}, \pi) \le \exp(-\alpha N h)\, W_2(\mu_0, \pi) + O\Big(\frac{\beta^2 d^{1/2} h}{\alpha^{3/2}} + \frac{\beta d^{1/2} h^{1/2}}{\alpha}\Big)\,.$$

The result follows from squaring. □

***Remark* 4.1.6.** By inspecting the proof, one can see that that the following stronger inequality holds. Let us denote by $\hat{P}^{\mathrm{LMC}}$ the transition kernel for one step of LMC, and P the transition kernel for the Langevin diffusion run for time h. Then, under the assumptions of Theorem 4.1.2, for any $x, y \in \mathbb{R}^d$,

$$W_2\big(\hat{P}^{\mathrm{LMC}}(x, \cdot),\, P(y, \cdot)\big) \le \exp(-\alpha h)\, \|x - y\| + O(\beta^2 h^2\, \|y\| + \beta d^{1/2} h^{3/2})\,. \tag{4.1.7}$$

If we square this inequality and apply Young's inequality, it implies

$$\begin{aligned} W_2^2\big(\hat{P}^{\mathsf{LMC}}(x,\cdot),\, P(y,\cdot)\big) &\le \exp(-2\alpha h)\,\|x-y\|^2 + O(\beta^4 h^4\,\|y\|^2 + \beta^2 d h^3) \\ &\quad + O\big(\|x-y\|\,(\beta^2 h^2\,\|y\| + \beta d^{1/2} h^{3/2})\big) \\ &\le \exp(-\alpha h)\,\|x-y\|^2 + O\Big(\frac{\beta^4 h^3\,\|y\|^2}{\alpha} + \frac{\beta^2 d h^2}{\alpha}\Big)\,. \end{aligned} \tag{4.1.8}$$

Iterating either (4.1.7) or (4.1.8), together with a coupling argument, imply back the guarantee of Theorem 4.1.2.

We finish by presenting the lemma we used in the proof of Theorem 4.1.2. The following proof is very typical of stochastic calculus arguments, so it is worth internalizing.

Lemma 4.1.9. *Let $(Z_t)_{t\ge 0}$ denote the Langevin diffusion and let $(\pi_t)_{t\ge 0}$ denote its law. Assume that $\nabla V(0) = 0$ and $\|\nabla^2 V\|_{\mathrm{op}} \le \beta$. Then, provided that $t \le \frac{1}{3\beta}$,*

$$\mathbb{E}[\|Z_t - Z_0\|^2] \le 8\beta^2 t^2\,\mathbb{E}[\|Z_0\|^2] + 8dt\,.$$

Proof. By definition,

$$\begin{aligned} \mathbb{E}[\|Z_t - Z_0\|^2] &= \mathbb{E}\Big[\Big\|-\int_0^t \nabla V(Z_s)\,\mathrm{d}s + \sqrt{2}\,B_t\Big\|^2\Big] \\ &\le 2t\int_0^t \mathbb{E}[\|\nabla V(Z_s)\|^2]\,\mathrm{d}s + 4\,\mathbb{E}[\|B_t\|^2]\,. \end{aligned}$$

Using the assumption that $\nabla V(0) = 0$ and that ∇V is β-Lipschitz, we have the bound $\|\nabla V(Z_s)\| \le \beta\,\|Z_s\|$. Thus,

$$\begin{aligned} \mathbb{E}[\|Z_t - Z_0\|^2] &\le 2\beta^2 t\int_0^t \mathbb{E}[\|Z_s\|^2]\,\mathrm{d}s + 4dt \\ &\le 4\beta^2 t\int_0^t \mathbb{E}[\|Z_s - Z_0\|^2]\,\mathrm{d}s + 4\beta^2 t^2\,\mathbb{E}[\|Z_0\|^2] + 4dt\,. \end{aligned}$$

Applying Grönwall's inequality (Lemma 1.1.21), it implies

$$\mathbb{E}[\|Z_t - Z_0\|^2] \le \{4\beta^2 t^2\,\mathbb{E}[\|Z_0\|^2] + 4dt\}\exp(4\beta^2 t^2)\,.$$

Finally, use the assumption $t \le \frac{1}{3\beta}$ to conclude. □

4.2 Proof via Interpolation Argument

We now give a guarantee for LMC that holds even when V is possibly non-convex.

Optimization Box 4.2.1. Suppose $V : \mathbb{R}^d \to \mathbb{R}$ is β-smooth but non-convex. Recall from Optimization Box 4.1.1 that along the iterates of GD,

$$V(x_{k+1}) - V(x_k) \le -\frac{h}{2} \|\nabla V(x_k)\|^2 ,$$

provided that $h \le 1/\beta$ (we only used smoothness to derive this inequality). This is known as the **descent lemma**. We now combine this with an assumption that V satisfies a **Polyak–Łojasiewicz (PL) inequality**

$$\|\nabla V(x)\|^2 \ge 2\alpha \{V(x) - V(x_\star)\} \qquad \text{for all } x \in \mathbb{R}^d .$$

The PL inequality is implied by α-strong convexity (see Section 1.4.2), but it is weaker and allows for non-convex V. We then obtain

$$\begin{aligned} V(x_{k+1}) - V(x_\star) &= V(x_{k+1}) - V(x_k) + V(x_k) - V(x_\star) \\ &\le -\frac{h}{2} \|\nabla V(x_k)\|^2 + V(x_k) - V(x_\star) \le (1 - \alpha h) \{V(x_k) - V(x_\star)\} . \end{aligned}$$

Setting $h = 1/\beta$, we can achieve $V(x_N) - V(x_\star) \le \varepsilon^2$ in $O(\kappa \log((V(x_0) - V(x_\star))/\varepsilon^2))$ iterations under PL and smoothness. Note that in this setting, convergence of the objective gap is the best we can hope for, since the PL inequality allows for multiple global minimizers.

Recall from Section 1.4.2 that the sampling analogue of the PL inequality is the log-Sobolev inequality (LSI), which naturally raises the question of whether the LSI is enough to obtain sampling guarantees. The next proof we give is from [VW19] (slightly refined using a lemma from [Che+21a]). Here, we mimic the continuous-time convergence proof in KL divergence by first defining a continuous-time interpolation of the LMC iterates. Upon differentiating the KL divergence along this interpolation, we discover two terms: the first is the Fisher information, and the second is a discretization error term. By controlling the latter, we prove a convergence result for LMC assuming only that π satisfies LSI and that ∇V is Lipschitz.

The interpolation of LMC is defined as follows: for $t \in [kh, (k+1)h]$, we set

$$X_t := X_{kh} - (t - kh) \nabla V(X_{kh}) + \sqrt{2} (B_t - B_{kh}) . \tag{4.2.2}$$

Proposition 4.2.3. *Let $(\mu_t)_{t\geq 0}$ be the law of the interpolated process (4.2.2). Then,*

$$\partial_t \mu_t = \operatorname{div}\left[\mu_t \left(\nabla \ln \frac{\mu_t}{\pi} + \mathbb{E}[\nabla V(X_{kh}) - \nabla V(X_t) \mid X_t = \cdot]\right)\right].$$

The proof is a little tricky to write out formally.

Proof. Let $\mu_{t|\mathscr{F}_{kh}}$ denote the law of X_t conditioned on the filtration $\mathscr{F}_{kh}$ at time kh. Then, $(\mu_{t|\mathscr{F}_{kh}})_{t\in[kh,(k+1)h]}$ satisfies the Fokker–Planck equation

$$\partial_t \mu_{t|\mathscr{F}_{kh}} = \Delta \mu_{t|\mathscr{F}_{kh}} + \operatorname{div}\left(\mu_{t|\mathscr{F}_{kh}} \nabla V(X_{kh})\right).$$

Next, we take the expectation of the above equation; since $\mathbb{E}\,\mu_{t|\mathscr{F}_{kh}} = \mu_t$,

$$\partial_t \mu_t = \Delta \mu_t + \operatorname{div} \mathbb{E}[\mu_{t|\mathscr{F}_{kh}} \nabla V(X_{kh})].$$

Write $\mathbb{P}_{\mathscr{F}_{kh}}$ for the probability measure on $\mathscr{F}_{kh}$, and write $\mathbb{P}_{\mathscr{F}_{kh}|t}$ to denote the conditional measure given X_t. Note that $\mu_{t|\mathscr{F}_{kh}}(x \mid \omega)\, \mathbb{P}_{\mathscr{F}_{kh}}(\mathrm{d}\omega) = \mu_t(x)\, \mathbb{P}_{\mathscr{F}_{kh}|t}(\mathrm{d}\omega \mid x)$. So,

$$\begin{aligned}
\mathbb{E}[\mu_{t|\mathscr{F}_{kh}}(x) \nabla V(X_{kh})] &= \int \mu_{t|\mathscr{F}_{kh}}(x \mid \omega) \nabla V\big(X_{kh}(\omega)\big)\, \mathbb{P}_{\mathscr{F}_{kh}}(\mathrm{d}\omega) \\
&= \mu_t(x) \int \mathbb{P}_{\mathscr{F}_{kh}|t}(\mathrm{d}\omega \mid x) \nabla V\big(X_{kh}(\omega)\big) \\
&= \mu_t(x)\, \mathbb{E}[\nabla V(X_{kh}) \mid X_t = x].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\partial_t \mu_t &= \Delta \mu_t + \operatorname{div}\big(\mu_t\, \mathbb{E}[\nabla V(X_{kh}) \mid X_t = \cdot]\big) \\
&= \operatorname{div}\left(\mu_t \nabla \ln \frac{\mu_t}{\pi}\right) + \operatorname{div}\big(\mu_t\, \mathbb{E}[\nabla V(X_{kh}) - \nabla V(X_t) \mid X_t = \cdot]\big).
\end{aligned}$$

□

Corollary 4.2.4. *Along the law $(\mu_t)_{t\geq 0}$ of the interpolated process (4.2.2),*

$$\partial_t \operatorname{KL}(\mu_t \parallel \pi) \leq -\frac{3}{4} \operatorname{FI}(\mu_t \parallel \pi) + \mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2].$$

Recall that the Fisher information is $\operatorname{FI}(\mu \parallel \pi) := \mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2]$ if μ has a smooth density with respect to π.

Proof. Using Proposition 4.2.3,

$$\begin{aligned}\partial_t \mathsf{KL}(\mu_t \parallel \pi) &= -\mathbb{E}_{\mu_t}\left\langle \nabla \ln \frac{\mu_t}{\pi}, \nabla \ln \frac{\mu_t}{\pi} + \mathbb{E}[\nabla V(X_{kh}) - \nabla V(X_t) \mid X_t = \cdot]\right\rangle \\ &= -\mathsf{FI}(\mu_t \parallel \pi) + \mathbb{E}_{\mu_t}\left\langle \nabla \ln \frac{\mu_t}{\pi}, \mathbb{E}[\nabla V(X_t) - \nabla V(X_{kh}) \mid X_t = \cdot]\right\rangle.\end{aligned}$$

Using Young's inequality,

$$\begin{aligned}&\mathbb{E}_{\mu_t}\left\langle \nabla \ln \frac{\mu_t}{\pi}, \mathbb{E}[\nabla V(X_t) - \nabla V(X_{kh}) \mid X_t = \cdot]\right\rangle \\ &\quad\le \frac{1}{4}\mathsf{FI}(\mu_t \parallel \pi) + \mathbb{E}\left[\|\mathbb{E}[\nabla V(X_t) - \nabla V(X_{kh}) \mid X_t]\|^2\right] \\ &\quad\le \frac{1}{4}\mathsf{FI}(\mu_t \parallel \pi) + \mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]\,.\end{aligned}$$

□

Before we give the convergence proof for LMC, we need one more lemma. Recall that for Theorem 4.1.2, we needed to control $\mathbb{E}[\|X_{kh}\|^2]$, which we accomplished via strong convexity of V (Lemma 4.1.5). Under the weaker assumption of an LSI, it is trickier to control the moments of the LMC iterates, but we have the following magic lemma.

Lemma 4.2.5 ([Che+21a, Lemma 16]). *Suppose that $\pi \propto \exp(-V)$ where ∇V is β-Lipschitz. Then, for any probability measure μ,*

$$\mathbb{E}_\mu[\|\nabla V\|^2] \le \mathsf{FI}(\mu \parallel \pi) + 2\beta d\,.$$

Proof. For the generator $\mathscr{L}$ of the Langevin diffusion (with potential V), we can calculate $\mathscr{L}V = \Delta V - \|\nabla V\|^2$. Also, since $\nabla^2 V \preceq \beta I_d$, then $\Delta V \le \beta d$. Thus, using the fundamental integration by parts identity (Theorem 1.2.14),

$$\begin{aligned}\mathbb{E}_\mu[\|\nabla V\|^2] = \mathbb{E}_\mu[\Delta V - \mathscr{L}V] &\le \beta d + \int (-\mathscr{L}V)\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\pi = \beta d + \int \left\langle \nabla V, \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right\rangle \mathrm{d}\pi \\ &= \beta d + \int \left\langle \nabla V, \nabla \ln \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right\rangle \mathrm{d}\mu \\ &\le \beta d + \frac{1}{2}\mathbb{E}_\mu[\|\nabla V\|^2] + \frac{1}{2}\mathsf{FI}(\mu \parallel \pi)\,.\end{aligned}$$

Rearranging the inequality yields the result. □

Also, recall that an LSI implies $\mathsf{KL}(\cdot \parallel \pi) \le \frac{C_{\mathsf{LSI}}}{2}\mathsf{FI}(\cdot \parallel \pi)$.

Theorem 4.2.6 ([VW19]). *For $k \in \mathbb{N}$, let μ_{kh} denote the law of the k-th iterate of LMC with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies LSI and that ∇V is β-Lipschitz. Then, for all $h \leq \frac{1}{4\beta}$, for all $N \in \mathbb{N}$,*

$$\mathsf{KL}(\mu_{Nh} \,\|\, \pi) \leq \exp\Big(-\frac{Nh}{C_{\mathsf{LSI}}}\Big)\, \mathsf{KL}(\mu_0 \,\|\, \pi) + O(C_{\mathsf{LSI}}\beta^2 dh + \beta^2 dh^2)\,.$$

In particular, for all $\varepsilon \in [0, C_{\mathsf{LSI}}\beta\sqrt{d}]$, if we take $h \asymp \frac{\varepsilon^2}{C_{\mathsf{LSI}}\beta^2 d}$, then we obtain the guarantee $\sqrt{\mathsf{KL}(\mu_{Nh} \,\|\, \pi)} \leq \varepsilon$ after

$$N = O\Big(\frac{C_{\mathsf{LSI}}^2 \beta^2 d}{\varepsilon^2} \log \frac{\mathsf{KL}(\mu_0 \,\|\, \pi)}{\varepsilon}\Big) \qquad \textit{iterations}\,.$$

Proof. In light of Corollary 4.2.4, we focus our attention on the discretization error term

$$\begin{aligned}\mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2] &\leq \beta^2\, \mathbb{E}[\|X_t - X_{kh}\|^2] \\ &= \beta^2\, (t-kh)^2\, \mathbb{E}[\|\nabla V(X_{kh})\|^2] + 2\beta^2\, \mathbb{E}[\|B_t - B_{kh}\|^2]\,.\end{aligned}$$

In order to apply Lemma 4.2.5, it is more convenient to have $\mathbb{E}[\|\nabla V(X_t)\|^2]$ instead of $\mathbb{E}[\|\nabla V(X_{kh})\|^2]$. So, we use

$$\mathbb{E}[\|\nabla V(X_{kh})\|^2] \leq 2\, \mathbb{E}[\|\nabla V(X_t)\|^2] + 2\, \mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]\,.$$

If $h \leq \frac{1}{2\beta}$, we can combine this inequality with the previous one and rearrange to obtain

$$\begin{aligned}\mathbb{E}[\|\nabla V(X_t) - \nabla V(X_{kh})\|^2] &\leq 4\beta^2\, (t-kh)^2\, \mathbb{E}[\|\nabla V(X_t)\|^2] + 4\beta^2\, \mathbb{E}[\|B_t - B_{kh}\|^2] \\ &\leq 4\beta^2\, (t-kh)^2\, \mathbb{E}[\|\nabla V(X_t)\|^2] + 4\beta^2 d\, (t-kh)\,.\end{aligned}$$

For the first term, we apply Lemma 4.2.5, yielding for $h \leq \frac{1}{4\beta}$

$$\begin{aligned}4\beta^2\, (t-kh)^2\, \mathbb{E}[\|\nabla V(X_t)\|^2] &\leq 4\beta^2 h^2\, \mathsf{FI}(\mu_t \,\|\, \pi) + 8\beta^3 d\, (t-kh)^2 \\ &\leq \frac{1}{4}\, \mathsf{FI}(\mu_t \,\|\, \pi) + 2\beta^2 d\, (t-kh)\,.\end{aligned}$$

Combining with our differential inequality from Corollary 4.2.4 and LSI,

$$\partial_t\, \mathsf{KL}(\mu_t \,\|\, \pi) \leq -\frac{1}{2}\, \mathsf{FI}(\mu_t \,\|\, \pi) + 6\beta^2 d\, (t-kh) \leq -\frac{1}{C_{\mathsf{LSI}}}\, \mathsf{KL}(\mu_t \,\|\, \pi) + 6\beta^2 d\, (t-kh)\,.$$

This implies that

$$\partial_t \Big[\exp\big(\frac{t-kh}{C_{\mathsf{LSI}}}\big)\, \mathsf{KL}(\mu_t \parallel \pi) \Big] \le 6\beta^2 d\,(t-kh)\exp\big(\frac{t-kh}{C_{\mathsf{LSI}}}\big)$$

and upon integration,

$$\mathsf{KL}(\mu_{(k+1)h} \parallel \pi) \le \exp\Big(-\frac{h}{C_{\mathsf{LSI}}}\Big)\, \mathsf{KL}(\mu_{kh} \parallel \pi) + 3\beta^2 dh^2\,.$$

Iterating and splitting into cases based on whether or not $h \le C_{\mathsf{LSI}}$,

$$\mathsf{KL}(\mu_{Nh} \parallel \pi) \le \exp\Big(-\frac{Nh}{C_{\mathsf{LSI}}}\Big)\, \mathsf{KL}(\mu_0 \parallel \pi) + O(\max\{C_{\mathsf{LSI}}\beta^2 dh, \beta^2 dh^2\})\,. \qquad \square$$

Recall from Theorem 2.2.15 that an LSI implies exponential decay in *every* Rényi divergence, not just the KL divergence. Working with Rényi divergences of order $q > 1$ introduces substantial new difficulties for the discretization analysis, which is why it is remarkable that the proof above can be adapted to the Rényi case with the introduction of some additional tricks; see Chapter 6.

4.3 Proof via Convex Optimization

Next, we turn towards an astonishing proof, due to [DMM19], which is inspired by convex optimization. This proof also yields the state-of-the-art dependence of LMC on the condition number κ of the target π.

Let the target be $\pi = \exp(-V)$ (for this proof, we are assuming that V is normalized so that $\int \exp(-V) = 1$; this just simplifies the notation but does not change the algorithm nor the analysis). We now view sampling as the composite optimization problem of minimizing the objective

$$\mathsf{KL}(\mu \parallel \pi) := \underbrace{\int V \,\mathrm{d}\mu}_{=:\mathcal{E}(\mu)} + \underbrace{\int \mu \ln \mu}_{=:\mathcal{H}(\mu)},$$

where the two terms are the *energy* and the (negative) *entropy*. Accordingly, we break up the iterates of LMC into the steps

$$\begin{aligned} X_{kh}^{+} &:= X_{kh} - h\,\nabla V(X_{kh})\,, \\ X_{(k+1)h} &:= X_{kh}^{+} + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,. \end{aligned}$$

The first step is simply a deterministic gradient descent update on the function V. If we write μ_{kh}^+ for the law of X_{kh}^+, then in the space of measures one can show that μ_{kh}^+ is obtained from μ_{kh} by taking a gradient step for the energy functional $\mathcal{E}$ w.r.t. the Wasserstein geometry.[1] On the other hand, the second step applies the heat flow; in the space of measures, this is a Wasserstein gradient flow for the entropy functional $\mathcal{H}$. Since the gradient descent algorithm is sometimes known as the "forward" method in optimization (as opposed to a proximal step which is the "backward" method), this has led to LMC being dubbed the "forward–flow" algorithm.

We refer to [Wib18] for more on this perspective. In particular, it suggests why the LMC scheme is biased: the "forward-flow" discretization scheme is biased for optimization as well! Generally speaking, when we split dynamics into its constituent parts and apply a discretization method to each part, this is known as a *splitting scheme*, and it is the cornerstone of numerical integration. Not all splitting schemes are born equal, however—as we have seen, it requires care to design one that is asymptotically unbiased. Recall from Exercise 1.18 that $\mathcal{E}$ is smooth if V is smooth, but $\mathcal{H}$ is non-smooth. Wisdom from optimization theory tells us that the appropriate scheme to use here is the "forward–backward" or "proximal gradient" method, but unfortunately the "backward" step for the entropy cannot be easily implemented.

This can be viewed as the blessing and curse of sampling. It is a blessing because for the non-smooth term in sampling—namely, the entropy $\mathcal{H}$—although we can implement neither the forward nor the backward discretizations, we *can* implement the exact gradient flow, by sampling a Gaussian, and the gradient flow thankfully succeeds even in the presence of non-smoothness. On the other hand, it is a curse because the "mismatch" of the forward and flow operations as a splitting scheme leads to asymptotic bias. We will revisit this issue of bias in Chapter 8 via the proximal sampler.

Nevertheless, we will leverage this splitting perspective to provide another analysis of LMC. The strategy of the proof is to show that the forward step of LMC dissipates the energy while not increasing the entropy too much, and that the flow step of LMC dissipates the entropy while not increasing the energy too much.

Optimization Box 4.3.1. Let $V : \mathbb{R}^d \to \mathbb{R}$ satisfy $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$, and recall from Optimization Box 4.1.1 that along GD,

$$\|x_{k+1} - y\|^2 \leq (1 - \alpha h) \|x_k - y\|^2 - 2h \{V(x_k) - V(y)\} + h^2 \|\nabla V(x_k)\|^2 . \quad (4.3.2)$$

[1] Technically this is only true if the step size h is chosen so that $h \|\nabla^2 V\|_{\mathrm{op}} \leq 1$. This is because on any Riemannian manifold in which geodesics cannot be extended indefinitely, the gradient descent steps must be short enough to ensure that the iterates are still travelling along geodesics.

Applying the descent lemma $\|\nabla V(x_k)\|^2 \le \frac{2}{h}\{V(x_k) - V(x_{k+1})\}$ for $h \le 1/\beta$,

$$\|x_{k+1} - y\|^2 \le (1 - \alpha h)\|x_k - y\|^2 - 2h\{V(x_{k+1}) - V(y)\}. \tag{4.3.3}$$

Inspired by [AGS08], we refer to this as an **evolution variational inequality (EVI)**. It is quite a flexible tool for optimization. For example, we can set $y = x_\star$, and if $\alpha > 0$, use the fact that $V(x_k) - V(x_\star) \ge 0$ to conclude that $\|x_{k+1} - x_\star\|^2 \le (1-\alpha h)\|x_k - y\|^2$, recovering Optimization Box 4.1.1. However, even when $\alpha = 0$, we can set $y = x_\star$, $h = 1/\beta$, and telescope this inequality to conclude that

$$V(x_N) - V(x_\star) \le \frac{1}{N}\sum_{k=0}^{N-1}\{V(x_{k+1}) - V(x_\star)\} \le \frac{\beta\|x_0 - x_\star\|^2}{2N},$$

where the first inequality follows from the descent lemma.

In analogy, we aim to prove the following key lemma.

Lemma 4.3.4. *Let $\pi = \exp(-V)$ be the target and assume that $0 \le \alpha I_d \le \nabla^2 V \le \beta I_d$. Let $(\mu_{kh})_{k\in\mathbb{N}}$ denote the iterates of LMC with step size $h \in [0, \frac{1}{\beta}]$. Then,*

$$2h\,\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) \le (1 - \alpha h)\, W_2^2(\mu_{kh}, \pi) - W_2^2(\mu_{(k+1)h}, \pi) + 2\beta d h^2. \tag{4.3.5}$$

From this, we deduce the following results.

Theorem 4.3.6 ([DMM19])**.** *Suppose that $\pi = \exp(-V)$ is the target distribution and that V satisfies $0 \le \alpha I_d \le \nabla^2 V \le \beta I_d$. Let $(\mu_{kh})_{k\in\mathbb{N}}$ denote the law of LMC.*

1. *(weakly convex case) Suppose that $\alpha = 0$. For any $\varepsilon \in [0, \sqrt{d}]$, if we take step size $h \asymp \frac{\varepsilon^2}{\beta d}$, then for the mixture distribution $\bar{\mu}_{Nh} := N^{-1}\sum_{k=1}^N \mu_{kh}$ it holds that $\sqrt{\mathsf{KL}(\bar{\mu}_{Nh} \,\|\, \pi)} \le \varepsilon$ after*

$$N = O\Big(\frac{\beta d\, W_2^2(\mu_0, \pi)}{\varepsilon^4}\Big) \qquad \textit{iterations}.$$

2. *(strongly convex case) Suppose that $\alpha > 0$ and let $\kappa := \beta/\alpha$ denote the condition number. Then, for any $\varepsilon \in [0, \sqrt{d}]$, with step size $h \asymp \frac{\varepsilon^2}{\beta d}$ we obtain*

$\sqrt{\alpha}\, W_2(\mu_{Nh}, \pi) \le \varepsilon$ *and* $\sqrt{\mathsf{KL}(\bar{\mu}_{Nh,2Nh} \,\|\, \pi)} \le \varepsilon$ *after*

$$N = O\Big(\frac{\kappa d}{\varepsilon^2} \log \frac{\sqrt{\alpha}\, W_2(\mu_0, \pi)}{\varepsilon}\Big) \qquad \textit{iterations},$$

where $\bar{\mu}_{Nh,2Nh} := N^{-1} \sum_{k=N+1}^{2N} \mu_{kh}$.

Proof. We use Lemma 4.3.4.

1. By summing the inequality (4.3.5) and using the convexity of the KL divergence,

$$\mathsf{KL}(\bar{\mu}_{Nh} \,\|\, \pi) \le \frac{1}{N} \sum_{k=1}^{N} \mathsf{KL}(\mu_{kh} \,\|\, \pi) \le \frac{W_2^2(\mu_0, \pi)}{2Nh} + \beta d h\,.$$

   The result follows from our choice of h and N.

2. First, we prove the W_2 guarantee. Using the fact that $\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) \ge 0$ and iterating the inequality (4.3.5) we obtain

$$\begin{aligned} W_2^2(\mu_{Nh}, \pi) &\le (1-\alpha h)^N\, W_2^2(\mu_0, \pi) + 2\beta d h^2 \sum_{k=0}^{N-1} (1-\alpha h)^k \\ &\le \exp(-\alpha N h)\, W_2^2(\mu_0, \pi) + O(\kappa d h)\,. \end{aligned}$$

   With our choice of h and N, we obtain $\sqrt{\alpha}\, W_2(\mu_{Nh}, \pi) \le \varepsilon$.

   Next, forget about the previous N iterations of LMC and consider μ_{Nh} to be the new initialization to LMC. Applying the weakly convex result now yields the KL guarantee $\sqrt{\mathsf{KL}(\bar{\mu}_{Nh,2Nh} \,\|\, \pi)} \le \varepsilon$. □

We next turn towards the proof of Lemma 4.3.4.

Proof of Lemma 4.3.4. We break the proof into three steps.

1. The forward step dissipates the energy. Let $Z \sim \pi$ be optimally coupled to X_{kh}. Then, $\mathcal{E}(\mu_{kh}^+) - \mathcal{E}(\pi) = \mathbb{E}[V(X_{kh}^+) - V(Z)]$. However, since X_{kh}^+ is obtained from X_{kh} via a gradient descent step on V, we can apply the EVI (4.3.3) to argue that

$$\begin{aligned} \mathcal{E}(\mu_{kh}^+) - \mathcal{E}(\pi) &\le \frac{1}{2h}\, \mathbb{E}[(1-\alpha h)\, \|X_{kh} - Z\|^2 - \|X_{kh}^+ - Z\|^2] \\ &\le \frac{1}{2h} \left\{(1-\alpha h)\, W_2^2(\mu_{kh}, \pi) - W_2^2(\mu_{kh}^+, \pi)\right\}. \end{aligned} \tag{4.3.7}$$

2. The flow step does not substantially increase the energy. Next, using the β-smoothness of V,

$$\begin{aligned}\mathcal{E}(\mu_{(k+1)h}) - \mathcal{E}(\mu^+_{kh}) &= \mathbb{E}[V(X_{(k+1)h}) - V(X^+_{kh})] \\ &\le \mathbb{E}\Big[\langle \nabla V(X^+_{kh}), X_{(k+1)h} - X^+_{kh}\rangle + \frac{\beta}{2}\,\|X_{(k+1)} - X^+_{kh}\|^2\Big] \\ &= \mathbb{E}[\sqrt{2}\,\langle \nabla V(X^+_{kh}), B_{(k+1)h} - B_{kh}\rangle + \beta\,\|B_{(k+1)h} - B_{kh}\|^2] \\ &= \beta d h\,. \end{aligned} \tag{4.3.8}$$

3. The flow step dissipates the entropy. Let $(Q_t)_{t\ge 0}$ denote the heat semigroup, i.e., $Q_t f(x) := \mathbb{E}\, f(x + \sqrt{2}\,B_t)$, so that $\mu_{(k+1)h} = \mu^+_{kh} Q_h$. Then, since the heat flow is the Wasserstein gradient flow of $\mathcal{H}$, and the Wasserstein gradient of $\mathcal{H}$ is $\nabla_{W_2}\mathcal{H}(\mu) = \nabla \ln \mu$, one can show that

$$\partial_t W_2^2(\mu^+_{kh} Q_t, \pi) \le 2\,\mathbb{E}\langle \nabla \ln \mu(X^+_{kh+t}), Z - X^+_{kh+t}\rangle$$

where $X^+_{kh+t} \sim \mu^+_{kh} Q_t$ and $Z \sim \pi$ are optimally coupled. This follows from the formula for the gradient of the squared Wasserstein distance (Theorem 1.4.11); it may be justified more rigorously using, e.g., [AGS08, Theorem 10.2.2].

On the other hand, we showed that $\mathcal{H}$ is geodesically convex (see (1.4.3)), so

$$\mathcal{H}(\pi) - \mathcal{H}(\mu^+_{kh} Q_t) \ge \mathbb{E}\langle \nabla \ln \mu(X^+_{kh+t}), Z - X^+_{kh+t}\rangle\,.$$

Using the fact that $t \mapsto \mathcal{H}(\mu^+_{kh} Q_t)$ is decreasing (which also follows because $t \mapsto \mu^+_{kh} Q_t$ is the gradient flow of $\mathcal{H}$), we then have

$$W_2^2(\mu_{(k+1)h}, \pi) - W_2^2(\mu^+_{kh}, \pi) \le 2h\,\{\mathcal{H}(\pi) - \mathcal{H}(\mu_{(k+1)h})\}\,. \tag{4.3.9}$$

Concluding the proof. Combine (4.3.7), (4.3.8), and (4.3.9) to obtain (4.3.5). □

Non-smooth case. The proof via convex optimization can also handle the *non-smooth* case in which we only assume that V is convex and Lipschitz. As before, we deduce the convergence result from a key one-step inequality.

Lemma 4.3.10. *Let $\pi = \exp(-V)$ be the target and assume that V is convex and L-Lipschitz. Let $(\mu_{kh})_{k\in\mathbb{N}}$ denote the iterates of* LMC *with step size $h > 0$. Then,*

$$2h\,\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) \le W_2^2(\mu^+_{kh}, \pi) - W_2^2(\mu^+_{(k+1)h}, \pi) + L^2 h^2\,.$$

Theorem 4.3.11 ([DMM19]). *Suppose that $\pi = \exp(-V)$ is the target distribution and that V is convex and L-Lipschitz. Let $(\mu_{kh})_{k\in\mathbb{N}}$ denote the law of LMC. For any $\varepsilon > 0$, if we take step size $h \asymp \frac{\varepsilon^2}{L^2}$, then for the mixture distribution $\bar{\mu}_{Nh} := N^{-1}\sum_{k=1}^{N} \mu_{kh}$ it holds that $\sqrt{\mathsf{KL}(\bar{\mu}_{Nh} \,\|\, \pi)} \le \varepsilon$ after*

$$N = O\Big(\frac{L^2\, W_2^2(\mu_0^+, \pi)}{\varepsilon^4}\Big) \qquad \textit{iterations}\,.$$

Proof. This follows from Lemma 4.3.10 in exactly the same way that the weakly convex case of Theorem 4.3.6 follows from Lemma 4.3.4. □

Proof of Lemma 4.3.10. The main task here is to obtain dissipation of the energy functional $\mathcal{E}$ under our new assumptions. Let $Z \sim \pi$ be optimally coupled to $X_{(k+1)h}$. From (4.3.2),

$$2h\,\{\mathcal{E}(\mu_{(k+1)h}) - \mathcal{E}(\pi)\} \le \mathbb{E}[\|X_{(k+1)h} - Z\|^2 - \|X^+_{(k+1)h} - Z\|^2 + h^2\,\|\nabla V(X_{(k+1)h})\|^2]\,.$$

We no longer have the descent lemma (which requires smoothness) at our disposal, but we can instead bound the last term by L^2h^2 using the Lipschitz assumption. Hence,

$$2h\,\{\mathcal{E}(\mu_{(k+1)h}) - \mathcal{E}(\pi)\} \le W_2^2(\mu_{(k+1)h}, \pi) - W_2^2(\mu^+_{(k+1)h}, \pi) + L^2h^2\,. \tag{4.3.12}$$

On the other hand, recall from (4.3.9) that

$$2h\,\{\mathcal{H}(\mu_{(k+1)h}) - \mathcal{H}(\pi)\} \le W_2^2(\mu^+_{kh}, \pi) - W_2^2(\mu_{(k+1)h}, \pi)\,.$$

Together with (4.3.12), this completes the proof. □

4.4 Proof via Girsanov's Theorem

The idea behind the next proof is to control the discretization error $\mathsf{KL}(\mu_{Nh} \,\|\, \pi_{Nh})$, where μ_{Nh} is the law of the LMC iterate X_{Nh} and π_{Nh} is the law of the Langevin diffusion Z_{Nh} initialized at μ_0. This is accomplished using Girsanov's theorem (see Section 3.2.2).

Unlike the previous proofs, which assumed strong log-concavity or LSI for π, controlling this discretization error will only require mild assumptions, such as smoothness of V and sub-Gaussianity of π. The point, however, is that control of $\mathsf{KL}(\mu_{Nh} \,\|\, \pi_{Nh})$ does not immediately yield quantitative convergence of $\mu_{Nh} \to \pi$; indeed, this also requires

quantitative convergence of $\pi_{Nh} \to \pi$, which does require some kind of assumption such as strong log-concavity or a functional inequality.

Another remark is that the preceding proofs all had the following structure: for a fixed step size $h > 0$, as the number of iterations $N \to \infty$, the error of LMC is at most a quantity depending on h, and which can be made as small as we like by taking h small. In particular, to achieve a desired error it suffices that the step size be sufficiently small and that the number of iterations be sufficiently large. In contrast, for the following proof we will only be able to establish a bound on $\mathsf{KL}(\mu_{Nh} \,\|\, \pi_{Nh})$ which grows with the iteration number N. Consequently, in our final sampling guarantee, we will not be able to take N too large; our guarantee will only imply that the error of LMC is small if N lies in some range. Conceptually, this is unsatisfying because "running the Markov chain too long" should not be a problem, and it only arises as an artefact of the proof. Nonetheless, it is worthwhile learning the proof because it is broadly applicable.

Historically, the Girsanov method was one of the first discretization techniques utilized in the modern quantitative study of sampling (see [DT12]). The argument we present here is similar in spirit to [DT12], although we have made a few refinements.

Discretization analysis. In the following theorem, we assume that π is strongly log-concave for simplicity.

Theorem 4.4.1. *Let $\pi \propto \exp(-V)$ be the target and assume $0 \preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$. Let $(\mu_{kh})_{t \geq 0}$ denote the law of* LMC*, and let $(\pi_t)_{t \geq 0}$ denote the law of the Langevin diffusion initialized at μ_0. Then, for $h \in [0, \frac{1}{\beta}]$,*

$$\mathsf{KL}(\mu_{Nh} \,\|\, \pi_{Nh}) \lesssim \beta^4 h^3 N \left(\int \|\cdot\|^2 \, \mathrm{d}\mu_0 + \frac{d}{\alpha} \right) + \beta^2 d h^2 N \,.$$

Proof. For $T \coloneqq Nh$, let $(B_t)_{t \in [0,T]}$ be our standard Brownian motion and consider the SDE

$$\mathrm{d}X_t = -\nabla V(X_t) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t \,, \qquad X_0 \sim \mu_0 \,.$$

Let $\mathbf{W}_T$ denote the Wiener measure on our path space, under which $(X_t)_{t \in [0,T]}$ becomes the Langevin diffusion started at $X_0 \sim \mu_0$. We would like to write

$$\mathrm{d}X_t = -\nabla V(X_{kh}) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\tilde{B}_t \,, \qquad t \in [kh, (k+1)h] \,,$$

and to find a path measure $\mathbf{P}_T$ under which $(\tilde{B}_t)_{t \in [0,T]}$ is a $\mathbf{P}_T$-Brownian motion. If so, then under $\mathbf{P}_T$, we see that $(X_t)_{t \in [0,T]}$ is the interpolated LMC process. Noting that $\mathrm{d}B_t =$

$\mathrm{d}\tilde{B}_t - \mathrm{d}[B, M]_t$ where $\mathrm{d}M_t := \frac{1}{\sqrt{2}} \langle \nabla V(X_{kh}) - \nabla V(X_t), \mathrm{d}B_t \rangle$, we consider the exponential martingale $\mathcal{E}(M)$ associated with M.

By the data-processing inequality (Theorem 1.5.3), $\mathsf{KL}(\mu_{Nh} \parallel \pi_{Nh}) \le \mathsf{KL}(\mathbf{P}_T \parallel \mathbf{W}_T)$, so it suffices to bound the latter. By Girsanov's theorem (Theorem 3.2.6),[2]

$$\begin{aligned}
\mathsf{KL}(\mathbf{P}_T \parallel \mathbf{W}_T) &= \mathbb{E}^{\mathbf{P}_T} \ln \frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{W}_T} = \mathbb{E}^{\mathbf{P}_T} \ln \mathcal{E}(M)_T \\
&= \mathbb{E}^{\mathbf{P}_T} \sum_{k=0}^{N-1} \Big(\frac{1}{\sqrt{2}} \int_{kh}^{(k+1)h} \langle \nabla V(X_{kh}) - \nabla V(X_t), \mathrm{d}B_t \rangle \\
&\qquad - \frac{1}{4} \int_{kh}^{(k+1)h} \|\nabla V(X_{kh}) - \nabla V(X_t)\|^2 \, \mathrm{d}t \Big) .
\end{aligned}$$

However, we must be cautious! Here, $(B_t)_{t \in [0,T]}$ is a $\mathbf{W}_T$-standard Brownian. As a sanity check, if $(B_t)_{t \in [0,T]}$ were a $\mathbf{P}_T$-standard Brownian motion, then the first term would vanish (since stochastic integrals have zero mean) and the KL divergence would be *negative*, which is absurd. Instead, we rewrite the above expression in terms of $\tilde{B}$, obtaining

$$\begin{aligned}
\mathsf{KL}(\mathbf{P}_T \parallel \mathbf{W}_T) &= \frac{1}{4} \sum_{k=0}^{N-1} \int_{kh}^{(k+1)h} \mathbb{E}^{\mathbf{P}_T} [\|\nabla V(X_t) - \nabla V(X_{kh})\|^2] \, \mathrm{d}t \\
&\le \frac{\beta^2}{4} \sum_{k=0}^{N-1} \int_{kh}^{(k+1)h} \mathbb{E}^{\mathbf{P}_T} [\|X_t - X_{kh}\|^2] \, \mathrm{d}t \\
&= \frac{\beta^2}{4} \sum_{k=0}^{N-1} \int_{kh}^{(k+1)h} \{(t-kh)^2 \, \mathbb{E}^{\mathbf{P}_T} [\|\nabla V(X_{kh})\|^2] + 2d\,(t-kh)\} \, \mathrm{d}t \\
&\le \frac{\beta^4 h^3}{12} \sum_{k=0}^{N-1} \mathbb{E}^{\mathbf{P}_T} [\|X_{kh}\|^2] + \frac{\beta^2 d h^2 N}{4} .
\end{aligned}$$

We can use the bound on the second moment of the LMC iterates (Lemma 4.1.5) to get

$$\mathsf{KL}(\mathbf{P}_T \parallel \mathbf{W}_T) \lesssim \beta^4 h^3 N \left(\mathbb{E}^{\mathbf{P}_T} [\|X_0\|^2] + \frac{d}{\alpha} \right) + \beta^2 d h^2 N . \qquad \square$$

The preceding argument is similar to the Wasserstein coupling proof (Theorem 4.1.2), and indeed in both proofs we used strong log-concavity. However, in the Wasserstein coupling proof, the strong log-concavity assumption is crucial because it implies contraction

[2] Actually, as noted in the discussion in Section 3.2.2, to obtain the first equality one should check Novikov's condition. However, since all we desire is an upper bound on the KL divergence, this can be avoided with a localization argument. We omit the details.

in the Wasserstein metric, whereas the preceding discretization argument only requires a bound on the second moment of the LMC iterates which can be obtained in other ways.

What sampling guarantee does Theorem 4.4.1 imply? Unfortunately, neither KL nor $\sqrt{\mathsf{KL}}$ satisfy the triangle inequality, which poses a difficulty for bounding the distance of μ_{Nh} from the target π. One way to skirt this difficulty is to simply use the fact that $\sqrt{\mathsf{KL}} \gtrsim \|\cdot\|_{\mathrm{TV}}$ (Pinsker's inequality, Exercise 2.13) and the fact that the total variation distance satisfies the triangle inequality.

Corollary 4.4.2. *Let $\pi \propto \exp(-V)$ be the target and assume $0 \preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$. Let $(\mu_{kh})_{t\geq 0}$ denote the law of LMC initialized at the distribution $\mu_0 = \mathsf{normal}(0, \beta^{-1} I_d)$. Then, for all $\varepsilon \in (0,1)$ and for $h = \widetilde{\Theta}(\frac{\alpha\varepsilon^2}{\beta^2 d})$, we obtain the guarantee $\|\mu_{Nh} - \pi\|_{\mathrm{TV}} \leq \varepsilon$ provided that*

$$N = \widetilde{\Theta}\Big(\frac{\kappa^2 d}{\varepsilon^2}\Big) \qquad \textit{iterations}\,.$$

Proof. Since strong log-concavity implies an LSI, which in turn implies exponential convergence of the Langevin diffusion to its target in KL divergence (Theorem 1.2.25 and Theorem 1.2.29), we obtain $\sqrt{\mathsf{KL}(\pi_{Nh} \,\|\, \pi)} \leq \frac{\varepsilon}{\sqrt{2}}$ provided $Nh \gtrsim \frac{1}{\alpha} \log \frac{\mathsf{KL}(\mu_0 \| \pi)}{\varepsilon^2}$. With our choice of initialization, $\mathsf{KL}(\mu_0 \,\|\, \pi) \lesssim d \log \kappa$ and $\int \|\cdot\|^2 \,\mathrm{d}\mu_0 \lesssim d/\beta \leq d/\alpha$.

We now take the number of iterations to satisfy $Nh \asymp \frac{1}{\alpha} \log \frac{\mathsf{KL}(\mu_0 \| \pi)}{\varepsilon^2}$. Then, with our choice of h, we obtain from Theorem 4.4.1 that $\sqrt{\mathsf{KL}(\mu_{Nh} \,\|\, \pi_{Nh})} \leq \frac{\varepsilon}{\sqrt{2}}$. By the triangle inequality and Pinsker's inequality,

$$\begin{aligned}\|\mu_{Nh} - \pi\|_{\mathrm{TV}} &\leq \|\mu_{Nh} - \pi_{Nh}\|_{\mathrm{TV}} + \|\pi_{Nh} - \pi\|_{\mathrm{TV}} \\ &\leq \sqrt{\frac{1}{2}\,\mathsf{KL}(\mu_{Nh} \,\|\, \pi_{Nh})} + \sqrt{\frac{1}{2}\,\mathsf{KL}(\pi_{Nh} \,\|\, \pi)} \leq \varepsilon\,.\end{aligned}$$

Finally, plugging in the choice of h into $Nh \asymp \frac{1}{\alpha} \log \frac{\mathsf{KL}(\mu_0 \| \pi)}{\varepsilon^2}$ yields the result. □

We remark that if one follows this Pinsker approach, then in the Girsanov bound above one could alternatively bound $\mathsf{KL}(\pi_{Nh} \,\|\, \mu_{Nh})$ if this turns out to be easier.

Although the quantitative dependence in Corollary 4.4.2 matches prior results (e.g., via the interpolation method in Theorem 4.2.6), the final result is unsatisfying because we have moved to a weaker metric (TV rather than KL) for a seemingly silly reason (the failure of the triangle inequality for the KL divergence). Indeed, we have a convergence result for the Langevin diffusion in KL, and our discretization bound is in KL, yet our final result is in TV. Can we remedy this?

To address this, we can introduce the Rényi divergences (defined in (2.2.14)); recall that $\mathsf{KL} = \mathcal{R}_1$. We have a continuous-time result for the Langevin diffusion in Rényi divergence (Theorem 2.2.15), and it turns out that with some additional tricks it is possible to extend the Girsanov discretization argument to any Rényi divergence. Moreover, the Rényi divergences satisfy a *weak triangle inequality*: for any $q > 1$, any $\lambda \in (0, 1)$, and any probability measures μ, ν, π:

$$\mathcal{R}_q(\mu \parallel \pi) \leq \frac{q-\lambda}{q-1} \mathcal{R}_{q/\lambda}(\mu \parallel \nu) + \mathcal{R}_{(q-\lambda)/(1-\lambda)}(\nu \parallel \pi) \,. \tag{4.4.3}$$

This allows us to combine a continuous-time Rényi result with a Rényi discretization argument to yield a Rényi sampling guarantee. We provide the details for this approach in Chapter 6.

Bibliographical Notes

Historically, the LMC algorithm, which is called *unadjusted* because of the lack of a Metropolis–Hastings filter, was only studied relatively recently in non-asymptotic settings. Before the work of [DT12], it was more common to study MALA (which we introduce and study in Chapter 7). The ideas which go into the basic W_2 coupling proof for Theorem 4.1.2 were developed in a series of works on strongly log-concave sampling: [DT12; Dal17a; Dal17b; DM17; DM19]. The Girsanov argument of Theorem 4.4.1 is also due to [DT12].

There are two other notable proof techniques that we have omitted from this chapter: **reflection coupling** [Ebe11; Ebe16] and mean squared analysis [Li+19; Li+22; LZT22]. Reflection coupling uses a carefully chosen coupling of the Brownian motions rather than just taking the two Brownian motions to be the same as we have done (the latter coupling is called the **synchronous coupling**). Mean squared analysis is a general framework which combines local errors (one-step discretization bounds) into global error bounds. These two methods are useful for performing discretization analysis under more general sets of assumptions, but they are limited to providing guarantees in W_1 or W_2.

Exercises

Proof via Wasserstein Coupling

▷ **Exercise 4.1** (explicit computations for a Gaussian target)
Suppose that the target distribution is a Gaussian, $\pi = \mathsf{normal}(0, \Sigma)$, and that LMC is initialized at a Gaussian. Can you write down the iterates and stationary distribution of LMC explicitly? What happens when $\Sigma = I_d$?

Perform some explicit computations for this example and compare them to the general results for LMC that we derived in this chapter.

⊵ **Exercise 4.2** (second moment bounds for LMC)
Prove the second moment bound in Lemma 4.1.5. (Recall that the generator $\mathscr{L}$ of the Langevin diffusion satisfies $\mathbb{E}_\pi \mathscr{L} f = 0$. Apply this with $f = \|\cdot\|^2/2$. For the LMC iterates, write a recursion for $\mathbb{E}[\|X_{(k+1)h}\|^2]$ in terms of $\mathbb{E}[\|X_{kh}\|^2]$. In order to prove the lemma for all step sizes $h \leq \frac{1}{\beta}$, you may need to appeal to coercivity of the gradient, Lemma 5.2.3.)

⊵ **Exercise 4.3** (W_2 guarantees from a one-step bound)
Write out the details for Remark 4.1.6.

⊵ **Exercise 4.4** (LMC with decaying step size)
Show that by considering LMC with a decaying step size $h_k \asymp \frac{1}{\alpha k}$, one can obtain an iteration complexity which removes the logarithmic factor in Theorem 4.1.2.

CHAPTER 5

Faster Low-Accuracy Samplers

We now move beyond the basic LMC algorithm and consider samplers with better dependence on the dimension and inverse accuracy. There are two main sources of improvement that we explore in this chapter. The first is to use a more sophisticated discretization method than the basic Euler–Maruyama discretization. The second is to consider a different stochastic process, called the *underdamped* Langevin diffusion. By combining these two ideas, we arrive at the state-of-the-art complexity bounds for low-accuracy samplers.

5.1 Randomized Midpoint Discretization

In this section, we study the **randomized midpoint discretization**, which was introduced in [SL19]. The application to the Langevin diffusion was carried out in [HBE20].

Consider the continuous-time Langevin diffusion from time kh to $(k+1)h$:

$$Z_{(k+1)h} = Z_{kh} - \int_{kh}^{(k+1)h} \nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,.$$

In the Euler discretization, we approximate the second term via $-h\,\nabla V(Z_{kh})$. However, if we want an *unbiased* estimator of the integral, then we can introduce an auxiliary random variable $u_k \sim \mathsf{uniform}[0,1]$ and use

$$Z_{(k+1)h} \approx Z_{kh} - h\,\nabla V(Z_{(k+u_k)h}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,.$$

To compute this approximation, however, we need to know $Z_{(k+u_k)h}$. We have

$$Z_{(k+u_k)h} = Z_{kh} - \int_{kh}^{(k+u_k)h} \nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,(B_{(k+u_k)h} - B_{kh})\,.$$

Note that we are in the same situation as before. In particular, if we desire, we can draw another uniform random variable u_k' and approximate the second term above via $-u_k h\,\nabla V(Z_{(k+u_k u_k')h})$. In principle, this procedure can be repeated indefinitely. However, we will see that just one step of this procedure suffices: further applications of this procedure do not improve the discretization error. Instead, we will simply approximate $Z_{(k+u_k)h}$ via an Euler–Maruyama step:

$$Z_{(k+u_k)h} \approx Z_{kh} - u_k h\,\nabla V(Z_{kh}) + \sqrt{2}\,(B_{(k+u_k)h} - B_{kh})\,.$$

To summarize, the randomized midpoint discretization of the Langevin diffusion, which we will call RM-LMC, is the following update:

$$\begin{aligned} X_{(k+1)h} &\coloneqq X_{kh} - h\,\nabla V(X_{(k+u_k)h}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,,\\ X_{(k+u_k)h} &\coloneqq X_{kh} - u_k h\,\nabla V(X_{kh}) + \sqrt{2}\,(B_{(k+u_k)h} - B_{kh})\,, \end{aligned} \tag{RM-LMC}$$

where $(u_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. $\mathsf{uniform}[0,1]$ random variables which are independent of X_0 and the Brownian motion. The algorithm uses two gradient evaluations per iteration. Also, when implementing this recursion, it is important to note that the two Brownian increments are *coupled*. To sample the Brownian increments, draw two i.i.d. standard Gaussians ξ_k and ξ_k', and set

$$\begin{aligned} B_{(k+u_k)h} - B_{kh} &\coloneqq \sqrt{u_k h}\,\xi_k\,,\\ B_{(k+1)h} - B_{kh} &\coloneqq \sqrt{u_k h}\,\xi_k + \sqrt{(1-u_k)h}\,\xi_k'\,. \end{aligned}$$

We now analyze the complexity of this algorithm following the Wasserstein coupling proof of Section 4.1.

Theorem 5.1.1. *For $k \in \mathbb{N}$, let μ_{kh} denote the law of the k-th iterate of RM-LMC with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies $\nabla V(0) = 0$ and $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, provided $h \lesssim \frac{1}{\beta\kappa^{1/2}}$, for all $N \in \mathbb{N}$,*

$$W_2^2(\mu_{Nh}, \pi) \le \exp(-\alpha N h)\, W_2^2(\mu_0, \pi) + O\Big(\frac{\beta^2 d h^2}{\alpha} + \frac{\beta^4 d h^3}{\alpha^2} + \frac{\beta^6 d h^4}{\alpha^3}\Big)\,.$$

In particular, if we initialize at $\mu_0 = \delta_0$ and take $h \asymp \frac{\varepsilon}{\beta d^{1/2}}\,(1 \wedge \frac{d^{1/4}}{\varepsilon^{1/2}\kappa^{1/2}})$, then for any

$\varepsilon \in [0, \sqrt{d}]$ we obtain the guarantee $\sqrt{\alpha}\, W_2(\mu_{Nh}, \pi) \le \varepsilon$ after

$$N = \widetilde{O}\Big(\frac{\kappa d^{1/2}}{\varepsilon}\,\big(1 \vee \frac{\varepsilon^{1/2}\kappa^{1/2}}{d^{1/4}}\big)\Big) \qquad \textit{iterations}\,.$$

Proof. Recall from the proof of Theorem 4.1.2 that we started with a one-step discretization bound, and then we derived a multi-step discretization bound. In particular, for the one-step bound, we showed that if π_h is the law of the continuous-time Langevin diffusion started at μ_0, then

$$W_2^2(\mu_h, \pi_h) \lesssim \beta^4 h^4\, \mathbb{E}[\|Z_0\|^2] + \beta^2 d h^3\,. \tag{5.1.2}$$

It turns out that (5.1.2) still holds for RM-LMC. Since the proof is almost the same as before, we leave it as an exercise (Exercise 5.1).

The benefits of the randomized midpoint discretization enter once we consider the multi-step discretization. As in Theorem 4.1.2, we let $X_{kh} \sim \mu_{kh}$ and $Z_{kh} \sim \pi$ be optimally coupled, and we let $(\bar{X}_t)_{t\in[kh,(k+1)h]}$ and $(Z_t)_{t\in[kh,(k+1)h]}$ denote continuous-time Langevin diffusions initialized at X_{kh} and Z_{kh} respectively; all of these processes are coupled by using the same Brownian motion to drive them. We bound

$$\begin{aligned} W_2^2(\mu_{(k+1)h}, \pi) &\le \mathbb{E}[\|X_{(k+1)h} - Z_{(k+1)h}\|^2] \\ &= \mathbb{E}[\|\bar{X}_{(k+1)h} - Z_{(k+1)h}\|^2] + \mathbb{E}[\|X_{(k+1)h} - \bar{X}_{(k+1)h}\|^2] \\ &\qquad + 2\,\mathbb{E}\langle \bar{X}_{(k+1)h} - Z_{(k+1)h}, X_{(k+1)h} - \bar{X}_{(k+1)h}\rangle\,. \end{aligned}$$

Now observe that in the cross term, the only quantity that depends on the uniform random variable u_k is $X_{(k+1)h}$. In particular, if we let $\mathscr{F}_{(k+1)h}$ denote the σ-algebra generated by X_{kh} and $(B_t)_{t\in[kh,(k+1)h]}$ and we apply Young's inequality, then for any $\lambda > 0$,

$$\begin{aligned} &2\,\mathbb{E}\langle \bar{X}_{(k+1)h} - Z_{(k+1)h}, X_{(k+1)h} - \bar{X}_{(k+1)h}\rangle \\ &\quad = 2\,\mathbb{E}\langle \bar{X}_{(k+1)h} - Z_{(k+1)h}, \mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] - \bar{X}_{(k+1)h}\rangle \\ &\quad \le \lambda\,\mathbb{E}[\|\bar{X}_{(k+1)h} - Z_{(k+1)h}\|^2] + \frac{1}{\lambda}\,\mathbb{E}\big[\|\mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] - \bar{X}_{(k+1)h}\|^2\big] \end{aligned}$$

and plugging this in,

$$\begin{aligned} W_2^2(\mu_{(k+1)h}, \pi) \le (1+\lambda)\,\mathbb{E}[\|\bar{X}_{(k+1)h} - Z_{(k+1)h}\|^2] + \mathbb{E}[\|X_{(k+1)h} - \bar{X}_{(k+1)h}\|^2] \\ + \frac{1}{\lambda}\,\mathbb{E}\big[\|\mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] - \bar{X}_{(k+1)h}\|^2\big]\,. \end{aligned}$$

Before jumping into the calculations, let us see what we have gained, focusing on the dependence on the step size. As before, we need to take $\lambda \asymp h$. Previously, in Remark 4.1.6

(see (4.1.8)), the error term was $O(h^2)$ due to the use of Young's inequality. In this calculation, we have split the error term into $\mathbb{E}[\|X_{(k+1)} - \bar{X}_{(k+1)h}\|^2] = O(h^3)$ as well as another error term, which has a factor of $O(\frac{1}{h})$ but also has the expectation over the uniform random variable *inside* the norm. If we can show that this expectation makes the error smaller order than before (the interpretation being that the randomized midpoint reduces the bias), then we obtain smaller discretization error overall.

The one-step discretization bound (5.1.2) yields

$$\mathbb{E}[\|X_{(k+1)h} - \bar{X}_{(k+1)h}\|^2] \lesssim \beta^4 h^4 \, \mathbb{E}[\|X_{kh}\|^2] + \beta^2 d h^3 \,.$$

Next,

$$\begin{aligned}\mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] &= X_{kh} - h \, \mathbb{E}[\nabla V(X_{(k+u_k)h}) \mid \mathscr{F}_{(k+1)h}] + \sqrt{2} \, (B_{(k+1)h} - B_{kh}) \\ &= X_{kh} - \int_{kh}^{(k+1)h} \nabla V(X_t) \, \mathrm{d}t + \sqrt{2} \, (B_{(k+1)h} - B_{kh}) \,.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}\big[\|\mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] - \bar{X}_{(k+1)h}\|^2\big] &= \mathbb{E}\Big[\Big\|\int_{kh}^{(k+1)h} \{\nabla V(X_t) - \nabla V(\bar{X}_t)\} \, \mathrm{d}t\Big\|^2\Big] \\ &\le h \int_{kh}^{(k+1)h} \mathbb{E}[\|\nabla V(X_t) - \nabla V(\bar{X}_t)\|^2] \, \mathrm{d}t \\ &\le \beta^2 h \int_{kh}^{(k+1)h} \mathbb{E}[\|X_t - \bar{X}_t\|^2] \, \mathrm{d}t \,.\end{aligned}$$

By definition, $X_t = X_{kh} - (t - kh) \, \nabla V(X_{kh}) + \sqrt{2} \, (B_t - B_{kh})$, so

$$\begin{aligned}&\mathbb{E}\big[\|\mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] - \bar{X}_{(k+1)h}\|^2\big] \\ &\quad\le \beta^2 h \int_{kh}^{(k+1)h} \mathbb{E}\Big[\Big\|\int_{kh}^{t} \{\nabla V(\bar{X}_{kh}) - \nabla V(X_s)\} \, \mathrm{d}s\Big\|^2\Big] \, \mathrm{d}t \\ &\quad\le \beta^2 h^2 \int_{kh}^{(k+1)h} \int_{kh}^{t} \mathbb{E}[\|\nabla V(\bar{X}_{kh}) - \nabla V(X_s)\|^2] \, \mathrm{d}s \, \mathrm{d}t \\ &\quad\le \beta^4 h^2 \int_{kh}^{(k+1)h} \int_{kh}^{t} \mathbb{E}[\|\bar{X}_{kh} - X_s\|^2] \, \mathrm{d}s \, \mathrm{d}t \lesssim \beta^4 h^4 \, \{\beta^2 h^2 \, \mathbb{E}[\|X_{kh}\|^2] + dh\}\end{aligned}$$

where the last inequality uses the movement bound for the Langevin process that we proved in Lemma 4.1.9.

Next, recall that by contraction of the Langevin diffusion under strong log-concavity (Theorem 1.4.10), $\mathbb{E}[\|\bar{X}_{(k+1)h} - Z_{(k+1)h}\|^2] \le \exp(-2\alpha h)\, W_2^2(\mu_{kh}, \pi)$. Choosing $\lambda = \frac{\alpha h}{2}$,

$$W_2^2(\mu_{(k+1)h}, \pi) \le \exp\Big(-\frac{3\alpha h}{2}\Big)\, W_2^2(\mu_{kh}, \pi) + O\Big(\beta^4 h^4\, \mathbb{E}[\|X_{kh}\|^2] + \beta^2 d h^3 + \frac{\beta^6 h^5}{\alpha}\, \mathbb{E}[\|X_{kh}\|^2] + \frac{\beta^4 d h^4}{\alpha}\Big).$$

At this point, we could bound $\mathbb{E}[\|X_{kh}\|^2]$ recursively, similarly to Lemma 4.1.5, but instead we will use a trick:

$$\mathbb{E}[\|X_{kh}\|^2] = W_2^2(\mu_{kh}, \delta_0) \lesssim W_2^2(\mu_{kh}, \pi) + W_2^2(\pi, \delta_0) \lesssim W_2^2(\mu_{kh}, \pi) + \frac{d}{\alpha}.$$

It implies that if we take $h \lesssim \frac{1}{\beta \kappa^{1/2}}$, then

$$W_2^2(\mu_{(k+1)h}, \pi) \le \exp(-\alpha h)\, W_2^2(\mu_{kh}, \pi) + O\Big(\beta^2 d h^3 + \frac{\beta^4 d h^4}{\alpha} + \frac{\beta^6 d h^5}{\alpha^2}\Big).$$

Unrolling the recursion,

$$W_2^2(\mu_{Nh}, \pi) \le \exp(-\alpha N h)\, W_2^2(\mu_0, \pi) + O\Big(\frac{\beta^2 d h^2}{\alpha} + \frac{\beta^4 d h^3}{\alpha^2} + \frac{\beta^6 d h^4}{\alpha^3}\Big).$$

From the analysis, it can be seen that the bottleneck term (at least for dimension dependence) is not from the term involving $\mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}]$. This justifies our earlier comment that one step of the randomized midpoint procedure already suffices. □

***Remark* 5.1.3.** From the proof, we obtain the following one-step bound for the kernel $\hat{P}^{\mathrm{RM-LMC}}$ of RM-LMC:

$$W_2^2\big(\hat{P}^{\mathrm{RM-LMC}}(x, \cdot),\, P(y, \cdot)\big) \le \exp(-\alpha h)\, \|x - y\|^2 + O(\beta^4 h^4\, \|y\|^2 + \beta^2 d h^3).$$

Note that this improves over (4.1.8) for LMC.

The complexity guarantee for RM-LMC is considerably better than that for LMC. In fact, it is known that the randomized midpoint method is essentially an optimal discretization method (which is not the same as saying that RM-LMC is an optimal sampling algorithm); see [CLW21]. Another optimal discretization, not covered in this book, is the shifted ODE method of [FLO21].

One notable downside of the randomized midpoint discretization is that the analysis seems specific to the Wasserstein coupling approach. In particular, it is currently not known how to obtain matching guarantees in KL divergence.

In the above result, we proved a slightly weaker complexity bound in order to streamline the proof. It is possible to improve the second term of $\kappa^{3/2} d^{1/4} / \varepsilon^{1/2}$ in the guarantee of Theorem 5.1.1; see Exercise 5.2.

5.2 Hamiltonian Monte Carlo

The next algorithm we introduce, known as **Hamiltonian Monte Carlo** (**HMC**), was popularized in the context of sampling by Neal [Nea11]. As the name suggests, it is inspired by Hamiltonian mechanics. Although this algorithm is usually combined with a Metropolis–Hastings filter, we defer a discussion of this until Chapter 7. In this section, we instead focus on an analysis of the ideal (i.e., continuous-time) dynamics.

5.2.1 Introduction to Ideal HMC

First, we augment the target distribution π to add a momentum variable p. Specifically, define the distribution $\boldsymbol{\pi}$ on *phase space* $\mathbb{R}^d \times \mathbb{R}^d$ via

$$\boldsymbol{\pi}(x,p) \propto \exp\left(-V(x) - \frac{1}{2}\|p\|^2\right).$$

The first marginal of $\boldsymbol{\pi}$ is $\pi \propto \exp(-V)$, so if we obtain a sample from $\boldsymbol{\pi}$ then upon projecting to the first coordinate we obtain a sample from π.

The augmented target can also be written as $\boldsymbol{\pi} \propto \exp(-H)$, where H is the **Hamiltonian** $H(x,p) := V(x) + \frac{1}{2}\|p\|^2$. In Hamiltonian mechanics, which is a reformulation of classical mechanics, the laws of motion are governed by **Hamilton's equations**, a system of coupled first-order ODEs:[1]

$$\begin{aligned}\dot{x}_t &= \nabla_p H(x_t,p_t) = p_t\,,\\ \dot{p}_t &= -\nabla_x H(x_t,p_t) = -\nabla V(x_t)\,.\end{aligned}$$

Introducing the antisymmetric matrix

$$\boldsymbol{J} := \begin{bmatrix} 0 & I_d \\ -I_d & 0 \end{bmatrix},$$

Hamilton's equations can be written succinctly as

$$(\dot{x}_t, \dot{p}_t) = \boldsymbol{J}\,\nabla H(x_t,p_t)\,.$$

Let $F_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ denote the **flow map**, i.e., $F_t(x_0,p_0)$ is the solution (x_t,p_t) to Hamilton's equations started from (x_0,p_0). Then, we show that F_t leaves the augmented target $\boldsymbol{\pi}$ invariant: $(F_t)_\#\boldsymbol{\pi} = \boldsymbol{\pi}$. Indeed, if $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a function on phase space and $(x_t,p_t)_{t\ge 0}$ evolve via Hamilton's equations started at $(x_0,p_0) \sim \boldsymbol{\pi}$,

$$\partial_t\big|_{t=0}\,\mathbb{E}\,f(x_t,p_t) = \mathbb{E}\langle \nabla f(x_0,p_0), (\dot{x}_0,\dot{p}_0)\rangle = \int \langle \nabla f(x_0,p_0), (\dot{x}_0,\dot{p}_0)\rangle\,\mathrm{d}\boldsymbol{\pi}(x_0,p_0)$$

[1] In contrast, Newton's law $\ddot{x}_t = -\nabla V(x_t)$ is a second-order ODE.

$$= -\int f \operatorname{div}\left(\begin{bmatrix} p \\ -\nabla V \end{bmatrix} \boldsymbol{\pi}\right)$$
$$= -\int f \left\{\operatorname{div}\begin{bmatrix} p \\ -\nabla V \end{bmatrix} + \left\langle \begin{bmatrix} p \\ -\nabla V \end{bmatrix}, \begin{bmatrix} \nabla_x \ln \boldsymbol{\pi} \\ \nabla_p \ln \boldsymbol{\pi} \end{bmatrix} \right\rangle\right\} \mathrm{d}\boldsymbol{\pi} = 0 \,.$$

Further properties of the Hamiltonian dynamics are explored in Exercise 5.4.

However, simply running Hamilton's equations does not yield a convergent sampling algorithm. For example, suppose that $V(x) = \frac{1}{2}\|x\|^2$; then, each flow map F_t is actually a diffeomorphism. This implies, for example, that $\mathsf{KL}((F_t)_{\#}\boldsymbol{\mu} \parallel (F_t)_{\#}\boldsymbol{\pi}) = \mathsf{KL}(\boldsymbol{\mu} \parallel \boldsymbol{\pi})$ for any initial distribution $\boldsymbol{\mu}$ on phase space. To get around this issue, we can "refresh" the momentum periodically. More specifically, we pick an integration time $T > 0$, and every T units of time we draw a new momentum vector from the standard Gaussian distribution (which is the distribution of the momentum under $\boldsymbol{\pi}$).

Ideal HMC: Pick an integration time $T > 0$ and draw $(X_0, P_0) \sim \boldsymbol{\mu}_0$. For each iteration $k = 0, 1, 2, \ldots$:

1. *Refresh* the velocity by drawing $P'_{kT} \sim \mathsf{normal}(0, I_d)$.
2. *Integrate* Hamilton's equations: set $(X_{(k+1)T}, P_{(k+1)T}) := F_T(X_{kT}, P'_{kT})$.

Since both steps of each iteration preserve $\boldsymbol{\pi}$, the entire algorithm preserves $\boldsymbol{\pi}$. At this stage, though, this algorithm is still idealized because it assumes the ability to exactly integrate Hamilton's equations. This may be possible for very special cases, but it is not in general, and certainly not within our oracle model. Nevertheless, it is instructive to first analyze the ideal algorithm.

5.2.2 Analysis of Ideal HMC

Theorem 5.2.1 (ideal HMC, [CV19]). *Assume that the target $\pi \propto \exp(-V)$ satisfies $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. For $k \in \mathbb{N}$, let π_{kT} denote the law of the k-th iterate X_{kT} of ideal HMC with integration time $T > 0$. Then, if we set $T = \frac{1}{2\sqrt{\beta}}$, we obtain*

$$W_2^2(\mu_{NT}, \pi) \le \exp\left(-\frac{N}{16\kappa}\right) W_2^2(\mu_0, \pi) \,.$$

It is known that the convergence rate in this theorem is optimal, see [CV19]. We now follow the proof, which is a purely deterministic analysis of Hamilton's equations. First, we need two lemmas.

Lemma 5.2.2 (a priori bound). *Let $(x_t, p_t)_{t\geq 0}$ and $(x'_t, p'_t)_{t\geq 0}$ denote two solutions to Hamilton's equations of motion with $p_0 = p'_0$ and a potential V satisfying $\nabla^2 V \preceq \beta I_d$. Then, for all $t \in [0, \frac{1}{2\sqrt{\beta}}]$, it holds that*

$$\frac{1}{2}\|x_0 - x'_0\|^2 \leq \|x_t - x'_t\|^2 \leq 2\|x_0 - x'_0\|^2 .$$

Proof. First, note that $\partial_t\|p_t - p'_t\| \leq \|\nabla V(x_t) - \nabla V(x'_t)\| \leq \beta\|x_t - x'_t\|$. It follows that $|\partial_t \|x_t - x'_t\|| \leq \|p_t - p'_t\| \leq \beta\int_0^t \|x_s - x'_s\|\,\mathrm{d}s$ and hence

$$\|x_t - x'_t\| \leq \|x_0 - x'_0\| + \beta\int_0^t\int_0^s \|x_r - x'_r\|\,\mathrm{d}r\,\mathrm{d}s .$$

Applying an ODE comparison lemma, one may deduce that $\|x_t - x'_t\| \leq \sqrt{2}\|x_0 - x'_0\|$. The lower bound is similar. □

Lemma 5.2.3 (coercivity). *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ satisfies $0 \preceq \nabla^2 f \preceq \beta I_d$. Then, for all $x, y \in \mathbb{R}^d$, it holds that*

$$\beta\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \|\nabla f(x) - \nabla f(y)\|^2 .$$

Proof. See Exercise 5.5. □

We now prove the contraction result for the Hamiltonian dynamics.

Proposition 5.2.4 (contraction of Hamilton's equations). *Consider any two solutions $(x_t, p_t)_{t\geq 0}$ and $(x'_t, p'_t)_{t\geq 0}$ to Hamilton's equations of motion with $p_0 = p'_0$ and a potential V satisfying $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, for all $t \in [0, \frac{1}{2\sqrt{\beta}}]$, it holds that*

$$\|x_t - x'_t\|^2 \leq \exp\Big(-\frac{\alpha t^2}{4}\Big)\|x_0 - x'_0\|^2 .$$

Proof. We compute

$$\frac{1}{2}\partial_t\|x_t - x'_t\|^2 = \langle x_t - x'_t, p_t - p'_t\rangle ,$$

$$\frac{1}{2}\partial_t^2\|x_t - x'_t\|^2 = \|p_t - p'_t\|^2 - \langle x_t - x'_t, \nabla V(x_t) - \nabla V(x'_t)\rangle = -\rho_t\|x_t - x'_t\|^2 + \|p_t - p'_t\|^2 ,$$

where we define

$$\rho_t := \frac{\langle \nabla V(x_t) - \nabla V(x_t'), x_t - x_t' \rangle}{\|x_t - x_t'\|^2}.$$

To bound $\|p_t - p_t'\|^2$, we use $|\partial_t \|p_t - p_t'\|| \le \|\nabla V(x_t) - \nabla V(x_t')\|$. Also, by the coercivity lemma (Lemma 5.2.3),

$$\|\nabla V(x_t) - \nabla V(x_t')\|^2 \le \beta \langle \nabla V(x_t) - \nabla V(x_t'), x_t - x_t' \rangle = \beta \rho_t \|x_t - x_t'\|^2 \le 2\beta\rho_t \|x_0 - x_0'\|^2$$

where we used Lemma 5.2.2. Hence, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\|p_t - p_t'\|^2 &\le \left| \int_0^t \left| \partial_s \|p_s - p_s'\| \right| \mathrm{d}s \right|^2 \le \left| \int_0^t \sqrt{2\beta\rho_s}\, \|x_0 - x_0'\| \, \mathrm{d}s \right|^2 \\ &\le 2\beta t \|x_0 - x_0'\|^2 \int_0^t \rho_s \, \mathrm{d}s .\end{aligned}$$

From this and Lemma 5.2.2, we deduce

$$\partial_t^2 \|x_t - x_t'\|^2 \le -\Big(\rho_t - 4\beta t \int_0^t \rho_s \, \mathrm{d}s \Big) \|x_0 - x_0'\|^2 .$$

Integrating and using $\rho \ge 0$,

$$\begin{aligned}\partial_t \|x_t - x_t'\|^2 &\le -\Big(\int_0^t \rho_s \, \mathrm{d}s - 4\beta \int_0^t s \int_0^r \rho_r \, \mathrm{d}r \, \mathrm{d}s \Big) \|x_0 - x_0'\|^2 \\ &\le -\Big(\int_0^t \rho_s \, \mathrm{d}s - 2\beta t^2 \int_0^t \rho_s \, \mathrm{d}s \Big) \|x_0 - x_0'\|^2 \\ &= -(1 - 2\beta t^2) \Big(\int_0^t \rho_s \, \mathrm{d}s \Big) \|x_0 - x_0'\|^2 \le -\frac{\alpha t}{2} \|x_0 - x_0'\|^2 .\end{aligned}$$

Integrating again then yields

$$\|x_t - x_t'\|^2 \le \|x_0 - y_0\|^2 - \frac{\alpha t^2}{4} \|x_0 - x_0'\|^2 \le \exp\Big(-\frac{\alpha t^2}{4}\Big) \|x_0 - x_0'\|^2 . \qquad \square$$

If we choose $T = \frac{1}{2\sqrt{\beta}}$, then the contraction factor is $\exp(-\frac{1}{16\kappa}) \le 1 - 1/(32\kappa)$. In particular, let P denote the transition kernel of one step of ideal HMC with integration time T. We have the W_1 contraction

$$W_1\big(P((x,p),\cdot), P((x',p'),\cdot)\big) \le \Big(1 - \frac{1}{32\kappa}\Big) \|(x,p) - (x',p')\| ,$$

which, in the language of Section 2.7, says that the coarse Ricci curvature of P is bounded below by $\kappa/32$. Hence, by Theorem 2.7.5, we immediately obtain the following corollary.

Corollary 5.2.5. *Assume that the target $\pi \propto \exp(-V)$ satisfies $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Let P be the Markov kernel for ideal HMC with integration time $T = \frac{1}{2\sqrt{\beta}}$. Then, P satisfies a Poincaré inequality with constant at most 32κ, where $\kappa := \beta/\alpha$.*

We conclude this section by observing that, since Hamilton's equations are a deterministic system of ODEs, we can approximately integrate them using any ODE solver; unlike for the Langevin diffusion, there is no need to consider any SDE discretization here. By following this approach, [CV19] also provide the following sampling guarantee.

Theorem 5.2.6 (unadjusted HMC, [CV19])**.** *Assume that the target $\pi \propto \exp(-V)$ satisfies $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$. Then, there is a sampling algorithm based on a discretization of ideal HMC which outputs μ satisfying $\sqrt{\alpha}\, W_2(\mu, \pi) \leq \varepsilon$ using*

$$\widetilde{O}\Big(\frac{\kappa^{3/2} d^{1/2}}{\varepsilon}\Big) \qquad \textit{gradient queries}\,.$$

5.3 The Underdamped Langevin Diffusion

In ideal HMC, the momentum is refreshed periodically. We now consider a variant in which the momentum is refreshed continuously. The **underdamped Langevin diffusion** is the solution to the SDE

$$\begin{aligned} \mathrm{d}X_t &= P_t \,\mathrm{d}t\,, \\ \mathrm{d}P_t &= -\nabla V(X_t)\,\mathrm{d}t - \gamma P_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,. \end{aligned}$$

Here, $\gamma > 0$ is a parameter known as the **friction parameter**; as the name suggests, the physical interpretation is that the Hamiltonian system is damped by friction.

The underdamped Langevin diffusion is motivated by the acceleration phenomenon in optimization, which we first recall in continuous time.

Optimization Box 5.3.1. Consider the following ODE system:

$$\begin{aligned} \dot{x}_t &= p_t\,, \\ \dot{p}_t &= -\nabla V(x_t) - \gamma p_t\,. \end{aligned}$$

This is the deterministic analogue of the underdamped Langevin diffusion. If we

assume that V is α-strongly convex, then one can show that with the choice $\gamma = 2\sqrt{\alpha}$,

$$V(x_t) - V(x_\star) \le 2\exp(-\sqrt{\alpha}\, t)\,\{V(x_0) - V(x_\star)\}\,, \tag{5.3.2}$$

see Exercise 5.6. Moreover, the ODE system is stable for integration times of order $t \asymp 1/\sqrt{\beta}$, where β is the smoothness of V, and hence one expects that the discretization of this system yields an algorithm for optimization with a *square root* dependence on the condition number κ. This is indeed the case, but discretization is subtle; see, e.g., [Nes18, §2.2] for an analysis of a discrete-time scheme which achieves $V(x_N) - V(x_\star) \le \varepsilon^2$ in $O(\sqrt{\kappa}\log(\kappa\,(V(x_0) - V(x_\star))/\varepsilon^2))$ iterations.

This phenomenon is known as **acceleration** in optimization. Historically, the development happened in the opposite order: Nesterov put forth his algorithm, now known as **Nesterov's accelerated gradient descent**, in [Nes83], and his algorithm is optimal amongst all first-order algorithms [NY83]. The continuous-time formulation of acceleration was introduced later, in [SBC16].

Motivated by the acceleration phenomenon in optimization, we undertake a detailed study of the underdamped Langevin diffusion to see if such a phenomenon also holds for log-concave sampling. At present, however, our understanding is inconclusive.

Unlike the Langevin diffusion, the underdamped Langevin diffusion is *not* a reversible Markov process. Moreover, it is an example of *hypocoercive dynamics*, which means that the Markov semigroup approach based on Poincaré and log-Sobolev inequalities fails, necessitating the use of more sophisticated PDE analysis.

5.3.1 Continuous-Time Considerations

It is illuminating to write the dynamics in the space of measures. By computing the generator of this Markov process and writing down the corresponding Fokker–Planck equation, we arrive at the PDE

$$\partial_t \boldsymbol{\pi}_t = \gamma\Delta_p \boldsymbol{\pi}_t + \operatorname{div}\left(\boldsymbol{\pi}_t \begin{bmatrix} -p \\ \nabla V + \gamma p \end{bmatrix}\right)$$

for the evolution of the law $(\boldsymbol{\pi}_t)_{t\ge 0}$ of the underdamped Langevin diffusion. We can write this as the continuity equation

$$\partial_t \boldsymbol{\pi}_t = \operatorname{div}\left(\boldsymbol{\pi}_t \begin{bmatrix} \nabla_p \ln \boldsymbol{\pi} \\ -\nabla_x \ln \boldsymbol{\pi} - \gamma\nabla_p \ln \boldsymbol{\pi} + \gamma\nabla_p \ln \boldsymbol{\pi}_t \end{bmatrix}\right)$$

where $\boldsymbol{\pi}(x,p) \propto \exp(-H(x,p)) = \exp(-V(x) - \frac{1}{2}\|p\|^2)$. However, taking advantage of the fact that

$$\operatorname{div}\left(\boldsymbol{\pi}_t \begin{bmatrix} -\nabla_p \ln \boldsymbol{\pi}_t \\ \nabla_x \ln \boldsymbol{\pi}_t \end{bmatrix}\right) = 0\,,$$

we have the more interpretable expression

$$\partial_t \boldsymbol{\pi}_t = \operatorname{div}\left(\boldsymbol{\pi}_t \boldsymbol{J}_\gamma \begin{bmatrix} \nabla_x \ln(\boldsymbol{\pi}_t/\boldsymbol{\pi}) \\ \nabla_p \ln(\boldsymbol{\pi}_t/\boldsymbol{\pi}) \end{bmatrix}\right), \qquad \boldsymbol{J}_\gamma := \begin{bmatrix} 0 & 1 \\ -1 & \gamma \end{bmatrix},$$

or

$$\partial_t \boldsymbol{\pi}_t = \operatorname{div}\left(\boldsymbol{\pi}_t \boldsymbol{J}_\gamma \left[\nabla_{W_2} \mathsf{KL}(\cdot \,\|\, \boldsymbol{\pi})\right](\boldsymbol{\pi}_t)\right). \tag{5.3.3}$$

This shows that the underdamped Langevin diffusion is not interpreted as a gradient flow of the KL divergence, but rather a "damped Hamiltonian flow" for the KL divergence.

We begin with a contraction result for the continuous-time process based on [Che+18]. Note that we use the same change of variables as in Exercise 5.6.

Theorem 5.3.4. *Let $(X_t^0, P_t^0)_{t\geq 0}$ and $(X_t^1, P_t^1)_{t\geq 0}$ be two copies of the underdamped Langevin diffusion, driven by the same Brownian motion. Assume that the potential satisfies $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, defining the modified norm*

$$|||(x,p)|||^2 := \left\|x + \sqrt{\frac{2}{\beta}}\, p\right\|^2 + \|x\|^2$$

and setting $\gamma = \sqrt{2\beta}$, we obtain the contraction

$$|||(X_t^1, P_t^1) - (X_t^0, P_t^0)||| \leq \exp\left(-\frac{\alpha t}{\sqrt{2\beta}}\right) |||(X_0^1, P_0^1) - (X_0^0, P_0^0)|||\,.$$

Proof. Write $\delta X_t := X_t^1 - X_t^0$ and $\delta P_t := P_t^1 - P_t^0$. Then, by Itô's formula (Theorem 1.1.18),

$$\begin{aligned}
\mathrm{d}(\delta X_t + \eta\, \delta P_t) &= \left[\delta P_t - \eta\left\{\nabla V(X_t^1) - \nabla V(X_t^0)\right\} - \gamma\eta\, \delta P_t\right] \mathrm{d}t \\
&= \Bigg[-(\gamma\eta - 1)\, \delta P_t - \eta \underbrace{\left(\int_0^1 \nabla^2 V\left((1-s)\, X_t^0 + s\, X_t^1\right) \mathrm{d}s\right)}_{=:H_t} \delta X_t\Bigg] \mathrm{d}t \\
&= \left[-\left(\gamma - \frac{1}{\eta}\right)(\delta X_t + \eta\, \delta P_t) + \left(\gamma - \frac{1}{\eta} - \eta H_t\right) \delta X_t\right] \mathrm{d}t
\end{aligned}$$

as well as

$$\mathrm{d}(\delta X_t) = \delta P_t \,\mathrm{d}t = \Big[\frac{1}{\eta}\,(\delta X_t + \eta\,\delta P_t) - \frac{1}{\eta}\,\delta X_t\Big]\,\mathrm{d}t$$

so that

$$\begin{aligned}&\frac{1}{2}\,\partial_t\{\|\delta X_t + \eta\,\delta P_t\|^2 + \|\delta X_t\|^2\}\\ &\qquad = -\Big\langle\begin{bmatrix}\delta X_t + \eta\,\delta P_t\\ X_t\end{bmatrix}, \begin{bmatrix}\gamma - \frac{1}{\eta} & \frac{1}{2}\,(\eta H_t - \gamma)\\ \frac{1}{2}\,(\eta H_t - \gamma) & \frac{1}{\eta}\end{bmatrix}\begin{bmatrix}\delta X_t + \eta\,\delta P_t\\ X_t\end{bmatrix}\Big\rangle\,.\end{aligned}$$

We now check that if $\gamma = \frac{2}{\eta}$ and $\eta = \sqrt{\frac{2}{\beta}}$, then the eigenvalues of the matrix above are lower bounded by $\alpha\eta/2 = \alpha/\sqrt{2\beta}$. □

We check that the new norm we defined is equivalent to the Euclidean norm.

Lemma 5.3.5. *For all $x, p \in \mathbb{R}^d$,*

$$\frac{1}{3}\left(\|x\|^2 + \frac{2}{\beta}\,\|p\|^2\right) \le |||(x,p)|||^2 \le 3\left(\|x\|^2 + \frac{2}{\beta}\,\|p\|^2\right).$$

Proof. The upper bound follows from

$$|||(x,p)|||^2 = \Big\|x + \sqrt{\frac{2}{\beta}}\,p\Big\|^2 + \|x\|^2 \le 2\left(\|x\|^2 + \frac{2}{\beta}\,\|p\|^2\right) + \|x\|^2\,.$$

The lower bound follows from

$$\frac{2}{\beta}\,\|p\|^2 \le 2\,\Big\|x + \sqrt{\frac{2}{\beta}}\,p\Big\|^2 + 2\,\|x\|^2\,.$$

□

Consequently, the contraction result in Theorem 5.3.4 implies

$$\|X_t^1 - X_t^0\|^2 + \frac{2}{\beta}\,\|P_t^1 - P_t^0\|^2 \le 9\exp\Big(-\frac{\sqrt{2}\,\alpha t}{\sqrt{\beta}}\Big)\left(\|X_0^1 - X_0^0\|^2 + \frac{2}{\beta}\,\|P_0^1 - P_0^0\|^2\right).$$

***Remark* 5.3.6.** If we compare Theorem 5.3.4 with Optimization Box 5.3.1, we see that we had to choose a larger value for the friction ($\gamma \asymp \sqrt{\beta}$ rather than $\gamma \asymp \sqrt{\alpha}$) and this leads to a slower exponential contraction with rate $\alpha/\sqrt{\beta}$ (instead of $\sqrt{\alpha}$). Thus, Theorem 5.3.4 can be considered an *unaccelerated* convergence rate.

5.3.2 Wasserstein Coupling Argument

We now discretize the underdamped Langevin diffusion. Of course, we could apply a simple Euler–Maruyama discretization to the SDE, but there is a slightly better discretization here. We observe that if we fix the value of the gradient term at time kh, then the rest of the SDE is a linear SDE, and can be integrated exactly. Namely, consider

$$\begin{aligned} \mathrm{d}X_t &= P_t \,\mathrm{d}t\,, \\ \mathrm{d}P_t &= -\nabla V(X_{kh})\,\mathrm{d}t - \gamma P_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,, \end{aligned} \qquad \text{for } t \in [kh, (k+1)h]\,. \tag{ULMC}$$

Then, the solution to the SDE is given explicitly in the following lemma.

Lemma 5.3.7. *Conditioned on (X_{kh}, P_{kh}), the law of $(X_{(k+1)h}, P_{(k+1)h})$ is explicitly given as* $\mathsf{normal}(M_{(k+1)h}, \Sigma)$ *where*

$$M_{(k+1)h} = \begin{bmatrix} X_{kh} + \gamma^{-1}\left(1 - \exp(-\gamma h)\right) P_{kh} - \gamma^{-1}\left(h - \gamma^{-1}\left(1 - \exp(-\gamma h)\right)\right) \nabla V(X_{kh}) \\ P_{kh}\exp(-\gamma h) - \gamma^{-1}\left(1 - \exp(-\gamma h)\right) \nabla V(X_{kh}) \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} \frac{2}{\gamma}\left\{h - \frac{2}{\gamma}\left(1 - \exp(-\gamma h)\right) + \frac{1}{2\gamma}\left(1 - \exp(-2\gamma h)\right)\right\} & * \\ \frac{1}{\gamma}\left\{1 - 2\exp(-\gamma h) + \exp(-2\gamma h)\right\} & 1 - \exp(-2\gamma h) \end{bmatrix} \otimes I_d\,.$$

The $*$ indicates that the entry is determined by symmetry.

The lemma is an exercise in stochastic calculus (Exercise 5.8). The point is that the discretization given as ULMC is implementable.

We now proceed to a discretization analysis based on [Che+18].

Theorem 5.3.8. *For $k \in \mathbb{N}$, let $\boldsymbol{\mu}_{kh}$ denote the law of the k-th iterate of ULMC with appropriately tuned step size $h > 0$ and friction parameter $\gamma > 0$. Also, let μ_{kh} denote the law of X_{kh}. Assume that the target $\pi \propto \exp(-V)$ satisfies $\nabla V(0) = 0$ and $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, we obtain the guarantee $\sqrt{\alpha}\, W_2(\mu_{Nh}, \pi) \le \varepsilon$ after*

$$N = \widetilde{O}\Big(\frac{\kappa^2 d^{1/2}}{\varepsilon}\Big) \qquad \textit{iterations}\,.$$

***Remark* 5.3.9.** This result is not the best possible. Indeed, a refined analysis by [DR20] obtains the iteration complexity $\widetilde{O}(\kappa^{3/2} d^{1/2}/\varepsilon)$ by improving the continuous-time contraction result of Theorem 5.3.4.

Proof. **One-step discretization bound.** As in Theorem 4.1.2, we start with a one-step bound. Let $(X_t, P_t)_{t\geq 0}$ denote ULMC and let $(\bar{X}_t, \bar{P}_t)_{t\geq 0}$ denote the continuous-time underdamped Langevin diffusion, both driven by the same Brownian motion and started at the same random pair. We want to bound the distance $\mathbb{E}[|||(X_h, P_h) - (\bar{X}_h, \bar{P}_h)|||^2]$. According to Lemma 5.3.5, it suffices to bound $\mathbb{E}[\|X_h - \bar{X}_h\|^2]$ and $\mathbb{E}[\|P_h - \bar{P}_h\|^2]$ separately.

First,

$$\mathbb{E}[\|X_h - \bar{X}_h\|^2] = \mathbb{E}\Big[\Big\|\int_0^h \{P_t - \bar{P}_t\}\,\mathrm{d}t\Big\|^2\Big] \leq h\int_0^h \mathbb{E}[\|P_t - \bar{P}_t\|^2]\,\mathrm{d}t\,.$$

Next,

$$\begin{aligned}\mathbb{E}[\|P_t - \bar{P}_t\|^2] &= \mathbb{E}\Big[\Big\|\int_0^t \{-\nabla V(\bar{X}_0) + \nabla V(\bar{X}_s) - \gamma\,(P_s - \bar{P}_s)\}\,\mathrm{d}s\Big\|^2\Big] \\ &\lesssim h\int_0^t \big(\mathbb{E}[\|\nabla V(\bar{X}_s) - \nabla V(\bar{X}_0)\|^2] + \gamma^2\,\mathbb{E}[\|P_s - \bar{P}_s\|^2]\big)\,\mathrm{d}s\,.\end{aligned}$$

By Grönwall's inequality, if $h \leq \frac{1}{\gamma} = \frac{1}{\sqrt{2\beta}}$, then

$$\mathbb{E}[\|P_t - \bar{P}_t\|^2] \lesssim h\int_0^t \mathbb{E}[\|\nabla V(\bar{X}_s) - \nabla V(\bar{X}_0)\|^2]\,\mathrm{d}s \leq \beta^2 h\int_0^t \mathbb{E}[\|\bar{X}_s - \bar{X}_0\|^2]\,\mathrm{d}t\,.$$

Again, we need a movement bound for the underdamped Langevin diffusion, which is done in Lemma 5.3.10. Substituting this in and assuming $h \lesssim \frac{1}{\beta}$,

$$\begin{aligned}\mathbb{E}[|||(X_h, P_h) - (\bar{X}_h, \bar{P}_h)|||^2] &\lesssim \mathbb{E}[\|X_h - \bar{X}_h\|^2] + \frac{1}{\beta}\,\mathbb{E}[\|P_h - \bar{P}_h\|^2] \\ &\lesssim \beta^2 h^4\,\mathbb{E}[|||(X_0, P_0)|||^2] + \beta^{3/2} d h^5\,.\end{aligned}$$

Multi-step discretization bound. Let $\mathcal{W}^2$ denote the coupling cost for the norm $|||\cdot|||^2$. Let $\hat{\boldsymbol{\mu}}_{(k+1)h}$ denote the law of the continuous-time underdamped Langevin diffusion started at $\boldsymbol{\mu}_{kh}$. Then, from Theorem 5.3.4 and the one-step discretization bound,

$$\begin{aligned}\mathcal{W}(\boldsymbol{\mu}_{(k+1)h}, \boldsymbol{\pi}) &\leq \mathcal{W}(\hat{\boldsymbol{\mu}}_{(k+1)h}, \boldsymbol{\pi}) + \mathcal{W}(\boldsymbol{\mu}_{(k+1)h}, \hat{\boldsymbol{\mu}}_{(k+1)h}) \\ &\leq \exp\Big(-\frac{\alpha h}{\sqrt{2\beta}}\Big)\,\mathcal{W}(\boldsymbol{\mu}_{kh}, \boldsymbol{\pi}) + O\big(\beta h^2\,\mathcal{W}(\boldsymbol{\mu}_{kh}, \boldsymbol{\delta}_0) + \beta^{3/4} d^{1/2} h^{5/2}\big)\,.\end{aligned}$$

Also, $\mathcal{W}(\boldsymbol{\mu}_{kh}, \boldsymbol{\delta}_0) \leq \mathcal{W}(\boldsymbol{\mu}_{kh}, \boldsymbol{\pi}) + \mathcal{W}(\boldsymbol{\pi}, \boldsymbol{\delta}_0) \lesssim \mathcal{W}(\boldsymbol{\mu}_{kh}, \boldsymbol{\pi}) + \sqrt{d/\alpha}$, where we used the moment bound in Lemma 4.1.5. If $h \lesssim \frac{1}{\beta^{1/2}\kappa}$, then we can absorb the $\mathcal{W}(\boldsymbol{\mu}_{kh}, \boldsymbol{\pi})$ term into

the contraction rate and deduce

$$\mathcal{W}(\boldsymbol{\mu}_{(k+1)h}, \boldsymbol{\pi}) \le \exp\Big(-\frac{\alpha h}{2\sqrt{\beta}}\Big)\mathcal{W}(\boldsymbol{\mu}_{kh}, \boldsymbol{\pi}) + O\Big(\frac{\beta d^{1/2}h^2}{\alpha^{1/2}} + \beta^{3/4}d^{1/2}h^{5/2}\Big)\,.$$

Iterating,

$$\mathcal{W}(\boldsymbol{\mu}_{Nh}, \boldsymbol{\pi}) \le \exp\Big(-\frac{\alpha N h}{2\sqrt{\beta}}\Big)\mathcal{W}(\boldsymbol{\mu}_{0}, \boldsymbol{\pi}) + O\Big(\frac{\beta^{3/2}d^{1/2}h}{\alpha^{3/2}} + \frac{\beta^{5/4}d^{1/2}h^{3/2}}{\alpha}\Big)\,.$$

Choosing the step size appropriately yields the result. □

The next lemma provides the movement bound (Exercise 5.9).

Lemma 5.3.10. *Let $(\bar{X}_t, \bar{P}_t)_{t\ge 0}$ denote the underdamped Langevin diffusion with potential V satisfying $\nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$. If $t \le \frac{1}{\gamma} \wedge \frac{1}{\sqrt{\beta}}$, then*

$$\mathbb{E}[\|\bar{X}_t - \bar{X}_0\|^2] \lesssim t^2\,\mathbb{E}[\|\bar{P}_0\|^2] + \gamma d t^3 + \beta^2 t^4\,\mathbb{E}[\|\bar{X}_0\|^2]\,.$$

5.3.3 Randomized Midpoint Discretization

The randomized midpoint method of Section 5.1 can be applied to the underdamped Langevin diffusion to yield an even better sampling guarantee. This was carried out in [SL19], and we state the final result here.

Theorem 5.3.11. *Assume that the target $\pi \propto \exp(-V)$ satisfies $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$. Then, the randomized midpoint discretization of the underdamped Langevin diffusion outputs μ such that $\sqrt{\alpha}\,W_2(\mu, \pi) \le \varepsilon$ using*

$$\widetilde{O}\Big(\frac{\kappa d^{1/3}}{\varepsilon^{2/3}}\,\big(1 \vee \frac{\varepsilon^{1/3}\kappa^{1/6}}{d^{1/6}}\big)\Big) \qquad \textit{gradient queries}\,.$$

With respect to the dimension dependence, this is the current state-of-the-art guarantee for sampling from strongly log-concave distributions.

TODO: Provide a proof.

Bibliographical Notes

In the analysis of the randomized midpoint discretization of the Langevin diffusion (Theorem 5.1.1), we have simplified the original proof of [HBE20] at the cost of proving a slightly weaker result. The sharper argument of [HBE20] is outlined in Exercise 5.2.

Besides the randomized midpoint method, the **shifted ODE discretization** [FLO21] also achieves a state-of-the-art iteration complexity of $\widetilde{O}(d^{1/3}/\varepsilon^{2/3})$ when applied to the underdamped Langevin diffusion.

HMC and its variants are some of the most popular algorithms employed in practice, especially the **no-U-turn sampler** (**NUTS**) [HG14] which adaptively sets the integration time. In terms of complexity analysis, the paper [CV19] (whose proof we followed in Theorem 5.2.1) provided the tight analysis of ideal HMC. Other complexity results obtained for HMC under various assumptions include [MV18; MS19; BEZ20].

The underdamped Langevin diffusion has been studied quantitatively in [Che+18; EGZ19; DR20; Ma+21].

TODO: Literature on hypocoercivity.

Exercises

Randomized Midpoint Discretization

▷ **Exercise 5.1** (one-step discretization bound for RM-LMC)
Prove the one-step discretization bound (5.1.2) for RM-LMC.

▷ **Exercise 5.2** (sharper rate for RM-LMC)
In this exercise we show how to obtain a slightly sharper guarantee for RM-LMC than the one in Theorem 5.1.1. Note that Theorem 5.1.1 provides the rate

$$N = \widetilde{O}\Big(\frac{\kappa d^{1/2}}{\varepsilon} \vee \frac{\kappa^{3/2} d^{1/4}}{\varepsilon^{1/2}}\Big) = \begin{cases} \kappa d^{1/2}/\varepsilon\,, & \kappa \le \sqrt{d}/\varepsilon\,, \\ \kappa^{3/2} d^{1/4}/\varepsilon^{1/2}\,, & \kappa \ge \sqrt{d}/\varepsilon\,, \end{cases} \tag{5.E.1}$$

whereas the rate we show in this exercise is

$$N = \widetilde{O}\Big(\frac{\kappa d^{1/2}}{\varepsilon} \vee \frac{\kappa^{4/3} d^{1/3}}{\varepsilon^{2/3}}\Big) = \begin{cases} \kappa d^{1/2}/\varepsilon\,, & \kappa \le \sqrt{d}/\varepsilon\,, \\ \kappa^{4/3} d^{1/3}/\varepsilon^{2/3}\,, & \kappa \ge \sqrt{d}/\varepsilon\,. \end{cases} \tag{5.E.2}$$

1. Check that the rate (5.E.2) is indeed better than (5.E.1).

The main idea behind the improved rate is that throughout the proof of Theorem 5.1.1, we used the inequality

$$\mathbb{E}[\|\nabla V(X_{kh})\|^2] \le \beta^2 \mathbb{E}[\|X_{kh}\|^2], \tag{5.E.3}$$

which is wasteful. Instead, we will show that

$$\frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}[\|\nabla V(X_{kh})\|^2] \lesssim \beta d. \tag{5.E.4}$$

Note that since we expect $\mathbb{E}[\|X_{kh}\|^2] \asymp d/\alpha$, then the new bound (5.E.4) is an improvement by a factor of κ.

2. Rewrite the proof of Theorem 5.1.1, avoiding the use of the inequality (5.E.3), and leaving the error bound in terms of $\sum_{k=0}^{N-1} \mathbb{E}[\|\nabla V(X_{kh})\|^2]$. As a sanity check, the result should imply the rate (5.E.2) once the key inequality (5.E.4) is proven.

3. Applying Itô's formula (Theorem 1.1.18) to $V(\bar{X}_t)$, write down an expression for $\mathbb{E}[V(\bar{X}_{(k+1)h}) - V(X_{kh})]$. By bounding the error terms carefully and assuming that $h \lesssim \frac{1}{\beta}$, prove that

$$\begin{aligned}&\mathbb{E}[V(X_{(k+1)h}) - V(\bar{X}_{(k+1)h})] \\ &\quad \ge \mathbb{E}[V(X_{(k+1)h}) - V(X_{kh})] + \frac{h}{4}\mathbb{E}[\|\nabla V(X_{kh})\|^2] - O(\beta d h).\end{aligned}$$

4. Using the smoothness inequality,

$$\begin{aligned}&\mathbb{E}[V(X_{(k+1)h}) \mid \mathscr{F}_{(k+1)h}] \\ &\quad \le V(\bar{X}_{(k+1)h}) + \langle \nabla V(\bar{X}_{(k+1)h}), \mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] - \bar{X}_{(k+1)h}\rangle \\ &\qquad + \frac{\beta}{2}\mathbb{E}[\|X_{(k+1)h} - \bar{X}_{(k+1)h}\|^2 \mid \mathscr{F}_{(k+1)h}].\end{aligned}$$

   Applying the Cauchy–Schwarz and Young's inequality to the middle term,

$$\begin{aligned}&\langle \nabla V(\bar{X}_{(k+1)h}), \mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] - \bar{X}_{(k+1)h}\rangle \\ &\quad \le \lambda \|\nabla V(\bar{X}_{(k+1)h})\|^2 + \frac{1}{\lambda}\|\mathbb{E}[X_{(k+1)h} \mid \mathscr{F}_{(k+1)h}] - \bar{X}_{(k+1)h}\|^2\end{aligned}$$

   for an appropriate choice of $\lambda > 0$. Use this to show that

$$\mathbb{E}[V(X_{(k+1)h}) - V(\bar{X}_{(k+1)h})] \lesssim \beta h^2 \mathbb{E}[\|\nabla V(X_{kh})\|^2] + \beta^3 d h^3.$$

5. Combining these inequalities, assuming that βh is sufficiently small and that we initialize with $X_0 = \arg\min V$, prove the key inequality (5.E.4).

Hamiltonian Monte Carlo

⊵ Exercise 5.3 (Gaussian calculations for HMC)
Suppose that the target π is a standard Gaussian distribution. Compute the flow map F_t for the Hamiltonian dynamics. Also, if we start at the initial distribution $\mu_0 = \text{normal}(0, \sigma^2 I_d)$, show that the distribution $\boldsymbol{\mu}_t$ over phase space at time t of ideal HMC is a Gaussian distribution, $\text{normal}(0, \Sigma_t)$, and compute $\Sigma_t \in \mathbb{R}^{2d \times 2d}$.

⊵ Exercise 5.4 (basic properties of Hamiltonian dynamics)
In this exercise, we explore some fundamental properties of the Hamiltonian dynamics.

1. (conservation of energy) Along the Hamiltonian dynamics, $(x_t, p_t)_{t \geq 0}$, show that $H(x_t, p_t) = H(x_0, p_0)$. In fact, for any function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ whose **Poisson bracket** with H vanishes, i.e.,
$$\{f, H\} \coloneqq \langle \nabla_x f, \nabla_p H \rangle - \langle \nabla_p f, \nabla_x H \rangle = 0\,,$$
it holds that $f(x_t, p_t) = f(x_0, p_0)$.

2. (conservation of volume) By differentiating $t \mapsto \det \nabla F_t(x, p)$ and using the flow map equation $\partial_t F_t(x, p) = \boldsymbol{J} \nabla H(F_t(x, p))$, prove that $\det \nabla F_t(x, p) = 1$ for all $t \geq 0$ and $x, p \in \mathbb{R}^d$. This shows that $F_t : \mathbb{R}^d \to \mathbb{R}^d$ is a volume-preserving map.

3. (time reversibility) Suppose that $(x_t, p_t)_{t \in [0,T]}$ solve Hamilton's equations. Show that $(x_{T-t}, -p_{T-t})_{t \in [0,T]}$ also solve Hamilton's equations. In other words, if R is the moment reversal operator, i.e.,
$$R = \begin{bmatrix} I_d & 0 \\ 0 & -I_d \end{bmatrix},$$
then $F_T^{-1} = R \circ F_T \circ R$.

⊵ Exercise 5.5 (coercivity)
Prove Lemma 5.2.3.

Hint: Let $z \coloneqq y - \frac{1}{\beta}\{\nabla f(y) - \nabla f(x)\}$. Apply the convexity inequality to $f(x) - f(z)$, and the smoothness inequality to $f(z) - f(y)$, in order to upper bound $f(x) - f(y)$. Combine this with the symmetric inequality for $f(y) - f(x)$.

The Underdamped Langevin Diffusion

⊵ Exercise 5.6 (Nesterov's algorithm in continuous time)
Consider the continuous-time formulation of Nesterov's algorithm, as given in Optimization Box 5.3.1. Assume that V is α-strongly convex. Prove the rate (5.3.2).

Hint: Let $z_t \coloneqq x_t + \frac{2}{\gamma} p_t$. Consider the Lyapunov functional

$$\mathcal{L}_t \coloneqq V(x_t) - V(x_\star) + \frac{\alpha}{2} \|z_t - x_\star\|^2 .$$

Prove that $\dot{\mathcal{L}}_t \leq 0$. You may find the following identity to be helpful: $\langle z - x, z - x^\star \rangle = \frac{1}{2} (\|z - x\|^2 + \|z - x^\star\|^2 - \|x - x^\star\|^2)$.

▷ **Exercise 5.7** (Fokker–Planck equation for the underdamped Langevin diffusion)
Compute the generator $\mathscr{L}$ of the underdamped Langevin diffusion and show that it can be written as $\mathscr{L} = \gamma \mathscr{L}_{\mathrm{OU}} + \mathscr{L}_{\mathrm{Ham}}$, where $\mathscr{L}_{\mathrm{OU}}$ is the generator of the Ornstein–Uhlenbeck process (Exercise 1.5) acting on the momentum coordinate,

$$\mathscr{L}_{\mathrm{OU}} f \coloneqq \Delta_p f - \langle p, \nabla_p f \rangle ,$$

and $\mathscr{L}_{\mathrm{Ham}}$ captures the Hamiltonian part of the dynamics,

$$\mathscr{L}_{\mathrm{Ham}} f \coloneqq \langle p, \nabla_x f \rangle - \langle \nabla V, \nabla_p f \rangle .$$

Then, check the various calculations leading up to the Fokker–Planck equation (5.3.3).

▷ **Exercise 5.8** (derivation of the ULMC updates)
Solve the SDE for ULMC to prove Lemma 5.3.7.

▷ **Exercise 5.9** (movement bound for the underdamped Langevin diffusion)
Prove the movement bound for underdamped Langevin (Lemma 5.3.10).

CHAPTER 6

Convergence in Rényi Divergence

In this chapter, we study convergence guarantees for the LMC algorithm in Rényi divergences. Recall that the Rényi divergences are a family of information divergences, indexed by a parameter q, such that the Rényi divergence of order $q = 1$ is the KL divergence, and the Rényi divergence of order 2 is related to the chi-squared divergence via $\mathcal{R}_2 = \ln(1 + \chi^2)$. We studied the continuous-time convergence of the Langevin diffusion in Rényi divergence under either a Poincaré inequality or a log-Sobolev inequality in Section 2.2.4. Here, we will build upon these results in order to study the discretized algorithm. Rényi divergence guarantees are stronger than the W_2 or KL guarantees from Chapter 4, and they have recently found use in areas such as differential privacy [Mir17].

6.1 Proof under LSI via Interpolation Argument

In this section, we follow [Che+21a], which generalizes the argument of Theorem 4.2.6 and provides a clean Rényi convergence proof for LMC under the assumption of a log-Sobolev inequality (LSI).

As in Section 4.2, we begin by writing a differential inequality for the Rényi divergence along the interpolation (4.2.2) of LMC. The proof is a combination of the proofs of Theorem 2.2.15 and Corollary 4.2.4, so it is left as Exercise 6.1. Throughout this section, let $q \geq 2$ be fixed.

Proposition 6.1.1. *Along the law $(\mu_t)_{t\geq 0}$ of the interpolated process (4.2.2),*

$$\partial_t \mathcal{R}_q(\mu_t \parallel \pi) \leq -\frac{3}{q}\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + q\,\mathbb{E}[\psi_t(X_t)\,\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]\,,$$

where $\rho_t := \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$ and $\psi_t := \rho_t^{q-1}/\mathbb{E}_\pi(\rho_t^q)$.

In analogy with the usual Fisher information, the quantity

$$\mathsf{FI}_q(\mu \parallel \pi) := \frac{4}{q}\frac{\mathbb{E}_\pi[\|\nabla(\rho^{q/2})\|^2]}{\mathbb{E}_\pi(\rho^q)}\,, \qquad \rho := \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,,$$

may be considered the "Rényi Fisher information". As in the proof of Theorem 2.2.15, under a log-Sobolev inequality, we have

$$\mathsf{FI}_q(\mu \parallel \pi) \geq \frac{2}{qC_{\mathsf{LSI}}}\,\mathcal{R}_q(\mu \parallel \pi)\,,$$

so the first term in Proposition 6.1.1 provides a decay in the Rényi divergence. In the discretization analysis, our task is to control the second term.

Note that

$$\mathbb{E}\,\psi_t(X_t) = \mathbb{E}_{\mu_t}\,\psi_t = \mathbb{E}_\pi\Big[\rho_t\,\frac{\rho_t^{q-1}}{\mathbb{E}_\pi(\rho_t^q)}\Big] = 1\,,$$

so $\psi_t(X_t)$ acts as a change of measure. The main difficulty of the proof is that whereas we know how to control the term $\|\nabla V(X_t) - \nabla V(X_{kh})\|^2$ under the original probability measure $\mathbb{P}$ (indeed, this is precisely what we accomplished in Theorem 4.2.6), it is not straightforward to control this term under the measure $\mathbf{P}$ defined by $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbb{P}} = \psi_t(X_t)$. Towards this end, we shall employ change of measure inequalities that allow us to relate expectations under $\mathbb{P}$ to expectations under $\mathbf{P}$. Note that $\psi_t = 1$ when $q = 1$, which is why these difficulties can be avoided when working with the KL divergence.

The main theorem that we wish to prove is as follows.

Theorem 6.1.2 ([Che+21a])**.** *For $k \in \mathbb{N}$, let μ_{kh} denote the law of the k-th iterate of LMC with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies LSI and that ∇V is β-Lipschitz. Also, for simplicity, assume that $C_{\mathsf{LSI}}, \beta \geq 1$. Then, for all $h \leq \frac{1}{192 C_{\mathsf{LSI}}\beta^2 q^2}$,*

for all $N \geq N_0$, it holds that

$$\mathcal{R}_q(\mu_{Nh} \,\|\, \pi) \leq \exp\Big(-\frac{(N-N_0)h}{4C_{\mathsf{LSI}}}\Big)\, \mathcal{R}_2(\mu_0 \,\|\, \pi) + \widetilde{O}(C_{\mathsf{LSI}}\beta^2 dhq)\,,$$

where $N_0 := \lceil \frac{2C_{\mathsf{LSI}}}{h} \ln(q-1) \rceil$. In particular, for all $\varepsilon \in [0, \sqrt{d/q}]$, if we choose the step size $h = \widetilde{\Theta}(\frac{\varepsilon^2}{\beta^2 dq C_{\mathsf{LSI}}})$, then we obtain the guarantee $\sqrt{\mathcal{R}_q(\mu_{Nh} \,\|\, \pi)} \leq \varepsilon$ after

$$N = \widetilde{O}\Big(\frac{C_{\mathsf{LSI}}^2 \beta^2 dq}{\varepsilon^2} \log \mathcal{R}_2(\mu_0 \,\|\, \pi)\Big) \qquad \textit{iterations}\,.$$

For clarity of exposition, we begin with a discretization analysis that incurs a worse dependence on q. Afterwards, we show how to improve the dependence on q via a hypercontractivity argument.

Proof of Theorem 6.1.2 with suboptimal dependence on q. As in the proof of Theorem 4.2.6, our aim is to control the error term $\mathbb{E}[\psi_t(X_t)\, \|\nabla V(X_t) - \nabla V(X_{kh})\|^2]$, where from the β-smoothness of V and from $h \leq \frac{1}{3\beta}$ we have

$$\|\nabla V(X_t) - \nabla V(X_{kh})\|^2 \leq 9\beta^2\, (t-kh)^2\, \|\nabla V(X_t)\|^2 + 6\beta^2\, (t-kh)\, \|B_t - B_{kh}\|^2\,.$$

There are two terms to control. For the first term, applying the duality lemma for the Fisher information (Lemma 4.2.5) to the measure $\psi_t \mu_t$,

$$\begin{aligned}\mathbb{E}_{\psi_t\mu_t}[\|\nabla V\|^2] &\leq \mathsf{FI}(\psi_t\mu_t \,\|\, \pi) + 2\beta d = \mathbb{E}_{\mu_t}\Big[\psi_t \Big\|\nabla \ln\big(\psi_t \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\big)\Big\|^2\Big] + 2\beta d \\ &= \frac{\mathbb{E}_\pi[\rho_t^q\, \|\nabla \ln(\rho_t^q)\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2\beta d = \frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2\beta d\,,\end{aligned}$$

where we used the identity

$$\mathbb{E}_{\mu_t}\Big[\psi_t \Big\|\nabla \ln\big(\psi_t \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\big)\Big\|^2\Big] = \frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} \tag{6.1.3}$$

which follows from the chain rule from calculus.

For the second error term, we must control the term $\|B_t - B_{kh}\|^2$ under the measure $\mathbf{P}$, where $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbb{P}} = \psi_t(X_t)$. The difficulty is that under $\mathbf{P}$, B is no longer a standard Brownian motion, so it is difficult to control this term directly. Instead, we apply the Donsker–Varadhan variational principle (Theorem 1.5.4) to relate the expectation under $\mathbf{P}$ (denoted $\mathbf{E}$) with the expectation under $\mathbb{P}$. For any random variable ζ, it yields

$$\mathbf{E}\zeta \leq \mathsf{KL}(\mathbf{P} \,\|\, \mathbb{P}) + \ln \mathbb{E} \exp \zeta\,.$$

Applying this to $\zeta := c\,(\|B_t - B_{kh}\| - \mathbb{E}\|B_t - B_{kh}\|)^2$ for a constant $c > 0$ to be chosen later, we obtain

$$\begin{aligned}\mathbf{E}[\|B_t - B_{kh}\|^2] &\le 2\,\mathbb{E}[\|B_t - B_{kh}\|^2] + \frac{2}{c}\,\mathbf{E}\zeta \\ &\le 2d\,(t-kh) + \frac{2}{c}\left\{\mathsf{KL}(\mathbf{P} \parallel \mathbb{P}) + \ln \mathbb{E}\exp\bigl(c\,(\|B_t - B_{kh}\| - \mathbb{E}\|B_t - B_{kh}\|)^2\bigr)\right\}.\end{aligned}$$

Under $\mathbb{P}$, $B_t - B_{kh} \sim \mathsf{normal}(0, (t-kh)\,I_d)$. Applying concentration of measure for the Gaussian distribution (see, e.g., Theorem 2.4.8), if $c \lesssim \frac{1}{t-kh}$, then

$$\mathbb{E}\exp\bigl(c\,(\|B_t - B_{kh}\| - \mathbb{E}\|B_t - B_{kh}\|)^2\bigr) \le 2\,.$$

In fact, it suffices to take $c = \frac{1}{8\,(t-kh)}$. Next, using the LSI for π,

$$\begin{aligned}\mathsf{KL}(\mathbf{P} \parallel \mathbb{P}) = \mathbb{E}_{\psi_t\mu_t} \ln \psi_t &= \mathbb{E}_{\psi_t\mu_t} \ln \frac{\rho_t^{q-1}}{\mathbb{E}_{\mu_t}(\rho_t^{q-1})} = \frac{q-1}{q}\,\mathbb{E}_{\psi_t\mu_t} \ln \frac{\rho_t^q}{\mathbb{E}_{\mu_t}(\rho_t^{q-1})^{q/(q-1)}} \\ &= \frac{q-1}{q}\left\{\mathbb{E}_{\psi_t\mu_t} \ln \frac{\rho_t^q}{\mathbb{E}_{\mu_t}(\rho_t^{q-1})} - \underbrace{\frac{1}{q-1}\ln \mathbb{E}_{\mu_t}(\rho_t^{q-1})}_{\ge 0}\right\} \\ &\le \frac{q-1}{q}\,\mathsf{KL}(\psi_t\mu_t \parallel \pi) \\ &\le \frac{(q-1)\,C_{\mathsf{LSI}}}{2q}\,\mathbb{E}_{\psi_t\mu_t}\Bigl[\Bigl\|\nabla \ln\bigl(\psi_t\,\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\bigr)\Bigr\|^2\Bigr] = \frac{2\,(q-1)\,C_{\mathsf{LSI}}}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,,\end{aligned}$$

where we applied the identity (6.1.3). Hence,

$$\begin{aligned}&\mathbb{E}[\psi_t(X_t)\,\|B_t - B_{kh}\|^2] \\ &\quad\le 2d\,(t-kh) + \frac{32 C_{\mathsf{LSI}} h\,(q-1)}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + (16\ln 2)\,(t-kh) \\ &\quad\le 14d\,(t-kh) + 32 C_{\mathsf{LSI}} h\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,.\end{aligned}$$

All in all, applying Proposition 6.1.1 and collecting the error terms,

$$\partial_t \mathcal{R}_q(\mu_t \parallel \pi) \le -\frac{3}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 9\beta^2 q\,(t-kh)^2\left\{\frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2\beta d\right\}$$

$$+ 6\beta^2 q \left\{ 14d\,(t - kh) + 32 C_{\mathsf{LSI}} h \, \frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} \right\}.$$

From $C_{\mathsf{LSI}}, \beta \geq 1$ and $h \leq \frac{1}{192 C_{\mathsf{LSI}} \beta^2 q^2}$, we can absorb some of the error terms into the decay term and apply the LSI for π (see Theorem 2.2.15), yielding

$$\begin{aligned} \partial_t \mathcal{R}_q(\mu_t \parallel \pi) &\leq -\frac{1}{q} \frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + O(\beta^3 d h^2 q + \beta^2 d h q) \\ &\leq -\frac{1}{2qC_{\mathsf{LSI}}} \mathcal{R}_q(\mu_t \parallel \pi) + O(\beta^2 d h q)\,. \end{aligned}$$

This implies the differential inequality

$$\partial_t \Big\{ \exp\Big(\frac{t - kh}{2qC_{\mathsf{LSI}}}\Big) \mathcal{R}_q(\mu_t \parallel \pi) \Big\} \lesssim \exp\Big(\frac{t - kh}{2qC_{\mathsf{LSI}}}\Big) \beta^2 d h q \lesssim \beta^2 d h q\,.$$

Integrating this over $t \in [kh, (k+1)h]$ yields

$$\mathcal{R}_q(\mu_{(k+1)h} \parallel \pi) \leq \exp\Big(-\frac{h}{2qC_{\mathsf{LSI}}}\Big) \mathcal{R}_q(\mu_{kh} \parallel \pi) + O(\beta^2 d h^2 q)\,.$$

Unrolling the recursion,

$$\mathcal{R}_q(\mu_{Nh} \parallel \pi) \leq \exp\Big(-\frac{Nh}{2qC_{\mathsf{LSI}}}\Big) \mathcal{R}_q(\mu_0 \parallel \pi) + O(C_{\mathsf{LSI}} \beta^2 d h q^2)\,. \qquad \square$$

We pause to reflect upon the proof. As discussed above, the key steps are to use change of measure inequalities in order to relate expectations under $\mathbf{P}$ to expectations under $\mathbb{P}$. This is accomplished via the Fisher information duality lemma (Lemma 4.2.5) and the Donsker–Varadhan variational principle (Theorem 1.5.4). These inequalities yield an additional error term of the form $\mathsf{FI}(\psi_t \mu_t \parallel \pi)$ (for the latter, this error term appears after an application of the LSI). The magical part of the calculation is that $\mathsf{FI}(\psi_t \mu_t \parallel \pi)$ is precisely equal to the Rényi Fisher information (up to constants), and when the step size h is sufficiently small it can be absorbed into the decay term of the differential inequality in Proposition 6.1.1.

The proof above implies an iteration complexity whose dependence on q scales as $N = O(q^3)$. In order to improve the dependence on q, we modify the differential inequality of Proposition 6.1.1 by making the parameter q time-dependent, similarly to the hypercontractivity principle (Exercise 2.7). The proof is left as Exercise 6.1.

Proposition 6.1.4 (hypercontractivity). *Suppose that π satisfies a log-Sobolev inequality. Along the law $(\mu_t)_{t\geq 0}$ of the interpolated process (4.2.2), if we define the parameter $q(t) := 1 + (q_0 - 1)\exp\frac{t}{2C_{\mathrm{LSI}}}$, then*

$$\partial_t\Big(\frac{1}{q(t)}\ln\int \rho_t^{q(t)}\,\mathrm{d}\pi\Big) \leq -\frac{2\,(q(t)-1)}{q(t)^2}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q(t)/2})\|^2]}{\mathbb{E}_\pi(\rho_t^{q(t)})} + \big(q(t)-1\big)\,\mathbb{E}[\psi_t(X_t)\,\|\nabla V(X_t) - \nabla V(X_{kh})\|^2]\,,$$

where $\rho_t := \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$ and $\psi_t := \rho_t^{q(t)-1}/\mathbb{E}_\pi(\rho_t^{q(t)})$.

Proof of Theorem 6.1.2 with improved dependence on q. Let $\bar{q} \geq 3$.

Initial waiting phase. We apply hypercontractivity (Proposition 6.1.4) with $q_0 = 2$ and for $t \leq N_0 h$, where $N_0 := \lceil \frac{2C_{\mathrm{LSI}}}{h}\ln(\bar{q}-1)\rceil$. Note that $\bar{q} \leq q(N_0 h) \leq 2\bar{q}$. The bound on the error term from the previous proof yields

$$\partial_t\Big(\frac{1}{q(t)}\ln\int \rho_t^{q(t)}\,\mathrm{d}\pi\Big) \lesssim \beta^2 dh q(t)\,.$$

Integrating this over $t \in [kh, (k+1)h]$ yields

$$\frac{1}{q((k+1)h)}\ln\int \rho_{(k+1)h}^{q((k+1)h)}\,\mathrm{d}\pi - \frac{1}{q(kh)}\ln\int \rho_{kh}^{q(kh)}\,\mathrm{d}\pi \lesssim \beta^2 dh^2\bar{q}\,.$$

Unrolling the recursion yields

$$\frac{1}{q(N_0 h)}\ln\int \rho_{N_0 h}^{q(N_0 h)}\,\mathrm{d}\pi - \frac{1}{2}\ln\int \rho_0^2\,\mathrm{d}\pi \lesssim \beta^2 dh^2\bar{q}N_0 \leq \widetilde{O}(C_{\mathrm{LSI}}\beta^2 dh\bar{q})\,.$$

Finishing the convergence analysis. Next, after shifting time indices and applying the previous proof of Theorem 6.1.2 with $q = 2$,

$$\begin{aligned}\mathcal{R}_{\bar{q}}(\mu_{(N+N_0)h} \,\|\, \pi) &\leq \frac{1}{q(N_0 h)-1}\ln\int \rho_{(N+N_0)h}^{q(N_0 h)}\,\mathrm{d}\pi \leq \frac{3}{4}\ln\int \rho_{Nh}^2\,\mathrm{d}\pi + \widetilde{O}(C_{\mathrm{LSI}}\beta^2 dh\bar{q})\\ &\leq \frac{3}{4}\exp\Big(-\frac{Nh}{4C_{\mathrm{LSI}}}\Big)\,\mathcal{R}_2(\mu_0 \,\|\, \pi) + \widetilde{O}(C_{\mathrm{LSI}}\beta^2 dh\bar{q})\,.\end{aligned}$$

This proves the desired result. □

The proof of Theorem 6.1.2 is rather specific to the LSI case because we use the LSI to bound the KL term $\mathsf{KL}(\psi_t \mu_t \,\|\, \pi)$ via the Rényi Fisher information, which is then absorbed into the differential inequality of Proposition 6.1.1. However, it turns out that rather than assuming an LSI for π, it suffices to have an LSI for μ_t for all $t \geq 0$ (possibly with an LSI constant that grows with t). One situation in which this holds is when we initialize LMC with a measure μ_0 that satisfies an LSI, and the potential V is *convex*. Note that this situation is not included in the case when π satisfies an LSI, because V may only have linear growth at infinity (whereas from Theorem 2.4.8, if $\pi \propto \exp(-V)$ satisfies an LSI, then V necessarily has quadratic growth at infinity). We explore this in Exercise 6.2.

Bibliographical Notes

Discretization of LMC in Rényi divergence was first considered in [VW19], which proved convergence of LMC to its biased stationary distribution. However, this does not lead to a sampling guarantee unless the size of the "Rényi bias" (the Rényi divergence between the biased stationary distribution and the true target distribution) can be estimated.

Motivated by applications to differential privacy, [GT20] provided the first Rényi sampling guarantees for LMC under strong log-concavity by using a technique based on the adaptive composition lemma for Rényi divergences. Then, [EHZ22] improved the analysis of [GT20] via a two-phase analysis that weakens the assumption of strong log-concavity to a dissipativity assumption and obtains a sharper bound, but which still relies on the adaptive composition lemma. Subsequently, building off the earlier work of [Che+21b] (which essentially does a one-step discretization argument in Rényi divergence), in [Che+21a] it was realized that the earlier arguments of [GT20; EHZ22] can be streamlined by replacing the adaptive composition lemma entirely with Girsanov's theorem. The proofs of Theorem 6.1.2 and Exercise 6.2 are also from [Che+21a].

Exercises

Proof under LSI via Interpolation Argument

▷ **Exercise 6.1** (Rényi differential inequality)
Prove Proposition 6.1.1 and Proposition 6.1.4.

▷ **Exercise 6.2** (Rényi discretization bound for log-concave targets)
Suppose that $\pi \propto \exp(-V)$ is log-concave, that $\nabla V(0) = 0$, and that ∇V is β-Lipschitz. Also, suppose that we initialize LMC at $\mu_0 = \mathsf{normal}(0, \beta^{-1} I_d)$. The goal of this exercise is to prove a Rényi discretization bound for LMC under these assumptions.

1. First, show that μ_t satisfies an LSI for all $t \geq 0$, and write down a bound for $C_{\mathsf{LSI}}(\mu_t)$ (the bound should grow linearly with t).

   Hint: See Section 2.3.

2. Follow the proof of Theorem 6.1.2 (the first proof which incurs a suboptimal dependence on q). Note that in Theorem 6.1.2, we bounded $\mathsf{KL}(\mathbf{P} \,\|\, \mathbb{P}) \leq \frac{q-1}{q} \mathsf{KL}(\psi_t \mu_t \,\|\, \pi)$ and we applied the LSI for π. This time, use $\mathsf{KL}(\mathbf{P} \,\|\, \mathbb{P}) = \mathsf{KL}(\psi_t \mu_t \,\|\, \mu_t)$ and apply the LSI for μ_t instead.

   Also, instead of using the decay of the Rényi divergence under a LSI, use the decay of the Rényi divergence under a PI (Theorem 2.2.15). (Since π is log-concave, it necessarily satisfies a Poincaré inequality with some constant C_{PI}, see the Bibliographical Notes to Chapter 2.)

   Prove that if $\varepsilon \leq \sqrt{1/q} \wedge \sqrt{C_{\mathsf{PI}} d/\beta}$, then with an appropriate choice of step size h and with $N = \widetilde{\Theta}(C_{\mathsf{PI}}^2 \beta^2 d^2 q^3/\varepsilon^2)$ iterations of LMC, we obtain $\sqrt{\mathcal{R}_q(\mu_{Nh} \,\|\, \pi)} \leq \varepsilon$. (Unlike the guarantee of Theorem 6.1.2, here the guarantee does not allow N to be too large, due to the growing LSI constant of the iterates.)

CHAPTER 7

High-Accuracy Samplers

So far, we have focused on discretizations of diffusions. Discretization of a continuous-time Markov process yields a discrete-time Markov chain whose stationary distribution is no longer equal to the target π; the algorithm is *biased*. Nevertheless, we showed that the size of the bias can be made smaller than any desired accuracy ε by choosing a small step size h, which then leads to quantitative sampling guarantees.

However, the number of iterations of the algorithm is proportional to the inverse step size $1/h$, and consequently the complexity of the algorithms scaled as $\mathrm{poly}(1/\varepsilon)$. In this section, we address the problem of designing *high-accuracy* samplers, i.e., samplers whose complexity scales as $\mathrm{polylog}(1/\varepsilon)$. To accomplish this, we must fix the bias of the sampling algorithm, which is accomplished via the Metropolis–Hastings filter.

7.1 Rejection Sampling

Before introducing the Metropolis–Hastings filter, we begin with a warm up and introduce the concept of rejection via the **rejection sampling** algorithm.

Rejection Sampling: Let π be the target distribution and let $\widetilde{\pi}$ be an unnormalized version of π, i.e., $\widetilde{\pi} \propto \pi$. Suppose we can sample from a distribution μ and that an unnormalized version $\widetilde{\mu}$ of μ satisfies $\widetilde{\mu} \geq \widetilde{\pi}$. Then, repeat until acceptance:

1. Draw $X \sim \mu$.
2. Accept X with probability $\widetilde{\pi}(X)/\widetilde{\mu}(X)$.

Unlike the other sampling algorithms we have considered thus far, rejection sampling always terminates with an *exact* sample from π.

Theorem 7.1.1. *The output of rejection sampling is a sample drawn exactly from π. Also, the number of samples drawn from μ until a sample is accepted follows a geometric distribution with mean Z_μ/Z_π, where $Z_\mu \coloneqq \int \widetilde{\mu}$ and $Z_\pi \coloneqq \int \widetilde{\pi}$.*

Proof. The probability of acceptance is

$$\mathbb{P}(\text{acceptance}) = \int \frac{\widetilde{\pi}}{\widetilde{\mu}} \, \mathrm{d}\mu = \frac{Z_\pi}{Z_\mu} \int \frac{\pi}{\mu} \, \mathrm{d}\mu = \frac{Z_\pi}{Z_\mu}$$

and clearly the number of samples drawn until acceptance is geometrically distributed.

To show that the output X of rejection sampling is drawn exactly according to π, let $(U_i)_{i=1}^\infty \overset{\text{i.i.d.}}{\sim} \mathsf{uniform}[0,1]$ and $(X_i)_{i=1}^\infty \overset{\text{i.i.d.}}{\sim} \mu$ be independent. Then, for any event A,

$$\begin{aligned}
\mathbb{P}(X \in A) &= \sum_{n=0}^\infty \mathbb{P}\Big(X_{n+1} \in A,\ U_i > \frac{\widetilde{\pi}(X_i)}{\widetilde{\mu}(X_i)}\ \forall i \in [n],\ U_{n+1} \le \frac{\widetilde{\pi}(X_{n+1})}{\widetilde{\mu}(X_{n+1})}\Big) \\
&= \sum_{n=0}^\infty \mathbb{P}\Big(X_{n+1} \in A,\ U_{n+1} \le \frac{\widetilde{\pi}(X_{n+1})}{\widetilde{\mu}(X_{n+1})}\Big)\, \mathbb{P}\Big(U_1 > \frac{\widetilde{\pi}(X_1)}{\widetilde{\mu}(X_1)}\Big)^n \\
&= \sum_{n=0}^\infty \Big(\int_A \frac{\widetilde{\pi}}{\widetilde{\mu}} \, \mathrm{d}\mu\Big) \Big(\int \big(1 - \frac{\widetilde{\pi}}{\widetilde{\mu}}\big) \, \mathrm{d}\mu\Big)^n = \frac{Z_\pi}{Z_\mu}\, \pi(A) \sum_{n=0}^\infty \Big(1 - \frac{Z_\pi}{Z_\mu}\Big)^n = \pi(A)\,. \qquad \square
\end{aligned}$$

The rejection sampling algorithm requires the construction of the upper envelope $\widetilde{\mu}$. We now demonstrate how to construct this envelope for our usual class of distributions. Namely, suppose that $\pi \propto \exp(-V)$ satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$, and that $\nabla V(0) = 0$. We can assume that our unnormalized version $\widetilde{\pi} = \exp(-V)$ of π satisfies $V(0) = 0$ (if not, replace V by $V - V(0)$). Then, by strong convexity of V, we see that $\widetilde{\pi} \le \exp(-\frac{\alpha}{2}\|\cdot\|^2)$, and we can take $\widetilde{\mu} \coloneqq \exp(-\frac{\alpha}{2}\|\cdot\|^2)$, which means that the normalized distribution is $\mu = \mathsf{normal}(0, \alpha^{-1} I_d)$. To understand the efficiency of rejection sampling, we need to bound the ratio Z_μ/Z_π of normalizing constants. By smoothness of V,

$$\frac{Z_\mu}{Z_\pi} = \frac{(2\pi/\alpha)^{d/2}}{\int \exp(-V)} \le \frac{(2\pi/\alpha)^{d/2}}{\int \exp(-\frac{\beta}{2}\|\cdot\|^2)} = \frac{(2\pi/\alpha)^{d/2}}{(2\pi/\beta)^{d/2}} = \kappa^{d/2},$$

with $\kappa \coloneqq \beta/\alpha$. We summarize this result in the following proposition.

Proposition 7.1.2. *Let the target $\pi \propto \exp(-V)$ on $\mathbb{R}^d$ satisfy $0 < \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$, $V(0) = 0$, and $\nabla V(0) = 0$. Then, rejection sampling with the envelope $\widetilde{\mu} := \exp(-\frac{\alpha}{2}\|\cdot\|^2)$ returns an exact sample from π with a number of iterations that is a geometric random variable with mean at most $\kappa^{d/2}$, where $\kappa := \beta/\alpha$.*

The rejection sampling guarantee can be formulated in one of two ways. We can think of the algorithm as returning an exact sample from π, with a random number of iterations (the number of iterations is geometrically distributed). Alternatively, if we place an upper bound N on the number of iterations of the algorithm and output "FAIL" if we have not terminated by iteration N, then the probability of "FAIL" is at most $\varepsilon := (1 - 1/\kappa^{d/2})^N$, and if μ_N denotes the law of the output of the algorithm, then $\|\mu_N - \pi\|_{\mathrm{TV}} \leq \varepsilon$. If we flip this around and fix the target accuracy ε, we see that the number of iterations required to achieve this guarantee is $N \geq \kappa^{d/2} \ln(1/\varepsilon)$.

Although this result is acceptable in low dimension, the complexity of this approach quickly becomes intractable even for moderately high-dimensional problems. In the next section, we will see that by combining the idea of rejection with *local* proposals, we can obtain tractable sampling algorithms in high dimension.

7.2 The Metropolis–Hastings Filter

A Metropolis–Hastings algorithm consists of *proposing* moves from a proposal kernel Q, and then *accepting or rejecting* each move with a carefully chosen probability which ensures that the resulting Markov chain has the desired stationary distribution π.

In more detail, let Q be a kernel on $\mathbb{R}^d \times \mathbb{R}^d$, that is: for each $x \in \mathbb{R}^d$, $Q(x, \cdot)$ is a probability measure on $\mathbb{R}^d$. We will mostly consider proposals such that each $Q(x, \cdot)$ has a density with respect to Lebesgue measure, and via an abuse of notation we will write $Q(x, y)$ for this density evaluated at y (an exception is when we consider MHMC below).

Starting from $X \in \mathbb{R}^d$, we tentatively propose a new point $Y \sim Q(X, \cdot)$. We then accept the point Y with probability $A(X, Y)$ (called the *acceptance probability*); otherwise, we stay at the old point X. Iterate this process until convergence.

There are different possible choices for the acceptance probability A, but the choice we consider here is the **Metropolis–Hastings filter**

$$A(x, y) := 1 \wedge \frac{\pi(y)\, Q(y, x)}{\pi(x)\, Q(x, y)}\,. \tag{7.2.1}$$

The overall algorithm is summarized as follows.

Metropolis–Hastings algorithm (with proposal Q): initialize at a point $X_0 \in \mathbb{R}^d$. Then, iterate the following steps for $k = 1, 2, 3, \ldots$:

1. Propose a new point $Y_k \sim Q(X_{k-1}, \cdot)$.
2. With probability $A(X_{k-1}, Y_k)$, set $X_k := Y_k$; otherwise, set $X_k := X_{k-1}$. Here, A is the acceptance probability defined via (7.2.1).

This algorithm defines a discrete-time Markov chain whose transition kernel T can be written explicitly as

$$T(x, \mathrm{d}y) = Q(x, \mathrm{d}y)\, A(x, y) + \underbrace{\Big(1 - \int Q(x, \mathrm{d}y')\, A(x, y')\Big)}_{\text{rejection probability}} \delta_x(\mathrm{d}y)\,. \tag{7.2.2}$$

A discrete-time Markov chain with transition kernel P is called **reversible** with respect to π if it holds that $\pi(\mathrm{d}x)\, P(x, \mathrm{d}y) = \pi(\mathrm{d}y)\, P(y, \mathrm{d}x)$. Similarly to our discussion in Section 1.2, discrete-time reversible Markov chains can be studied via spectral theory.

Theorem 7.2.3. *The Metropolis–Hastings algorithm with proposal Q is reversible with respect to π.*

Proof. We want to check that $\pi(x)\, T(x, y) = \pi(y)\, T(y, x)$ for all $x, y \in \mathbb{R}^d$ with $x \neq y$. We can write

$$\begin{aligned} \pi(x)\, T(x, y) &= \pi(x)\, Q(x, y)\, A(x, y) = \pi(x)\, Q(x, y) \min\Big\{1, \frac{\pi(y)\, Q(y, x)}{\pi(x)\, Q(x, y)}\Big\} \\ &= \min\{\pi(x)\, Q(x, y), \pi(y)\, Q(y, x)\} \end{aligned}$$

and this expression is symmetric in x and y. □

Take note of the flexibility of the Metropolis–Hastings algorithm! Regardless of the choice of proposal kernel Q, the filter always makes the algorithm unbiased. Of course, the choice of Q will be crucial later in order to guarantee rapid convergence to stationarity.

Implementability of the Metropolis–Hastings algorithm. To implement the algorithm, the proposal Q must be simple enough such that (1) we can sample from $Q(x, \cdot)$ easily, and (2) we can compute the density $Q(x, y)$ easily (which is required to compute the acceptance probability). Note that although the target density π appears in the expression (7.2.1) for the acceptance probability, it only appears as a *ratio*, and in particular we

do not need to know the normalization constant of π. Hence, the Metropolis–Hastings filter can be implemented using queries to the density of π up to normalization, which are "zeroth-order queries" (unlike, e.g., LMC, which uses first-order information through queries to the gradient ∇V).

Metropolis–Hastings as a projection. There is a nice geometric interpretation of the Metropolis–Hastings filter as a projection, due to [BD01]. Given a proposal kernel Q, let $T(Q)$ denote the Metropolis–Hastings kernel obtained from Q (see (7.2.2)). Then, the mapping $Q \mapsto T(Q)$ is a projection of the proposal kernel Q onto the space of reversible Markov chains with stationary distribution π with respect to an L^1 notion of distance.

The distance is defined as follows:

$$\mathrm{d}(T, T') := \int_{(\mathbb{R}^d \times \mathbb{R}^d) \setminus \mathrm{diag}} |T(x,y) - T'(x,y)| \, \pi(\mathrm{d}x) \, \mathrm{d}y \tag{7.2.4}$$

where $\mathrm{diag} := \{(x,x) \mid x \in \mathbb{R}^d\}$ is the diagonal in $\mathbb{R}^d \times \mathbb{R}^d$.

Theorem 7.2.5 ([BD01]). *Let $\mathscr{R}(\pi)$ denote the space of kernels T which are reversible with respect to π, and such that for each $x \in \mathbb{R}^d$, $T(x,\cdot)$ admits a density with respect to Lebesgue measure (except possibly having an atom at x). Then,*

$$T(Q) \in \operatorname*{arg\,min}_{T \in \mathscr{R}(\pi)} \mathrm{d}(Q, T) \, .$$

Proof. Let $T \in \mathscr{R}(\pi)$, and let $S := \{(x,y) \in \mathbb{R}^d \mid \pi(x)\, Q(x,y) > \pi(y)\, Q(y,x)\}$. Then,

$$\begin{aligned}
\mathrm{d}(Q,T) &= \int_{(\mathbb{R}^d \times \mathbb{R}^d) \setminus \mathrm{diag}} |Q(x,y) - T(x,y)| \, \pi(\mathrm{d}x) \, \mathrm{d}y \\
&= \int_S |Q(x,y) - T(x,y)| \, \pi(\mathrm{d}x) \, \mathrm{d}y + \int_{(\mathbb{R}^d \times \mathbb{R}^d) \setminus (S \cup \mathrm{diag})} |Q(x,y) - T(x,y)| \, \pi(\mathrm{d}x) \, \mathrm{d}y \\
&= \int_S |Q(x,y) - T(x,y)| \, \pi(\mathrm{d}x) \, \mathrm{d}y + \int_S |Q(y,x) - T(y,x)| \, \pi(\mathrm{d}y) \, \mathrm{d}x \, .
\end{aligned}$$

Using reversibility of T, the second term is

$$\begin{aligned}
\int_S |Q(y,x) - T(y,x)| \, \pi(\mathrm{d}y) \, \mathrm{d}x &= \int_S |\pi(x)\, Q(y,x) - \pi(x)\, T(x,y)| \, \mathrm{d}x \, \mathrm{d}y \\
&\geq \int_S |\pi(x)\, Q(x,y) - \pi(y)\, Q(y,x)| \, \mathrm{d}x \, \mathrm{d}y - \int_S |Q(x,y) - T(x,y)| \, \pi(\mathrm{d}x) \, \mathrm{d}y \, .
\end{aligned} \tag{7.2.6}$$

Putting this together, $\mathrm{d}(Q,T) \geq \int_S |\pi(x)\,Q(x,y) - \pi(y)\,Q(y,x)|\,\mathrm{d}x\,\mathrm{d}y$, so we have obtained a lower bound which does not depend on T. On the other hand, we can check that the only inequality (7.2.6) that we used is an equality for $T = T(Q)$. □

7.3 An Overview of High-Accuracy Samplers

As we already discussed, the Metropolis–Hastings framework is quite flexible: by instantiating it with different choices for the proposal Q, we obtain several different algorithms.

Metropolized random walk (MRW). Perhaps the simplest proposal is to simply take $Q(x,\cdot) = \mathsf{normal}(x, h\,I_d)$, which yields the **Metropolized random walk (MRW)** algorithm. This corresponds to simply taking a random walk around the state space, where some steps are occasionally rejected. Note that since the proposal is independent of the target π, the overall algorithm only uses queries to the density of π up to normalization (to implement the filter); thus, it is the only algorithm we have discussed so far (besides rejection sampling) which uses only a zeroth-order oracle for the potential V.

Metropolis-adjusted Langevin algorithm (MALA). A better choice of proposal is

$$Q(x,\cdot) = \mathsf{normal}\big(x - h\,\nabla V(x), 2h\,I_d\big)$$

which is simply one step of the LMC algorithm; this yields the **Metropolis-adjusted Langevin algorithm (MALA)**. We will carefully study the convergence guarantees for MALA in this chapter.

Metropolized Hamiltonian Monte Carlo (MHMC). Recall the Hamiltonian Monte Carlo (HMC) algorithm that we introduced in Section 5.2. The ideal HMC algorithm is not implementable because it requires the ability to exactly integrate Hamilton's equations, and this is generally not possible outside of a few special cases.

We now consider approximately implementing Hamilton's equations through the use of a numerical integrator. Although several choices are available, for Hamilton's equations it is preferable to use a **symplectic integrator**.[1] We will focus on the simplest and most well-known such integrator, called the **leapfrog integrator**.

[1]When placed within the framework of geometry, Hamiltonian mechanics is encoded via *symplectic geometry*, which is the study of manifolds equipped with a symplectic 2-form. The flow map for Hamilton's equations preserves this symplectic form, and is therefore known as a *symplectomorphism*. Symplectic integrators are special integrators which also preserve the symplectic form. This property leads to stability, especially for long integration times.

Leapfrog Integrator: Pick a step size $h > 0$ and a total number of iterations K, corresponding to the total integration time via $T = Kh$. Let (x_0, p_0) be the initial point. For $k = 0, 1, 2, \ldots, K-1$:

1. Set $p_{(k+\frac{1}{2})h} \coloneqq p_{kh} - \frac{h}{2} \nabla V(x_{kh})$.
2. Set $x_{(k+1)h} \coloneqq x_{kh} + h\, p_{(k+\frac{1}{2})h}$.
3. Set $p_{(k+1)h} \coloneqq p_{(k+\frac{1}{2})h} - \frac{h}{2} \nabla V(x_{(k+1)h})$.

Once we apply the leapfrog integrator to HMC, we obtain a discrete-time sampling algorithm which is once again biased. We then correct the bias through the use of the Metropolis–Hastings filter. Specifically, for an integration time $T = Kh$, let

$$F_{\text{leap}}(x, p) = \text{output } x_T \text{ of the leapfrog integrator with } K \text{ steps,} \\ \text{started at } (x, p)\,.$$

Remarkably, the acceptance probability can be computed in closed form, and this relies on specific properties of the leapfrog integrator. The full algorithm is summarized as follows.

Metropolized Hamiltonian Monte Carlo (MHMC): Initialize at $X_0 \sim \mu_0$. For iterations $k = 0, 1, 2, \ldots$:

1. *Refresh* the momentum: draw $P_k \sim \text{normal}(0, I_d)$.
2. *Propose* a trajectory: let $(X'_k, P'_k) \coloneqq F_{\text{leap}}(X_k, P_k)$.
3. *Accept* the trajectory with probability $1 \wedge \exp\{H(X_k, P_k) - H(X'_k, P'_k)\}$. If the trajectory is accepted, set $X_{k+1} \coloneqq X'_k$; otherwise, we set $X_{k+1} \coloneqq X_k$.

It turns out that when $K = 1$, the MHMC algorithm reduces to MALA (Exercise 7.1).

We next justify why the MHMC algorithm leaves $\boldsymbol{\pi}$ invariant. Actually, although we have written down the MHMC algorithm in the form which is easiest to implement, it obscures the underlying structure of the algorithm. The proof of the next theorem will clarify this point.

Theorem 7.3.1. *The augmented target distribution $\boldsymbol{\pi} \propto \exp(-H)$ is invariant for the MHMC algorithm.*

Proof. First, we note that the step of refreshing the momentum leaves $\boldsymbol{\pi}$ invariant, so it suffices to study the proposal and acceptance steps.

For the moment, let us pretend that the proposal actually uses the true flow map F_T (which exactly integrates Hamilton's equations for time T) rather than the leapfrog integrator F_{leap}. Then, the proposal kernel is deterministic, $Q((x,p),\cdot) = \delta_{F_T(x,p)}$. Up until now, we have been assuming that the proposal kernel admits a density w.r.t. Lebesgue measure, which certainly does not hold here, but we will brush over this technicality as it is not the key point here.

If we naïvely apply the Metropolis–Hastings filter, the probability of accepting a proposal (x',p') starting from (x,p) involves a ratio $Q((x',p'),(x,p))/Q((x,p),(x',p'))$, but this ratio is ill-defined in our setting. The problem is that if $(x',p') = F_T(x,p)$, then it is not the case that $(x,p) = F_T(x',p')$; the proposal is not reversible. Hence, we would be led to reject every single trajectory.

To fix this, recall from Exercise 5.4 that we have the following time reversibility property: if R denotes the momentum flip operator $(x,p) \mapsto (x,-p)$, then it holds that $F_T^{-1} = R \circ F_T \circ R$. It implies that $F_T \circ R = R \circ F_T^{-1} = (F_T \circ R)^{-1}$, so $F_T \circ R$ is idempotent. In other words, if we use the proposal $F_T \circ R$ (i.e., first flip the momentum before integrating Hamilton's equations), then the proposal would be reversible and the above issue does not arise, as the ratio $Q((x',p'),(x,p))/Q((x,p),(x',p'))$ would equal 1. Observe also that using $F_T \circ R$ instead of F_T does not change the algorithm since we refresh the momentum at each step (and if $P_k \sim \text{normal}(0, I_d)$, then $-P_k \sim \text{normal}(0, I_d)$ as well).

Once we use the proposal $(x',p') = (F_T \circ R)(x,p) = F_T(x,-p)$, the Metropolis–Hastings acceptance probability is calculated to be

$$1 \wedge \frac{\boldsymbol{\pi}(x',p')}{\boldsymbol{\pi}(x,p)} = 1 \wedge \exp\{H(x,p) - H(x',p')\}\,. \tag{7.3.2}$$

When we use the exact flow map F_T, then the Hamiltonian is conserved (Exercise 5.4) so the above probability is one; every trajectory is accepted. However, the above expression is indeed meaningful if we instead use the leapfrog integrator F_{leap}.

So far, we have motivated the expression (7.3.2) based on the exact flow map F_T, but clearly the above argument holds just as well for the leapfrog integrator F_{leap} as soon as we verify the property $F_{\text{leap}}^{-1} = R \circ F_{\text{leap}} \circ R$, and this is where we use the specific form of the leapfrog integrator. We leave the verification as Exercise 7.2. □

***Remark* 7.3.3.** The proof shows that the proposal of MHMC should really be thought of as $F_{\text{leap}} \circ R$, instead of F_{leap}. In fact, if we did not refresh the momentum, then repeatedly applying the idempotent operator $F_{\text{leap}} \circ R$ would just cause the algorithm to jump back and forth between two points (x,p) and (x',p'), which is silly; hence one should also apply another momentum flip after the filter. In symbols, if MH denotes the Metropolis–Hastings filter step, and Refresh denotes the momentum refreshment step, we should

think of MHMC as the composition

$$\mathsf{MHMC} = R \circ \mathsf{MH}(F_{\text{leap}} \circ R) \circ \mathsf{Refresh}\,.$$

This is simplified to

$$\mathsf{MHMC} = \mathsf{MH}(F_{\text{leap}} \circ R) \circ \mathsf{Refresh}\,.$$

because $\mathsf{Refresh} \circ R = \mathsf{Refresh}$.

Lazy chains. Technically, many of the convergence results actually hold for *lazy* versions of the Markov chain. Specifically, for $\ell \in [0,1]$, the **ℓ-lazy** version of a Markov chain replaces its transition kernel T with the modified kernel T_ℓ given by

$$T_\ell(x, \mathrm{d}y) = (1-\ell)\, T(x, \mathrm{d}y) + \ell\, \delta_x(\mathrm{d}y)\,.$$

The laziness condition is familiar from the study of discrete-time Markov chains on discrete state spaces, in which laziness is useful for avoiding periodic behavior. For the remainder of this section, we will generally be considering $\frac{1}{2}$-lazy versions of the Metropolis–Hastings chains without explicitly mentioning this. In any case, this modification only multiplies the mixing time by a factor of 2, so it does not significantly alter the results.

Feasible start vs. warm start. When discussing Metropolis–Hastings algorithms, we must distinguish between convergence rates when initialized at a **feasible start**, vs. a **warm start**. These terms are not precisely defined, but loosely speaking a feasible start refers to an easily computable distribution which works well uniformly over the class of target distributions under consideration. In this section, a feasible start usually refers to the $\mathsf{normal}(0, \beta^{-1} I_d)$ distribution, where β is the smoothness of V and we assume that the minimizer of V is 0. On the other hand, a *warm start* is a distribution which is already somewhat close to the target π; for this section, it can be taken to mean a distribution μ_0 such that $\chi^2(\mu_0 \,\|\, \pi) = O(1)$. Unsurprisingly, the rates are faster with a warm start.

The situation at hand is similar to the discussion in Section 1.5. Basically, the simplest way to study a Metropolis–Hastings chain is via spectral theory, which is related to Poincaré inequalities and hence to the chi-squared divergence at initialization. We also know that Poincaré inequalities tend to yield poor convergence guarantees in continuous time, which can be remedied via stronger inequalities (such as a log-Sobolev inequality). To an extent, this is also possible for Metropolis–Hastings algorithms. However, it is a fairly recent[2] finding that for MALA there is an *intrinsic* and substantial difference in

[2]The phenomenon described here is anticipated, at least qualitatively, from older work on Markov chains. The recent part of this story is the quantitative study of this effect in the context of MALA.

convergence rates for the feasible and warm start cases, even under the assumption of strong log-concavity. This is unlike the case of LMC; e.g., the guarantee of Theorem 4.2.6 is not significantly improved by assuming $\mathsf{KL}(\mu_0 \,\|\, \pi) = O(1)$.

State-of-the-art results. We now give the current state-of-the-art convergence guarantees for the Metropolis–Hastings algorithms that we have introduced.

Theorem 7.3.4 (feasible start case, [Dwi+19; Che+20a; LST20]). *Suppose that the target $\pi \propto \exp(-V)$ satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$. Consider the following Metropolis–Hastings algorithms initialized at $\mathsf{normal}(0, \beta^{-1} I_d)$ and with an appropriately tuned choice of parameters.*

1. *MRW outputs a measure μ_N satisfying $\sqrt{\chi^2(\mu_N \,\|\, \pi)} \le \varepsilon$ after*

$$N = \widetilde{O}\big(\kappa^2 d \operatorname{polylog} \frac{1}{\varepsilon}\big) \qquad \textit{iterations}.$$

2. *MALA outputs a measure μ_N satisfying $\sqrt{\chi^2(\mu_N \,\|\, \pi)} \le \varepsilon$ after*

$$N = \widetilde{O}\big(\kappa d \operatorname{polylog} \frac{1}{\varepsilon}\big) \qquad \textit{iterations}.$$

3. *Assume in addition that $\nabla^3 V$ is bounded and that $\kappa \ll \sqrt{d}$. Then, MHMC outputs a measure μ_N satisfying $\sqrt{\chi^2(\mu_N \,\|\, \pi)} \le \varepsilon$ after*

$$N = \widetilde{O}\big(\kappa^{3/4} d \operatorname{polylog} \frac{1}{\varepsilon}\big) \qquad \textit{gradient queries}.$$

Note that the result for MHMC is not directly comparable because it makes a stronger second-order smoothness assumption.

Next, we present the results under a warm start.

Theorem 7.3.5 (warm start case, [Che+21b; WSC21]). *Suppose that $\pi \propto \exp(-V)$ satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Consider MALA initialized at a distribution satisfying $\chi^2(\mu_0 \,\|\, \pi) = O(1)$. Then, MALA outputs a measure μ_N satisfying $\sqrt{\mathsf{KL}(\mu_N \,\|\, \pi)} \le \varepsilon$ (or*

$\sqrt{\mathcal{R}_q(\mu_N \,\|\, \pi)} \le \varepsilon$ *for any* $1 \le q < 2$*) after*

$$N = O\big(\kappa d^{1/2} \operatorname{polylog} \frac{1}{\varepsilon}\big) \qquad \textit{iterations}\,.$$

Moreover, it is known that the results for MALA in both the feasible and warm start cases are sharp in a suitable sense. The goal for the rest of the chapter is to prove these MALA convergence results (up to some technical details).

7.4 Markov Chains in Discrete Time

As discussed in the introduction to this chapter, the key advantage of Metropolis–Hastings algorithms is that they are *unbiased* and hence lead to high-accuracy algorithms. In order to prove complexity bounds that scale as $\operatorname{polylog}(1/\varepsilon)$, where ε is the target accuracy, it is important that we do not simply bound the distance between MALA and, e.g., the continuous-time Langevin diffusion, as we did in Chapter 4. This is not to say that tools from Chapter 4 are completely irrelevant, only that we must first develop some new techniques for studying discrete-time Markov chains.

7.4.1 Markov Semigroup Theory

Let P be a Markov kernel. It generates a discrete-time semigroup $(P^k)_{k\in\mathbb{N}}$, and some of the ideas from Markov semigroup theory (Section 1.2) can be adapted to the present context.

Generator. We define the **generator** of the semigroup to be the operator $\mathscr{L} := P - \mathrm{id}$, acting on $L^2(\pi)$ via $Pf(x) := \int f(y)\, P(x, \mathrm{d}y)$ (where π is the stationary distribution for P). Note that since the operator norm of P is at most 1, then $P - \mathrm{id}$ is always a negative operator (similarly to the infinitesimal generator $\mathscr{L}$ from Section 1.2).

Reversibility. We defined reversibility in Section 7.2 and showed that Metropolis–Hastings algorithms are reversible w.r.t. the target distribution π. For the rest of the section, we will focus on reversible Markov chains.

Spectral gap. The **spectral gap** of P is the largest $\lambda > 0$ such that for all $f \in L^2(\pi)$ with $\mathbb{E}_\pi f = 0$,

$$\langle f, (-\mathscr{L})f\rangle_{L^2(\pi)} \ge \lambda \|f\|^2_{L^2(\pi)}\,.$$

Equivalently, if (X_0, X_1) are two successive iterates of the chain started at stationarity, it is equivalent to require

$$2\lambda \operatorname{var} f(X_0) \le \mathbb{E}[|f(X_1) - f(X_0)|^2] \,. \tag{7.4.1}$$

In analogy with Section 1.2, we also say that P satisfies a **Poincaré inequality** with constant $1/\lambda$. We already saw in Section 2.7 that a Poincaré inequality is implied by a lower bound on the coarse Ricci curvature.

The right-hand side of (7.4.1) can be interpreted as a **Dirichlet energy**,

$$\mathscr{E}(f, f) := \langle f, (-\mathscr{L})f \rangle_{L^2(\pi)} \,,$$

and the Markov chain can be viewed as an $L^2(\pi)$ gradient descent on the Dirichlet energy; see Exercise 7.3. We have the following convergence result.

Theorem 7.4.2. *Suppose that the spectral gap of P is $\lambda > 0$. Then, for the law $(\mu_k)_{k\in\mathbb{N}}$ of the iterates of the $\frac{1}{2}$-lazy version of P, we have*

$$\chi^2(\mu_N \,\|\, \pi) \le \exp(-\lambda N)\, \chi^2(\mu_0 \,\|\, \pi) \,.$$

Modified log-Sobolev inequality. We say that P satisfies a **modified log-Sobolev inequality** (**MLSI**) with constant C_{MLSI} if for all $f \in L^2(\pi)$ with $f \ge 0$,

$$\operatorname{ent}_\pi f \le \frac{C_{\mathrm{MLSI}}}{2} \mathscr{E}(f, \ln f) \,.$$

We have already encountered this inequality as Definition 1.2.24, although there we simply called it the log-Sobolev inequality. In the context of discrete Markov processes, however, since the chain rule fails and the different variants of the log-Sobolev inequality are no longer equivalent, it is worth being careful about the terminology.

It is trickier to deduce entropy decay from the MLSI in discrete time, and to avoid this issue we shall work in continuous time instead. The Markov kernel P gives rise to the generator $\mathscr{L} := P - \mathrm{id}$, which in turn generates a *continuous-time* semigroup $(P_t)_{t\ge 0}$ via $P_t := \exp(t\mathscr{L})$. Note that the generator of $(P_t)_{t\ge 0}$ is also $\mathscr{L}$ and hence the Dirichlet energy for $(P_t)_{t\ge 0}$ coincides with the Dirichlet energy for $(P^k)_{k\in\mathbb{N}}$. Now, if we apply the calculation (1.2.23) to the semigroup $(P_t)_{t\ge 0}$, we find that under an MLSI,

$$\mathsf{KL}(\mu P_t \,\|\, \pi) \le \exp\Big(-\frac{2t}{C_{\mathrm{MLSI}}}\Big)\, \mathsf{KL}(\mu \,\|\, \pi) \,,$$

see Theorem 1.2.25.

Moreover, the continuous-time semigroup $(P_t)_{t\geq 0}$ can be simulated. Namely, let $(\tau_k)_{k\in\mathbb{N}^+} \overset{\text{i.i.d.}}{\sim} \text{exponential}(1)$, $T_k := \sum_{j=1}^{k} \tau_j$, and consider the following algorithm. Initialize at $X_0 \sim \mu_0$, and for $k = 0, 1, 2, \ldots$, let $X_{T_{k+1}} \sim P(X_{T_k}, \cdot)$, so that $(X_{T_k})_{k\in\mathbb{N}}$ are the iterates of the discrete-time Markov chain with kernel P. Also, for $t \geq 0$, if $T_k \leq t < T_{k+1}$, then set $X_t := X_{T_k}$. This yields a continuous-time Markov process $(X_t)_{t\geq 0}$, and one can check that the associated Markov semigroup is exactly $(P_t)_{t\geq 0}$. Moreover, by concentration of i.i.d. sums, it holds that $T_k \approx k$, so that if the semigroup $(P_t)_{t\geq 0}$ requires time T_{mix} in order to mix to a desired level of accuracy, then the algorithm which simulates $(X_t)_{t\geq 0}$ requires $\approx T_{\text{mix}}$ iterations to reach the same level of mixing.

This argument can even be made rigorous, using concentration inequalities for the Poisson random variable, in order to argue that a MLSI for P implies a mixing time bound (in total variation distance, say) for the discrete-time chain $(P^k)_{k\in\mathbb{N}}$. We omit the details and content ourselves with the knowledge that an MLSI for P at least leads to the existence of an implementable algorithm (simulating $(X_t)_{t\geq 0}$) with good mixing.

We leave the converse implication (that entropy decay for the discrete-time Markov chain generated by P implies an MLSI for P) as Exercise 7.4.

7.4.2 Conductance

Unfortunately, it is usually quite challenging to prove either a Poincaré inequality or a modified log-Sobolev inequality for discrete-time Markov chains, which motivates the use of conductance.

The **conductance** of P is the greatest number $\mathfrak{c} > 0$ such that for all events $A \subseteq \mathbb{R}^d$,

$$\int_A P(x, A^c)\, \pi(\mathrm{d}x) \geq \mathfrak{c}\, \pi(A)\, \pi(A^c)\,.$$

A small conductance implies the presence of *bottlenecks* in the space: subsets A of the state space from which it is difficult for the Markov chain to exit. On the other hand, it is a remarkable fact that once the presence of these bottlenecks is eliminated, then there is a positive spectral gap. This is the content of a celebrated result of Cheeger.

Theorem 7.4.3 (Cheeger's inequality, [LS88]). *The conductance $\mathfrak{c}$ and the spectral gap λ satisfy the inequalities*

$$\frac{1}{8}\,\mathfrak{c}^2 \leq \lambda \leq \mathfrak{c}\,.$$

Both inequalities are sharp up to constants. The upper bound on λ is fairly immediate (see Exercise 7.3), so we focus on the lower bound. We begin by reformulating the conductance as a functional inequality.

Lemma 7.4.4. *Let the conductance of the chain be $\mathfrak{c} > 0$. Then, for all $f \in L^1(\pi)$,*

$$\mathbb{E}_\pi |f - \mathbb{E}_\pi f| \le \frac{1}{\mathfrak{c}} \mathbb{E}|f(X_1) - f(X_0)|, \tag{7.4.5}$$

where (X_0, X_1) are two successive iterates of the chain started at stationarity.

Proof. Let X_0' be an i.i.d. copy of X_0. Then,

$$\mathbb{E}_\pi |f - \mathbb{E}_\pi f| = \mathbb{E}|f(X_0') - \mathbb{E} f(X_0)| \le \mathbb{E}|f(X_0') - f(X_0)|.$$

On the other hand, by reversibility,

$$\begin{aligned}
\mathbb{E}|f(X_1) - f(X_0)| &= \iint |f(x_1) - f(x_0)| \, P(x_0, \mathrm{d}x_1) \, \pi(\mathrm{d}x_0) \\
&= 2 \iint \mathbb{1}\{f(x_1) > f(x_0)\} \, [f(x_1) - f(x_0)] \, P(x_0, \mathrm{d}x_1) \, \pi(\mathrm{d}x_0) \\
&= 2 \iiint \mathbb{1}\{f(x_1) > t \ge f(x_0)\} \, P(x_0, \mathrm{d}x_1) \, \pi(\mathrm{d}x_0) \, \mathrm{d}t \\
&= 2 \int \Big(\int_{\{f \le t\}} P(x_0, \{f \le t\}^c) \, \pi(\mathrm{d}x_0) \Big) \, \mathrm{d}t \\
&\ge 2\mathfrak{c} \int \pi(\{f \le t\}) \, \pi(\{f > t\}) \, \mathrm{d}t \\
&= 2\mathfrak{c} \iiint \mathbb{1}\{f(x_0') > t \ge f(x_0)\} \, \pi(\mathrm{d}x_0) \, \pi(\mathrm{d}x_0') \, \mathrm{d}t \\
&= 2\mathfrak{c} \iint \mathbb{1}\{f(x_0') > f(x_0)\} \, [f(x_0') - f(x_0)] \, \pi(\mathrm{d}x_0) \, \pi(\mathrm{d}x_0') \\
&= \mathfrak{c} \iint |f(x_0') - f(x_0)| \, \pi(\mathrm{d}x_0) \, \pi(\mathrm{d}x_0') = \mathfrak{c} \, \mathbb{E}|f(X_0') - f(X_0)|.
\end{aligned}$$

□

Compare this with the relationship between the Cheeger isoperimetric inequality and the L^1–L^1 Poincaré inequality in Theorem 2.5.14. Indeed, the trick above of passing to the level sets of f is the discrete version of the coarea inequality (Theorem 2.5.12).

Recall also that an L^1–L^1 Poincaré inequality implies an L^2–L^2 Poincaré inequality with $C_{2,2} \lesssim C_{1,1}$, see Proposition 2.5.17. On the other hand, $C_{2,2}$ is the square root of the

usual Poincaré constant, $C_{2,2} = 1/\sqrt{\lambda}$, where λ is the spectral gap. To prove Cheeger's inequality, we are going to follow the same principle in discrete time. This is exactly the source of the square in the lower bound $\lambda \gtrsim \mathfrak{c}^2$ of Cheeger's inequality.

Proof of Cheeger's inequality (Theorem 7.4.3). We will prove the lower bound on the spectral gap with a worse constant than $\frac{1}{8}$ in order to make the proof more straightforward; see [LS88] for a proof with the constant $\frac{1}{8}$.

Let $f : \mathbb{R}^d \to \mathbb{R}$ have 0 as a median and let $g := f^2 \operatorname{sgn} f$, so that 0 is a median of g as well. Assume that the chain has conductance $\mathfrak{c} > 0$. Then, recalling the equivalence between the mean and the median (Lemma 2.4.4) and using Lemma 7.4.4 on g,

$$\begin{aligned}
\mathbb{E}_\pi[|f - \mathbb{E}_\pi f|^2] &\asymp \mathbb{E}_\pi[|f - \operatorname{med}_\pi f|^2] = \mathbb{E}_\pi|g - \operatorname{med}_\pi g| \asymp \mathbb{E}_\pi|g - \mathbb{E}_\pi g| \\
&\le \frac{1}{\mathfrak{c}}\,\mathbb{E}|g(X_1) - g(X_0)| = \frac{1}{\mathfrak{c}}\,\mathbb{E}|f(X_1)^2 - f(X_0)^2| \\
&= \frac{1}{\mathfrak{c}}\,\mathbb{E}[|f(X_1) - f(X_0)|\,|f(X_0) + f(X_1)|] \\
&\lesssim \frac{1}{\mathfrak{c}}\,\sqrt{\mathbb{E}[|f(X_1) - f(X_0)|^2]\,\mathbb{E}_\pi[f^2]} \\
&= \frac{1}{\mathfrak{c}}\,\sqrt{\mathbb{E}[|f(X_1) - f(X_0)|^2]\,\mathbb{E}_\pi[|f - \operatorname{med}_\pi f|^2]} \\
&\asymp \frac{1}{\mathfrak{c}}\,\sqrt{\mathbb{E}[|f(X_1) - f(X_0)|^2]\,\mathbb{E}_\pi[|f - \mathbb{E}_\pi f|^2]}
\end{aligned}$$

and rearranging this inequality proves the result. □

A lower bound on the conductance via overlaps. At this stage, it may not seem that we have gained anything by moving from the spectral gap to the conductance. We now introduce a key lemma, which provides a tractable lower bound on the conductance in terms of two geometric quantities: a Cheeger isoperimetric inequality for target π (we introduced this inequality in Section 2.5.2), and overlap bounds on the Markov chain. Recall that an α-strongly log-concave measure π satisfies the Cheeger isoperimetric inequality with $\mathsf{Ch} \lesssim 1/\sqrt{\alpha}$ (Corollary 2.5.19).

Lemma 7.4.6. *Assume the following:*

1. *The target π satisfies a Cheeger isoperimetric inequality with constant $\mathsf{Ch} > 0$.*
2. *There exists $r \in [0, \mathsf{Ch}]$ such that for any points $x, y \in \mathbb{R}^d$ with $\|x - y\| \le r$, it holds that $\|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}} \le \frac{1}{2}$.*

Then, $\mathfrak{c} \geq r/(64\,\mathsf{Ch})$.

Proof. Let $A_0 \subseteq \mathbb{R}^d$; for symmetry of notation, write $A_1 := A_0^{\mathrm{c}}$. By reversibility,

$$\int_{A_0} P(x, A_1)\,\pi(\mathrm{d}x) = \int_{A_1} P(y, A_0)\,\pi(\mathrm{d}y) = \frac{1}{2}\left(\int_{A_0} P(x, A_1)\,\pi(\mathrm{d}x) + \int_{A_1} P(y, A_0)\,\pi(\mathrm{d}y)\right).$$

We want to lower bound this by a constant times $\pi(A_0)\,\pi(A_1)$.

Define bad sets and a good set:

$$\begin{aligned} B_0 &:= \left\{x \in A_0 \;\middle|\; P(x, A_1) < \frac{1}{4}\right\}, \\ B_1 &:= \left\{y \in A_1 \;\middle|\; P(y, A_0) < \frac{1}{4}\right\}, \\ G &:= \mathbb{R}^d \setminus (B_0 \cup B_1)\,. \end{aligned}$$

We can assume that $\pi(B_0) \geq \pi(A_0)/2$ and $\pi(B_1) \geq \pi(A_1)/2$. Indeed, if we have, e.g., $\pi(B_0) \leq \pi(A_0)/2$, then $\pi(A_0 \setminus B_0) \geq \pi(A_0)/2$, and

$$\int_{A_0} P(x, A_1)\,\pi(\mathrm{d}x) \geq \int_{A_0 \setminus B_0} P(x, A_1)\,\pi(\mathrm{d}x) \geq \frac{1}{4}\,\pi(A_0 \setminus B_0) \geq \frac{1}{8}\,\pi(A_0)\,.$$

Next, suppose that $x \in B_0$ and $y \in B_1$. Then, $P(x, A_0) \geq \frac{3}{4}$, whereas $P(y, A_0) < \frac{1}{4}$. It follows that $\|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}} > \frac{1}{2}$. By our second assumption, $\|x - y\| > r$. This shows that $B_1 \subseteq (B_0^r)^{\mathrm{c}}$, or $B_1^{\mathrm{c}} \supseteq B_0^r$. On the other hand, $G = B_0^{\mathrm{c}} \cap B_1^{\mathrm{c}} \supseteq B_0^r \setminus B_0$. The integral form of the isoperimetric inequality in (2.5.11) shows that

$$\pi(G) \geq \pi(B_0^r) - \pi(B_0) \geq \frac{r}{2\,\mathsf{Ch}}\,\pi(B_0)\,\pi(B_1) \geq \frac{r}{8\,\mathsf{Ch}}\,\pi(A_0)\,\pi(A_1)\,.$$

Hence,

$$\begin{aligned} &\frac{1}{2}\left(\int_{A_0} P(x, A_1)\,\pi(\mathrm{d}x) + \int_{A_1} P(y, A_0)\,\pi(\mathrm{d}y)\right) \\ &\quad\geq \frac{1}{2}\left(\int_{A_0 \cap G} P(x, A_1)\,\pi(\mathrm{d}x) + \int_{A_1 \cap G} P(y, A_0)\,\pi(\mathrm{d}y)\right) \\ &\quad\geq \frac{1}{8}\left(\pi(A_0 \cap G) + \pi(A_1 \cap G)\right) = \frac{1}{8}\,\pi(G) \geq \frac{r}{64\,\mathsf{Ch}}\,\pi(A_0)\,\pi(A_1)\,. \end{aligned}$$

□

From conductance to s-conductance. Unfortunately, the framework that we have developed so far is not flexible enough to study MALA. In particular, requiring that the second condition in Lemma 7.4.6 hold for *all* pairs of points $x, y \in \mathbb{R}^d$ is rather restrictive, especially because there are many points which we are unlikely to ever visit in the course of running the sampling algorithm. To address these issues, many variants of conductance have been proposed in the literature. Here we will introduce only one other variant, the *s-conductance*, which seems reasonably flexible.

For $s \in [0, 1]$, the **s-conductance** of T is the largest $\mathfrak{c}_s > 0$ such that for all events $A \subseteq \mathbb{R}^d$, it holds that

$$\int_A P(x, A^c)\, \pi(\mathrm{d}x) \geq \mathfrak{c}_s \left(\pi(A) - s\right) \left(\pi(A^c) - s\right).$$

Observe that if $\pi(A) \leq s$, then the above inequality holds trivially. Hence, this definition allows us to restrict attention to events which are reasonably probable under π.

For the conductance, we had Cheeger's inequality which relates conductance to the spectral gap and ultimately to convergence. For the s-conductance, the following theorem is an appropriate substitute.

Theorem 7.4.7 ([LS93, Corollary 1.6]). *For any $0 < s \leq \frac{1}{2}$, let*

$$\Delta_s := \sup\{|\mu_0(A) - \pi(A)| : A \subseteq \mathbb{R}^d,\ \pi(A) \leq s\}.$$

Then, the law μ_N of the N-th iterate of a Markov chain with s-conductance $\mathfrak{c}_s$ and initialized at μ_0 satisfies

$$\|\mu_N - \pi\|_{\mathrm{TV}} \leq \Delta_s + \frac{\Delta_s}{s} \exp\left(-\frac{\mathfrak{c}_s^2 N}{2}\right).$$

In particular,

$$\|\mu_N - \pi\|_{\mathrm{TV}} \leq \sqrt{s\, \chi^2(\mu_0 \,\|\, \pi)} + \sqrt{\frac{\chi^2(\mu_0 \,\|\, \pi)}{s}} \exp\left(-\frac{\mathfrak{c}_s^2 N}{2}\right).$$

Proof. The first statement is from [LS93, Corollary 1.6] and the proof is omitted, as the proof is not particularly straightforward.

The second statement follows from the first: indeed, for $A \subseteq \mathbb{R}^d$ with $\pi(A) \leq s$,

$$|\mu_0(A) - \pi(A)| = \left|\int \mathbb{1}_A\, \mathrm{d}(\mu_0 - \pi)\right| = \left|\int \mathbb{1}_A \left(\frac{\mu_0}{\pi} - 1\right) \mathrm{d}\pi\right| \leq \sqrt{\pi(A)\, \chi^2(\mu_0 \,\|\, \pi)}$$

so that $\Delta_s \le \sqrt{s\,\chi^2(\mu_0 \,\|\, \pi)}$. □

This result says that if $s = \varepsilon^2/(4\,\chi^2(\mu_0 \,\|\, \pi))$, then we obtain $\|\mu_N - \pi\|_{\mathrm{TV}} \le \varepsilon$ after

$$N = O\Big(\frac{1}{\mathfrak{c}_s^2} \log \frac{\chi^2(\mu_0 \,\|\, \pi)}{\varepsilon^2}\Big) \qquad \text{iterations}\,.$$

Unfortunately, if $\chi^2(\mu_0 \,\|\, \pi) = \exp O(d)$, then the logarithmic term incurs additional dimension dependence, which is why we can expect better mixing time bounds under the warm start condition $\chi^2(\mu_0 \,\|\, \pi) = O(1)$.

The key advantage of the s-conductance is that it allows for a version of the key lemma with weaker assumptions; the proof is left as Exercise 7.5.

Lemma 7.4.8. *Assume the following:*

1. *The target π satisfies a Cheeger isoperimetric inequality with constant $\mathsf{Ch} > 0$.*
2. *There exists $r \in [0, \mathsf{Ch}]$ and an event $E \subseteq \mathbb{R}^d$ with probability $\pi(E) \ge 1 - \frac{rs}{16\,\mathsf{Ch}}$ such that*

$$\forall x, y \in E,\ \|x - y\| \le r \implies \|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}} \le \frac{1}{2}\,.$$

Then, $\mathfrak{c}_s \gtrsim r/\mathsf{Ch}$.

7.5 Analysis of MALA for a Feasible Start

Using the tools we have developed, we now proceed to analyze the mixing time of MALA under the assumptions of Theorem 7.3.4. However, we will not prove the full strength of the result in Theorem 7.3.4; at the end of this section, we will indicate the extra steps needed to reach Theorem 7.3.4.

Basic decomposition. The overall plan is to lower bound the s-conductance using the key lemma (Lemma 7.4.8), which then upper bounds the mixing time via Theorem 7.4.7. By strong log-concavity of π, the first hypothesis of Lemma 7.4.8 is verified, so it remains to bound the overlaps. For a kernel T, we use the shorthand $T_x := T(x, \cdot)$.

By the triangle inequality, we have the decomposition

$$\|T_x - T_y\|_{\mathrm{TV}} \le \|Q_x - T_x\|_{\mathrm{TV}} + \|Q_x - Q_y\|_{\mathrm{TV}} + \|Q_y - T_y\|_{\mathrm{TV}}\,. \tag{7.5.1}$$

The middle term $\|Q_x - Q_y\|_{\mathrm{TV}}$ measures the overlap for the proposal kernel, and we will shortly see that this term is easy to bound. Then, controlling the first and third terms essentially amounts to lower bounding the acceptance probability of MALA, since the only difference between Q and T is the Metropolis–Hastings filter.

Overlap of the proposal kernel.

Lemma 7.5.2. *For $x \in \mathbb{R}^d$, let $Q_x := \mathsf{normal}(x - h\nabla V(x), 2h\, I_d)$, and assume that $\|\nabla^2 V\|_{\mathrm{op}} \le \beta$. Then, provided $h \le \frac{1}{\beta}$, we have*

$$\|Q_x - Q_y\|_{\mathrm{TV}} \le \frac{\|x - y\|}{\sqrt{2h}}\,.$$

Proof. By Pinsker's inequality (Exercise 2.13),

$$\|Q_x - Q_y\|_{\mathrm{TV}}^2 \le \frac{1}{2}\,\mathsf{KL}(Q_x \,\|\, Q_y) = \frac{\|x - h\nabla V(x) - y + h\nabla V(y)\|^2}{8h} \le \frac{\|x - y\|^2}{2h}$$

where the last inequality uses the fact that $\mathrm{id} - h\nabla V$ is 2-Lipschitz. □

Control of the acceptance probability. Next, consider the term $\|Q_x - T_x\|_{\mathrm{TV}}$. Computing this term is slightly tricky because T_x has an atom at x, but in the end we obtain

$$\begin{aligned}
\|Q_x - T_x\|_{\mathrm{TV}} &= \frac{1}{2}\Big[\underbrace{1 - \int Q(x, \mathrm{d}y)\, A(x, y)}_{\text{from the atom of } T_x} + \int_{\mathbb{R}^d \setminus \{x\}} |Q(x, y) - T(x, y)|\,\mathrm{d}y\Big] \\
&= \frac{1}{2}\Big[1 - \int Q(x, \mathrm{d}y)\, A(x, y) + \int Q(x, \mathrm{d}y)\,\{1 - A(x, y)\}\,\mathrm{d}y\Big] \\
&= 1 - \int Q(x, \mathrm{d}y)\, A(x, y)\,. \qquad (7.5.3)
\end{aligned}$$

This has a very clear interpretation: it is the probability that the proposed move starting at x is rejected. If we let $\xi \sim \mathsf{normal}(0, I_d)$ and $Y := x - h\nabla V(x) + \sqrt{2h}\,\xi$, we want a lower bound on the quantity $\mathbb{E}\, A(x, Y)$, which comes from Markov's inequality:

$$\mathbb{E}\, A(x, Y) = \mathbb{E}\min\Big\{1, \frac{\pi(Y)\, Q(Y, x)}{\pi(x)\, Q(x, Y)}\Big\} \ge \lambda\, \mathbb{P}\Big\{\frac{\pi(Y)\, Q(Y, x)}{\pi(x)\, Q(x, Y)} \ge \lambda\Big\} \qquad \text{for all } 0 < \lambda < 1\,.$$

The approach now is to write out the ratio more explicitly, and then carefully group together and bound the terms. (Unfortunately, this is not the most enlightening.)

Explicitly, we have

$$\frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} = \exp\Big(-V(Y) - \frac{\|x - Y + h\,\nabla V(Y)\|^2}{4h} + V(x) + \frac{\|Y - x + h\,\nabla V(x)\|^2}{4h}\Big)\,.$$

After some careful algebra,

$$4 \ln \frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} = h\,\{\|\nabla V(x)\|^2 - \|\nabla V(Y)\|^2\} \tag{7.5.4}$$

$$- 2\,\{V(Y) - V(x) - \langle \nabla V(x), Y - x\rangle\} \tag{7.5.5}$$

$$+ 2\,\{V(x) - V(Y) - \langle \nabla V(Y), x - Y\rangle\}\,. \tag{7.5.6}$$

Note that the terms are grouped to more easily apply the strong convexity and smoothness of V. It yields

$$(7.5.5) \ge -\beta\,\|x - Y\|^2 \qquad \text{and} \qquad (7.5.6) \ge \alpha\,\|x - Y\|^2 \ge 0\,.$$

Also, for $h \le \frac{1}{\beta}$,

$$\begin{aligned}(7.5.4) &= h\,\langle \nabla V(x) - \nabla V(Y), \nabla V(x) + \nabla V(Y)\rangle \\ &\ge -h\,\|\nabla V(x) - \nabla V(Y)\|\,\|\nabla V(x) + \nabla V(Y)\| \\ &\ge -\beta h\,\|x - Y\|\,(2\,\|\nabla V(x)\| + \beta\,\|x - Y\|) \ge -\beta h^2\,\|\nabla V(x)\|^2 - 2\beta\,\|x - Y\|^2\,.\end{aligned}$$

Therefore,

$$\ln \frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} \gtrsim -\beta h^2\,\|\nabla V(x)\|^2 - \beta\,\|x - Y\|^2 \gtrsim -\beta h^2\,\|\nabla V(x)\|^2 - \beta h\,\|\xi\|^2\,.$$

At this stage, observe that we cannot lower bound this quantity (with high probability) *uniformly* over x, since $\|\nabla V(x)\| \to \infty$ as $\|x\| \to \infty$. This is why it is helpful to restrict to x belonging to some high-probability event E, which is ultimately achieved by working with s-conductance rather than conductance.

By standard concentration bounds, $\|\xi\|^2 \le 2d$ with probability at least $1 - \exp(-d/2)$. Also, let $E_R := \{x \in \mathbb{R}^d : \|\nabla V(x)\| \le \sqrt{\beta}\,R\}$. It follows that for all $x \in E_R$, if we take $h \lesssim \frac{1}{\beta\,(d \vee R)}$ with a sufficiently small constant, then

$$\mathbb{E}\,A(x,Y) \ge \frac{11}{12}\,\mathbb{P}\Big\{\frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} \ge \frac{11}{12}\Big\} \ge \frac{11}{12}\,\Big(1 - \exp\big(-\frac{d}{2}\big)\Big) \ge \frac{5}{6}\,,$$

for sufficiently large d. Hence, for $x \in E_R$, we have $\|Q_x - T_x\|_{\mathrm{TV}} \le \frac{1}{6}$.

Completing the analysis. We have shown: if the step size is $h \lesssim \frac{1}{\beta(d \vee R)}$, then for all $x, y \in E_R$ with $\|x - y\| \le r$,

$$\|T_x - T_y\|_{\mathrm{TV}} \le \|Q_x - T_x\|_{\mathrm{TV}} + \|Q_x - Q_y\|_{\mathrm{TV}} + \|Q_y - T_y\|_{\mathrm{TV}} \le \frac{1}{6} + \frac{r}{\sqrt{2h}} + \frac{1}{6} \le \frac{1}{2}$$

provided we take $r = \sqrt{2h}/6$. Applying Lemma 7.4.8 (assuming $h \lesssim \frac{1}{\alpha}$), we deduce that $\mathfrak{c}_s \gtrsim \sqrt{\alpha h}$ provided $\pi(E_R) \ge 1 - c_0 s\sqrt{\alpha h}$, where $c_0 > 0$ is a universal constant. Since we want the step size h to be as large as possible, we take $h \asymp \frac{1}{\beta(d \vee R)}$, where R is chosen to satisfy $\pi(E_R^{\mathrm{c}}) \lesssim s\sqrt{\frac{1}{\kappa(d \vee R)}}$ and $s \asymp \varepsilon^2/\chi^2(\mu_0 \,\|\, \pi)$. The final mixing time bound implied by Theorem 7.4.7 is then $O(\kappa\,(d \vee R)\log(\chi^2(\mu_0 \,\|\, \pi)/\varepsilon^2))$ iterations.

Up until this point, the analysis is largely similar to [Dwi+19].

Gradient concentration. The bound involves the parameter R. By definition, R is such that the norm $\|\nabla V\|$ of the gradient under π is typically of size $\sqrt{\beta}\,R$. Recall from, e.g., Lemma 4.2.5 that $\mathbb{E}_\pi\|\nabla V\| \lesssim \sqrt{\beta d}$, which suggests that we can take $R \lesssim \sqrt{d}$. However, we need a high-probability bound on $\|\nabla V\|$, not a bound in expectation. Unfortunately, a naïve application of the fact that $\|\nabla V\|$ is β-Lipschitz, together with sub-Gaussian concentration of Lipschitz functions (Theorem 2.4.8), only shows that the fluctuations of $\|\nabla V\|$ around its expectation are of size $\sqrt{\beta\kappa}$ (exercise!). When $\kappa \gg d$, this does not recover the promised rate of $\widetilde{O}(\kappa d)$ (ignoring the dependence on initialization and accuracy). To resolve this issue, [LST20] introduced a new concentration inequality for $\|\nabla V\|$ via the Brascamp–Lieb inequality (Theorem 2.2.8).

Lemma 7.5.7 ([LST20]). *Suppose that $\pi \propto \exp(-V)$ and that $0 \preceq \nabla^2 V \preceq \beta I_d$. Then, for all $t \ge 0$,*

$$\pi\{\|\nabla V\| \ge \mathbb{E}_\pi\|\nabla V\| + t\} \le 3\exp\Big(-\frac{t}{\sqrt{\beta}}\Big).$$

This shows that the fluctuations of $\|\nabla V\|$ around its expectation are only of size $\sqrt{\beta}$.

From warm start to feasible start. The factor of $\log \chi^2(\mu_0 \,\|\, \pi)$ in the bound incurs additional dimension dependence under a feasible start, since with a Gaussian initialization we can only show $\chi^2(\mu_0 \,\|\, \pi) \le \kappa^{d/2}$. The problem is that the conductance-based analysis relies upon Poincaré-type inequalities, instead of log-Sobolev inequalities. To address this issue, we can replace the assumption of a Cheeger isoperimetric inequality with a

Gaussian isoperimetric inequality (see Section 2.5.4). The essential difference is that under a Cheeger isoperimetric inequality, as $p := \pi(A) \searrow 0$ we have $\pi^+(A) \gtrsim p$, whereas under a Gaussian isoperimetric inequality we have $\pi^+(A) \gtrsim p\sqrt{\log \frac{1}{p}}$. Using this stronger assumption, [Che+20a] show that the dependence on the initialization can be improved to $\log\log \chi^2(\mu_0 \,\|\, \pi)$. A similar effect can be achieved via the *blocking conductance* [KLM06], which was used in [LST20]. We omit the details.

Lower bound. Finally, the analysis of MALA in Theorem 7.3.4 is tight, as shown in the following lower bound.

Theorem 7.5.8 ([LST21a]). *For every choice of step size $h > 0$, there exists a target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^d$ with $I_d \preceq \nabla^2 V \preceq \kappa I_d$, as well as an initialization μ_0 with $\chi^2(\mu_0 \,\|\, \pi) \le \exp d$, such that the number of iterations required for MALA to reach total variation at most $\frac{1}{4}$ from π is at least $\widetilde{\Omega}(\kappa d)$.*

This theorem is a lower bound in the sense that all of the known proofs for MALA do not use any property of the initialization μ_0 except through $\chi^2(\mu_0 \,\|\, \pi)$. Thus, in order to improve the analysis of MALA under a feasible start, one must use more specific properties of the initialization, or use some other modification that bypasses the lower bound (e.g., random step sizes).

7.6 Analysis of MALA for a Warm Start

We next turn towards the warm start case (Theorem 7.3.5). The improvement under a warm start was first shown in [Che+21b], which obtained a rate of $\widetilde{O}(\kappa^{3/2} d^{1/2} \operatorname{polylog}(1/\varepsilon))$. This result was improved in [WSC21] which obtained the sharp rate of $\widetilde{O}(\kappa d^{1/2} \operatorname{polylog}(1/\varepsilon))$ via completely different techniques. In this section, we follow [Che+21b] because the proof is more conceptual. (Anyway, we will see in Chapter 8 how to boost the condition number dependence to κ using the proximal sampler.)

We still follow the s-conductance framework of the previous section, including the basic decomposition (7.5.1). The main difference lies in the control of $\|Q_x - T_x\|_{\mathrm{TV}}$, which was previously accomplished by lower bounding the acceptance probability. Surprisingly, the following proof never works directly with the acceptance probability, despite the fact that $\|Q_x - T_x\|_{\mathrm{TV}}$ is precisely the rejection probability at x (see (7.5.3)).

Using the projection property. The key insight is to use projection characterization of the Metropolis–Hastings filter (Theorem 7.2.5): the MALA kernel T is the closest kernel

to the proposal Q (in an appropriate L^1 distance) among all reversible Markov chains with stationary distribution π. Concretely, for any other kernel $\bar{Q}$ which is reversible w.r.t. π,

$$\iint_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus\mathrm{diag}} |Q(x,y) - T(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y \le \iint_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus\mathrm{diag}} |Q(x,y) - \bar{Q}(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y\,.$$

Now supposing that $\bar{Q}$ has no atoms, this inequality is the same as

$$\int \|Q_x - T_x\|_{\mathrm{TV}}\,\pi(\mathrm{d}x) \le 2\int \|Q_x - \bar{Q}_x\|_{\mathrm{TV}}\,\pi(\mathrm{d}x)\,. \tag{7.6.1}$$

Thus, we can indirectly bound $\|Q_x - T_x\|_{\mathrm{TV}}$, at least on average. Moreover, there is a very natural choice of $\bar{Q}$ here: since Q is obtained from a discretization of the Langevin diffusion, we can take $\bar{Q}$ to be the continuous-time Langevin diffusion run for time h, which is indeed reversible with respect to π. The right-hand side of the above expression then simply measures the discretization error, which we have already studied in detail.

Pointwise projection property. The projection property is not enough for our purposes, however, since it only bounds $\|Q_x - T_x\|_{\mathrm{TV}}$ in average, whereas we really need high-probability bounds. Thankfully, we can extend the projection property.

Theorem 7.6.2 (pointwise projection property, [Che+21b, Theorem 6])**.** *Let Q be an atomless proposal kernel and let T be the corresponding Metropolis–Hastings kernel with target π. Then, for any atomless kernel $\bar{Q}$ which is reversible with respect to π, and for every $x \in \mathbb{R}^d$,*

$$\|Q_x - T_x\|_{\mathrm{TV}} \le 2\,\|Q_x - \bar{Q}_x\|_{\mathrm{TV}} + \int \frac{\pi(y)\,\bar{Q}(y,x)}{\pi(x)}\Big|\frac{Q(y,x)}{\bar{Q}(y,x)} - 1\Big|\,\mathrm{d}y\,.$$

Consequently, for any convex increasing function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$,

$$\begin{aligned}\int \Phi(\|Q_x - T_x\|_{\mathrm{TV}})\,\pi(\mathrm{d}x) \le \frac{1}{2}\int \Phi(4\,\|Q_x - \bar{Q}_x\|_{\mathrm{TV}})\,\pi(\mathrm{d}x) \\ + \frac{1}{2}\iint \Phi\Big(2\Big|\frac{Q(x,y)}{\bar{Q}(x,y)} - 1\Big|\Big)\,\bar{Q}(x,\mathrm{d}y)\,\pi(\mathrm{d}x)\,.\end{aligned} \tag{7.6.3}$$

We will not need the inequality (7.6.3), so the proof is left as Exercise 7.8. The reason why (7.6.3) is included in the theorem is because it makes it clear why we can expect the pointwise projection property to imply high-probability bounds for $\|Q_x - T_x\|_{\mathrm{TV}}$. Note that when we integrate the projection property w.r.t. $\pi(\mathrm{d}x)$, we recover (7.6.1) with a factor of 4 on the right-hand side instead of 2.

Proof. We can write

$$\begin{aligned}\|Q_x - T_x\|_{\mathrm{TV}} = 1 - \int Q(x, \mathrm{d}y)\, A(x, y) &= \int \left[1 - \left(1 \wedge \frac{\pi(y)\, Q(y, x)}{\pi(x)\, Q(x, y)}\right)\right] Q(x, \mathrm{d}y) \\ &\le \int \left|1 - \frac{\pi(y)\, Q(y, x)}{\pi(x)\, Q(x, y)}\right| Q(x, \mathrm{d}y) \\ &\le \int \left|1 - \frac{\pi(y)\, \bar{Q}(y, x)}{\pi(x)\, Q(x, y)}\right| Q(x, \mathrm{d}y) + \int \frac{\pi(y)\, \bar{Q}(y, x)}{\pi(x)} \left|\frac{Q(y, x)}{\bar{Q}(y, x)} - 1\right| \mathrm{d}y\,.\end{aligned}$$

Using reversibility of $\bar{Q}$, the first term is

$$\int \left|1 - \frac{\pi(y)\, \bar{Q}(y, x)}{\pi(x)\, Q(x, y)}\right| Q(x, \mathrm{d}y) = \int |Q(x, y) - \bar{Q}(x, y)|\, \mathrm{d}y = 2\, \|Q_x - \bar{Q}_x\|_{\mathrm{TV}}\,,$$

which completes the proof. □

Applying the pointwise projection property. Our goal is to bound $\|Q_x - T_x\|_{\mathrm{TV}}$ for all x which lies in an event E of very high probability under π. We proceed by controlling the two terms in the pointwise projection property separately.

We will omit many of the calculations from this point forwards. The calculations are actually fairly straightforward (once one has some familiarity with stochastic calculus), but are somewhat tedious. Moreover, the best way to learn these particular calculations is to try them for oneself. We refer to [Che+21b] for details. Moreover, to simplify the exposition, we will only focus on the dependence on d and h.

The first term, $\|Q_x - \bar{Q}_x\|_{\mathrm{TV}}$, is more straightforward. It is helpful to apply Pinsker's inequality, leaving us to control $\mathsf{KL}(\bar{Q}_x \,\|\, Q_x)$. This is precisely the kind of discretization error that we controlled via Girsanov's theorem in Section 4.4. In particular, it is possible to show that $\|Q_x - \bar{Q}_x\|_{\mathrm{TV}} \lesssim h\sqrt{d + \|x\|^2}$. Since this is Lipschitz in x, we can then apply sub-Gaussian concentration under π (Theorem 2.4.8) to obtain a high-probability bound for this term under π. In particular, we expect a step size of $h \lesssim \frac{1}{\sqrt{d}}$ to control this term.

To control the second term with high probability, it suffices to control the moments of this quantity under π: for $p \ge 1$,

$$\int \left| \int \frac{\pi(y)\, \bar{Q}(y, x)}{\pi(x)} \left|\frac{Q(y, x)}{\bar{Q}(y, x)} - 1\right| \mathrm{d}y \right|^p \pi(\mathrm{d}x) \le \iint \left|\frac{Q(y, x)}{\bar{Q}(y, x)} - 1\right|^p \bar{Q}(y, \mathrm{d}x)\, \pi(\mathrm{d}y)\,.$$

Let $\bar{\boldsymbol{Q}}_x$ denote the measure on path space $C([0, h]; \mathbb{R}^d)$ of the Langevin diffusion started at x. Similarly, let $\boldsymbol{Q}_x$ denote the same for the interpolation of LMC. By the data-processing

inequality, we obtain

$$\iint \Big|\frac{Q(y,x)}{\bar{Q}(y,x)} - 1\Big|^p \, \bar{Q}(y, \mathrm{d}x)\, \pi(\mathrm{d}y) \le \int \mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Big[\Big|\frac{\mathrm{d}\boldsymbol{Q}_x}{\mathrm{d}\bar{\boldsymbol{Q}}_x} - 1\Big|^p\Big]\, \pi(\mathrm{d}x)\,.$$

We have a formula for the Radon–Nikodym derivative $\frac{\mathrm{d}\bar{\boldsymbol{Q}}_x}{\mathrm{d}\boldsymbol{Q}_x}$ thanks to Girsanov's theorem, so again we can approach this via stochastic calculus. Controlling this term is slightly more involved than controlling the KL divergence (essentially we are controlling a Rényi divergence instead) but nevertheless we can bound the term when $h \lesssim \frac{1}{\sqrt{d}}$.

With these high probability bounds, we can then return to the s-conductance analysis, which implies that the mixing time is of order $\frac{1}{h}$. Hence, under a warm start, the mixing time improves from d to $\sqrt{d}$.

TODO: Flesh out the calculations in this section.

Lower bound. Under a warm start, [Che+21b] showed a lower bound of roughly $\widetilde{\Omega}(\sqrt{d})$, which was improved in [WSC21] to $\widetilde{\Omega}(\kappa\sqrt{d})$. We state the result here.

Theorem 7.6.4 ([WSC21]). *For every choice of step size $h > 0$, there exists a target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^d$ with $I_d \preceq \nabla^2 V \preceq \kappa I_d$, as well as an initialization μ_0 with $\chi^2(\mu_0 \,\|\, \pi) \lesssim 1$, such that the number of iterations required for MALA to reach total total variation ε from π is at least $\widetilde{\Omega}(\kappa\sqrt{d}\log(1/\varepsilon))$.*

Bibliographical Notes

TODO: Fill in.

Exercises

An Overview of High-Accuracy Samplers

▷ **Exercise 7.1** (MALA is a special case of MHMC)
Show that when $K = 1$, the MHMC algorithm reduces to MALA.

▷ **Exercise 7.2** (MH filter for the leapfrog integrator)
For the leapfrog integrator F_{leap}, verify that $F_{\text{leap}}^{-1} = R \circ F_{\text{leap}} \circ R$.

Hint: First show that it suffices to consider $T = h$ (i.e., $K = 1$).

Markov Chains in Discrete Time

▷ **Exercise 7.3** (reversible Markov chains as gradient descent on the Dirichlet energy)
Consider the setting of Section 7.4.

1. Show that $\mathscr{E}(f,f) = \frac{1}{2}\mathbb{E}[|f(X_1) - f(X_0)|^2]$, where (X_0, X_1) are two successive iterates of the Markov chain started at stationarity. We also write $\mathscr{E}(f) := \mathscr{E}(f,f)$ as a useful shorthand.

2. Show that if $(\mu_k)_{k\in\mathbb{N}}$ are the laws of the iterates of the ℓ-lazy version of P, then the relative densities $(\frac{\mu_k}{\pi})_{k\in\mathbb{N}}$ are the iterates of gradient descent on the Dirichlet energy $\mathscr{E}$ in $L^2(\pi)$. How does the laziness parameter ℓ relate to the step size of the gradient descent?

3. Observe that $\mathscr{E}$ is a convex quadratic functional; show that $0 \preceq \nabla^2_{L^2(\pi)}\mathscr{E} \preceq 2$. What does the theory of convex optimization suggest for the value of the laziness parameter ℓ?

4. Next, prove a generalization of Theorem 7.4.2 for any value of the laziness parameter $\ell \in [\frac{1}{2}, 1]$ by showing that the spectral gap condition is equivalent to strong convexity of $\mathscr{E}$. Why do we want $\ell \geq \frac{1}{2}$ here?

5. Show that the conductance of the chain can also be described as the largest $\mathfrak{c} > 0$ such that for all events $A \subseteq \mathbb{R}^d$, it holds that $\mathscr{E}(\mathbb{1}_A) \geq \mathfrak{c}\,\|\mathbb{1}_A - \pi(A)\|^2_{L^2(\pi)}$. Hence, conductance can be viewed as a restricted strong convexity condition (restricting the space of functions to indicators of events). In particular, show the bound $\lambda \leq \mathfrak{c}$ in Cheeger's inequality (Theorem 7.4.3).

▷ **Exercise 7.4** (entropy decay implies MLSI)
Suppose that P satisfies the following entropy decay condition: there exists $c \in (0,1)$ such that for all probability measures P,

$$\mathsf{KL}(\mu P \,\|\, \pi) \leq (1-c)\,\mathsf{KL}(\mu \,\|\, \pi)\,.$$

Prove that P satisfies a MLSI with constant $C_{\mathsf{MLSI}} \leq 2/c$.

▷ **Exercise 7.5** (s-conductance lemma)
Prove the s-conductance lemma (Lemma 7.4.8).

Analysis of MALA for a Feasible Start

⊵ **Exercise 7.6** (analysis of MRW)
Follow the analysis in this section and adapt it to the Metropolized random walk (MRW) algorithm. What mixing time bound can you prove?

⊵ **Exercise 7.7** (mixing time for a Gaussian target)
Adapt the analysis in this section to the case when the target distribution is the standard Gaussian. Here, it is possible to do a much more refined analysis; see if you can show that the mixing time of MALA is $\widetilde{O}(d^{1/3}\operatorname{polylog}(1/\varepsilon))$ from a warm start. See [Che+21b, Appendix C] for hints.

Analysis of MALA for a Warm Start

⊵ **Exercise 7.8** (pointwise projection property)
Prove (7.6.3) from the pointwise projection property.

CHAPTER 8

The Proximal Sampler

In this chapter, we discuss the proximal sampler, which was introduced in [LST21c]. The applications of the proximal sampler include improving the condition number dependence of high-accuracy samplers and providing new state-of-the-art sampling guarantees for various classes of target distributions. Besides these applications, the proximal sampler is interesting in its own right due to its remarkable convergence analysis and its connections with the proximal point method in optimization.

8.1 Introduction to the Proximal Sampler

Let $\pi \propto \exp(-V)$ denote the target distribution. We fix $h > 0$ and define the augmented target distribution

$$\boldsymbol{\pi}(x, y) \propto \exp\Big(-V(x) - \frac{\|y - x\|^2}{2h}\Big).$$

To avoid confusion, we will explicitly write $\pi^X = \pi$ for the X-marginal, and π^Y for the Y-marginal. Similarly, $\pi^{X|Y}$ and $\pi^{Y|X}$ denote the conditional distributions.

The proximal sampler applies Gibbs sampling to the augmented target. Explicitly, the updates of the proximal sampler are as follows.

Proximal Sampler: Initialize $X_0 \sim \mu_0$. For $k = 0, 1, 2, \ldots$:

1. Draw $Y_k \sim \pi^{Y|X}(\cdot \mid X_k) = \mathsf{normal}(X_k, hI_d)$.

2. Draw $X_{k+1} \sim \pi^{X|Y}(\cdot \mid Y_k)$.

Since Gibbs sampling always forms a reversible Markov chain with respect to the target distribution, we conclude that the proximal sampler is *unbiased*: its stationary distribution of the proximal sampler is $\boldsymbol{\pi}$. As written, however, the proximal sampler is an idealized algorithm because it is not yet clear how to implement the second step of sampling from $\pi^{X|Y}$. Note that

$$\pi^{X|Y}(x \mid y) \propto_x \exp\left(-V(x) - \frac{\|y-x\|^2}{2h}\right).$$

Also, recall that in optimization, we wish to minimize the function V, whereas in sampling we want to sample from $\pi \propto \exp(-V)$. Via this correspondence, we see that the optimization analogue of sampling from $\pi^{X|Y}$ is computing the *proximal map*

$$\mathrm{prox}_{hV}(y) \coloneqq \arg\min_{x \in \mathbb{R}^d} \left\{ V(x) + \frac{\|y-x\|^2}{2h} \right\}.$$

The distribution $\pi^{X|Y}$ is known as the **restricted Gaussian oracle (RGO)**, and the proximal sampler can be viewed as the analogue of the proximal point method for sampling. See Exercise 8.1 for another connection between the proximal sampler and the proximal point method from optimization.

Implementability of the RGO. In order to obtain an actual algorithm from the proximal sampler, an implementation of the RGO must be provided. As we will see in Section 8.6, the RGO can be implemented by using an auxiliary high-accuracy sampler such as MALA. Although this may seem circular (if we need to use an auxiliary sampler to implement the RGO, then why not use the auxiliary sampler in the first place without bothering with the proximal sampler?), we will see there are benefits to the overall scheme. Namely, the proximal sampler can boost the condition number dependence of the auxiliary sampler, and it can be used to sample from a larger class of distributions.

For now, we will consider a simple implementation of the RGO based on rejection sampling, which we studied in Section 7.1. Suppose that the potential V is β-smooth. Then, for $V_y(x) \coloneqq V(x) + \frac{1}{2h}\|y-x\|^2$ we have $(\frac{1}{h} - \beta)\, I_d \preceq \nabla^2 V_y \preceq (\frac{1}{h} + \beta)\, I_d$. In particular, if $h < \frac{1}{\beta}$, then the RGO is strongly log-concave. Note that the condition number of V_y is $\kappa = (\frac{1}{h} + \beta)/(\frac{1}{h} - \beta)$. If we now choose $h = \frac{1}{\beta d}$, we can check that $\kappa \le \exp(4/d)$ for $d \ge 2$. By Proposition 7.1.2, if we have access to the minimizer of V_y (which is equivalent to being able to compute the proximal operator for hV), we can construct an upper envelope for which the average number of iterations of rejection sampling is bounded by $\kappa^{d/2} \le \exp(2) \le 8$. We summarize this discussion as follows.

Implementing the RGO via Rejection Sampling: To sample from $\pi^{X|Y}(\cdot \mid y)$, where V is β-smooth, we proceed via the following steps.

1. Compute the minimizer $x_y^\star$ of V_y defined via $V_y(x) := V(x) + \frac{1}{2h}\|y - x\|^2$, and compute the minimum value $V_y^\star$. This can be done exactly if we assume access to the proximal mapping of V (which is a natural assumption when designing a proximal algorithm for sampling); otherwise, if $h < \frac{1}{\beta}$, then this is a strongly convex optimization problem and can be implemented using standard algorithms.

2. Let $\widetilde{\pi}^{X|Y}(\cdot \mid y) := \exp\{-(V_y - V_y^\star)\}$ and $\widetilde{\mu}_y := \exp(-\frac{1/h-\beta}{2}\|\cdot - x_y^\star\|^2)$. Use rejection sampling to sample from $\pi^{X|Y}$ with the envelope $\widetilde{\mu}_y$.

Each iteration of rejection sampling requires one call to an evaluation oracle for V (in order to compute the acceptance probability). We summarize the guarantees for this implementation of the RGO in the following theorem.

Theorem 8.1.1. *Assume that V is β-smooth. Then, if $h \le \frac{1}{\beta d}$, rejection sampling implements the RGO for π^X exactly using one computation of the proximal map for V and $O(1)$ expected calls to an evaluation oracle for V.*

Notation. We write μ_k^X for the law of X_k and μ_k^Y for the law of Y_k for the iterates of the proximal sampler. Observe that if $(Q_t)_{t\ge 0}$ denotes the standard heat semigroup, i.e. $\mu Q_t = \mu * \mathrm{normal}(0, tI_d)$, then $\mu_k^Y = \mu_k^X Q_h$ and $\pi^Y = \pi^X Q_h$.

We also abbreviate $\pi^{X|Y}(\cdot \mid y)$ as $\pi^{X|Y=y}$.

8.2 Convergence under Strong Log-Concavity

One of the most remarkable features of the proximal sampler is that its convergence analysis closely mirrors the continuous-time theory for the Langevin diffusion. In this section, we initiate this study starting with the strongly log-concave case.

Recall that under strong log-concavity, we have contraction of the Langevin diffusion (Theorem 1.4.10). We prove the analogue of this fact for the proximal sampler.

Theorem 8.2.1. *Assume that the target π^X is α-strongly log-concave. Also, let $(\mu_k^X)_{k\in\mathbb{N}}$*

and $(\bar{\mu}_k^X)_{k\in\mathbb{N}}$ denote two runs of the proximal sampler with target π^X. Then,

$$W_2(\mu_k^X, \bar{\mu}_k^X) \le \frac{W_2(\mu_0^X, \bar{\mu}_0^X)}{(1+\alpha h)^k}\,.$$

The contraction factor matches the contraction for the proximal point method in optimization, see Exercise 8.2. Since π^X is left invariant by the proximal sampler, the contraction result also implies a convergence result in W_2.

We will give two proofs of this theorem. First, note that it suffices to consider one iteration and to prove $W_2(\mu_1^X, \bar{\mu}_1^X) \le \frac{1}{1+\alpha h} W_2(\mu_0^X, \bar{\mu}_0^X)$. Next, since the heat flow is a Wasserstein contraction (which follows from (1.4.9) but can also be proven by a straightforward coupling), it holds that $W_2(\mu_0^Y, \bar{\mu}_0^Y) \le W_2(\mu_0^X, \bar{\mu}_0^X)$, so it suffices to show $W_2(\mu_1^X, \bar{\mu}_1^X) \le \frac{1}{1+\alpha h} W_2(\mu_0^Y, \bar{\mu}_0^Y)$.

We will use the following coupling lemma.

Lemma 8.2.2. *Suppose that for all $y, \bar{y} \in \mathbb{R}^d$, we have*

$$W_2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \le C\,\|y - \bar{y}\|\,. \tag{8.2.3}$$

Then, $W_2(\mu_1^X, \bar{\mu}_1^X) \le C\, W_2(\mu_0^Y, \bar{\mu}_0^Y)$.

The intuition is that since μ_1^X and $\bar{\mu}_1^X$ are obtained from μ_0^Y and $\bar{\mu}_0^Y$ by sampling from the RGO $\pi^{X|Y}$, the contraction statement in (8.2.3) can be used to bound $W_2(\mu_1^X, \bar{\mu}_1^X)$. The proof of the lemma is relatively straightforward and good practice for working with couplings, so it is left as Exercise 8.3.

The first proof we present is from [LST21b].

Proof of Theorem 8.2.1 via functional inequalities. To prove (8.2.3), we note that $\pi^{X|Y}(\cdot \mid \bar{y})$ is $(\alpha + \frac{1}{h})$-strongly log-concave. Recall that by the Bakry–Émery theorem (Theorem 1.2.29) and the Otto–Villani theorem (Exercise 1.16) this implies the log-Sobolev inequality (1.4.7) and Talagrand's T_2 inequality (1.4.8). Applying these inequalities,

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \le \frac{2}{\alpha + \frac{1}{h}} \mathsf{KL}(\pi^{X|Y=y} \parallel \pi^{X|Y=\bar{y}}) \le \frac{1}{(\alpha + \frac{1}{h})^2} \mathsf{FI}(\pi^{X|Y=y} \parallel \pi^{X|Y=\bar{y}})\,.$$

We can compute the Fisher information explicitly. Indeed,

$$\nabla \ln \frac{\pi^{X|Y=y}}{\pi^{X|Y=\bar{y}}} = \nabla\Big(\frac{\|\bar{y} - \cdot\|^2}{2h} - \frac{\|y - \cdot\|^2}{2h}\Big) = \frac{y - \bar{y}}{h}$$

so that

$$\mathsf{FI}(\pi^{X|Y=y} \,\|\, \pi^{X|Y=\bar{y}}) = \mathbb{E}_{\pi^{X|Y=y}}\Big[\Big\|\nabla \ln \frac{\pi^{X|Y=y}}{\pi^{X|Y=\bar{y}}}\Big\|^2\Big] = \frac{\|y-\bar{y}\|^2}{h^2}\,.$$

Hence,

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \le \frac{1}{\left(\alpha + \frac{1}{h}\right)^2} \frac{\|y-\bar{y}\|^2}{h^2} = \frac{1}{(1+\alpha h)^2}\,\|y-\bar{y}\|^2\,. \qquad \square$$

The next proof, from [Che+22a], directly uses strong convexity in Wasserstein space.

Proof of Theorem 8.2.1 via Wasserstein calculus. This proof rests on the following interpretation of the RGO. Let $\mathcal{F}(\mu) \coloneqq \mathsf{KL}(\mu \,\|\, \pi^X)$. Then, by Exercise 8.1,

$$\pi^{X|Y=y} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \Big\{\mathcal{F}(\mu) + \frac{1}{2h}\, W_2^2(\mu, \delta_y)\Big\} =: \operatorname{prox}_{h\mathcal{F}}(\delta_y)\,.$$

The first-order optimality conditions on Wasserstein space [AGS08, Lemma 10.1.2] reads

$$0 \in \partial\mathcal{F}(\pi^{X|Y=y}) + \frac{1}{h}\,(\mathrm{id} - y)\,, \qquad \pi^{X|Y=y}\text{-a.s.}$$

where $\partial\mathcal{F}$ is the subdifferential of $\mathcal{F}$ on Wasserstein space.

Using this, we obtain

$$\begin{aligned} \mathrm{id} &\in y - h\,\partial\mathcal{F}(\pi^{X|Y=y})\,, && \pi^{X|Y=y}\text{-a.s.} \\ \mathrm{id} &\in \bar{y} - h\,\partial\mathcal{F}(\pi^{X|Y=\bar{y}})\,, && \pi^{X|Y=\bar{y}}\text{-a.s.} \end{aligned}$$

Let T be the optimal transport map from $\pi^{X|Y=y}$ to $\pi^{X|Y=\bar{y}}$. The second condition above can then be rewritten as

$$T \in \bar{y} - h\,\partial\mathcal{F}(\pi^{X|Y=\bar{y}}) \circ T\,, \qquad \pi^{X|Y=y}\text{-a.s.}$$

We now abuse notation and write $\partial\mathcal{F}(\pi^{X|Y=y})$ for a particular element of the subdifferential and similarly for $\partial\mathcal{F}(\pi^{X|Y=\bar{y}})$. Then, $\pi^{X|Y=y}$-a.s.,

$$\begin{aligned} \|T - \mathrm{id}\|^2 = \|\bar{y} - y\|^2 - 2h\,\langle \partial\mathcal{F}(\pi^{X|Y=\bar{y}}) \circ T - \partial\mathcal{F}(\pi^{X|Y=y}), T - \mathrm{id}\rangle \\ - h^2\,\|\partial\mathcal{F}(\pi^{X|Y=\bar{y}}) \circ T - \partial\mathcal{F}(\pi^{X|Y=y})\|^2\,. \end{aligned}$$

We now integrate w.r.t. $\pi^{X|Y=y}$ and apply strong convexity of $\mathcal{F}$ in Wasserstein space:

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \le \|y-\bar{y}\|^2 - 2\alpha h\, W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) - \alpha^2 h^2\, W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}})$$

and hence

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \le \frac{1}{(1+\alpha h)^2}\,\|y-\bar{y}\|^2\,. \qquad \square$$

The point of the second proof is that, although it uses some heavy machinery, it is just a translation of a Euclidean optimization proof into the language of Wasserstein space (see Exercise 8.2).

8.3 Simultaneous Heat Flow and Time Reversal

We now introduce two new techniques in order to further analyze the proximal sampler.

Simultaneous heat flow. The first technique is based on the observation that in going from μ_k^X to μ_k^Y, and from π^X to π^Y, we are applying the heat flow. Given any f-divergence $\mathcal{D}_f(\cdot \,\|\, \cdot)$, we will compute its time derivative when both arguments undergo simultaneous heat flow. Remarkably, the result will be almost the same as the time derivative of the f-divergence to the target along the continuous-time Langevin diffusion, in a sense to be made precise. The upshot is that the analysis of the proximal sampler closely resembles the analysis of the continuous-time Langevin diffusion.

The simultaneous heat flow calculation is inspired by [VW19], and was carried out at this level of generality in [Che+22a].

Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function with $f(1) = 0$, and let $\mathcal{D}_f$ be the associated f-divergence (see Section 1.5). We begin with a quick computation of the time derivative of the f-divergence along the Langevin diffusion.

Theorem 8.3.1. *Let $(\pi_t)_{t \geq 0}$ denote the law of the continuous-time Langevin diffusion with target π. Then, for any f-divergence $\mathcal{D}_f$, it holds that*

$$\partial_t \mathcal{D}_f(\pi_t \,\|\, \pi) = -\mathcal{J}_f(\pi_t \,\|\, \pi)\,,$$

where

$$\mathcal{J}_f(\mu \,\|\, \pi) \coloneqq \mathbb{E}_\mu \Big\langle \nabla\big(f' \circ \frac{\mu}{\pi}\big), \nabla \ln \frac{\mu}{\pi} \Big\rangle\,. \tag{8.3.2}$$

Proof. Using the Fokker–Planck equation,

$$\begin{aligned}\partial_t \mathcal{D}_f(\pi_t \,\|\, \pi) = \partial_t \int f\big(\frac{\pi_t}{\pi}\big)\,\mathrm{d}\pi = \int f'\big(\frac{\pi_t}{\pi}\big)\,\partial_t \pi_t = \int f'\big(\frac{\pi_t}{\pi}\big)\,\mathrm{div}\big(\pi_t \nabla \ln \frac{\pi_t}{\pi}\big) \\ = -\int \Big\langle \nabla\big(f' \circ \frac{\pi_t}{\pi}\big), \nabla \ln \frac{\pi_t}{\pi} \Big\rangle\,\mathrm{d}\pi_t\,.\end{aligned}$$

□

Next, we compute the time derivative of the f-divergence when both arguments simultaneously evolve according to the heat flow.

Theorem 8.3.3. *Let $(Q_t)_{t\geq 0}$ denote the standard heat semigroup. Then,*

$$\partial_t \mathcal{D}_f(\mu Q_t \parallel \pi Q_t) = -\frac{1}{2}\,\mathcal{J}_f(\mu Q_t \parallel \pi Q_t)\,,$$

where $\mathcal{J}_f$ is defined in (8.3.2).

Proof. For brevity, write $\mu_t := \mu Q_t$ and $\pi_t := \pi Q_t$. Since $\partial_t \mu_t = \frac{1}{2}\Delta\mu_t = \frac{1}{2}\operatorname{div}(\mu_t \nabla \ln \mu_t)$ and similarly for $\partial_t \pi_t$, we compute

$$\begin{aligned}
2\,\partial_t \mathcal{D}_f(\mu_t \parallel \pi_t) &= 2\,\partial_t \int f\big(\frac{\mu_t}{\pi_t}\big)\,\mathrm{d}\pi_t = 2\int f'\big(\frac{\mu_t}{\pi_t}\big)\Big(\partial_t\mu_t - \frac{\mu_t}{\pi_t}\,\partial_t\pi_t\Big) + 2\int f\big(\frac{\mu_t}{\pi_t}\big)\,\partial_t\pi_t \\
&= \int f'\big(\frac{\mu_t}{\pi_t}\big)\Big(\operatorname{div}(\mu_t\nabla\ln\mu_t) - \frac{\mu_t}{\pi_t}\operatorname{div}(\pi_t\nabla\ln\pi_t)\Big) + \int f\big(\frac{\mu_t}{\pi_t}\big)\operatorname{div}(\pi_t\nabla\ln\pi_t) \\
&= -\int \Big\langle \nabla\big(f'\circ\frac{\mu_t}{\pi_t}\big), \nabla\ln\mu_t\Big\rangle\,\mathrm{d}\mu_t + \int\Big\langle \nabla\big[f'\big(\frac{\mu_t}{\pi_t}\big)\,\frac{\mu_t}{\pi_t}\big], \nabla\ln\pi_t\Big\rangle\,\mathrm{d}\pi_t \\
&\qquad\qquad - \int\Big\langle \nabla\big(f\circ\frac{\mu_t}{\pi_t}\big), \nabla\ln\pi_t\Big\rangle\,\mathrm{d}\pi_t \\
&= -\int \Big\langle \nabla\big(f'\circ\frac{\mu_t}{\pi_t}\big), \nabla\ln\frac{\mu_t}{\pi_t}\Big\rangle\,\mathrm{d}\mu_t + \int\Big\langle \nabla\frac{\mu_t}{\pi_t}, \nabla\ln\pi_t\Big\rangle\, f'\big(\frac{\mu_t}{\pi_t}\big)\,\mathrm{d}\pi_t \\
&\qquad\qquad - \int\Big\langle \nabla\frac{\mu_t}{\pi_t}, \nabla\ln\pi_t\Big\rangle\, f'\big(\frac{\mu_t}{\pi_t}\big)\,\mathrm{d}\pi_t \\
&= -\mathcal{J}_f(\mu_t \parallel \pi_t)\,.
\end{aligned}$$

□

Although this theorem is already enough to prove new convergence results for the proximal sampler, the rates will be slightly suboptimal. The reason for this is because we have only considered one step of the proximal sampler, in which the algorithm goes from μ_k^X to μ_k^Y (and the target goes from π^X to π^Y). In order to obtain the sharp convergence rates, we also need to consider the second step, in which we go from μ_k^Y to μ_{k+1}^X (and the target returns from π^Y to π^X). For reasons that will become clear shortly, we refer to these steps as the "forwards step" and the "backwards step" respectively.

First, consider the evolution of the target along the heat semigroup $t \mapsto \pi^X Q_t$, so that at time h we arrive at π^Y. The stochastic process representation of this evolution is $\mathrm{d}Z_t = \mathrm{d}B_t$, with $Z_0 \sim \pi^X$ and $Z_h \sim \pi^Y$, thus describing the forward step. So far so good, but how should we think about the backwards step? By definition, π^X is obtained from π^Y by the relation $\pi^X = \int \pi^{X|Y=y}\,\mathrm{d}\pi^Y(y)$, but this is not as helpful because we lose the stochastic process view which allows us to apply calculus. Instead, we will think of π^X as being obtained from π^Y by the *time reversal* of the diffusion $(Z_t)_{t\in[0,h]}$.

Time reversal. We now apply the result of Section 3.3.2, which implies that the time reversal of the SDE

$$\mathrm{d}Z_t = \mathrm{d}B_t\,, \qquad Z_0 \sim \pi^X$$

is given by the SDE

$$\mathrm{d}Z_t^{\leftarrow} = \nabla \ln(\pi^X Q_{h-t})(Z_t^{\leftarrow})\,\mathrm{d}t + \mathrm{d}B_t \tag{8.3.4}$$

in the sense that if we initialize the SDE at $Z_0^{\leftarrow} = y$, then $Z_h^{\leftarrow} \sim \pi^{X|Y=y}$.

In particular, if we initialize the process at $Z_0^{\leftarrow} \sim \pi^Y$, then $Z_h^{\leftarrow} \sim \pi^X$. On the other hand, if we initialize the process at $Z_0^{\leftarrow} \sim \mu_k^Y$, then the law of $Z_h^{\leftarrow}$ is $\int \pi^{X|Y=y}\,\mathrm{d}\mu_k^Y(y) = \mu_{k+1}^X$. Thus, we have successfully exhibited a stochastic process representation which takes us from μ_k^Y to μ_{k+1}^X. For any measure μ, write $\mu Q_t^{\leftarrow}$ for the law of $Z_t^{\leftarrow}$ initialized at $Z_0^{\leftarrow} \sim \mu$.

Simultaneous backwards heat flow. Next, we will show that the time derivative of the f-divergence along the simultaneous *backwards* heat flow also behaves the same way as the simultaneous forwards heat flow. This leads to a pleasing symmetry between the forwards and backwards steps of the proximal sampler.

Theorem 8.3.5. *Let $(Q_t^{\leftarrow})_{t\in[0,h]}$ denote the construction described above by reversing the heat flow started at π^X. Then,*

$$\partial_t \mathcal{D}_f(\mu Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow}) = -\frac{1}{2}\,\mathcal{J}_f(\mu Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow})\,,$$

where $\mathcal{J}_f$ is defined in (8.3.2).

Proof. For brevity, write $\mu_t^{\leftarrow} := \mu Q_t^{\leftarrow}$ and $\pi_t^{\leftarrow} := \pi^Y Q_t^{\leftarrow}$. By construction of the reversed process, $\pi_t^{\leftarrow} = \pi^X Q_{h-t}$. Then, by the Fokker–Planck equation,

$$\begin{aligned}
\partial_t \pi_t^{\leftarrow} &= -\operatorname{div}(\pi_t^{\leftarrow} \nabla \ln \pi_t^{\leftarrow}) + \frac{1}{2}\,\Delta \pi_t^{\leftarrow} = -\frac{1}{2}\,\Delta \pi_t^{\leftarrow}\,,\\
\partial_t \mu_t^{\leftarrow} &= -\operatorname{div}(\mu_t^{\leftarrow} \nabla \ln \pi_t^{\leftarrow}) + \frac{1}{2}\,\Delta \mu_t^{\leftarrow} = \operatorname{div}\big(\mu_t^{\leftarrow} \nabla \ln \frac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big) - \frac{1}{2}\,\Delta \mu_t^{\leftarrow}\,.
\end{aligned}$$

Note that the fact that $(\pi_t^{\leftarrow})_{t\in[0,h]}$ satisfies the *backwards* heat equation is completely natural in light of our construction via the reversed process.

Hence, we compute

$$2\,\partial_t \mathcal{D}_f(\mu_t^{\leftarrow} \parallel \pi_t^{\leftarrow}) = 2\int f'\big(\frac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\Big(\partial_t \mu_t^{\leftarrow} - \frac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\,\partial_t \pi_t^{\leftarrow}\Big) + 2\int f\big(\frac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\,\partial_t \pi_t^{\leftarrow}$$

$$
\begin{aligned}
&= \int f'\big(\tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\Big(2\operatorname{div}\big(\mu_t^{\leftarrow}\nabla \ln \tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big) - \Delta\mu_t^{\leftarrow} + \tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\Delta\pi_t^{\leftarrow}\Big) - \int f\big(\tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\Delta\pi_t^{\leftarrow} \\
&= 2\int f'\big(\tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\operatorname{div}\big(\mu_t^{\leftarrow}\nabla \ln \tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big) \\
&\qquad \underbrace{- \Big[\int f'\big(\tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\Big(\Delta\mu_t^{\leftarrow} - \tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\Delta\pi_t^{\leftarrow}\Big) + \int f\big(\tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big)\Delta\pi_t^{\leftarrow}\Big]}_{(\star)} .
\end{aligned}
$$

The term $(\star)$ is exactly the same kind of term we encountered in the proof of Theorem 8.3.3, and by the same calculations it equals $-\mathcal{J}_f(\mu_t^{\leftarrow} \parallel \pi_t^{\leftarrow})$. Therefore,

$$
\begin{aligned}
2\,\partial_t \mathcal{D}_f(\mu_t^{\leftarrow} \parallel \pi_t^{\leftarrow}) &= -2\int \Big\langle \nabla\big(f' \circ \tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}}\big), \nabla \ln \tfrac{\mu_t^{\leftarrow}}{\pi_t^{\leftarrow}} \Big\rangle \mathrm{d}\mu_t^{\leftarrow} + \mathcal{J}_f(\mu_t^{\leftarrow} \parallel \pi_t^{\leftarrow}) \\
&= -\mathcal{J}_f(\mu_t^{\leftarrow} \parallel \pi_t^{\leftarrow}) .
\end{aligned}
$$

□

8.4 Convergence under Log-Concavity

Next, we present a convergence proof for the proximal sampler under log-concavity, following [Che+22a]. The proof can be compared to the $1/t$ convergence rate for the Langevin diffusion under log-concavity (1.4.12), which was obtained via a Lyapunov function argument.

Theorem 8.4.1. *Assume that the target π^X is log-concave. Then, for the law μ_k^X of the k-th iterate of the proximal sampler,*

$$
\mathsf{KL}(\mu_k^X \parallel \pi^X) \le \frac{W_2^2(\mu_0^X, \pi^X)}{kh} .
$$

Proof. **Forwards step**. Along the simultaneous heat flow, Theorem 8.3.3 shows that

$$
\partial_t \,\mathsf{KL}(\mu_0^X Q_t \parallel \pi^X Q_t) = -\frac{1}{2}\,\mathsf{FI}(\mu_0^X Q_t \parallel \pi^X Q_t)
$$

so we need to lower bound the Fisher information. Also, log-concavity is preserved by convolution, so $\pi^X Q_t$ is log-concave. [TODO: Justify this fact.] Hence, by convexity of the KL divergence to a log-concave target along Wasserstein geodesics (Theorem 1.4.5),

$$
0 = \mathsf{KL}(\pi^X Q_t \parallel \pi^X Q_t) \ge \mathsf{KL}(\mu_0^X Q_t \parallel \pi^X Q_t) + \mathbb{E}_{\mu_0^X Q_t}\Big\langle \nabla \ln \frac{\mu_0^X Q_t}{\pi^X Q_t}, T_{\mu_0^X Q_t \to \pi^X Q_t} - \mathrm{id} \Big\rangle .
$$

Rearranging this and using the Cauchy–Schwarz inequality,

$$\underbrace{\mathbb{E}_{\mu_0^X Q_t}\left[\left\|\nabla \ln \frac{\mu_0^X Q_t}{\pi^X Q_t}\right\|^2\right]}_{\mathsf{FI}(\mu_0^X Q_t \| \pi^X Q_t)} W_2^2(\mu_0^X Q_t, \pi^X Q_t) \geq \mathsf{KL}(\mu_0^X Q_t \parallel \pi^X Q_t)^2 .$$

Combining this with the fact that the Wasserstein distance is decreasing along the simultaneous heat flow,

$$\partial_t \, \mathsf{KL}(\mu_0^X Q_t \parallel \pi^X Q_t) \leq -\frac{1}{2} \frac{\mathsf{KL}(\mu_0^X Q_t \parallel \pi^X Q_t)^2}{W_2^2(\mu_0^X, \pi^X)} .$$

Solving this differential inequality,

$$\frac{1}{\mathsf{KL}(\mu_0^Y \parallel \pi^Y)} = \frac{1}{\mathsf{KL}(\mu_0^X Q_h \parallel \pi^X Q_h)} \geq \frac{1}{\mathsf{KL}(\mu_0^X \parallel \pi^X)} + \frac{h}{2\, W_2^2(\mu_0^X, \pi^X)} .$$

Backwards step. Along the simultaneous backwards heat flow, Theorem 8.3.5 gives

$$\partial_t \, \mathsf{KL}(\mu_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow}) = -\frac{1}{2} \mathsf{FI}(\mu_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow}) .$$

Since $\pi^Y Q_t^{\leftarrow} = \pi^X Q_{h-t}$ is log-concave and $t \mapsto W_2^2(\mu_0^Y Q_t^{\leftarrow}, \pi^Y Q_t^{\leftarrow})$ is decreasing (which is checked via a coupling argument using the diffusion (8.3.4)), a similar calculation as the forwards step leads to the inequality

$$\frac{1}{\mathsf{KL}(\mu_1^X \parallel \pi^X)} = \frac{1}{\mathsf{KL}(\mu_0^Y Q_h^{\leftarrow} \parallel \pi^Y Q_h^{\leftarrow})} \geq \frac{1}{\mathsf{KL}(\mu_0^Y \parallel \pi^Y)} + \frac{h}{2\, W_2^2(\mu_0^X, \pi^X)} .$$

We iterate these inequalities, using the fact that $W_2(\mu_k^X, \pi^X) \leq W_2(\mu_0^X, \pi^X)$ for all $k \in \mathbb{N}$ (which follows from Theorem 8.2.1) to obtain

$$\frac{1}{\mathsf{KL}(\mu_k^X \parallel \pi^X)} \geq \frac{1}{\mathsf{KL}(\mu_0^X \parallel \pi^X)} + \frac{kh}{W_2^2(\mu_0^X, \pi^X)}$$

or

$$\mathsf{KL}(\mu_k^X \parallel \pi^X) \leq \frac{\mathsf{KL}(\mu_0^X \parallel \pi^X)}{1 + kh\, \mathsf{KL}(\mu_0^X \parallel \pi^X)/W_2^2(\mu_0^X, \pi^X)} \leq \frac{W_2^2(\mu_0^X, \pi^X)}{kh} . \qquad \square$$

8.5 Convergence under Functional Inequalities

We now prove convergence guarantees for the proximal sampler when the target satisfies either a Poincaré inequality or a log-Sobolev inequality, following [Che+22a].

Theorem 8.5.1. *Suppose that the target π^X satisfies a Poincaré inequality with constant C_{PI}. Then, for the law μ_k^X of the k-th iterate of the proximal sampler,*

$$\chi^2(\mu_k^X \parallel \pi^X) \le \frac{\chi^2(\mu_0^X \parallel \pi^X)}{(1 + h/C_{\mathsf{PI}})^{2k}}.$$

Proof. **Forwards step**. For the chi-squared divergence, we can check that the dissipation functional is given by $\mathcal{J}(\mu \parallel \pi) = 2\,\mathbb{E}_\pi[\|\nabla(\mu/\pi)\|^2]$. Along the simultaneous heat flow, by Theorem 8.3.3,

$$\partial_t \chi^2(\mu_0^X Q_t \parallel \pi^X Q_t) = -\mathcal{J}(\mu_0^X Q_t \parallel \pi^X Q_t) = -\,\mathbb{E}_{\pi^X Q_t}\left[\left\|\nabla \frac{\mu_0^X Q_t}{\pi^X Q_t}\right\|^2\right].$$

Since π^X satisfies a Poincaré inequality with constant C_{PI}, by subadditivity of the Poincaré constant under convolution (Proposition 2.3.7), $\pi^X Q_t$ satisfies a Poincaré inequality with constant at most $C_{\mathsf{PI}} + t$. It therefore yields

$$\partial_t \chi^2(\mu_0^X Q_t \parallel \pi^X Q_t) \le -\frac{1}{C_{\mathsf{PI}} + t} \operatorname{var}_{\pi^X Q_t} \frac{\mu_0^X Q_t}{\pi^X Q_t} = -\frac{1}{C_{\mathsf{PI}} + t}\, \chi^2(\mu_0^X Q_t \parallel \pi^X Q_t)$$

and hence

$$\chi^2(\mu_0^Y \parallel \pi^Y) = \chi^2(\mu_0^X Q_h \parallel \pi^X Q_h) \le \exp\left(-\int_0^h \frac{1}{C_{\mathsf{PI}} + t}\,\mathrm{d}t\right) \chi^2(\mu_0^X \parallel \pi^X) = \frac{\chi^2(\mu_0^X \parallel \pi^X)}{1 + h/C_{\mathsf{PI}}}.$$

Backwards step. Along the simultaneous backwards heat flow, Theorem 8.3.5 yields

$$\partial_t \chi^2(\mu_0^Y Q_t^{\leftarrow} \parallel \pi^Y Q_t^{\leftarrow}) = -\,\mathbb{E}_{\pi^Y Q_t^{\leftarrow}}\left[\left\|\nabla \frac{\mu_0^Y Q_t^{\leftarrow}}{\pi^Y Q_t^{\leftarrow}}\right\|^2\right].$$

Using the fact that $\pi^Y Q_t^{\leftarrow} = \pi^X Q_{h-t}$ satisfies the Poincaré inequality with constant at most $C_{\mathsf{PI}} + h - t$, we deduce similarly that

$$\chi^2(\mu_1^X \parallel \pi^X) = \chi^2(\mu_0^Y Q_h^{\leftarrow} \parallel \pi^Y Q_h^{\leftarrow}) \le \frac{\chi^2(\mu_0^Y \parallel \pi^Y)}{1 + h/C_{\mathsf{PI}}}.$$

Iterating this pair of inequalities yields the result. □

A similar result holds for the log-Sobolev inequality; since the proof is entirely analogous, we leave it as Exercise 8.5.

Theorem 8.5.2. *Suppose that the target π^X satisfies a log-Sobolev inequality with constant C_{LSI}. Then, for the law μ_k^X of the k-th iterate of the proximal sampler,*

$$\mathsf{KL}(\mu_k^X \parallel \pi^X) \le \frac{\mathsf{KL}(\mu_0^X \parallel \pi^X)}{(1 + h/C_{\mathsf{LSI}})^{2k}} .$$

Recall that if π^X is α-strongly log-concave, then it satisfies a log-Sobolev inequality with constant $C_{\mathsf{LSI}} \le 1/\alpha$ (Theorem 1.2.29). Thus, the contraction factor of $\frac{1}{(1+\alpha h)^2}$ in KL divergence matches the contraction factor in W_2^2 distance (Theorem 8.2.1). To get this sharp result, it is necessary to utilize the backwards step.

Similarly to Theorem 2.2.15, it is also possible to obtain guarantees for Rényi divergences, see Exercise 8.6.

***Remark* 8.5.3.** It is a curious observation that in the W_2 guarantee of Theorem 8.2.1, the contraction factor of $\frac{1}{(1+\alpha h)^2}$ occurs solely in the backwards step, whereas in Theorem 8.5.2 the forwards and backwards steps each contribute a contraction factor of $\frac{1}{1+\alpha h}$.

8.6 Applications

The original application of the proximal sampler was for sampling from certain families of structured log-concave distributions [LST21c]. Since then, the proximal sampler has been used to provide new guarantees for non-smooth and weakly smooth potentials [GLL22; LC22a; LC22b]. We will restrict ourselves to applications which are more or less immediate corollaries of our present analysis.

New guarantees for sampling from smooth potentials. When the potential V is β-smooth, as discussed in Section 8.1, the RGO can be implemented via rejection sampling. We obtain the following corollaries.

Corollary 8.6.1. *Let $\pi^X \propto \exp(-V)$, where V is β-smooth. Take $h = \frac{1}{\beta d}$ and assume we have an oracle to V which evaluates V and the proximal operator for V. Let μ_N^X denote the law of the N-th iterate of the proximal sampler, in which the RGO is implemented via rejection sampling.*

1. *(Theorem 8.2.1) If in addition V is α-strongly convex with $\alpha > 0$, then writing $\kappa := \beta/\alpha$ we obtain $\sqrt{\alpha}\, W_2(\mu_N^X, \pi^X) \le \varepsilon$ using $O(\kappa d \log \frac{W_2(\mu_0^X, \pi^X)}{\varepsilon})$ queries to the oracle in expectation.*

2. *(Theorem 8.4.1) If V is convex, we obtain the guarantee $\sqrt{\mathsf{KL}(\mu_N^X \parallel \pi^X)} \le \varepsilon$ using $O(\frac{\beta d\, W_2^2(\mu_0^X, \pi^X)}{\varepsilon^2})$ queries to the oracle in expectation.*

3. *(Theorem 8.5.1) If V satisfies a Poincaré inequality with constant C_{PI}, we obtain the guarantee $\sqrt{\chi^2(\mu_N^X \parallel \pi^X)} \le \varepsilon$ using $O(C_{\mathsf{PI}} \beta d \log \frac{\chi^2(\mu_0^X \| \pi)}{\varepsilon^2})$ queries to the oracle in expectation.*

4. *(Theorem 8.5.2) If V satisfies a log-Sobolev inequality with constant C_{LSI}, we obtain the guarantee $\sqrt{\mathsf{KL}(\mu_N^X \parallel \pi^X)} \le \varepsilon$ using $O(C_{\mathsf{LSI}} \beta d \log \frac{\mathsf{KL}(\mu_0^X \| \pi)}{\varepsilon^2})$ queries to the oracle in expectation.*

Note that for the strongly log-concave case, these results are competitive with the state-of-the-art results for MALA under a feasible start (Theorem 7.3.4)!

Improving the condition number dependence of high-accuracy samplers. The next application we present is the original use of the proximal sampler in [LST21c]. Namely, suppose that $\pi^X \propto \exp(-V)$ is such that $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ with condition number $\kappa := \beta/\alpha$. Suppose we have a high-accuracy sampler which, given any target satisfying these assumptions, outputs a sample from a probability measure μ with $\|\mu - \pi^X\|_{\mathrm{TV}} \le \varepsilon$ using $\widetilde{O}(f(\kappa)\, d^c \operatorname{polylog}(1/\varepsilon))$ queries, where $f : \mathbb{R}_+ \to \mathbb{R}_+$ is some increasing function. Then, by combining this high-accuracy sampler with the proximal sampler, we can obtain a new sampler whose complexity is only $\widetilde{O}(\kappa d^c \operatorname{polylog}(\kappa/\varepsilon))$, i.e., we have improved the dependence on the condition number to near linear.

To see how this works, observe that if we choose the step size $h = \frac{1}{\beta}$ for the proximal sampler, then the RGO $\pi^{X|Y=y}$ has condition number $O(1)$. Thus, the high-accuracy sampler can obtain a δ-approximate sample from the RGO using $\widetilde{O}(d^c \operatorname{polylog}(1/\delta))$ queries. On the other hand, with this choice of step size, we know from Theorem 8.5.2 that with a *perfect* implementation of the RGO the number of iterations required for the proximal sampler to output μ_N^X with $\|\mu_N^X - \pi\|_{\mathrm{TV}} \le \varepsilon$ is $\widetilde{O}(\kappa \log(d/\varepsilon^2))$. To complete the analysis, we need to analyze how the error propagates due to the imperfect implementation of the RGO. This is handled via a coupling argument (Exercise 8.8).

Lemma 8.6.2. *Let μ_N^X denote the law of the N-th iterate of the proximal sampler with perfect implementation of the RGO. Suppose that instead, in each step of the proximal sampler, we use a sample from a distribution which is δ-close to the RGO in total variation distance; let $\hat{\mu}_N^X$ denote the law of the N-th iterate of the proximal sampler with imperfect implementation of the RGO. Then,*

$$\|\hat{\mu}_N^X - \mu_N^X\|_{\mathrm{TV}} \le N\delta\,.$$

Since $N = \widetilde{O}(\kappa \log(d/\varepsilon^2))$, we can take $\delta \asymp \varepsilon/N$. The total complexity of the proximal sampler (the number of iterations N of the proximal sampler multiplied by the cost of approximately implementing the RGO with the high-accuracy sampler) is then $\widetilde{O}(\kappa d^c \operatorname{polylog}(1/\varepsilon))$ as claimed.

In particular, applying this to the Metropolized random walk (MRW) algorithm (Theorem 7.3.4) improves the complexity from $\widetilde{O}(\kappa^2 d \operatorname{polylog}(1/\varepsilon))$ to $\widetilde{O}(\kappa d \operatorname{polylog}(1/\varepsilon))$.

Zeroth-order algorithms for sampling. The example above shows that boosting the MRW algorithm with the proximal sampler leads to an algorithm whose complexity is competitive with that of MALA. Moreover, unlike MALA, the algorithm based on MRW only uses zeroth-order information, which is crucial for certain applications such as Bayesian inverse problems in which gradient information is prohibitively expensive.

Similarly, implementing the RGO using rejection sampling only uses zeroth-order information, except possibly for computing the minimizer of the potential V_y.

Lack of discretization analysis. Finally, we mention that the results in Corollary 8.6.1 are state-of-the-art under the various assumptions. A key reason why the proximal sampler yields powerful complexity guarantees is because there is no "discretization analysis". For example, consider the sampling from a target distribution satisfying a Poincaré inequality. Since a Poincaré inequality implies convergence in chi-squared divergence, it is natural to perform a χ^2 analysis of LMC, but this leads to substantial new technical hurdles (see Chapter 6). Moreover, under a Poincaré inequality it becomes non-trivial even to prove moment bounds for the LMC iterates. All of this is handled via a careful analysis in [Che+21a], but the results there have worse dependence on d, $C_{\mathsf{PI}}\beta$, and ε^{-1}. In contrast, Corollary 8.6.1 bypasses all of these difficulties because the proximal sampler reduces the task of sampling from distributions satisfying a Poincaré inequality to the task of sampling from strongly log-concave distributions for the implementation of the RGO, and even this is made straightforward via rejection sampling provided that we take a small enough step size for the proximal sampler.

Bibliographical Notes

The reader is encouraged to read the original paper [LST21b] on the proximal sampler, which contains applications to sampling from composite densities $\pi \propto \exp(-(f+g))$, where f is well-conditioned and g admits an implementable RGO, as well as to sampling from log-concave finite sums $\pi \propto \exp(-F)$ where $F \coloneqq n^{-1}\sum_{i=1}^n f_i$ is well-conditioned and the complexity is measured via the number of oracle calls to the individual functions $(f_i)_{i\in[n]}$. The proximal sampler has also been used to sample from weakly smooth and non-smooth potentials [LC22a; LC22b], and it has been applied to the problem of differentially private convex optimization [GLL22].

The optimization results in Exercise 8.2 obtained in analogy with the proximal sampler are given in [Che+22a].

Exercises

Introduction to the Proximal Sampler

⊵ Exercise 8.1 (RGO as a proximal operator on the Wasserstein space)
Given a functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}\cup\{\infty\}$, the proximal operator for $\mathcal{F}$ on the Wasserstein space is defined via

$$\mathrm{prox}_{\mathcal{F}}(\mu) \coloneqq \operatorname*{arg\,min}_{\mu'\in\mathcal{P}_2(\mathbb{R}^d)} \left\{\mathcal{F}(\mu') + \frac{1}{2}W_2^2(\mu,\mu')\right\}.$$

The proximal operator was used in the seminal work [JKO98] in order to rigorously make sense of gradient flows on the Wasserstein space. Prove that the RGO satisfies

$$\pi^{X|Y=y} = \mathrm{prox}_{h\,\mathsf{KL}(\cdot\|\pi)}(\delta_y).$$

Hence, the assumption that we can implement the RGO is the same as assuming that we can evaluate the proximal operator for the KL divergence on any Dirac measure.

Convergence under Strong Log-Concavity

⊵ Exercise 8.2 (comparison with optimization results)
This exercise compares the results for the proximal sampler with the proximal point method in optimization.

1. Suppose that V is α-strongly convex. Prove that prox_{hV} is $\frac{1}{1+\alpha h}$-Lipschitz.

   Hint: Show that $\mathrm{prox}_{hV} = (\mathrm{id} + h\nabla V)^{-1}$. Argue via convex duality by considering the convex conjugate $(\frac{\|\cdot\|^2}{2} + hV)^*$.

2. Suppose that V is α-strongly convex. Translate both of the proofs in Section 8.2 to Euclidean optimization.

3. Suppose that V satisfies the gradient domination condition

$$\|\nabla V(x)\|^2 \geq 2\alpha \{V(x) - \inf V\}, \qquad \text{for all } x \in \mathbb{R}^d.$$

Also, let $x' := \mathrm{prox}_{hV}(x)$. Inspired by Theorem 8.5.2, we can ask whether or not it holds that

$$V(x') - \inf V \leq \frac{1}{(1+\alpha h)^2} \{V(x) - \inf V\}.$$

Prove that this is indeed the case.

Hint: Define $V_{t,x}(z) := V(z) + \frac{1}{2t}\|z-x\|^2$ and let $x_t := \arg\min V_{t,x}$; then, differentiate $t \mapsto V_{t,x}(x_t)$.

▷ **Exercise 8.3** (first coupling lemma)
Prove Lemma 8.2.2.

Simultaneous Heat Flow and Time Reversal

▷ **Exercise 8.4** (non-negativity of the dissipation functional)
By the data-processing inequality, the f-divergence to the target is always decreasing along the Langevin diffusion and hence the functional $\mathcal{J}_f$ defined in (8.3.2) is always non-negative. Prove this more directly from the expression for $\mathcal{J}_f$.

▷ **Exercise 8.5** (convergence under LSI)
Verify that the result under LSI (Theorem 8.5.2) holds.

▷ **Exercise 8.6** (convergence in Rényi divergence)
In [VW19], Vempala and Wibisono showed convergence of the Langevin diffusion in Rényi divergence under a Poincaré or log-Sobolev inequality (see Theorem 2.2.15). Similarly, extend Theorem 8.5.1 and Theorem 8.5.2 to provide Rényi divergence guarantees.

▷ **Exercise 8.7** (Gaussian case)
In this exercise, we consider the Gaussian case for intuition.

1. Suppose that $\pi^X = \mathsf{normal}(0, I_d)$ and $\mu_0^X = \mathsf{normal}(0, \sigma_0^2 I_d)$. Show that the iterates of the proximal sampler all have Gaussian distributions, and explicitly compute the variances. Use this to show that the contraction factors in Theorem 8.2.1 and Theorem 8.5.2 are sharp.

2. Next, suppose that $\pi^X = \text{normal}(0, \Sigma)$ and that $\mu_0^X = \text{normal}(m_0, \Sigma_0)$. Show that the next iterate of the proximal sampler is $\mu_1^X = \text{normal}(m_1, \Sigma_1)$, where the mean satisfies $m_1 = \text{prox}_{hV}(m_0)$ and $V(x) := \frac{1}{2}\langle x, \Sigma^{-1} x\rangle$. In other words, the mean of the iterate of the proximal sampler evolves according to the proximal point method.

Applications

▷ **Exercise 8.8** (second coupling lemma)
Prove Lemma 8.6.2.

CHAPTER 9

Lower Bounds for Sampling

In order to determine if our sampling guarantees are optimal, we need to pair them with lower bounds. However, the problem of establishing query complexity lower bounds for sampling is challenging and the work on this topic is nascent. Here, we will give an overview of the current progress in this direction.

9.1 A Query Complexity Result in One Dimension

In this section, we follow [Che+22b], which established a sharp query complexity result for sampling strongly log-concave distributions in one dimension. Define the class

$$\Pi_\kappa := \{\pi \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}) \mid \pi \propto \exp(-V),\ 1 \le V'' \le \kappa,\ V'(0) = 0\}\,.$$

In applications of sampling, one may first need to use an optimization algorithm to find the minimizer of V before applying the sampling algorithm. In our definition of Π_κ, however, we have enforced the requirement $V'(0) = 0$ in order to cleanly separate out the complexity of optimization (finding the minimizer of V) from the intrinsic complexity of sampling. Our goal is to understand the minimum number of queries required by an algorithm to output an approximate sample from any target $\pi \in \Pi_\kappa$.

Theorem 9.1.1 ([Che+22b]). *The query complexity of outputting a sample which is $\frac{1}{64}$ close in total variation distance to the target π, uniformly over the choice of $\pi \in \Pi_\kappa$, is*

$\Theta(\log\log\kappa)$.

In what follows, we will make this theorem more precise and give a proof.

Lower bound. The lower bound will hold for any local oracle. Loosely speaking, a local oracle accepts as an input a point $x \in \mathbb{R}$ and outputs some information about the target π such that if $\hat{\pi}$ is another possible target and $\pi \propto \hat{\pi}$ in some neighborhood of x, then the output of the oracle is the same for both π and $\hat{\pi}$. This just formalizes the idea that the oracle only outputs information about π "near the point x". To simplify the discussion, however, we will suppose for concreteness that we have access to a second-order oracle: given $x \in \mathbb{R}^d$, it outputs the triple $(V(x), V'(x), V''(x))$, where we recall that V is only specified up to an additive constant. (If this is confusing, you may instead suppose that the oracle outputs the triple $(V(x) - V(0), V'(x), V''(x))$ where $\pi = \exp(-V)$.)

The lower bound will proceed in two stages.

1. First, we reduce the sampling problem to a statistical testing problem. Namely, we will construct a family $\pi_1, \dots, \pi_m \in \Pi_\kappa$, and suppose that $\boldsymbol{i} \sim \text{uniform}([m])$ is drawn randomly. The statistical testing problem is defined as follows: given query access to π_i (through the oracle), guess the value of $\boldsymbol{i}$.

   We will show that an algorithm to sample from $\pi_{\boldsymbol{i}}$ can be used to solve the statistical testing problem; thus, "sampling is harder than testing".

2. Next, we will prove a lower bound on the number of queries required to solve the statistical testing problem: "testing is hard". This relies on standard information-theoretic techniques for proving minimax lower bounds for statistical problems. The main difference between this problem and the usual statistical setting is that rather than having i.i.d. samples from some data distribution, we instead have query access and the algorithm is allowed to be adaptive.

Combining the two steps then yields our query complexity lower bound for sampling. We begin with the construction of $\pi_1, \dots, \pi_m$, which is slightly tricky.

Let m be the largest integer such that $\exp(-\frac{2^{2m-2}}{2\kappa}) \geq \frac{1}{2}$ (and note that $m = \Theta(\log\kappa)$). We define two auxiliary functions

$$\phi(x) := \begin{cases} \kappa, & \frac{1}{2} \leq x < 1, \\ 1, & 1 \leq x < 2, \\ \kappa, & 2 \leq x < \frac{5}{2}, \\ 0, & \text{otherwise}, \end{cases} \qquad \psi(x) := \begin{cases} 1, & \frac{5}{2} \leq x < 4, \\ \kappa, & 4 \leq x < 5, \\ 0, & \text{otherwise}. \end{cases}$$

We define a family $(V_i)_{i\in[m]}$ of 1-strongly convex and κ-smooth potentials as follows. We require that $V_i(0) = V_i'(0) = 0$ and that V_i be an even function, so it suffices to specify V_i'' on $\mathbb{R}_+$. The second derivative is given by

$$V_i''(x) := \mathbb{1}\{x \le \kappa^{-\frac{1}{2}} 2^{i-1}\} + \phi\Big(\frac{x}{\kappa^{-\frac{1}{2}} 2^i}\Big) + \sum_{j=i}^{m-1} \psi\Big(\frac{x}{\kappa^{-\frac{1}{2}} 2^j}\Big) + \mathbb{1}\{x \ge 5\kappa^{-\frac{1}{2}} 2^{m-1}\}, \qquad x \ge 0\,.$$

Observe that all of the terms in the above summation have disjoint supports. Although the construction seems complicated, the basic idea is to make V_i'' oscillate between its minimum and maximum allowable values 1 and κ; see Figure 9.1 for a visual.

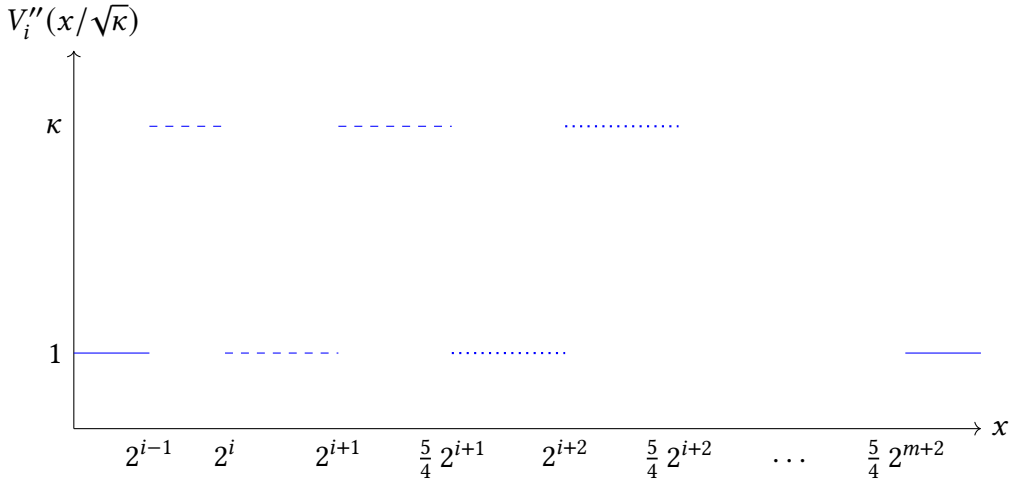

Figure 9.1: The dashed lines correspond to ϕ and the dotted lines correspond to ψ. Here, the horizontal axis is distorted for clarity.

There are two key properties of this construction. First, we will show in Lemma 9.1.2 that each π_i places a substantial amount of mass on the interval $(\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}]$. This implies that if we can sample from π_i, it is likely that the sample will land in this interval, which is used to reduce the sampling task to the statistical testing task. Then, we will show in Lemma 9.1.4 that V_i and V_{i+1} agree exactly outside of a small interval which is approximately located at $\kappa^{-\frac{1}{2}} 2^i$. This implies that for any given value $x \in \mathbb{R}^d$, there are only $O(1)$ possible values of $(V_i(x), V_i'(x), V_i''(x))$ as i ranges in $[m]$, which in turn will be used to show that the oracle is not very informative (and hence prove a lower bound for the statistical testing task).

The intuition behind the following lemma is that at $\kappa^{-\frac{1}{2}} 2^{i-1}$, $V_i'' = \kappa$ for the first time and so the density π_i drops off rapidly after this point.

Lemma 9.1.2. *For each $i \in [m]$,*

$$\pi_i\big((\kappa^{-\frac{1}{2}}\,2^{i-2}, \kappa^{-\frac{1}{2}}\,2^{i-1}]\big) \geq \frac{1}{32}\,.$$

Proof. According to the definition of π_i, we have

$$\pi_i\big((\kappa^{-\frac{1}{2}}\,2^{i-2}, \kappa^{-\frac{1}{2}}\,2^{i-1}]\big) = \frac{\int_{\kappa^{-\frac{1}{2}}\,2^{i-2}}^{\kappa^{-\frac{1}{2}}\,2^{i-1}} \exp(-x^2/2)\,\mathrm{d}x}{Z_{\pi_i}}\,, \qquad Z_{\pi_i} \coloneqq \int \exp(-V_i)\,.$$

Recalling that m is chosen so that $\exp(-x^2/2) \geq 1/2$ whenever $|x| \leq \kappa^{-\frac{1}{2}}\,2^{m-1}$,

$$\int_{\kappa^{-\frac{1}{2}}\,2^{i-2}}^{\kappa^{-\frac{1}{2}}\,2^{i-1}} \exp\big(-\frac{x^2}{2}\big)\,\mathrm{d}x \geq \frac{1}{2}\,\kappa^{-\frac{1}{2}}\,2^{i-2}\,.$$

For the normalizing constant, observe that

$$\int_0^\infty \exp(-V_i) = \int_0^{\kappa^{-\frac{1}{2}}\,2^i} \exp(-V_i) + \int_{\kappa^{-\frac{1}{2}}\,2^i}^\infty \exp(-V_i) \leq \kappa^{-\frac{1}{2}}\,2^i + \int_{\kappa^{-\frac{1}{2}}\,2^i}^\infty \exp(-V_i)\,.$$

Since $V_i'' = \kappa$ on $[\kappa^{-\frac{1}{2}}\,2^{i-1}, \kappa^{-\frac{1}{2}}\,2^i]$, it follows that $V_i'(\kappa^{-\frac{1}{2}}\,2^i) \geq \kappa^{-\frac{1}{2}}\,2^{i-1}$, and so

$$V_i(x) \geq \kappa^{-\frac{1}{2}}\,2^{i-1}\,(x - \kappa^{-\frac{1}{2}}\,2^i) + \frac{(x - \kappa^{-\frac{1}{2}}\,2^i)^2}{2}\,, \qquad x \geq \kappa^{-\frac{1}{2}}\,2^i\,.$$

Therefore,

$$\begin{aligned}\int_{\kappa^{-\frac{1}{2}}\,2^i}^\infty \exp(-V_i) &\leq \int_{\kappa^{-\frac{1}{2}}\,2^i}^\infty \exp\Big(-\kappa^{-\frac{1}{2}}\,2^{i-1}\,(x - \kappa^{-\frac{1}{2}}\,2^i) - \frac{(x - \kappa^{-\frac{1}{2}}\,2^i)^2}{2}\Big)\,\mathrm{d}x \\ &\leq \frac{1}{\kappa^{-\frac{1}{2}}\,2^{i-1}} \leq \frac{1}{\sqrt{\kappa}}\,,\end{aligned}$$

where we applied a standard tail estimate for Gaussian densities (Lemma 9.1.3). Then,

$$\pi_i\big((\kappa^{-\frac{1}{2}}\,2^{i-2}, \kappa^{-\frac{1}{2}}\,2^{i-1}]\big) \geq \frac{2^{i-3}}{2\,(2^i+1)} \geq \frac{1}{32}\,,$$

which proves the result. □

In the above proof, we used the following lemma (see Exercise 9.1).

Lemma 9.1.3. *Let $a, x_0 > 0$. Then,*

$$\int_{x_0}^{\infty} \exp\bigl(-a\,(x - x_0) - \frac{1}{2}\,(x - x_0)^2\bigr)\,\mathrm{d}x \le \frac{1}{a}\,.$$

The next lemma is the main reason why we used an oscillating construction for V_i''.

Lemma 9.1.4. *We have the equalities*

$$V_i = V_{i+1}\,, \qquad V_i' = V_{i+1}'\,, \qquad V_i'' = V_{i+1}''\,,$$

outside of the set $\{x \in \mathbb{R} : \kappa^{-\frac{1}{2}}\,2^{i-1} \le |x| \le \frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\}$.

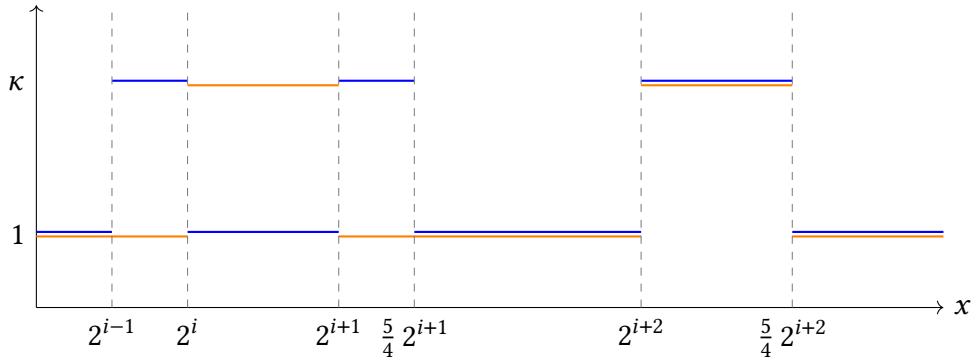

Figure 9.2: We plot $V_i''(x)$ (in blue) and $V_{i+1}''(x)$ (in orange). In this figure, we do not distort the horizontal axis lengths to make it easier to visually compare the relative lengths of intervals on which the second derivatives are constant.

Proof. Refer to Figure 9.2 for a visual aid for the proof.

Clearly the potentials and derivatives match when $|x| \le \kappa^{-\frac{1}{2}}\,2^{i-1}$. Since the second derivatives match when $|x| \ge \frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}$, it suffices to show that

$$V_i'\bigl(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\bigr) = V_{i+1}'\bigl(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\bigr) \qquad \text{and} \qquad V_i\bigl(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\bigr) = V_{i+1}\bigl(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\bigr)\,.$$

To that end, note that for $x \ge 0$,

$$V_{i+1}''(x) - V_i''(x) = \mathbb{1}\{\kappa^{-\frac{1}{2}}\,2^{i-1} < x \le \kappa^{-\frac{1}{2}}\,2^{i}\} - \phi\bigl(\frac{x}{\kappa^{-\frac{1}{2}}\,2^{i}}\bigr) + \phi\bigl(\frac{x}{\kappa^{-\frac{1}{2}}\,2^{i+1}}\bigr) - \psi\bigl(\frac{x}{\kappa^{-\frac{1}{2}}\,2^{i}}\bigr)$$

$$= \begin{cases} -(\kappa-1)\,, & \kappa^{-\frac{1}{2}}\,2^{i-1} \le x \le \kappa^{-\frac{1}{2}}\,2^{i}\,, \\ +(\kappa-1)\,, & \kappa^{-\frac{1}{2}}\,2^{i} \le x \le \kappa^{-\frac{1}{2}}\,2^{i+1}\,, \\ -(\kappa-1)\,, & \kappa^{-\frac{1}{2}}\,2^{i+1} \le x \le \frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

A little algebra shows that the above expression integrates to zero, hence we deduce the equality $V_i'(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}) = V_{i+1}'(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1})$. Also, by integrating this expression twice,

$$\begin{aligned} V_{i+1}\big(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\big) - V_i\big(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\big) = \quad & \underbrace{-\frac{\kappa-1}{2}\,(\kappa^{-\frac{1}{2}}\,2^{i-1})^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}\,2^{i-1},\,\kappa^{-\frac{1}{2}}\,2^{i}]} \\ & \underbrace{-\,(\kappa-1)\,\kappa^{-\frac{1}{2}}\,2^{i-1}\,\kappa^{-\frac{1}{2}}\,2^{i} + \frac{\kappa-1}{2}\,(\kappa^{-\frac{1}{2}}\,2^{i})^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}\,2^{i},\,\kappa^{-\frac{1}{2}}\,2^{i+1}]} \\ & \underbrace{+\,(\kappa-1)\,\kappa^{-\frac{1}{2}}\,2^{i-1}\,\frac{1}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1} - \frac{\kappa-1}{2}\,\big(\frac{1}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\big)^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}\,2^{i+1},\,\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}]} \\ = \quad & \frac{\kappa-1}{\kappa}\,\{-2^{2i-3} - 2^{2i-1} + 2^{2i-1} + 2^{2i-2} - 2^{2i-3}\} \\ = \quad & 0\,. \end{aligned}$$

□

We need one final ingredient: **Fano's inequality**, which is the standard tool for establishing information-theoretic lower bounds.

Theorem 9.1.5 (Fano's inequality). *Let* $\boldsymbol{i} \sim \mathsf{uniform}([m])$. *Then, for any estimator* $\widehat{\boldsymbol{i}}$ *of* $\boldsymbol{i}$, *where* $\widehat{\boldsymbol{i}}$ *is measurable with respect to some data* Y,

$$\mathbb{P}\big\{\widehat{\boldsymbol{i}} \neq \boldsymbol{i}\big\} \ge 1 - \frac{\mathsf{I}(\boldsymbol{i};Y) + \ln 2}{\ln m}\,,$$

where I *is the* ***mutual information*** $\mathsf{I}(\boldsymbol{i};Y) := \mathsf{KL}(\mathrm{law}(\boldsymbol{i},Y) \,\|\, \mathrm{law}(\boldsymbol{i}) \otimes \mathrm{law}(Y))$.

Proof. Let $\mathsf{H}(\cdot)$ denote the **entropy** of a discrete random variable, i.e., if X has law p on a discrete alphabet $\mathcal{X}$, then $\mathsf{H}(X) = \sum_{x\in\mathcal{X}} p(x)\ln(1/p(x))$. We refer to [CT06, Chapter 2] for the basic properties of entropy (and related quantities).

Let $E \coloneqq \mathbb{1}\{\widehat{\boldsymbol{i}} \neq \boldsymbol{i}\}$ denote the indicator of an error. Using the chain rule for entropy in two different ways,

$$\begin{aligned}\mathsf{H}(\boldsymbol{i}, E \mid \widehat{\boldsymbol{i}}) &= \mathsf{H}(\boldsymbol{i} \mid \widehat{\boldsymbol{i}}) + \underbrace{\mathsf{H}(E \mid \boldsymbol{i}, \widehat{\boldsymbol{i}})}_{=0} \\ &= \mathsf{H}(E \mid \widehat{\boldsymbol{i}}) + \mathsf{H}(\boldsymbol{i} \mid E, \widehat{\boldsymbol{i}})\,.\end{aligned}$$

Since conditioning reduces entropy, $\mathsf{H}(E \mid \widehat{\boldsymbol{i}}) \le \mathsf{H}(E) \le \ln 2$. Also,

$$\mathsf{H}(\boldsymbol{i} \mid E, \widehat{\boldsymbol{i}}) = \mathbb{P}\{\widehat{\boldsymbol{i}} = \boldsymbol{i}\} \underbrace{\mathsf{H}(\boldsymbol{i} \mid \widehat{\boldsymbol{i}}, E = 0)}_{=0} + \mathbb{P}\{\widehat{\boldsymbol{i}} \neq \boldsymbol{i}\}\, \mathsf{H}(\boldsymbol{i} \mid \widehat{\boldsymbol{i}}, E = 1) \le \mathbb{P}\{\widehat{\boldsymbol{i}} \neq \boldsymbol{i}\} \ln m\,.$$

Hence,

$$\mathbb{P}\{\widehat{\boldsymbol{i}} \neq \boldsymbol{i}\} \ln m + \ln 2 \ge \mathsf{H}(\boldsymbol{i} \mid \widehat{\boldsymbol{i}}) = \mathsf{H}(\boldsymbol{i}) - \mathsf{I}(\boldsymbol{i}; \widehat{\boldsymbol{i}}) \ge \ln m - \mathsf{I}(\boldsymbol{i}; Y)$$

where the last inequality is the data-processing inequality. Rearranging the inequality completes the proof of Fano's inequality. □

Proof of Theorem 9.1.1, lower bound. We follow the general outline described above.

1. Reduction to statistical testing. Let $\boldsymbol{i} \sim \mathsf{uniform}([m])$ and suppose that for each $i \in [m]$, $\hat{\pi}_i$ is a distribution with $\|\hat{\pi}_i - \pi_i\|_{\mathrm{TV}} \le \frac{1}{64}$. Suppose that we have a sample $X \sim \hat{\pi}_{\boldsymbol{i}}$ (more precisely, this means that conditioned on $\boldsymbol{i} = i$, we have $X \sim \hat{\pi}_i$). In light of Lemma 9.1.2, a good candidate estimator $\widehat{\boldsymbol{i}}$ for $\boldsymbol{i}$ is

$$\widehat{\boldsymbol{i}} \coloneqq i \in \mathbb{N} \quad \text{such that} \quad X \in (\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}] \quad \text{if such an } i \text{ exists}\,.$$

The probability that the estimator is correct is at least

$$\begin{aligned}\mathbb{P}\big\{\widehat{\boldsymbol{i}} = \boldsymbol{i}\big\} &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\big\{\widehat{\boldsymbol{i}} = i \,\big|\, \boldsymbol{i} = i\big\} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\big\{X \in (\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}] \,\big|\, \boldsymbol{i} = i\big\} \\ &= \frac{1}{m}\sum_{i=1}^{m} \hat{\pi}_i\big((\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}]\big) \ge \frac{1}{m}\sum_{i=1}^{m} \pi_i\big((\kappa^{-\frac{1}{2}} 2^{i-2}, \kappa^{-\frac{1}{2}} 2^{i-1}]\big) - \frac{1}{64} \ge \frac{1}{64}\,.\end{aligned} \tag{9.1.6}$$

Hence, a sampling can be used to solve the statistical testing problem.

2. A lower bound for the statistical testing problem. Next, we want to show for any algorithm which uses n queries to the oracle for $\pi_{\boldsymbol{i}}$ and outputs an estimator $\widehat{\boldsymbol{i}}$ of $\boldsymbol{i}$, there is a lower bound for the probability of error $\mathbb{P}\{\widehat{\boldsymbol{i}} \neq \boldsymbol{i}\}$.

First, suppose that the algorithm is *deterministic*, i.e., we assume that each query point x_j of the algorithm is a deterministic function of the previous query points and query values. Let $\mathcal{O}_i(x) := (V_i(x), V_i'(x), V_i''(x))$ denote the output of the oracle on input x when the target is π_i. Since the estimator $\widehat{\boldsymbol{i}}$ is a function of $\{x_j, \mathcal{O}_{\boldsymbol{i}}(x_j)\}_{j\in[n]}$, then Fano's inequality (Theorem 9.1.5) yields

$$\mathbb{P}\{\widehat{\boldsymbol{i}} \neq \boldsymbol{i}\} \geq 1 - \frac{\mathsf{I}(\boldsymbol{i}; \{x_j, \mathcal{O}_{\boldsymbol{i}}(x_j)\}_{j\in[n]}) + \ln 2}{\ln m}.$$

By the chain rule for mutual information,

$$\mathsf{I}\big(\boldsymbol{i}; \{x_j, \mathcal{O}_{\boldsymbol{i}}(x_j)\}_{j\in[n]}\big) = \sum_{j=1}^{n} \mathsf{I}\big(\boldsymbol{i}; x_j, \mathcal{O}_{\boldsymbol{i}}(x_j) \,\big|\, \{x_{j'}, \mathcal{O}_{\boldsymbol{i}}(x_{j'})\}_{j'\in[j-1]}\big). \tag{9.1.7}$$

By our assumption, conditioned on $\{x_{j'}, \mathcal{O}_{\boldsymbol{i}}(x_{j'})\}_{j'\in[j-1]}$, the query point x_j is deterministic. Also, for a fixed point x_j, Lemma 9.1.4 implies that $\mathcal{O}_i(x_j)$ can only take on a constant number of possible values as i ranges over $[m]$ (the careful reader can check that the number of possible values for $\mathcal{O}_i(x_j)$ is at most 5). Together with (9.1.7),

$$\mathsf{I}\big(\boldsymbol{i}; \{x_j, \mathcal{O}_{\boldsymbol{i}}(x_j)\}_{j\in[n]}\big) \leq n \ln 5.$$

Fano's inequality then yields

$$\mathbb{P}\{\widehat{\boldsymbol{i}} \neq \boldsymbol{i}\} \geq 1 - \frac{n \ln 5 + \ln 2}{\ln m}. \tag{9.1.8}$$

In general, for a possibly randomized algorithm, we can still deduce (9.1.8) by applying the previous argument conditioned on the random seed of the algorithm (which is independent of $\boldsymbol{i}$).

3. Finishing the argument. By combining together (9.1.6) and (9.1.8), and recalling that $m = \Theta(\log \kappa)$, we have shown that $n \gtrsim \log\log \kappa$. □

The argument above makes rigorous the following intuition: since there are m distributions in our lower bound construction, there are $\log_2 m$ bits of information to learn. On the other hand, Lemma 9.1.4 implies that each oracle query only reveals $O(1)$ bits of information. Hence, the number of queries required is at least $\Omega(\log m) = \Omega(\log\log \kappa)$.

Upper bound. To show that the lower bound is tight, we exhibit an algorithm, based on rejection sampling, which achieves the lower complexity bound. As per our discussion in Section 7.1, to implement rejection sampling we must specify the construction of an upper envelope $\widetilde{\mu} \geq \widetilde{\pi}$, where $\pi \in \Pi_\kappa$.

Without loss of generality, we assume that $V(0) = 0$ (if not, replace the output $V(x)$ of an oracle query with $V(x) - V(0)$). The upper bound algorithm only requires a zeroth-order oracle, and it is as follows.

1. Find the first index $i_- \in \{0, 1, \ldots, \lceil \frac{1}{2} \log_2 \kappa \rceil\}$ such that $V(-2^{i_-}/\sqrt{\kappa}) \geq \frac{1}{2}$.
2. Find the first index $i_+ \in \{0, 1, \ldots, \lceil \frac{1}{2} \log_2 \kappa \rceil\}$ such that $V(+2^{i_+}/\sqrt{\kappa}) \geq \frac{1}{2}$.
3. Set $x_- := -2^{i_-}/\sqrt{\kappa}$ and $x_+ := +2^{i_+}/\sqrt{\kappa}$; then, set

$$\widetilde{\mu}(x) := \begin{cases} \exp\left(-\dfrac{x - x_-}{2x_-} - \dfrac{(x - x_-)^2}{2}\right), & x \leq x_-, \\ 1, & x_- \leq x \leq x_+, \\ \exp\left(-\dfrac{x - x_+}{2x_+} - \dfrac{(x - x_+)^2}{2}\right), & x \geq x_+. \end{cases}$$

To see why i_- and i_+ exist, from $V'' \geq 1$ and $V(0) = V'(0) = 0$ we have $V(x) \geq x^2/2$. Hence, if $|x| = 2^i/\sqrt{\kappa}$ where $i \geq \frac{1}{2} \ln \kappa$, we have $V(x) \geq 1/2$.

Since V is decreasing (resp. increasing) on $\mathbb{R}_-$ (resp. $\mathbb{R}_+$), the first two steps can be implemented by running binary search over arrays of size $O(\log \kappa)$, which therefore only requires $O(\log \log \kappa)$ queries. We will prove that $\widetilde{\mu}$ is a valid upper envelope for the unnormalized target $\widetilde{\pi} := \exp(-V)$, and that $Z_\mu / Z_\pi \lesssim 1$. In turn, Theorem 7.1.1 shows that once $\widetilde{\mu}$ is constructed, an exact sample can be drawn from π using $O(1)$ additional queries in expectation.

Alternatively, if we require that the algorithm use a fixed (non-random) number of iterations, then note that in order to make the failure probability (the probability that rejection sampling fails to terminate within the allotted number of iterations) at most ε, it suffices to run rejection sampling for $O(\log(1/\varepsilon))$ steps. Combining this with the cost of constructing $\widetilde{\mu}$, we conclude that we can output a sample whose law is ε-close to π in total variation distance using $O(\log \log \kappa + \log(1/\varepsilon))$ queries.

Proof of Theorem 9.1.1, upper bound. First, we prove that $\widetilde{\mu}$ is a valid upper envelope. Since $\widetilde{\pi}$ is decreasing on $\mathbb{R}_+$ with $\widetilde{\pi}(0) = \exp(-V(0)) = 1$, then $\widetilde{\pi} \leq 1 \leq \widetilde{\mu}$ on $[0, x_+]$. Next, since $V(x_+) \geq 1/2$ (by the definition of x_+), convexity of V yields

$$V'(x_+) \geq \frac{V(x_+) - V(0)}{x_+} \geq \frac{1}{2x_+}.$$

Thus, for $x \geq x_+$,

$$V(x) \geq V(x_+) + V'(x_+)\,(x - x_+) + \frac{1}{2}\,(x - x_+)^2 \geq \frac{1}{2x_+}\,(x - x_+) + \frac{1}{2}\,(x - x_+)^2,$$

which shows that $\widetilde{\pi}(x) \le \widetilde{\mu}(x)$. By a symmetric argument on $\mathbb{R}_-$, we conclude that $\widetilde{\pi} \le \widetilde{\mu}$.

By Theorem 7.1.1, it suffices to bound Z_μ/Z_π. First, we claim that $\int_0^{x_+} \widetilde{\mu} \gtrsim x_+$. When $i_+ = 0$, this holds

$$\int_0^{x_+} \widetilde{\mu} = \int_0^{1/\sqrt{\kappa}} \exp(-V) \ge \int_0^{1/\sqrt{\kappa}} \exp\Big(-\frac{\kappa x^2}{2}\Big)\,\mathrm{d}x \ge \frac{1}{3\sqrt{\kappa}} = \frac{x_+}{3}\,.$$

When $i_+ > 0$, then by the definition of i_+ we have $V(x_+/2) \le 1/2$, so

$$\int_0^{x_+} \widetilde{\mu} \ge \int_0^{x_+/2} \exp(-V) \ge \frac{x_+}{4}\,.$$

On the other hand, by Lemma 9.1.3,

$$\int_{\mathbb{R}_+} \widetilde{\mu} = \int_0^{x_+} \widetilde{\mu} + \int_{x_+}^{\infty} \widetilde{\mu} \le x_+ + \int_{x_+}^{\infty} \exp\Big(-\frac{1}{2x_+}\,(x - x_+) - \frac{1}{2}\,(x - x_+)^2\Big)\,\mathrm{d}x \le 3x_+\,.$$

Hence, $\int_{\mathbb{R}_+} \widetilde{\mu} \le 3x_+ \le 12 \int_{\mathbb{R}_+} \widetilde{\pi}$, and similarly $\int_{\mathbb{R}_-} \widetilde{\mu} \le 12 \int_{\mathbb{R}_-} \widetilde{\pi}$. Therefore, $Z_\mu/Z_\pi \le 12$. □

Discussion. Although this query complexity result only pertains to one-dimensional targets, there are still some useful takeaways. For instance, the lower bound proof shows that information theoretic arguments can indeed be adapted to the context of sampling, and it may serve as a template for further results in this direction.

The obtained complexity $\Theta(\log\log\kappa)$ is surprisingly small; in particular, the upper bound uses a tailor-made algorithm based on rejection sampling, rather than any of the other existing algorithms (such as those based on Langevin dynamics). This is perhaps the best case scenario for a lower bound: it helps us to determine if our existing algorithms are optimal, and if not, it gives guidance on how to design a better one. On the other hand, the specific complexity is likely due to the one-dimensional structure; in high dimension, it is conjectured that the dependence on the condition number is polynomial.

9.2 Other Approaches

In this section, we discuss alternative approaches and partial progress towards obtaining lower bounds for sampling.

Lower bounds for particular algorithms. As already discussed in Section 7.3, the works [Che+21b; LST21a; WSC21] obtain lower bounds for the complexity of MALA,

culminating in a precise understanding of the runtime of MALA both from feasible and warm start initializations.

The paper [CLW21] provides an approach to proving lower bounds for discretization schemes. In their setup, there is a stochastic process $(Z_t)_{t\geq 0}$, driven by some underlying Brownian motion $(B_t)_{t\geq 0}$; for example, the process $(Z_t)_{t\geq 0}$ could be the Langevin diffusion or the underdamped Langevin diffusion (Section 5.3). The algorithm is allowed to make queries to the potential V, as well as certain queries to the driving Brownian motion $(B_t)_{t\geq 0}$, and the goal of the algorithm is to output a point $\widehat{Z}_T$ which is close to Z_T in mean squared error: $\mathbb{E}[\|Z_T - \widehat{Z}_T\|^2] \leq \varepsilon^2$. Within this framework, they prove that the randomized midpoint discretization (introduced in Section 5.1) is optimal for simulating the underdamped Langevin dynamics (see Theorem 5.3.11 for the upper bound).

Estimating the normalizing constant. In [RV08], the authors consider the number of membership queries needed to estimate the volume of a convex body $K \subseteq \mathbb{R}^d$ such that $\mathrm{B}(0,1) \subseteq K \subseteq \mathrm{B}(0, O(d^8))$ to within a small multiplicative constant; their lower bound for this problem is $\widetilde{\Omega}(d^2)$. In comparison, the state-of-the-art upper bound for volume computation is $\widetilde{O}(d^3)$ (see [CV18; Jia+21]).

In [GLL20], the authors consider the problem of estimating the normalizing constant $Z_\pi := \int \widetilde{\pi}$ from queries to the unnormalized density $\widetilde{\pi}$. Based on a multilevel Monte Carlo scheme, they show that sampling algorithms can be turned into approximation algorithms for the normalizing constant, with the cost of an extra $O(d)$ dimension dependence in the reduction. By combining this with the randomized midpoint discretization of the underdamped Langevin diffusion (Theorem 5.3.11), they show that a $1 \pm \varepsilon$ multiplicative approximation to Z_π can be obtained using $\widetilde{O}((d^{4/3}\kappa + d^{7/6}\kappa^{7/6})/\varepsilon^2)$ queries (in the strongly log-concave case).

They then prove that $\Omega(d^{1-o(1)}/\varepsilon^{2-o(1)})$ queries are necessary to obtain a $1 \pm \varepsilon$ multiplicative approximation to Z_π. Unfortunately, due to the $O(d)$ loss in the reduction from estimating the normalizing constant to sampling, this does not imply a non-trivial lower bound for the task of sampling.

Lower bound for a stochastic oracle. In [CBL22], the authors obtain a lower bound on the complexity of sampling using a *stochastic* oracle. Namely, in order to output an ε-approximate sample (in TV distance) from an α-strongly log-concave and β-log-smooth distribution whose mean lies in the ball $\mathrm{B}(0, 1/\alpha)$, with an oracle that given $x \in \mathbb{R}^d$ outputs $\nabla V(x) + \xi$ with $\xi \sim \mathrm{normal}(0, \Sigma)$ and $\operatorname{tr}\Sigma \leq \sigma^2 d$, the number of queries required is at least $\Omega(\sigma^2 d/\varepsilon^2)$. On the other hand, when $\alpha, \sigma \asymp 1$, this complexity is achieved via stochastic gradient Langevin Monte Carlo.

Bibliographical Notes

Since the theory of lower bounds for sampling is still early in its development, there are not too many works yet in this direction. Section 9.2 contains a brief survey.

Recently, the paper [GLL22] obtains a query complexity bound for the following class of target distributions:

$$\Pi_{\alpha,L} := \left\{ \pi \in \mathcal{P}_{\mathrm{ac}}\big(\mathrm{B}(0,1)\big) \;\middle|\; \pi \propto \exp\Big(-\sum_{i=1}^{\infty} f_i - \frac{\alpha}{2}\|\cdot\|^2\Big),\ f_i : \mathrm{B}(0,1) \to \mathbb{R} \text{ is } L\text{-Lipschitz} \right\}.$$

They show that the minimum number of queries to the individual functions $(f_i)_{i=1}^{\infty}$ required to obtain a sample which is ε-close to a target $\pi \in \Pi_{\alpha,L}$ is, in the regime $d \ll L^2/\alpha$, of the order $\widetilde{\Theta}(L^2/\alpha)$. The upper bound is based on the proximal sampler (Section 8.1), whereas the lower bound, which in this context reduces sampling to an optimization task, relies on information-theoretic arguments.

Exercises

A Query Complexity Result in One Dimension

⊵ Exercise 9.1 (Gaussian tail bound)
For $x > 0$, show that $\int_x^{\infty} \exp(-t^2/2)\,\mathrm{d}t \le x^{-1}\exp(-x^2/2)$. Use this to prove Lemma 9.1.3.

CHAPTER 10

Structured Sampling

So far, we have only considered sampling within the black-box model, in which we only have access to oracle queries to the potential and its gradient. We will now consider several new sampling algorithms which go beyond the black-box model.

10.1 Coordinate Langevin

10.2 Mirror Langevin

The *mirror descent* method in optimization changes the geometry of the algorithm via the use of a **mirror map** $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Here, ϕ is a convex function and we denote $\mathcal{X} := \operatorname{int} \operatorname{dom} \phi$; we assume that $\mathcal{X} \neq \varnothing$, that ϕ is strictly convex and differentiable on $\mathcal{X}$, and that ϕ is a barrier for $\mathcal{X}$ in the sense that $\|\nabla \phi(x_k)\| \to \infty$ whenever $(x_k)_{k \in \mathbb{N}} \subseteq \mathcal{X}$ converges to a point on $\partial \mathcal{X}$. Then, rather than following the gradient descent iteration

$$x_{k+1} := x_k - h \nabla V(x_k), \qquad k = 0, 1, 2, \ldots \tag{10.2.1}$$

we can instead consider the **mirror descent** iteration

$$\nabla \phi(x_{k+1}) := \nabla \phi(x_k) - h \nabla V(x_k), \qquad k = 0, 1, 2, \ldots \tag{10.2.2}$$

The assumptions on the mirror map ϕ ensure that the iteration (10.2.2) is well-defined. When $\phi = \frac{\|\cdot\|^2}{2}$, then mirror descent coincides with gradient descent.

Historically, mirror descent was introduced by Nemirovsky and Yudin [NY83] with the following intuition. Suppose we are optimizing a function V which is not defined over the Euclidean space $\mathbb{R}^d$, but rather over a Banach space $\mathcal{B}$. Then, the gradient of V is not an element of $\mathcal{B}$ but rather of the dual space $\mathcal{B}^*$, and so the gradient descent iteration (10.2.1) does not even make sense. On the other hand, the mirror descent iteration (10.2.2) works because the primal point $x_k \in \mathcal{B}$ is first mapped to the dual space $\mathcal{B}^*$ via the mapping $\nabla\phi$. This reasoning is not so esoteric as it may seem, because even for a function V defined over $\mathbb{R}^d$ its natural geometry may correspond to a different norm (e.g., the ℓ_1 norm), in which case $\mathbb{R}^d$ is better viewed as a Banach space.

Our aim is to understand the sampling analogue of mirror descent, known as *mirror Langevin*. We will keep in mind the key example of **constrained sampling**. Here, the potential $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ has domain $\mathcal{X} \subsetneq \mathbb{R}^d$. In this case, the standard Langevin algorithm leaves the constraint set $\mathcal{X}$ which is undesirable; in particular, it is not possible to obtain guarantees in metrics such as KL divergence because the law of the iterate of the algorithm is not absolutely continuous with respect to the target. Projecting the iterates onto $\mathcal{X}$ does not solve this issue because the law of the iterate will then have positive mass on the boundary $\partial\mathcal{X}$. Besides, projection may not adapt well to the shape of the constraint set $\mathcal{X}$. Instead, the use of a mirror map ϕ which is a barrier for $\mathcal{X}$ can *automatically* enforce the constraint.

10.2.1 Continuous-Time Considerations

In continuous time, the mirror Langevin diffusion $(Z_t)_{t\geq 0}$ is the solution to the stochastic differential equation

$$Z_t^* = \nabla\phi(Z_t)\,, \qquad \mathrm{d}Z_t^* = -\nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\left[\nabla^2\phi(Z_t)\right]^{1/2}\mathrm{d}B_t\,. \tag{10.2.3}$$

Here, the diffusion term is no longer an isotropic Brownian motion but rather involves the matrix $[\nabla^2\phi(Z_t)]^{1/2}$; this is necessary in order to ensure that the stationary distribution is π. Also, we have given the SDE in the dual space. Using Itô's formula (Theorem 1.1.18), one can write down an SDE for $(Z_t)_{t\geq 0}$ in the primal space, but it is more complicated, involving the third derivative tensor of ϕ (see Exercise 10.1), and as such we prefer to work with the representation (10.2.3).

Using (10.2.3), we can compute the generator $\mathscr{L}$, the carré du champ Γ, and the Dirichlet energy $\mathscr{E}$ of the mirror Langevin diffusion (see Section 1.2 and Exercise 10.2):

$$\Gamma(f,g) = \langle \nabla f, [\nabla^2\phi]^{-1}\nabla g\rangle\,, \qquad \mathscr{E}(f,g) = \int \langle \nabla f, [\nabla^2\phi]^{-1}\nabla g\rangle\,\mathrm{d}\pi\,. \tag{10.2.4}$$

The expression shows that the mirror Langevin diffusion is reversible with respect to π. Also, if π_t denotes the law of Z_t, then

$$\int f\,\partial_t \pi_t = \partial_t \int f\,\mathrm{d}\pi_t = \int \mathscr{L} f\,\frac{\pi_t}{\pi}\,\mathrm{d}\pi = -\int \left\langle \nabla f, \left[\nabla^2\phi\right]^{-1} \nabla \frac{\pi_t}{\pi} \right\rangle \mathrm{d}\pi \tag{10.2.5}$$

$$= -\int \left\langle \nabla f, \left[\nabla^2\phi\right]^{-1} \nabla \ln \frac{\pi_t}{\pi} \right\rangle \mathrm{d}\pi_t = \int f \operatorname{div}\Big(\pi_t \left[\nabla^2\phi\right]^{-1} \nabla \ln \frac{\pi_t}{\pi}\Big) \tag{10.2.6}$$

from which we deduce the Fokker–Planck equation

$$\partial_t \pi_t = \operatorname{div}\Big(\pi_t \left[\nabla^2\phi\right]^{-1} \nabla \ln \frac{\pi_t}{\pi}\Big)\,.$$

From the interpretation as a continuity equation (see Theorem 1.3.17), we deduce that $(\pi_t)_{t\geq 0}$ describes the evolution of a particle which travels according to the family of vector fields $t \mapsto -\left[\nabla^2\phi\right]^{-1} \nabla \ln(\pi_t/\pi)$. Recalling that $\nabla \ln(\pi_t/\pi)$ is the Wasserstein gradient of $\mathsf{KL}(\cdot \parallel \pi)$ at π_t, we can interpret the mirror Langevin diffusion as a "mirror flow" of the KL divergence in Wasserstein space.

Alternatively, we can equip $\mathcal{X}$ with the Riemannian metric induced by $\nabla^2\phi$, i.e., we set $\langle u, v\rangle_x := \langle u, \nabla^2\phi(x)\,v\rangle$. Then, the mirror Langevin diffusion becomes the Wasserstein gradient flow of the KL divergence over the Riemannian manifold $\mathcal{X}$ (see Section 2.6.1).

The Newton Langevin diffusion. In the special case when the mirror map ϕ is chosen to be the same as the potential V, we arrive at a sampling analogue of Newton's algorithm, and hence we call it the **Newton Langevin diffusion**. The equation for the Newton Langevin diffusion can be written (in the dual space) as

$$\mathrm{d}Z_t^* = -Z_t^*\,\mathrm{d}t + \sqrt{2}\left[\nabla^2 V^*(Z_t^*)\right]^{-1/2} \mathrm{d}B_t\,, \tag{10.2.7}$$

see Exercise 10.3.

Convergence in continuous time. In optimization, Newton's algorithm has many favorable properties. For example, at least locally, it is known that Newton's algorithm converges *quadratically* rather than linearly, which means that the error at iteration k scales as $\exp(-c_1 \exp(c_2 k))$ for constants $c_1, c_2 > 0$. Also, Newton's algorithm is *affine-invariant*, meaning that if A is any invertible matrix and we instead apply Newton's algorithm to the function $\widetilde{V}(x) := V(Ax)$, then the iterates $(\widetilde{x}_k)_{k\in\mathbb{N}}$ are related to the iterates $(x_k)_{k\in\mathbb{N}}$ of Newton's algorithm on the original function V via the transformation $\widetilde{x}_k = A^{-1}x_k$ (Exercise 10.4). Consequently, the convergence speed of Newton's algorithm should not be badly affected by poor conditioning of V.

Can we expect similar properties to hold for the Newton Langevin diffusion? At least for the property of affine invariance, we have the **Brascamp–Lieb inequality** (Theorem 2.2.8): if $\pi \propto \exp(-V)$ is strictly log-concave, then for all $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathrm{var}_\pi(f) \leq \mathbb{E}_\pi \langle \nabla f, [\nabla^2 V]^{-1} \nabla f \rangle \,.$$

Below, we will also give an alternative proof of the Bregman transport inequality (Theorem 2.2.10) based on Wasserstein calculus, which implies the Brascamp–Lieb inequality. Note that in the strongly convex case $\nabla^2 V \succeq \alpha I_d$, it implies a Poincaré inequality (in the sense of Example 1.2.22) for π with constant $1/\alpha$. However, in our present context with Dirichlet energy given by (10.2.4), we instead interpret the Brascamp–Lieb inequality as a Poincaré inequality (in the sense of Definition 1.2.19) for the Newton–Langevin diffusion. Then, the Poincaré constant is 1, independent of the strong convexity of V.

We also obtain a Poincaré inequality for the mirror Langevin diffusion under the condition of relative strong convexity.

Definition 10.2.8. Let $\phi, V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex functions, and assume that $\mathcal{X} = \operatorname{int} \operatorname{dom} \phi = \operatorname{int} \operatorname{dom} V$. Then:

1. V is α-**relatively convex** (w.r.t. ϕ) if for all $x \in \mathcal{X}$,
$$\nabla^2 V(x) \geq \alpha \nabla^2 \phi(x) \,.$$
2. V is β-**relatively smooth** (w.r.t. ϕ) if for all $x \in \mathcal{X}$,
$$\nabla^2 V(x) \leq \beta \nabla^2 \phi(x) \,.$$

Observe that when $\phi = \frac{\|\cdot\|^2}{2}$, these definitions reduce to the usual definitions of strong convexity and smoothness. Recall from Definition 2.2.9 that the Bregman divergence D_ϕ associated with ϕ is the mapping $D_\phi(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$D_\phi(x, y) \coloneqq \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle \,.$$

The Bregman divergence plays an important role in the analysis of mirror Langevin because it is the correct substitute for the Euclidean distance $(x, y) \mapsto \frac{1}{2} \|x - y\|^2$ in this context. Note the following observations: (1) D_ϕ is non-negative due to convexity of ϕ, and if ϕ is strictly convex then it equals 0 if and only if its two arguments are equal; (2) since $D_\phi(x, y)$ is defined by subtracting the first-order Taylor expansion of ϕ at y, it

behaves infinitesimally like a squared distance; in particular,

$$D_\phi(x,y) \sim \frac{1}{2}\langle y-x, \nabla^2\phi(x)\,(y-x)\rangle \qquad \text{as } y \to x\,;$$

(3) when $\phi = \frac{\|\cdot\|^2}{2}$, then D_ϕ is precisely one-half times the squared Euclidean distance.

Using this definition, we have the following reformulations of relative convexity and relative smoothness (Exercise 10.5).

Lemma 10.2.9. *V is α-relatively convex w.r.t. ϕ if and only if*

$$D_V \geq \alpha\, D_\phi\,.$$

Similarly, V is β-relatively smooth w.r.t. ϕ if and only if

$$D_V \leq \beta\, D_\phi\,.$$

Returning to the mirror Langevin diffusion, the following corollary is an immediate consequence of the Brascamp–Lieb inequality and the definition of relative convexity.

Corollary 10.2.10 (mirror Poincaré inequality, [Che+20b])**.** *Suppose that the potential V is α-relatively convex w.r.t. ϕ. Then, the mirror Langevin diffusion satisfies the following Poincaré inequality: for all $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{var}_\pi f \leq \frac{1}{\alpha}\,\mathbb{E}_\pi\langle \nabla f, [\nabla^2\phi]^{-1}\,\nabla f\rangle = \frac{1}{\alpha}\,\mathscr{E}(f,f)\,.$$

So far so good: we have defined relative convexity, which is a natural generalization of strong convexity and well-studied in the optimization literature; and we have shown that it implies a Poincaré inequality for the mirror Langevin diffusion.

However, the analogies with the standard Langevin diffusion stop here. There are counterexamples which show that relative convexity does *not* imply a log-Sobolev inequality for the mirror Langevin diffusion. This may seem to contradict the Bakry–Émery theorem (Theorem 1.2.29), which holds for any Markov diffusion. The issue here is that the assumption of relative convexity is *not* a curvature-dimension condition. Indeed, in order to properly formulate a curvature-dimension condition for the Hessian manifold $\mathcal{X}$ equipped with the Riemannian metric $\mathfrak{g}$ induced by $\nabla^2\phi$, one must check the $\mathrm{CD}(\alpha,\infty)$ condition $\nabla^2\widetilde{V} + \mathrm{Ric} \geq \alpha$. Here, ∇^2 denotes the Riemannian Hessian and $\widetilde{V}$ is the potential corresponding to the relative density of π w.r.t. the Riemannian volume measure on $(\mathcal{X},\mathfrak{g})$;

see Section 2.6. Such a calculation was performed in, e.g., [Kol14], which implies that if $(\nabla\phi)_{\#}\pi$ is log-concave, then the Newton Langevin diffusion satisfies $\mathrm{CD}(\frac{1}{2}, \infty)$. However, it is not clear under what conditions $(\nabla\phi)_{\#}\pi$ is log-concave.

Here is another consequence of the fact that relative convexity is not a curvature-dimension condition: relative convexity (apparently) does not seem to imply contraction properties for the mirror Langevin diffusion with respect to an appropriately defined Wasserstein metric.

To summarize: either we can assume the curvature-dimension condition $\mathrm{CD}(\alpha, \infty)$, which imposes complicated conditions on ϕ and V, or we can adopt the more interpretable relative convexity assumption, which in turn only implies a Poincaré inequality (Corollary 10.2.10). We will follow the latter approach.

Upon reflection, the curvature-dimension approach for studying the mirror Langevin diffusion is arguably the less natural one. Indeed, the curvature-dimension approach is based on viewing the mirror Langevin diffusion from the lens of Riemannian geometry, but the mirror descent algorithm in optimization is not typically studied via Riemannian geometry. Instead, the study of mirror descent is based on ideas from convex analysis, centered around the Bregman divergence. So it seems prudent at this stage to abandon the Riemannian interpretation of mirror Langevin in favor of convex analysis tools, and this is indeed how our discretization proof will go. In fact, in lieu of using the Poincaré inequality in Corollary 10.2.10, we will *directly* use relative convexity.

10.2.2 Discretization Preliminaries

Following [AC21], we consider the following discretization of (10.2.3).

$$\begin{aligned} \nabla\phi(X^+_{kh}) &:= \nabla\phi(X_{kh}) - h\,\nabla V(X_{kh})\,, \\ X_{(k+1)h} &:= \nabla\phi^*(X^*_{(k+1)h})\,, \end{aligned} \tag{MLMC}$$

where

$$X^*_t = \nabla\phi(X^+_{kh}) + \sqrt{2}\int_{kh}^{t} [\nabla^2\phi^*(X^*_s)]^{-1/2}\,\mathrm{d}B_s \qquad \text{for } t \in [kh, (k+1)h]\,. \tag{10.2.11}$$

Note that when $\phi = \frac{\|\cdot\|^2}{2}$, this reduces to the standard LMC algorithm. When generalizing LMC to different mirror maps, this discretization is chosen to preserve the "forward-flow" interpretation of [Wib18] (see Section 4.3). In particular, the update from X_{kh} to X^+_{kh} is a mirror descent step, while the update from X^+_{kh} to $X_{(k+1)h}$ follows a "Wasserstein mirror flow" of the (negative) entropy.

However, implementing MLMC requires the exact simulation of a diffusion process, so is it truly a "discretization"? To address this, note that simulating the mirror diffusion (10.2.11) does not require additional queries to the potential V, since it only depends on the mirror map ϕ (except for the initialization). To an extent, any algorithm based on mirror maps requires the implementation of certain primitive operations involving ϕ, such as computation of $\nabla\phi$ or inversion of $\nabla\phi$; in practice, this requires ϕ to have a "simple" structure such that these operations have closed-form expressions, or are at least cheap enough to be negligible relative to the cost of computing the gradient of V. In our consideration of MLMC, we take the diffusion (10.2.11) to be another primitive operation associated with the mirror map ϕ. This is indeed appropriate for many applications, e.g., when ϕ is a separable function $\phi(x) = \sum_{i=1}^{d} \phi_i(x_i)$, and it will streamline our technical analysis. Nevertheless, implementation of MLMC remains a key obstacle to its practicality and necessitates further research in this direction.

Our approach to studying MLMC is to adapt the convex optimization approach introduced in Section 4.3.

Key technical results. We now establish the analogues of the various facts that we invoked in the study of LMC.

First, in the standard gradient descent analysis, if $x_+ := x - h\nabla V(x)$, then we have the key inequality

$$\langle \nabla V(x), x^+ - z\rangle = \frac{1}{2h}\{\|x-z\|^2 - \|x^+ - z\|^2 - \|x^+ - x\|^2\} \qquad \text{for all } z \in \mathbb{R}^d\,.$$

Remarkably, there is an analogue of this fact for mirror descent, which follows from the following identity (which can be checked by simple algebra, see Exercise 10.5):

$$\langle \nabla\phi(\tilde{x}) - \nabla\phi(x), x - z\rangle = D_\phi(\tilde{x}, x) + D_\phi(z, \tilde{x}) - D_\phi(z, x)\,, \qquad \text{for all } x, \tilde{x}, z \in \mathcal{X}\,. \tag{10.2.12}$$

Lemma 10.2.13 (Bregman proximal lemma, [CT93]). *For $x \in \mathcal{X}$, let x^+ be defined via $\nabla\phi(x^+) = \nabla\phi(x) - h\nabla V(x)$. Then, for all $z \in \mathcal{X}$,*

$$\langle \nabla V(x), x^+ - z\rangle = \frac{1}{h}\{D_\phi(z, x) - D_\phi(z, x^+) - D_\phi(x^+, x)\}\,.$$

Proof. Note that

$$\langle \nabla V(x), x^+ - z\rangle = -\frac{1}{h}\langle \nabla\phi(x^+) - \nabla\phi(x), x^+ - z\rangle\,.$$

Substituting the identity (10.2.12) into the above equation proves the result. □

Unlike the case of LMC, the presence of a non-constant diffusion matrix involving $\nabla^2\phi$ introduces another source of discretization error. To address this, we introduce a condition on the third derivative of ϕ. Note however that a uniform bound on the operator norm of $\nabla^3\phi$ is not compatible with the assumption that ϕ tends to $+\infty$ on $\partial\mathcal{X}$. The solution to this issue was discovered by Nesterov and Nemirovsky [NN94]: we can ask that $\nabla^3\phi$ is bounded *with respect to the geometry induced by* ϕ. This approach also has the benefit of being consistent with the affine invariance of Newton's method. The precise definition of the third derivative condition is as follows.

Definition 10.2.14 (self-concordance)**.** The mirror map ϕ is said to be M_ϕ**-self-concordant** if for all $x \in \mathcal{X}$ and all $v \in \mathbb{R}^d$,

$$\nabla^3\phi(x)[v,v,v] \le 2M_\phi \|v\|^3_{\nabla^2\phi(x)} := 2M_\phi \langle v, \nabla^2\phi(x)\, v\rangle^{3/2}.$$

The norm $\|v\|_{\nabla^2\phi(x)} := \sqrt{\langle v, \nabla^2\phi(x)\, v\rangle}$ is called the *local norm*, and it is the tangent space norm for the Riemannian metric induced by ϕ.

The definition implies the following result, stated without proof:[1]

Lemma 10.2.15 ([Nes18, Corollary 5.1.1])**.** *Suppose that ϕ is M_ϕ-self-concordant. Then, for all $x \in \mathcal{X}$ and $u \in \mathbb{R}^d$,*

$$\nabla^3\phi(x)\, u \le 2M_\phi \|u\|_{\nabla^2\phi(x)}\, \nabla^2\phi(x).$$

Self-concordant functions are well-studied due to their central role in the theory of interior-point methods for optimization, see the monograph [NN94]. A key example of a self-concordant mirror map is when the constraint set is a polytope,

$$\mathcal{X} = \{x \in \mathbb{R}^d : \langle a_i, x\rangle < b_i \text{ for all } i \in [N]\},$$

in which case $\phi(x) = \ln(1/\sum_{i=1}^N (b_i - \langle a_i, x\rangle))$ is self-concordant with $M_\phi = 1$.

Finally, a key step in our analysis of LMC was to use the convexity of the entropy functional along W_2 geodesics. In our analysis of MLMC, we will replace the W_2 distance with the Bregman transport cost $\mathcal{D}_V$ (recall Definition 2.2.9). To study these costs, we first state an analogue of Brenier's theorem (Theorem 1.3.8).

[1] The proof is surprisingly difficult. The reader can try to prove the result with a worse constant factor.

Theorem 10.2.16 (Brenier's theorem for the Bregman transport cost)**.** *Suppose that $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Then, the unique optimal Bregman transport coupling (X, Y) for μ and ν is of the form*

$$\nabla\phi(Y) = \nabla\phi(X) - \nabla h(X)\,,$$

where $h : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ is such that $\phi - h$ is convex.

Proof sketch. We need facts about optimal transport with general costs c (Exercise 1.11). Namely, the optimal pair of dual potentials $(f^\star, g^\star)$ are *c-conjugates*, meaning that

$$f^\star(x) = \inf_{y\in\mathbb{R}^d} \{c(x,y) - g^\star(y)\}\,,$$
$$g^\star(y) = \inf_{x\in\mathbb{R}^d} \{c(x,y) - f^\star(x)\}\,.$$

For $\gamma^\star$-a.e. (x, y), it holds that $f^\star(x) + g^\star(y) = c(x, y)$. If c is smooth and such that $\nabla_x c(x, \cdot)$ is injective for all $x \in \mathbb{R}^d$, then from the definition of $g^\star$ it suggests that we have $\nabla_x c(x, y) = \nabla f^\star(x)$ for $\gamma^\star$-a.e. (x, y). See [Vil09, Theorem 10.28] for a rigorous statement and proof of these results.

Applying this to our cost function $c = D_\phi$, it yields the existence of D_ϕ-conjugates $h, \tilde{h} : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ such that $\nabla_x D_\phi(x, y) = \nabla h(x)$ under the optimal plan $\gamma^\star$. Hence,

$$\nabla\phi(y) = \nabla\phi(x) - \nabla h(x)\,, \qquad \text{for } \gamma^\star\text{-a.e. } (x, y)$$

and

$$h(x) = \inf_{y\in\mathbb{R}^d} \{D_\phi(x, y) - \tilde{h}(y)\}\,.$$

By expanding out the definition of D_ϕ, we rewrite this as

$$\phi(x) - h(x) = \sup_{y\in\mathcal{X}} \{\langle \nabla\phi(y), x - y\rangle + \tilde{h}(y) + \phi(y)\}\,.$$

As a supremum of affine functions, it follows that $\phi - h$ is convex. □

Recall that the usual W_2 geodesic from μ_0 to μ_1 is given as follows: there is a convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that for $X_0 \sim \mu_0$, the pair $(X_0, X_1) := (X_0, \nabla\varphi(X_0))$ is an optimal coupling. By taking the linear interpolation $X_t := (1 - t)\, X_0 + t\, X_1$ and setting $\mu_t := \mathrm{law}(X_t)$, we obtain the W_2 geodesic.

Now consider the above theorem. If we let $W \coloneqq \nabla\phi(Y) = \nabla\phi(X) - \nabla h(X)$, then we have the coupling $(X_0, X_1) \coloneqq (\nabla(\phi - h)^*(W), \nabla\phi^*(W))$ of $\mu_0 \coloneqq \mu$ and $\mu_1 \coloneqq \nu$. Then, we can interpolate by setting $X_t \coloneqq (1-t)\,X_0 + t\,X_1$ and $\mu_t \coloneqq \mathrm{law}(X_t)$, which defines an alternative path joining μ_0 to μ_1. Note that (X_0, W) and (X_1, W) are both optimally coupled for the W_2 distance (since $\phi - h$ is convex). This is a special case of the following.

Definition 10.2.17. A curve $(\mu_t)_{t\in[0,1]} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ is a **generalized geodesic** if there exists another measure $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ such that

$$\mu_t = \mathrm{law}(X_t)\,, \qquad X_t \coloneqq (1-t)\,X_0 + t\,X_1\,,$$

and (X_0, W), (X_1, W) are both optimally coupled for the W_2 metric, where $W \sim \rho$.

It is left as Exercise 10.6 to check that the entropy functional is also convex along generalized geodesics.

Theorem 10.2.18. *Let $\mathcal{H}(\mu) \coloneqq \int \mu \ln \mu$. Then, for any generalized geodesic $(\mu_t)_{t\in[0,1]}$,*

$$t \mapsto \mathcal{H}(\mu_t) \qquad \textit{is convex}\,.$$

In our context, it implies that if $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and (X, Y) is an optimal coupling for the Bregman cost $\mathcal{D}_\phi(\mu, \nu)$, then

$$\mathcal{H}(\nu) \geq \mathcal{H}(\mu) + \mathbb{E}\langle \nabla \ln \mu(X), Y - X\rangle\,. \tag{10.2.19}$$

As an aside, we remark that this observation leads to another proof of the Bregman transport inequality.

Proof of the Bregman transport inequality (Theorem 2.2.10). Let $\pi = \exp(-V)$, where V is strictly convex, and let $\mu \in \mathcal{P}(\mathbb{R}^d)$. Let $X \sim \mu$, $Z \sim \pi$ be optimally coupled for the Bregman transport cost. Then,

$$\mathsf{KL}(\mu \,\|\, \pi) = \mathbb{E}\,V(X) + \mathcal{H}(\mu)\,.$$

For the first term, by the definition of D_V,

$$\mathbb{E}\,V(X) = \mathbb{E}\,V(Z) + \mathbb{E}\,D_V(X, Z) + \mathbb{E}\langle \nabla V(Z), X - Z\rangle\,.$$

For the second term, (10.2.19) implies

$$\mathcal{H}(\mu) \geq \mathcal{H}(\pi) + \mathbb{E}\langle \nabla \ln \pi(Z), X - Z\rangle\,.$$

Hence,

$$\begin{aligned}\mathsf{KL}(\mu \parallel \pi) &\geq \underbrace{\mathbb{E}\, V(Z) + \mathcal{H}(\pi)}_{=\mathsf{KL}(\pi\|\pi)=0} + \mathbb{E}\langle \underbrace{\nabla V(Z) + \nabla \ln \pi(Z)}_{=0}, X - Z\rangle + \mathbb{E}\, D_V(X, Z) \\ &= \mathcal{D}_V(\mu \parallel \pi)\,,\end{aligned}$$

which is what we wanted to show. □

10.2.3 Discretization Analysis

Analysis for the smooth case. We now prove the following result.

Theorem 10.2.20 ([AC21]). *Suppose that $\pi = \exp(-V)$ is the target distribution and that ϕ is the mirror map. Assume:*

- *V is α-relatively convex and β-relatively smooth w.r.t. ϕ.*
- *ϕ is M_ϕ-self-concordant.*
- *V is L-relatively Lipschitz w.r.t. ϕ, i.e., $\|\nabla V(x)\|_{[\nabla^2\phi(x)]^{-1}} \leq L$ for all $x \in \mathcal{X}$.*

Let $(\mu_{kh})_{k\in\mathbb{N}}$ denote the law of MLMC and let $\beta' := \beta + 2LM_\phi$.

1. *(weakly convex case) Suppose that $\alpha = 0$. For any $\varepsilon \in [0, \sqrt{d}]$, if we take step size $h \asymp \frac{\varepsilon^2}{\beta' d}$, then for the mixture distribution $\bar{\mu}_{Nh} := N^{-1}\sum_{k=1}^{N} \mu_{kh}$ it holds that $\sqrt{\mathsf{KL}(\bar{\mu}_{Nh} \parallel \pi)} \leq \varepsilon$ after*

$$N = O\Big(\frac{\beta' d\, \mathcal{D}_\phi(\pi, \mu_0)}{\varepsilon^4}\Big) \qquad \text{iterations}\,.$$

2. *(strongly convex case) Suppose that $\alpha > 0$ and let $\kappa := \beta'/\alpha$ denote the "condition number". Then, for any $\varepsilon \in [0, \sqrt{d}]$, with step size $h \asymp \frac{\alpha\varepsilon^2}{\beta' d}$ we obtain $\sqrt{\alpha}\, \mathcal{D}_\phi(\pi, \mu_{Nh}) \leq \varepsilon$ and $\sqrt{\mathsf{KL}(\bar{\mu}_{Nh,2Nh} \parallel \pi)} \leq \varepsilon$ after*

$$N = O\Big(\frac{\kappa d}{\varepsilon^2} \log \frac{\alpha\, \mathcal{D}_\phi(\pi, \mu_0)}{\varepsilon^2}\Big) \qquad \text{iterations}\,,$$

where $\bar{\mu}_{Nh,2Nh} := N^{-1}\sum_{k=N+1}^{2N} \mu_{kh}$.

Similarly to Theorem 4.3.6, the theorem follows from a key recursion.

Lemma 10.2.21. *Under the assumptions of Theorem 10.2.20, if $h \in [0, \frac{1}{\beta}]$, then*

$$h\,\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) \le (1 - \alpha h)\,\mathcal{D}_\phi(\pi, \mu_{kh}) - \mathcal{D}_\phi(\pi, \mu_{(k+1)h}) + \beta' d h^2 .$$

We proceed to prove the lemma.

Proof. We follow the proof of Theorem 4.3.6, indicating the changes necessary to adapt the proof to MLMC. Recall that $\mathcal{E}(\mu) := \int V \,\mathrm{d}\mu$.

1. The forward step dissipates the energy. Let $Z \sim \pi$ be optimally coupled to X_{kh}. Then, applying the relative convexity and relative smoothness of V,

$$\begin{aligned}
\mathcal{E}(\mu_{kh}^+) - \mathcal{E}(\pi) &= \mathbb{E}[V(X_{kh}^+) - V(X_{kh}) + V(X_{kh}) - V(Z)] \\
&\le \mathbb{E}\big[\langle \nabla V(X_{kh}), X_{kh}^+ - X_{kh}\rangle + \beta\, D_\phi(X_{kh}^+, X_{kh}) \\
&\qquad + \langle \nabla V(X_{kh}), X_{kh} - Z\rangle - \alpha\, D_\phi(Z, X_{kh})\big] \\
&= \mathbb{E}\big[\langle \nabla V(X_{kh}), X_{kh}^+ - Z\rangle + \beta\, D_\phi(X_{kh}^+, X_{kh}) - \alpha\, D_\phi(Z, X_{kh})\big] . \qquad (10.2.22)
\end{aligned}$$

Next, by the Bregman proximal lemma (Lemma 10.2.13),

$$\langle \nabla V(X_{kh}), X_{kh}^+ - Z\rangle = \frac{1}{h}\left\{D_\phi(Z, X_{kh}) - D_\phi(Z, X_{kh}^+) - D_\phi(X_{kh}^+, X_{kh})\right\} .$$

Substituting this into (10.2.22) and using $h \le \frac{1}{\beta}$, it yields

$$\mathcal{E}(\mu_{kh}^+) - \mathcal{E}(\pi) \le \frac{1}{h}\left\{(1 - \alpha h)\,\mathcal{D}_\phi(\mu_{kh}, \pi) - \mathcal{D}_\phi(\mu_{kh}^+, \pi)\right\} . \qquad (10.2.23)$$

2. The flow step does not substantially increase the energy. We write

$$\mathcal{E}(\mu_{(k+1)h}) - \mathcal{E}(\mu_{kh}^+) = \mathbb{E}\big[V\big(\nabla\phi^*(X_{(k+1)h}^*)\big) - V\big(\nabla\phi^*(X_{kh}^*)\big)\big] .$$

Let $f(x) := V(\nabla\phi^*(x))$ and apply Itô's formula. Note that

$$\begin{aligned}
\nabla f(x) &= \nabla V(\nabla\phi^*(x))^\mathsf{T} \nabla^2\phi^*(x) = \nabla V(\nabla\phi^*(x))^\mathsf{T} [\nabla^2\phi(\nabla\phi^*(x))]^{-1} , \\
\nabla^2 f(x) &= [\nabla^2 V(\nabla\phi^*(x))]\,[\nabla^2\phi(\nabla\phi^*(x))]^{-1}\,[\nabla^2\phi^*(x)] \\
&\qquad + \nabla V(\nabla\phi^*(x))^\mathsf{T}\,[\nabla^2\phi(\nabla\phi^*(x))]^{-1}\,[\nabla^3\phi(\nabla\phi^*(x))]\,[\nabla^2\phi(\nabla\phi^*(x))]^{-2} .
\end{aligned}$$

Itô's formula decomposes $f(X_{(k+1)h}^*) - f(X_{kh}^*)$ into the sum of an integral and a stochastic integral; since the latter has mean zero, we focus on the first term.

$$\mathbb{E}[f(X_{(k+1)h}^*) - f(X_{kh}^*)]$$

$$= \mathbb{E} \int_{kh}^{(k+1)h} \left\langle \nabla^2 V(X_t) \left[\nabla^2 \phi(X_t)\right]^{-2}, \nabla^2 \phi(X_t) \right\rangle \mathrm{d}t$$
$$+ \mathbb{E} \int_{kh}^{(k+1)h} \left\langle \nabla V(X_t)^\mathsf{T} \left[\nabla^2 \phi(X_t)\right]^{-1} \left[\nabla^3 \phi(X_t)\right] \left[\nabla^2 \phi(X_t)\right]^{-2}, \nabla^2 \phi(X_t) \right\rangle \mathrm{d}t$$
$$= \mathbb{E} \int_{kh}^{(k+1)h} \left\langle \nabla^2 V(X_t), \left[\nabla^2 \phi(X_t)\right]^{-1} \right\rangle \mathrm{d}t \tag{10.2.24}$$
$$+ \mathbb{E} \int_{kh}^{(k+1)h} \mathrm{tr}\left(\nabla V(X_t)^\mathsf{T} \left[\nabla^2 \phi(X_t)\right]^{-1} \left[\nabla^3 \phi(X_t)\right] \left[\nabla^2 \phi(X_t)\right]^{-1}\right) \mathrm{d}t\,. \tag{10.2.25}$$

By relative smoothness, since $\nabla^2 V \preceq \beta \nabla^2 \phi$,

$$(10.2.24) \le \beta d h\,.$$

For (10.2.25), we use Lemma 10.2.15, which implies

$$(10.2.25) \le 2M_\phi \int_{kh}^{(k+1)h} \mathbb{E}\left[\left\|\left[\nabla^2 \phi(X_t)\right]^{-1} \nabla V(X_t)\right\|_{\nabla^2 \phi(X_t)} \mathrm{tr}\left(\left[\nabla^2 \phi(X_t)\right] \left[\nabla^2 \phi(X_t)\right]^{-1}\right)\right] \mathrm{d}t$$
$$\le 2M_\phi d \int_{kh}^{(k+1)h} \mathbb{E}\left[\|\nabla V(X_t)\|_{[\nabla^2 \phi(X_t)]^{-1}}\right] \mathrm{d}t \le 2LM_\phi dh\,.$$

Hence, we have proven

$$\mathcal{E}(\mu_{(k+1)h}) - \mathcal{E}(\mu^+_{kh}) \le \beta' dh\,. \tag{10.2.26}$$

3. The flow step dissipates the entropy. Let $\mu_t := \mathrm{law}(X_t) = \mathrm{law}(\nabla \phi^*(X_t^*))$. The mirror diffusion (10.2.11) evolves according to the vector field $-\left[\nabla^2 \phi\right]^{-1} \nabla \ln \mu_t$. Also, note that $\nabla_y D_\phi(x, y) = -\nabla^2 \phi(x)\,(y - x)$. Using these, one can show that

$$\partial_t \mathcal{D}_\phi(\pi, \mu_t) \le \mathbb{E}\langle \left[\nabla^2 \phi(X_t)\right]^{-1} \nabla \ln \mu(X_t), \left[\nabla^2 \phi(X_t)\right](Z - X_t)\rangle = \mathbb{E}\langle \nabla \ln \mu(X_t), (Z - X_t)\rangle\,,$$

where (Z, X_t) is an optimal coupling for $\mathcal{D}_\phi(\pi, \mu_t)$. Using the convexity of $\mathcal{H}$ along generalized geodesics (Theorem 10.2.18),

$$\mathcal{H}(\pi) - \mathcal{H}(\mu_t) \ge \mathbb{E}\langle \nabla \ln \mu(X_t), (Z - X_t)\rangle\,.$$

Using the fact that $t \mapsto \mathcal{H}(\mu_t)$ is decreasing (prove this from the Fokker–Planck equation for the mirror diffusion!), we then have

$$\mathcal{D}_\phi(\pi, \mu_{(k+1)h}) - \mathcal{D}_\phi(\pi, \mu^+_{kh}) \le h\left\{\mathcal{H}(\pi) - \mathcal{H}(\mu_{(k+1)h})\right\}. \tag{10.2.27}$$

Concluding the proof. Combine (10.2.23), (10.2.26), and (10.2.27) to conclude. □

To apply Theorem 10.2.20 to the problem of constrained sampling, we can choose ϕ to be a logarithmic barrier for $\mathcal{X}$, which is self-concordant. In the special case when $V = \phi$, the condition that ϕ is L-relatively Lipschitz with respect to itself is commonly expressed as saying that ϕ is a *self-concordant barrier* with parameters (L, M_ϕ). Self-concordant barriers are also a core part of the theory of interior point methods; in particular, it is known that every convex body in $\mathbb{R}^d$ admits a $(d, 2)$-self-concordant barrier, and that this is optimal (see [Che21b; LY21]). However, this situation is "cheating" because if we want to sample from $\pi \propto \exp(-\phi)$, it does not make sense to assume we can exactly simulate the mirror diffusion associated with ϕ.

Result for the non-smooth case. Although the preceding result applies when ϕ is a logarithmic barrier, it does not apply to perhaps one of the most classical applications of mirror descent: namely, $\mathcal{X}$ is the probability simplex in $\mathbb{R}^d$ and $\phi(x) \coloneqq \sum_{i=1}^d x_i \ln x_i$ is the entropy. The next result we formulate adopts assumptions which precisely match the usual ones for mirror descent in this context.

Theorem 10.2.28 ([AC21]). *Suppose that $\pi = \exp(-V)$ is the target distribution and that ϕ is the mirror map. Let $|||\cdot|||$ be a norm on $\mathbb{R}^d$. Assume:*

- *V is convex and L-Lipschitz w.r.t. the dual norm $|||\cdot|||_*$, in the sense that*

$$|||\nabla V(x)|||_* \le L \qquad \text{for all } x \in \mathcal{X}.$$

- *ϕ is 1-strongly convex w.r.t. $|||\cdot|||$.*

Let $(\mu_{kh})_{k\in\mathbb{N}}$ denote the law of MLMC. For any $\varepsilon > 0$, if we take step size $h \asymp \frac{\varepsilon^2}{L^2}$, then for the mixture $\bar{\mu}_{Nh} \coloneqq N^{-1} \sum_{k=1}^N \mu_{kh}$ it holds that $\sqrt{\mathsf{KL}(\bar{\mu}_{Nh} \,\|\, \pi)} \le \varepsilon$ after

$$N = O\Big(\frac{L^2\, \mathcal{D}_\phi(\pi, \mu_0^+)}{\varepsilon^4}\Big) \qquad \text{iterations}.$$

For example, it is a classical fact that the entropy is strongly convex w.r.t. the ℓ_1 norm. We leave the proof of the non-smooth case as Exercise 10.7.

10.3 Proximal Langevin

10.4 Stochastic Gradient Langevin

Bibliographical Notes

In the context of optimization, self-concordant barriers play a key role in interior point methods for constrained optimization [NN94; Bub15; Nes18]. Relative convexity and relative smoothness were introduced in [BBT17; LFN18].

The first use of mirror maps with the Langevin diffusion was via the *mirrored* Langevin algorithm (which is different from the *mirror* Langevin diffusion) in [Hsi+18]. The mirror Langevin diffusion was introduced in an earlier draft of [Hsi+18], as well as in [Zha+20]. In [Zha+20], Zhang et al. also studied the Euler–Maruyama discretization of the mirror Langevin diffusion (which differs from MLMC in that it discretizes the diffusion step as well), but they were unable to prove convergence of the algorithm; they were only able to prove convergence to a Wasserstein ball of non-vanishing radius around π, even as the step size tends to zero. They also conjectured that the non-vanishing bias of the algorithm is unavoidable. Subsequently, [Che+20b] studied the mirror Langevin diffusion in continuous time, and [AC21] introduced and studied the MLMC discretization, which *does* lead to vanishing bias (as $h \searrow 0$).

Since then, there have been further studying the non-vanishing bias issue: [Jia21] studied both the Euler–Maruyama and MLMC discretizations under a "mirror log-Sobolev inequality" and was only able to prove vanishing bias for the later discretization; [Li+22] showed that the Euler–Maruyama discretization has vanishing bias under stronger assumptions; and [GV22] studied MLMC as a special case of more general Riemannian Langevin algorithms. The bias issue is still not settled, and it is certainly of interest to obtain guarantees for fully discretized algorithms. Nevertheless, in our presentation, we have stuck with the analysis of [AC21] because it is the cleanest, and because it relies on assumptions which are well-motivated from convex optimization.

Exercises

Mirror Langevin

⊵ Exercise 10.1 (the mirror Langevin diffusion in the primal space)
Use Itô's formula (Theorem 1.1.18) to show that the mirror Langevin diffusion (10.2.3) in

the primal space solves the SDE

$$\mathrm{d}Z_t = \{-[\nabla^2\phi(Z_t)]^{-1}\,\nabla V(Z_t) - [\nabla^2\phi(Z_t)]^{-1}\,\nabla^3\phi(Z_t)\,[\nabla^2\phi(Z_t)]^{-1}\}\,\mathrm{d}t + \sqrt{2}\,[\nabla^2\phi(Z_t)]^{-1/2}\,\mathrm{d}B_t\,.$$

▷ **Exercise 10.2** (Markov semigroup theory for the mirror Langevin diffusion)
Here, we introduce the Markov semigroup perspective on the mirror Langevin diffusion.

1. Compute the generator of the mirror Langevin diffusion. Use this to show that π is stationary for the diffusion, and verify the equations (10.2.4) for the carré du champ and Dirichlet energy.

2. Let $\mathscr{L}_{\mathrm{dual}}$ denote the generator for $(Z_t^*)_{t\geq 0}$ (we write $\mathscr{L}_{\mathrm{dual}}$ instead of $\mathscr{L}^*$ to avoid confusion with the adjoint of $\mathscr{L}$). By computing $\partial_t\,\mathbb{E}\,f(Z_t^*)$, show that

$$\mathscr{L}_{\mathrm{dual}} f = \mathscr{L}(f\circ\nabla\phi)\,.$$

   Then, via a similar calculation to (10.2.5) and (10.2.6), show that the Dirichlet energy for $(Z_t^*)_{t\geq 0}$ can be expressed as

$$\mathscr{E}_{\mathrm{dual}}(f,g) = \int \langle \nabla f, [\nabla^2\phi^*]^{-1}\,\nabla g\rangle\,\mathrm{d}\pi^*\,,$$

   where $\pi^* := (\nabla\phi)_{\#}\pi$ is the stationary distribution of $(Z_t^*)_{t\geq 0}$.

3. Show that the mirror Poincaré inequality in Corollary 10.2.10 implies the following Poincaré inequality in the dual space: for all $f:\mathbb{R}^d\to\mathbb{R}$,

$$\mathrm{var}_{\pi^*} f \leq \frac{1}{\alpha}\,\mathbb{E}_{\pi^*}\langle \nabla f, [\nabla^2\phi^*]^{-1}\,\nabla f\rangle = \frac{1}{\alpha}\,\mathscr{E}_{\mathrm{dual}}(f,f)\,.$$

▷ **Exercise 10.3** (Newton Langevin diffusion)
Verify the SDE (10.2.7) for the Newton Langevin diffusion. What happens to the mirror descent iteration (10.2.2) when $\phi = V$?

▷ **Exercise 10.4** (affine invariance of Newton's method)
Verify the affine invariance of Newton's algorithm.

▷ **Exercise 10.5** (properties of the Bregman divergence)
In this exercise, we check basic properties of the Bregman divergence.

1. Prove the alternative definition of relative convexity/smoothness (Lemma 10.2.9).

2. If ϕ^* is the convex conjugate of ϕ, prove that $D_\phi(x, x') = D_{\phi^*}(\nabla\phi(x'), \nabla\phi(x))$.

3. Check the identity (10.2.12).

⊵ **Exercise 10.6** (generalized geodesic convexity of the entropy)
Generalize the proof of (1.4.3) to prove the convexity of entropy along generalized geodesics (Theorem 10.2.18).

⊵ **Exercise 10.7** (non-smooth guarantee for MLMC)
Adapt the proof of Theorem 4.3.11 using the techniques of this chapter to prove the non-smooth guarantee for MLMC (Theorem 10.2.28).

CHAPTER 11

Non-Log-Concave Sampling

In this chapter, we study the problem of sampling from a smooth but non-log-concave target. Although some results from previous chapters also cover some non-log-concave targets (such as targets satisfying a Poincaré or log-Sobolev inequality), these results do not encompass the full breadth of the non-log-concave sampling problem.

In general, one cannot hope for polynomial-time guarantees from sampling from non-log-concave targets in usual metrics such as total variation distance. Instead, taking inspiration from the literature on non-convex optimization, we will develop a notion of approximate first-order stationarity for sampling, and show that this goal can achieved via an averaged version of the LMC algorithm. This is based on the work [Bal+22].

11.1 Approximate First-Order Stationarity via Fisher Information

Suppose that V is smooth, but non-convex. In general, optimization lower bounds show that finding an approximate global minimizer of V is computationally intractable, i.e., the oracle complexity scales exponentially in the dimension d. To circumvent this, the notion of *approximate first-order stationarity* has arisen as the performance metric of choice in the non-convex optimization literature. Under this metric, we seek to find the minimal number of queries required to output a point x such that $\|\nabla V(x)\| \leq \varepsilon$.

Of course, in practice we may desire stronger guarantees, but first-order stationarity

is often a useful first step towards more detailed analysis, and it has the advantage that we can develop a general theory surrounding this notion. Note that in the convex case, finding a global minimizer is equivalent to finding a first-order stationarity point, so stationary point analysis can be viewed as a natural generalization of the convex optimization analysis to non-convex settings.

To develop a sampling analogue of this concept, we recall that the Langevin diffusion is the gradient flow of the KL divergence $\mathsf{KL}(\cdot \parallel \pi)$ w.r.t. the Wasserstein geometry (Section 1.4). Moreover, the gradient of the KL divergence at μ is $\nabla \ln(\mu/\pi)$, and the squared norm of the gradient is the Fisher information $\mathsf{FI}(\mu \parallel \pi) = \mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2]$. Hence, a reasonable definition of finding an approximate first-order stationary point in sampling is to output a sample from μ satisfying $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \le \varepsilon$. We will show shortly that it is indeed possible to achieve this goal in polynomially many queries to ∇V as soon as ∇V is Lipschitz, thereby establishing a framework for stationarity analysis in non-log-concave sampling. Before doing so, however, we pause to gain intuition for this solution concept.

Lack of spurious stationary points. An interesting feature of the Fisher information is that, unlike for general non-convex optimization, if π satisfies some mild regularity conditions (e.g., π has a smooth and positive density on $\mathbb{R}^d$), then there are no spurious stationary points: $\mathsf{FI}(\mu \parallel \pi) = 0$ implies $\mu = \pi$. This is a specific feature of the sampling problem.[1] The intuition behind the proof is straightforward: if $\mathsf{FI}(\mu \parallel \pi) = 0$, then $\nabla \ln(\mu/\pi) = 0$ (π-a.e.), so the density μ is proportional to π. Since μ is a probability measure, then μ must equal π. (See however the technical remark below.)

This might suggest that our goal of obtaining $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \le \varepsilon$ is too ambitious, because obtaining a small value of the Fisher information would solve the general problem of non-log-concave sampling. This is in fact not the case, and the devil is in the details: it is true that a small value of $\mathsf{FI}(\mu \parallel \pi)$ implies μ is close to π, but how small must $\mathsf{FI}(\mu \parallel \pi)$ be? For highly non-log-concave targets π, typically the Fisher information should be *exponentially small* in order for μ to be close to π in total variation distance.

We illustrate this point with an example: suppose that the target distribution π is a mixture of Gaussians in one dimension, $\pi = \frac{1}{2}\pi_- + \frac{1}{2}\pi_+$, where $\pi_\mp := \mathsf{normal}(\mp m, 1)$. Also, suppose that μ is a mixture of the same two Gaussians, but with the wrong mixing weights: $\mu = \frac{3}{4}\pi_- + \frac{1}{4}\pi_+$. Then, we leave the following computation to the reader (Exercise 11.1).

[1] In fact, the KL divergence $\mathsf{KL}(\cdot \parallel \pi)$ is always strictly convex with respect to taking convex combinations of measures, and hence always has a unique global minimum.

Proposition 11.1.1. *Let $m > 0$ and let $\pi_{\mp} := \mathsf{normal}(\mp m, 1)$. Let $\mu := \frac{3}{4}\,\pi_- + \frac{1}{4}\,\pi_+$ and $\pi := \frac{1}{2}\,\pi_- + \frac{1}{2}\,\pi_+$. Then, it holds that*

$$\liminf_{m\to\infty} \|\mu - \pi\|_{\mathrm{TV}} > 0$$

whereas

$$\mathsf{FI}(\mu \parallel \pi) \lesssim m^2 \exp\Big(-\frac{m^2}{2}\Big) \to 0 \qquad \text{as } m \to \infty\,.$$

Metastability. The example in Proposition 11.1.1 also provides an interpretation of the Fisher information. If we try to sample from the mixture of Gaussians π, then for $m \gg 1$ it takes an exponentially long time for the Langevin diffusion to jump to one mode from the other; this is the main reason behind the slow mixing of Langevin. Since it is hard to jump between the modes, it is difficult for the Langevin diffusion to "learn" the global mixing weights $(\frac{1}{2}, \frac{1}{2})$ of π. On the other hand, Proposition 11.1.1 shows that even with the wrong mixing weights $(\frac{3}{4}, \frac{1}{4})$, the Fisher information $\mathsf{FI}(\mu \parallel \pi)$ is small, demonstrating that the Fisher information is insensitive to the global weights.

The example in Proposition 11.1.1 therefore paints a cartoon picture of the behavior of the Langevin diffusion with target π, initialized at $\frac{3}{4}\,\delta_{-m} + \frac{1}{4}\,\delta_{+m}$: we expect that the Langevin diffusion quickly explores and captures the local structure of the modes but fails to jump between the modes, arriving at a distribution which resembles μ; it is this local mixing that a Fisher information bound captures. Meanwhile, the Langevin diffusion only obtains the correct global weights after an exponentially long waiting time.

The state μ is not truly stable for the Langevin diffusion: given enough time, the diffusion will eventually move away from μ and reach π. However, since states like μ persist for a very long period of time, they are usually called *metastable* in the statistical physics literature. A Fisher information bound can be interpreted as a way of quantitatively measuring the metastability phenomenon.

Technical remark. One has to be slightly careful with the definition of the Fisher information. For example, suppose that π is the standard Gaussian, and suppose that μ is the Gaussian restricted to the unit ball. Then, it is tempting to argue that the density of μ is proportional to that of π on the unit ball, and hence $\nabla \ln(\mu/\pi) = 0$ (μ-a.e.); from the expression $\mathsf{FI}(\mu \parallel \pi) = \mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2]$, it suggests that $\mathsf{FI}(\mu \parallel \pi) = 0$, and in particular μ is a spurious stationary point. However, this argument is *not* correct.

The reason is that μ does not have enough regularity w.r.t. π in order to apply the

formula $\mathsf{FI}(\mu \parallel \pi) = \mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2]$. Indeed, in order to apply the formula, we must require that the density $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}$ lie in an appropriate Sobolev space w.r.t. π (more precisely, $\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}$ should lie in the domain of the Dirichlet energy functional). If this does not hold, then we define the Fisher information to be infinite: $\mathsf{FI}(\mu \parallel \pi) = \infty$.

In our theorem below, the Fisher information bound should be interpreted as follows: $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \le \varepsilon$ means that μ has enough regularity w.r.t. π *and* $\mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2] \le \varepsilon^2$.

11.2 Fisher Information Bound

As before, we consider the interpolation of LMC,

$$X_t = X_{kh} - (t - kh)\,\nabla V(X_{kh}) + \sqrt{2}\,(B_t - B_{kh})\,, \qquad t \in [kh, (k+1)h]\,.$$

The Fisher information bound in the next theorem will be proven using the interpolation technique (§4.2).

Theorem 11.2.1 ([Bal+22]). *Let $(\mu_t)_{t\ge 0}$ denote the law of the interpolation of LMC with step size $h > 0$. Assume that $\pi \propto \exp(-V)$ where ∇V is β-Lipschitz. Then, for any step size $0 < h \le \frac{1}{4\beta}$, for all $N \in \mathbb{N}$,*

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \parallel \pi)\,\mathrm{d}t \le \frac{2\,\mathsf{KL}(\mu_0 \parallel \pi)}{Nh} + 6\beta^2 dh\,.$$

In particular, if $\mathsf{KL}(\mu_0 \parallel \pi) \le K_0$ and we choose $h = \sqrt{K_0}/(2\beta\sqrt{dN})$, then provided that $N \ge 9K_0/d$,

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \parallel \pi)\,\mathrm{d}t \le \frac{8\beta\sqrt{dK_0}}{\sqrt{N}}\,.$$

In order to translate the result into a more useful form, we recall that the Fisher information is convex in its first argument.

Lemma 11.2.2. *The Fisher information functional $\mathsf{FI}(\cdot \parallel \pi)$ is convex.*

Proof. Let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)$ be such that $\mathsf{FI}(\mu_0 \parallel \pi) \vee \mathsf{FI}(\mu_1 \parallel \pi) < \infty$. For $t \in (0,1)$, let

$\mu_t \coloneqq (1-t)\,\mu_0 + t\,\mu_1$, and write $f_t \coloneqq \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} = (1-t)\,f_0 + t\,f_1$. Then,

$$\begin{aligned}\mathsf{FI}(\mu_t \parallel \pi) = \int \frac{\|\nabla f_t\|^2}{f_t}\,\mathrm{d}\pi &\le (1-t)\int \frac{\|\nabla f_0\|^2}{f_0}\,\mathrm{d}\pi + t\int \frac{\|\nabla f_1\|^2}{f_1}\,\mathrm{d}\pi \\ &= (1-t)\,\mathsf{FI}(\mu_0 \parallel \pi) + t\,\mathsf{FI}(\mu_1 \parallel \pi)\end{aligned}$$

follows from the joint convexity of $(a,b) \mapsto \|a\|^2/b$ on $\mathbb{R}^d \times \mathbb{R}_{>0}$. □

Hence, for the averaged measure $\bar{\mu}_{Nh} \coloneqq \frac{1}{Nh}\int_0^{Nh} \mu_t\,\mathrm{d}t$, we have

$$\mathsf{FI}(\bar{\mu}_{Nh} \parallel \pi) \le \frac{1}{Nh}\int \mathsf{FI}(\mu_t \parallel \pi)\,\mathrm{d}t \tag{11.2.3}$$

and the guarantees of the theorem translate into guarantees for $\bar{\mu}_{Nh}$. Moreover, we can output a sample from $\bar{\mu}_{Nh}$ via the following procedure:

1. Pick a time $t \in [0, Nh]$ uniformly at random.
2. Let k be the largest integer such that $kh \le t$, and let X_{kh} denote the k-th iterate of the LMC algorithm.
3. Perform a partial LMC update

$$X_t = X_{kh} - (t-kh)\,\nabla V(X_{kh}) + \sqrt{2}\,(B_t - B_{kh})$$

and output X_t.

Combined with Theorem 11.2.1 and (11.2.3), and assuming that $\mathsf{KL}(\mu_0 \parallel \pi) = O(d)$, we conclude that it is possible to algorithmically obtain a sample from a measure μ with $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \le \varepsilon$ using $O(\beta^2 d^2/\varepsilon^4)$ queries to ∇V.

We now give the proof of Theorem 11.2.1, which combines the usual stationary point analysis in non-convex optimization with the interpolation argument.

Proof of Theorem 11.2.1. Recall from the proof of Theorem 4.2.6 that

$$\partial_t\,\mathsf{KL}(\mu_t \parallel \pi) \le -\frac{1}{2}\,\mathsf{FI}(\mu_t \parallel \pi) + 6\beta^2 d\,(t-kh)\,.$$

This inequality was obtained under the sole assumption that ∇V is β-Lipschitz. In Theorem 4.2.6, we proceeded to apply a log-Sobolev inequality, but here we will instead telescope this inequality. By integrating over $t \in [kh, (k+1)h]$,

$$\mathsf{KL}(\mu_{(k+1)h} \parallel \pi) - \mathsf{KL}(\mu_{kh} \parallel \pi) \le -\frac{1}{2}\int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \parallel \pi)\,\mathrm{d}t + 3\beta^2 d h^2\,.$$

Summing over $k = 0, 1, \dots, N-1$ and dividing by Nh,

$$\frac{1}{Nh} \int_0^{Nh} \mathsf{FI}(\mu_t \parallel \pi) \, \mathrm{d}t \le \frac{2 \, \mathsf{KL}(\mu_0 \parallel \pi)}{Nh} + 6\beta^2 dh \, .$$

This proves the first statement; the second statement is obtained by optimizing over the choice of h. □

11.3 Applications of the Fisher Information Bound

Asymptotic convergence of averaged LMC. Since Theorem 11.2.1 holds under very weak assumptions (only smoothness of V) and implies that the Fisher information can be driven to zero, and $\mathsf{FI}(\mu \parallel \pi) = 0$ implies that $\mu = \pi$, then putting these facts together leads to a straightforward proof of the asymptotic convergence of averaged LMC.

For a sequence of positive step sizes $(h_k)_{k \in \mathbb{N}^+}$, let $\tau_n := \sum_{k=1}^n h_k$ and consider the interpolation

$$X_t = X_{\tau_{n-1}} - (t - \tau_{n-1}) \nabla V(X_{\tau_{n-1}}) + \sqrt{2} \, (B_t - B_{\tau_{n-1}}) \, , \qquad t \in [\tau_{n-1}, \tau_n] \, . \tag{11.3.1}$$

Theorem 11.3.2. *Let $(\mu_t)_{t \ge 0}$ denote the law of the interpolation (11.3.1) of LMC, and suppose that the target is $\pi \propto \exp(-V)$ where ∇V is β-Lipschitz. Suppose that LMC is initialized at a measure μ_0 with $\mathsf{KL}(\mu_0 \parallel \pi) < \infty$, and that the sequence of step sizes satisfies $0 < h_k < \frac{1}{4\beta}$ for all $k \in \mathbb{N}^+$ together with the conditions*

$$\sum_{k=1}^{\infty} h_k = \infty \, , \qquad \text{and} \qquad \sum_{k=1}^{\infty} h_k^2 < \infty \, .$$

Write $\bar{\mu}_{\tau_n} := \frac{1}{\tau_n} \int_0^{\tau_n} \mu_t \, \mathrm{d}t$. Then, $\bar{\mu}_{\tau_n} \to \pi$ weakly.

Proof. By repeating the proof of Theorem 11.2.1 but incorporating time-varying step sizes, we obtain for $t \in [\tau_n, \tau_{n+1}]$

$$\partial_t \, \mathsf{KL}(\mu_t \parallel \pi) \le -\frac{1}{2} \, \mathsf{FI}(\mu_t \parallel \pi) + 6\beta^2 d \, (t - \tau_n) \, . \tag{11.3.3}$$

By integrating this inequality and summing,

$$\mathsf{KL}(\mu_{\tau_n} \parallel \pi) \le \mathsf{KL}(\mu_0 \parallel \pi) - \frac{1}{2} \int_0^{\tau_n} \mathsf{FI}(\mu_t \parallel \pi) \, \mathrm{d}t + 3\beta^2 d \sum_{k=1}^n h_k^2 \, . \tag{11.3.4}$$

Rearranging and using the convexity of the Fisher information, it yields

$$\mathsf{FI}(\bar{\mu}_{\tau_n} \,\|\, \pi) \le \frac{1}{\tau_n} \int_0^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\, \mathrm{d}t \le \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{\tau_n} + \frac{6\beta^2 d}{\tau_n} \sum_{k=1}^{\infty} h_k^2 \,. \tag{11.3.5}$$

On the other hand, if $t \in [\tau_n, \tau_{n+1}]$, then integrating (11.3.3) and combining with (11.3.4) yields

$$\mathsf{KL}(\mu_t \,\|\, \pi) \le \mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) + 3\beta^2 d\,(t - \tau_n)^2 \le \mathsf{KL}(\mu_0 \,\|\, \pi) + 6\beta^2 d \sum_{k=1}^{\infty} h_k^2 < \infty \,.$$

Therefore, $\{\mathsf{KL}(\mu_t \,\|\, \pi) \mid t \ge 0\}$ is bounded, and the convexity of the KL divergence implies that $\{\mathsf{KL}(\bar{\mu}_{\tau_n} \,\|\, \pi) \mid n \in \mathbb{N}^+\}$ is bounded. Since the sublevel sets of $\mathsf{KL}(\cdot \,\|\, \pi)$ are compact, to prove the theorem it suffices to show that every weak limit of $(\bar{\mu}_{\tau_n})_{n \in \mathbb{N}^+}$ is equal to π. Consider a subsequence of $(\bar{\mu}_{\tau_n})_{n \in \mathbb{N}^+}$ converging to a weak limit $\bar{\mu}$.

Taking $n \to \infty$ in (11.3.5) and noting that $\tau_n \to \infty$, we have $\mathsf{FI}(\bar{\mu}_{\tau_n} \,\|\, \pi) \to 0$ and thus along the subsequence as well. It is known that $\mathsf{FI}(\cdot \,\|\, \pi)$ is weakly lower semicontinuous, so $\mathsf{FI}(\bar{\mu} \,\|\, \pi) = 0$. However, since ∇V is Lipschitz, then π has a continuous and strictly positive density on $\mathbb{R}^d$, so $\mathsf{FI}(\bar{\mu} \,\|\, \pi) = 0$ entails $\bar{\mu} = \pi$ as desired. □

Convergence in total variation distance under a Poincaré inequality. As the example in Proposition 11.1.1 shows, for general non-log-concave targets a Fisher information bound does not translate into guarantees in other metrics. However, this can be carried out if π satisfies appropriate functional inequalities. For example, by a definition a log-Sobolev inequality for π states that

$$\mathsf{KL}(\mu \,\|\, \pi) \lesssim \mathsf{FI}(\mu \,\|\, \pi) \qquad \text{for all } \mu \in \mathcal{P}(\mathbb{R}^d) \,,$$

and in this case a Fisher information guarantee readily translates into a KL divergence guarantee; however, this is not very interesting because we have obtained a sharper KL divergence guarantee for targets π satisfying a log-Sobolev inequality in Theorem 4.2.6. Instead, we will show that under the weaker assumption of a Poincaré inequality, a Fisher information guarantee implies a total variation guarantee.

The key observation is the following implication of a Poincaré inequality.

Proposition 11.3.6. *Suppose that π satisfies a Poincaré inequality with constant C_{PI}.*

Then, for all $\mu \in \mathcal{P}(\mathbb{R}^d)$,

$$\|\mu - \pi\|_{\mathrm{TV}}^2 \le \frac{C_{\mathrm{PI}}}{4} \mathsf{FI}(\mu \parallel \pi) .$$

Proof. We can assume $\mu \ll \pi$; let $f := \frac{\mathrm{d}\mu}{\mathrm{d}\pi}$. The total variation distance has the expressions

$$\|\mu - \pi\|_{\mathrm{TV}} = \frac{1}{2} \int |f - 1| \, \mathrm{d}\pi = \frac{1}{2} \int \{(f \vee 1) - (f \wedge 1)\} \, \mathrm{d}\pi$$

which yields $\int (f \wedge 1) \, \mathrm{d}\pi = 1 - \|\mu - \pi\|_{\mathrm{TV}}$ and $\int (f \vee 1) \, \mathrm{d}\pi = 1 + \|\mu - \pi\|_{\mathrm{TV}}$. Using this,

$$\begin{aligned} \int \sqrt{f} \, \mathrm{d}\pi &= \int \sqrt{(f \wedge 1)\,(f \vee 1)} \, \mathrm{d}\pi \le \sqrt{\int (f \wedge 1) \, \mathrm{d}\pi \int (f \vee 1) \, \mathrm{d}\pi} \\ &= \sqrt{(1 - \|\mu - \pi\|_{\mathrm{TV}})\,(1 + \|\mu - \pi\|_{\mathrm{TV}})} = \sqrt{1 - \|\mu - \pi\|_{\mathrm{TV}}^2} . \end{aligned}$$

Therefore,

$$\|\mu - \pi\|_{\mathrm{TV}}^2 \le 1 - \left(\int \sqrt{f} \, \mathrm{d}\pi \right)^2 = \mathrm{var}_\pi \sqrt{f} .$$

This is sometimes called **Le Cam's inequality**; in statistics, the right-hand side is often written as $H^2(\mu, \pi)\,(1 - \frac{1}{4} H^2(\mu, \pi))$, where H^2 denotes the squared **Hellinger distance**.

Applying the Poincaré inequality,

$$\|\mu - \pi\|_{\mathrm{TV}}^2 \le C_{\mathrm{PI}} \, \mathbb{E}_\pi[\|\nabla \sqrt{f}\|^2] = \frac{C_{\mathrm{PI}}}{4} \mathsf{FI}(\mu \parallel \pi) . \qquad \square$$

Corollary 11.3.7. *Let $(\mu_t)_{t \ge 0}$ denote the law of the interpolation of* LMC *with step size $h > 0$. Assume that $\pi \propto \exp(-V)$, where ∇V is β-Lipschitz and π satisfies the Poincaré inequality with constant C_{PI}. Then, if $\mathsf{KL}(\mu_0 \parallel \pi) \le K_0$ and we choose step size $h = \sqrt{K_0}/(2\beta\sqrt{dN})$, then*

$$\|\bar{\mu}_{Nh} - \pi\|_{\mathrm{TV}}^2 := \left\| \frac{1}{Nh} \int_0^t \mu_t \, \mathrm{d}t - \pi \right\|_{\mathrm{TV}}^2 \le \frac{2 C_{\mathrm{PI}} \beta \sqrt{d K_0}}{\sqrt{N}} .$$

Bibliographical Notes

Exercises

⊵ **Exercise 11.1** (Fisher information for mixtures of Gaussians)
Prove Proposition 11.1.1.

Bibliography

[AC21] Kwangjun Ahn and Sinho Chewi. "Efficient constrained sampling via the mirror-Langevin algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, and Y. Dauphin. Vol. 34. Curran Associates, Inc., 2021, pp. 28405–28418.

[AB15] David Alonso-Gutiérrez and Jesús Bastero. *Approaching the Kannan-Lovász-Simonovits and variance conjectures*. Vol. 2131. Lecture Notes in Mathematics. Springer, Cham, 2015, pp. x+148.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Second. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2008, pp. x+334.

[AGS15] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. "Bakry–Émery curvature-dimension condition and Riemannian Ricci curvature bounds". In: *Ann. Probab.* 43.1 (2015), pp. 339–404.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552.

[Bal+22] Krishna Balasubramanian, Sinho Chewi, Murat A. Erdogdu, Adil Salim, and Matthew Zhang. "Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim

Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 2896–2923.

[Bar+18] Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, and Pierre-André Zitt. "Functional inequalities for Gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence". In: *Bernoulli* 24.1 (2018), pp. 333–353.

[BC13] Franck Barthe and Dario Cordero-Erausquin. "Invariances in variance estimates". In: *Proc. Lond. Math. Soc. (3)* 106.1 (2013), pp. 33–64.

[BBT17] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. "A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications". In: *Math. Oper. Res.* 42.2 (2017), pp. 330–348.

[BB99] Jean-David Benamou and Yann Brenier. "A numerical method for the optimal time-continuous mass transport problem and related problems". In: *Monge Ampère equation: applications to geometry and optimization (Deerfield Beach, FL, 1997)*. Vol. 226. Contemp. Math. Amer. Math. Soc., Providence, RI, 1999, pp. 1–11.

[BD01] Louis J. Billera and Persi Diaconis. "A geometric interpretation of the Metropolis–Hastings algorithm". In: *Statist. Sci.* 16.4 (2001), pp. 335–339.

[BB18] Adrien Blanchet and Jérôme Bolte. "A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions". In: *J. Funct. Anal.* 275.7 (2018), pp. 1650–1673.

[BL00] Sergey G. Bobkov and Michel Ledoux. "From Brunn–Minkowski to Brascamp–Lieb and to logarithmic Sobolev inequalities". In: *Geom. Funct. Anal.* 10.5 (2000), pp. 1028–1052.

[BH97] Serguei G. Bobkov and Christian Houdré. "Some connections between isoperimetric and Sobolev-type inequalities". In: *Mem. Amer. Math. Soc.* 129.616 (1997), pp. viii+111.

[BGG18] François Bolley, Ivan Gentil, and Arnaud Guillin. "Dimensional improvements of the logarithmic Sobolev, Talagrand and Brascamp–Lieb inequalities". In: *Ann. Probab.* 46.1 (2018), pp. 261–301.

[BV05] François Bolley and Cédric Villani. "Weighted Csiszár–Kullback–Pinsker inequalities and applications to transportation inequalities". In: *Ann. Fac. Sci. Toulouse Math. (6)* 14.3 (2005), pp. 331–352.

[Bor75] Christer Borell. "The Brunn–Minkowski inequality in Gauss space". In: *Invent. Math.* 30.2 (1975), pp. 207–216.

[Bor85] Christer Borell. "Geometric bounds on the Ornstein–Uhlenbeck velocity process". In: *Z. Wahrsch. Verw. Gebiete* 70.1 (1985), pp. 1–13.

[Bor00] Christer Borell. "Diffusion equations and geometric inequalities". In: *Potential Anal.* 12.1 (2000), pp. 49–71.

[BEZ20] Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. "Coupling and convergence for Hamiltonian Monte Carlo". In: *Ann. Appl. Probab.* 30.3 (2020), pp. 1209–1250.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481.

[Bub15] Sébastien Bubeck. "Convex optimization: algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[BBI01] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*. Vol. 33. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2001, pp. xiv+415.

[Caf00] Luis A. Caffarelli. "Monotonicity properties of optimal transportation and the FKG and related inequalities". In: *Comm. Math. Phys.* 214.3 (2000), pp. 547–563.

[CLL19] Yu Cao, Jianfeng Lu, and Yulong Lu. "Exponential decay of Rényi divergence under Fokker–Planck equations". In: *J. Stat. Phys.* 176.5 (2019), pp. 1172–1184.

[CLW21] Yu Cao, Jianfeng Lu, and Lihan Wang. "Complexity of randomized algorithms for underdamped Langevin dynamics". In: *Commun. Math. Sci.* 19.7 (2021), pp. 1827–1853.

[Car92] Manfredo P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Translated from the second Portuguese edition by Francis Flaherty. Birkhäuser Boston, Inc., Boston, MA, 1992, pp. xiv+300.

[CV21] José A. Carrillo and Urbain Vaes. "Wasserstein stability estimates for covariance-preconditioned Fokker–Planck equations". In: *Nonlinearity* 34.4 (Feb. 2021), pp. 2275–2295.

[Cha04] Djalil Chafai. "Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities". In: *J. Math. Kyoto Univ.* 44.2 (2004), pp. 325–363.

[CBL22] Niladri S. Chatterji, Peter L. Bartlett, and Philip M. Long. "Oracle lower bounds for stochastic gradient sampling algorithms". In: *Bernoulli* 28.2 (2022), pp. 1074–1092.

[CT93] Gong Chen and Marc Teboulle. “Convergence analysis of a proximal-like minimization algorithm using Bregman functions”. In: *SIAM J. Optim.* 3.3 (1993), pp. 538–543.

[CCN21] Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. “Dimension-free log-Sobolev inequalities for mixture distributions”. In: *Journal of Functional Analysis* 281.11 (2021), p. 109236.

[Che+22a] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. “Improved analysis for a proximal algorithm for sampling”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 2984–3014.

[CGP21] Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. “Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge”. In: *SIAM Rev.* 63.2 (2021), pp. 249–313.

[Che21a] Yuansi Chen. “An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture”. In: *Geom. Funct. Anal.* 31.1 (2021), pp. 34–61.

[Che+20a] Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. “Fast mixing of Metropolized Hamiltonian Monte Carlo: benefits of multi-step gradients”. In: *J. Mach. Learn. Res.* 21 (2020), Paper No. 92, 71.

[CV19] Zongchen Chen and Santosh S. Vempala. “Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions”. In: *Approximation, randomization, and combinatorial optimization. Algorithms and techniques*. Vol. 145. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2019, Art. No. 64, 12.

[Che+18] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. “Underdamped Langevin MCMC: a non-asymptotic analysis”. In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 300–323.

[Che21b] Sinho Chewi. “The entropic barrier is n-self-concordant”. In: *arXiv e-prints*, arXiv:2112.10947 (2021).

[Che+21a] Sinho Chewi, Murat A. Erdogdu, Mufan B. Li, Ruoqi Shen, and Matthew Zhang. “Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev”. In: *arXiv e-prints*, arXiv:2112.12662 (2021).

[Che+22b] Sinho Chewi, Patrik R. Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. "The query complexity of sampling from strongly log-concave distributions in one dimension". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 2041–2059.

[Che+20b] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. "Exponential ergodicity of mirror-Langevin diffusions". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 19573–19585.

[Che+21b] Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. "Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 1260–1300.

[CP23] Sinho Chewi and Aram-Alexandre Pooladian. "An entropic generalization of Caffarelli's contraction theorem via covariance inequalities". In: *Reports. Mathematical* 361 (2023), pp. 1471–1482.

[Cor02] Dario Cordero-Erausquin. "Some applications of mass transport to Gaussian-type inequalities". In: *Arch. Ration. Mech. Anal.* 161.3 (2002), pp. 257–269.

[Cor17] Dario Cordero-Erausquin. "Transport inequalities for log-concave measures, quantitative forms, and applications". In: *Canad. J. Math.* 69.3 (2017), pp. 481–501.

[CFM04] Dario Cordero-Erausquin, Matthieu Fradelizi, and Bernard Maurey. "The (B) conjecture for the Gaussian measure of dilates of symmetric convex sets and related problems". In: *J. Funct. Anal.* 214.2 (2004), pp. 410–427.

[CV18] Ben Cousins and Santosh Vempala. "Gaussian cooling and $O^*(n^3)$ algorithms for volume and Gaussian volume". In: *SIAM J. Comput.* 47.3 (2018), pp. 1237–1273.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006, pp. xxiv+748.

[Cut13] Marco Cuturi. "Sinkhorn distances: lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger. Vol. 26. Curran Associates, Inc., 2013.

[Dal17a] Arnak S. Dalalyan. "Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, July 2017, pp. 678–689.

[Dal17b] Arnak S. Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 79.3 (2017), pp. 651–676.

[DR20] Arnak S. Dalalyan and Lionel Riou-Durand. "On sampling from a log-concave density using kinetic Langevin diffusions". In: *Bernoulli* 26.3 (2020), pp. 1956–1988.

[DT12] Arnak S. Dalalyan and Alexandre B. Tsybakov. "Sparse regression learning by aggregation and Langevin Monte-Carlo". In: *J. Comput. System Sci.* 78.5 (2012), pp. 1423–1443.

[DZ10] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Vol. 38. Stochastic Modelling and Applied Probability. Corrected reprint of the second (1998) edition. Springer-Verlag, Berlin, 2010, pp. xvi+396.

[DMM19] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. "Analysis of Langevin Monte Carlo via convex optimization". In: *J. Mach. Learn. Res.* 20 (2019), Paper No. 73, 46.

[DM17] Alain Durmus and Éric Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". In: *Ann. Appl. Probab.* 27.3 (2017), pp. 1551–1587.

[DM19] Alain Durmus and Éric Moulines. "High-dimensional Bayesian inference via the unadjusted Langevin algorithm". In: *Bernoulli* 25.4A (2019), pp. 2854–2882.

[Dwi+19] Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. "Log-concave sampling: Metropolis–Hastings algorithms are fast". In: *Journal of Machine Learning Research* 20.183 (2019), pp. 1–42.

[Ebe11] Andreas Eberle. "Reflection coupling and Wasserstein contractivity without convexity". In: *C. R. Math. Acad. Sci. Paris* 349.19-20 (2011), pp. 1101–1104.

[Ebe16] Andreas Eberle. "Reflection couplings and contraction rates for diffusions". In: *Probab. Theory Related Fields* 166.3-4 (2016), pp. 851–886.

[EGZ19] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. "Couplings and quantitative contraction rates for Langevin dynamics". In: *Ann. Probab.* 47.4 (2019), pp. 1982–2010.

[Eld13] Ronen Eldan. "Thin shell implies spectral gap up to polylog via a stochastic localization scheme". In: *Geom. Funct. Anal.* 23.2 (2013), pp. 532–569.

[Eld15] Ronen Eldan. "A two-sided estimate for the Gaussian noise stability deficit". In: *Invent. Math.* 201.2 (2015), pp. 561–624.

[EM12] Matthias Erbar and Jan Maas. "Ricci curvature of finite Markov chains via convexity of the entropy". In: *Arch. Ration. Mech. Anal.* 206.3 (2012), pp. 997–1038.

[EHZ22] Murat A. Erdogdu, Rasa Hosseinzadeh, and Shunshi Zhang. "Convergence of Langevin Monte Carlo in chi-squared and Rényi divergence". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, 28–30 Mar 2022, pp. 8151–8175.

[Eva10] Lawrence C. Evans. *Partial differential equations*. Second. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2010, pp. xxii+749.

[FS18] Max Fathi and Yan Shu. "Curvature and transport inequalities for Markov chains in discrete spaces". In: *Bernoulli* 24.1 (2018), pp. 672–698.

[Fol99] Gerald B. Folland. *Real analysis*. Second. Pure and Applied Mathematics (New York). Modern techniques and their applications, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999, pp. xvi+386.

[Föl85] Hans Föllmer. "An entropy approach to the time reversal of diffusion processes". In: *Stochastic differential systems (Marseille-Luminy, 1984)*. Vol. 69. Lect. Notes Control Inf. Sci. Springer, Berlin, 1985, pp. 156–163.

[FLO21] James Foster, Terry Lyons, and Harald Oberhauser. "The shifted ODE method for underdamped Langevin MCMC". In: *arXiv e-prints*, arXiv:2101.03446 (2021).

[GT20] Arun Ganesh and Kunal Talwar. "Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 7222–7233.

[GM96] Wilfrid Gangbo and Robert J. McCann. "The geometry of optimal transportation". In: *Acta Math.* 177.2 (1996), pp. 113–161.

[GV22] Khashayar Gatmiry and Santosh S. Vempala. "Convergence of the Riemannian Langevin algorithm". In: *arXiv e-prints*, arXiv:2204.10818 (2022).

[GLL20] Rong Ge, Holden Lee, and Jianfeng Lu. "Estimating normalizing constants for log-concave distributions: algorithms and lower bounds". In: *STOC '20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, [2020] ©2020, pp. 579–586.

[Gen+20] Ivan Gentil, Christian Léonard, Luigia Ripani, and Luca Tamanini. "An entropic interpolation proof of the HWI inequality". In: *Stochastic Process. Appl.* 130.2 (2020), pp. 907–923.

[GLL22] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. "Private convex optimization via exponential mechanism". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 1948–1989.

[Goz+14] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. "Displacement convexity of entropy and related inequalities on graphs". In: *Probab. Theory Related Fields* 160.1-2 (2014), pp. 47–94.

[Gro07] Misha Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. English. Modern Birkhäuser Classics. Based on the 1981 French original, With appendices by M. Katz, P. Pansu and S. Semmes, Translated from the French by Sean Michael Bates. Birkhäuser Boston, Inc., Boston, MA, 2007, pp. xx+585.

[GM11] Olivier Guédon and Emanuel Milman. "Interpolating thin-shell and sharp large-deviation estimates for isotropic log-concave measures". In: *Geom. Funct. Anal.* 21.5 (2011), pp. 1043–1068.

[Han16] Ramon van Handel. *Probability in high dimension*. 2016.

[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*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 7366–7376.

[HS94] Bernard Helffer and Johannes Sjöstrand. "On the correlation for Kac-like models in the convex case". In: *J. Statist. Phys.* 74.1-2 (1994), pp. 349–409.

[HG14] Matthew D. Hoffman and Andrew Gelman. "The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo". In: *J. Mach. Learn. Res.* 15 (2014), pp. 1593–1623.

[Hsi+18] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. "Mirrored Langevin dynamics". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[Hsu02] Elton P. Hsu. *Stochastic analysis on manifolds*. Vol. 38. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2002, pp. xiv+281.

[IM12] Marcus Isaksson and Elchanan Mossel. "Maximally stable Gaussian partitions with discrete applications". In: *Israel J. Math.* 189 (2012), pp. 347–396.

[Jia+21] He Jia, Aditi Laddha, Yin Tat Lee, and Santosh Vempala. "Reducing isotropy and volume to KLS: an $O^*(n^3\psi^2)$ volume algorithm". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 961–974.

[Jia21] Qijia Jiang. "Mirror Langevin Monte Carlo: the case under isoperimetry". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 715–725.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM J. Math. Anal.* 29.1 (1998), pp. 1–17.

[KLM06] Ravi Kannan, László Lovász, and Ravi Montenegro. "Blocking conductance and mixing in random walks". In: *Combin. Probab. Comput.* 15.4 (2006), pp. 541–570.

[KLS95] Ravi Kannan, László Lovász, and Miklós Simonovits. "Isoperimetric problems for convex bodies and a localization lemma". In: *Discrete Comput. Geom.* 13.3-4 (1995), pp. 541–559.

[KKO18] Guy Kindler, Naomi Kirshner, and Ryan O'Donnell. "Gaussian noise sensitivity and Fourier tails". In: *Israel J. Math.* 225.1 (2018), pp. 71–109.

[Kla23] Bo'az Klartag. "Logarithmic bounds for isoperimetry and slices of convex sets". In: *Ars Inven. Anal.* (2023), Paper No. 4, 17.

[Kla+16] Bo'az Klartag, Gady Kozma, Peter Ralli, and Prasad Tetali. "Discrete curvature and abelian groups". In: *Canad. J. Math.* 68.3 (2016), pp. 655–674.

[KL22] Bo'az Klartag and Joseph Lehec. "Bourgain's slicing problem and KLS isoperimetry up to polylog". In: *arXiv e-prints*, arXiv:2203.15551 (2022).

[KP21] Bo'az Klartag and Eli Putterman. "Spectral monotonicity under Gaussian convolution". In: *arXiv preprint 2107.09496* (2021).

[Kol14] Alexander V. Kolesnikov. "Hessian metrics, $CD(K, N)$-spaces, and optimal transportation of log-concave measures". In: *Discrete Contin. Dyn. Syst.* 34.4 (2014), pp. 1511–1532.

[LO00] Rafał Latała and Krzysztof Oleszkiewicz. "Between Sobolev and Poincaré". In: *Geometric aspects of functional analysis*. Vol. 1745. Lecture Notes in Math. Springer, Berlin, 2000, pp. 147–168.

[LS88] Gregory F. Lawler and Alan D. Sokal. "Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality". In: *Trans. Amer. Math. Soc.* 309.2 (1988), pp. 557–580.

[Le 16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. French. Vol. 274. Graduate Texts in Mathematics. Springer, [Cham], 2016, pp. xiii+273.

[Led00] Michel Ledoux. "The geometry of Markov diffusion generators". In: vol. 9. 2. Probability theory. 2000, pp. 305–366.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. Vol. 89. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001, pp. x+181.

[LST20] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 2565–2597.

[LST21a] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Lower bounds on Metropolized sampling methods for well-conditioned distributions". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 18812–18824.

[LST21b] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Structured logconcave sampling with a restricted Gaussian oracle". In: *arXiv e-prints*, arXiv:2010.03106 (2021).

[LST21c] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Structured logconcave sampling with a restricted Gaussian oracle". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 2993–3050.

[LV17] Yin Tat Lee and Santosh S. Vempala. "Eldan's stochastic localization and the KLS hyperplane conjecture: an improved lower bound for expansion". In: *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*. IEEE Computer Soc., Los Alamitos, CA, 2017, pp. 998–1007.

[LY21] Yin Tat Lee and Man-Chung Yue. "Universal barrier is n-self-concordant". In: *Math. Oper. Res.* 46.3 (2021), pp. 1129–1148.

[Leh13] Joseph Lehec. "Representation formula for the entropy and functional inequalities". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 49.3 (2013), pp. 885–899.

[Léo17] Christian Léonard. "On the convexity of the entropy along entropic interpolations". In: *Measure theory in non-smooth spaces*. Partial Differ. Equ. Meas. Theory. De Gruyter Open, Warsaw, 2017, pp. 194–242.

[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*. Ed. by Sanjoy Dasgupta and Nika Haghtalab. Vol. 167. Proceedings of Machine Learning Research. PMLR, 29 Mar–01 Apr 2022, pp. 718–742.

[LZT22] Ruilin Li, Hongyuan Zha, and Molei Tao. "Sqrt(d) dimension dependence of Langevin Monte Carlo". In: *International Conference on Learning Representations*. 2022.

[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*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[LC22a] Jiaming Liang and Yongxin Chen. "A proximal algorithm for sampling". In: *arXiv e-prints*, arXiv:2202.13975 (2022).

[LC22b] Jiaming Liang and Yongxin Chen. "A proximal algorithm for sampling from non-smooth potentials". In: *arXiv e-prints*, arXiv:2110.04597 (2022).

[LV09] John Lott and Cédric Villani. "Ricci curvature for metric-measure spaces via optimal transport". In: *Ann. of Math. (2)* 169.3 (2009), pp. 903–991.

[LS93] László Lovász and Miklós Simonovits. "Random walks in a convex body and an improved volume algorithm". In: *Random Structures Algorithms* 4.4 (1993), pp. 359–412.

[LFN18] Haihao Lu, Robert M. Freund, and Yurii Nesterov. "Relatively smooth convex optimization by first-order methods, and applications". In: *SIAM J. Optim.* 28.1 (2018), pp. 333–354.

[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.

[Maa11] Jan Maas. "Gradient flows of the entropy for finite Markov chains". In: *J. Funct. Anal.* 261.8 (2011), pp. 2250–2292.

[MS19] Oren Mangoubi and Aaron Smith. "Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: numerical integrators". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 586–595.

[MV18] Oren Mangoubi and Nisheeth Vishnoi. "Dimensionally tight bounds for second-order Hamiltonian Monte Carlo". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[Mar96] Katalin Marton. "A measure concentration inequality for contracting Markov chains". In: *Geom. Funct. Anal.* 6.3 (1996), pp. 556–571.

[Mie13] Alexander Mielke. "Geodesic convexity of the relative entropy in reversible Markov chains". In: *Calc. Var. Partial Differential Equations* 48.1-2 (2013), pp. 1–31.

[Mil09] Emanuel Milman. "On the role of convexity in isoperimetry, spectral gap and concentration". In: *Invent. Math.* 177.1 (2009), pp. 1–43.

[Mir17] Ilya Mironov. "Rényi differential privacy". In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. 2017, pp. 263–275.

[MN15] Elchanan Mossel and Joe Neeman. "Robust optimality of Gaussian noise stability". In: *J. Eur. Math. Soc. (JEMS)* 17.2 (2015), pp. 433–482.

[Nea11] Radford M. Neal. "MCMC using Hamiltonian dynamics". In: *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handb. Mod. Stat. Methods. CRC Press, Boca Raton, FL, 2011, pp. 113–162.

[NY83] Arkadii S. Nemirovsky and David B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. Translated from the Russian and with a preface by E. R. Dawson. John Wiley & Sons, Inc., New York, 1983, pp. xv+388.

[Nes83] Yu. E. Nesterov. "A method for solving the convex programming problem with convergence rate $O(1/k^2)$". In: *Dokl. Akad. Nauk SSSR* 269.3 (1983), pp. 543–547.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*. Vol. 137. Springer Optimization and Its Applications. Springer, Cham, 2018, pp. xxiii+589.

[NN94] Yurii Nesterov and Arkadii S. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. Vol. 13. SIAM Studies in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994, pp. x+405.

[NW20] Richard Nickl and Sven Wang. "On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms". In: *arXiv e-prints*, arXiv:2009.05298 (2020).

[NW22] Marcel Nutz and Johannes Wiesel. "Entropic optimal transport: convergence of potentials". In: *Probab. Theory Related Fields* 184.1-2 (2022), pp. 401–424.

[Oll07] Yann Ollivier. "Ricci curvature of metric spaces". In: *C. R. Math. Acad. Sci. Paris* 345.11 (2007), pp. 643–646.

[Oll09] Yann Ollivier. "Ricci curvature of Markov chains on metric spaces". In: *J. Funct. Anal.* 256.3 (2009), pp. 810–864.

[OV12] Yann Ollivier and Cédric Villani. "A curved Brunn–Minkowski inequality on the discrete hypercube, or: what is the Ricci curvature of the discrete hypercube?" In: *SIAM J. Discrete Math.* 26.3 (2012), pp. 983–996.

[Ott01] Felix Otto. "The geometry of dissipative evolution equations: the porous medium equation". In: *Comm. Partial Differential Equations* 26.1-2 (2001), pp. 101–174.

[OV00] Felix Otto and Cédric Villani. "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality". In: *J. Funct. Anal.* 173.2 (2000), pp. 361–400.

[RV08] Luis Rademacher and Santosh Vempala. "Dispersion of mass and the complexity of randomized geometric algorithms". In: *Adv. Math.* 219.3 (2008), pp. 1037–1069.

[RS15] Firas Rassoul-Agha and Timo Seppäläinen. *A course on large deviations with an introduction to Gibbs measures*. Vol. 162. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2015, pp. xiv+318.

[San15] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Vol. 87. Progress in Nonlinear Differential Equations and their Applications. Calculus of variations, PDEs, and modeling. Birkhäuser/Springer, Cham, 2015, pp. xxvii+353.

[SL19] Ruoqi Shen and Yin Tat Lee. "The randomized midpoint method for log-concave sampling". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[Ste01] J. Michael Steele. *Stochastic calculus and financial applications*. Vol. 45. Applications of Mathematics (New York). Springer-Verlag, New York, 2001, pp. x+300.

[Stu06a] Karl-Theodor Sturm. "On the geometry of metric measure spaces. I". In: *Acta Math.* 196.1 (2006), pp. 65–131.

[Stu06b] Karl-Theodor Sturm. "On the geometry of metric measure spaces. II". In: *Acta Math.* 196.1 (2006), pp. 133–177.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights". In: *J. Mach. Learn. Res.* 17 (2016), Paper No. 153, 43.

[SC74] Vladimir N. Sudakov and Boris S. Cirel'son. "Extremal properties of half-spaces for spherically invariant measures". In: *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* 41 (1974). Problems in the theory of probability distributions, II, pp. 14–24, 165.

[Tal91] Michel Talagrand. "A new isoperimetric inequality and the concentration of measure phenomenon". In: *Geometric aspects of functional analysis (1989–90)*. Vol. 1469. Lecture Notes in Math. Springer, Berlin, 1991, pp. 94–124.

[Tal96] Michel Talagrand. "A new look at independence". In: *Ann. Probab.* 24.1 (1996), pp. 1–34.

[VW19] Santosh Vempala and Andre Wibisono. "Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8094–8106.

[Ver18] Roman Vershynin. *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018, pp. xiv+284.

[Vil03] Cédric Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370.

[Vil09] Cédric Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973.

[Wan11] Feng-Yu Wang. "Equivalent semigroup properties for the curvature-dimension condition". In: *Bull. Sci. Math.* 135.6-7 (2011), pp. 803–815.

[Wib18] Andre Wibisono. "Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 2093–3027.

[WSC21] Keru Wu, Scott Schmidler, and Yuansi Chen. "Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling". In: *arXiv e-prints*, arXiv:2109.13055 (2021).

[Zha+20] Kelvin S. Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. "Wasserstein control of mirror Langevin Monte Carlo". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 3814–3841.

Index

c-concavity, 60
c-conjugate potentials, 59
f-divergence, 53, 248

absolute continuity, 38
abstract Wiener space, 144
acceleration, 197
Alexandrov curvature, 114
Azuma–Hoeffding inequality, 133

Bakry–Émery theorem, 29
Benamou–Brenier formula, 61
blow-up, 88
Bobkov–Götze theorem, 93
Bochner identity, 68
Bonnet–Myers diameter bound, 117
bounded differences inequality, 133
Brascamp–Lieb inequality, 69, 276
Bregman divergence, 71, 276
Bregman proximal lemma, 279
Bregman transport cost, 71, 280
Bregman transport inequality, 71, 127, 282
Brenier potential, 36
Brenier's theorem, 33
    Bregman transport costs, 281
Brownian bridge, 149
Brownian motion, 4
Brunn–Minkowski inequality, 130

Caffarelli's contraction theorem, 79, 158
Cameron–Martin
    space, 143
    theorem, 143
carré du champ, 23
    iterated operator, 29, 68
Cheeger's inequality, 227
chi-squared divergence, 27
coarea inequality, 101
concentration function, 89, 97
conductance, 227
    s-conductance, 231
continuity equation, 39
covariant derivative, 109
Cramér–Rao inequality, 159
curvature

Gaussian, 111
Ricci, 112
coarse, 123
synthetic, 119
Riemann, 111
scalar, 112
sectional, 112
curvature-dimension condition, 29, 68, 116, 118

data-processing inequality, 54
descent lemma, 171
diffusion coefficient, 9
diffusion semigroup, 29, 73
Dirichlet energy, 23, 226
displacement interpolation, 44
divergence, 110
Donsker–Varadhan variational principle, 54
Doob martingale, 56
Doob's maximal inequality, 17
L^p version, 18
Doob's transform, 147
drift coefficient, 9

Efron–Stein inequality, 57
elementary process, 5, 15
entropy, 266
Euler–Maruyama discretization, 165
evolution variational inequality (EVI), 177
exponential map, 44

Föllmer drift, 151
Fano's inequality, 266
feasible start, 223
finite variation, 137
first variation, 45
Fisher information, 28, 74
flow map, 192
Fokker–Planck equation, 21
friction, 196

generalized geodesic, 282
geodesic, 42, 110, 114
geodesic convexity, 44
Girsanov's theorem, 146, 180
Gozlan's theorem, 96
Grönwall's lemma, 11
gradient, 108
gradient descent, 166
accelerated, 197
Gromov–Hausdorff convergence, 119
measured, 120

Hörmander's L^2 method, 69
Hamilton's equations of motion, 192
Hamiltonian, 192
Hamiltonian Monte Carlo (HMC)
ideal, 192
Metropolized (MHMC), 220
Hamilton–Jacobi equation, 61
Helffer–Sjöstrand identity, 128
Hellinger distance, 298
Herbst argument, 91
Hoeffding's lemma, 133
Holley–Stroock perturbation, 77
Hopf–Lax semigroup, 61
hypercontractivity, 129, 211

infinitesimal generator, 19, 225
isoperimetric profile, 99
isoperimetry
Cheeger, 101, 229
Gaussian, 98, 107, 236
sphere, 98, 134
Itô integral, 4, 16
Itô isometry, 7

Itô process, 9
Itô's formula, 10, 141

Jacobi equation, 121

Kannan–Lovász–Simonovits (KLS) conjecture, 125
Kantorovich problem, 31
Kazamaki's condition, 147
Kolmogorov's backward equation, 20
Kolmogorov's forward equation, 21
Kullback–Leibler (KL) divergence, 27
  chain rule, 54

Langevin diffusion, 3
Langevin Monte Carlo (LMC), 165
Laplace–Beltrami operator, 111
Latała–Oleszkiewicz inequality (LOI), 125
lazy chain, 223
Le Cam's inequality, 298
leapfrog integrator, 220
length, 113
Levi–Civita connection, 109
Lichnerowicz inequality, 134
Lie bracket, 111
local martingale, 8
localization, 7
localizing sequence, 7
log-Sobolev inequality (LSI), 28, 65, 171, 207
  defective, 82
  modified, 122, 226
logarithmic map, 44

Malliavin calculus, 160
manifold, 42
  Hessian, 109
Markov semigroup, 19
martingale, 5
  exponential, 146
Marton's tensorization, 85
McCann's interpolation, 44
mean squared analysis, 184
mesh, 137
metastability, 293
metric derivative, 38
metric geometry, 113
Metropolis-adjusted Langevin algorithm (MALA), 220
Metropolis–Hastings (MH) filter, 217
Metropolized random walk (MRW), 220
Minkowski content, 99
mirror descent, 273
mirror Langevin, 273
mirror Langevin Monte Carlo (MLMC), 278
mirror map, 273
model space, 114
Monge problem, 31
Monge–Ampère equation, 126
mutual information, 266

Newton Langevin diffusion, 275
no-U-turn sampler (NUTS), 203
Novikov's condition, 147

optimal transport, 30
  dual, 32
  entropic regularization, 156
  fundamental theorem, 33
  optimal transport plan, 30
Orlicz function, 90
Orlicz norm, 90
Ornstein–Uhlenbeck (OU) process, 57, 206
Otto calculus, 47
Otto–Villani theorem, 50

parallel transport, 110

Picard iteration, 13
Pinsker's inequality, 53, 133
Poincaré inequality (PI), 26, 65, 226
  L^p–L^q, 103
  local, 72
  mirror, 277
Poisson bracket, 205
Poisson equation, 70, 127
Polyak–Łojasiewicz (PL) inequality, 49, 171, 258
Prékopa–Leindler inequality, 129
progressive process, 15
proximal sampler, 243

quadratic variation, 139

Rényi divergence, 75, 207
randomized midpoint discretization, 187
reflection coupling, 184
rejection sampling, 215
relative convexity, 276
relative smoothness, 276
restricted Gaussian oracle (RGO), 244
reversibility, 22, 218
Ricci curvature, 68
Riemannian metric, 42, 108
Rothaus lemma, 82

Sanov's theorem, 96
Schrödinger bridge, 155
self-concordance, 280
semimartingale, 139
shifted ODE discretization, 203
Sobolev norm, 127
  inverse, 127
square integrable process, 15
stochastic calculus, 3
stochastic differential equation (SDE), 11
stochastic localization, 153
stopping time, 7
submartingale, 17
symplectic integrator, 220
synchronous coupling, 184

Talagrand's T_1 inequality, 66
Talagrand's T_2 inequality, 50, 65
tangent bundle, 108
tangent space, 42, 108
time reversal, 150, 250
total variation, 138
  total variation (TV) distance, 52, 138
  total variation norm, 138

underdamped Langevin diffusion, 196

vector field, 109
volume measure, 111

warm start, 223
Wasserstein gradient flow, 46
Wasserstein metric, 31, 60
Wiener measure, 142