Log-Concave Sampling

Sinho Chewi

This Version: February 22, 2026

Contents

Preface vi

Part 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 10
1.3 The Geometry of Optimal Transport 19
1.4 The Langevin SDE as a Wasserstein Gradient Flow 31
1.5 Overview of the Convergence Results 36
Bibliographical Notes 40
Exercises 41

2 Functional Inequalities 48
2.1 Overview of the Inequalities 48
2.2 Proofs via Markov Semigroup Theory 50
2.3 Operations Preserving Functional Inequalities 60
2.4 Concentration of Measure and Isoperimetry 67
2.5 Riemannian Manifolds 77
2.6 Discrete Space and Time 83
Bibliographical Notes 85
Exercises 86

3 Additional Topics in Stochastic Analysis 95
3.1 Quadratic Variation 95
3.2 Change of Measure in Path Space 99
3.3 Doob’s Transform 103
3.4 Föllmer Drift 107
3.5 Schrödinger Bridge 109
Bibliographical Notes 113

Exercises 114

Part II Complexity of Sampling 119

4 Analysis of Langevin Monte Carlo 121
4.1 Proof via Wasserstein Coupling 122
4.2 Proof via Interpolation Argument 125
4.3 Proof via Convex Optimization 129
4.4 Proof via Girsanov’s Theorem 133
Bibliographical Notes 135
Exercises 136

5 Faster Low-Accuracy Samplers 139
5.1 Randomized Midpoint Discretization 139
5.2 Hamiltonian Monte Carlo 143
5.3 The Underdamped Langevin Diffusion 147
Bibliographical Notes 161
Exercises 162

6 Convergence in Rényi Divergence 166
6.1 Analysis of LMC via Interpolation Argument 166
6.2 Analysis of LMC via Girsanov’s Theorem 170
6.3 Analysis of ULMC via Girsanov’s Theorem 175
Bibliographical Notes 178
Exercises 179

7 High-Accuracy Samplers 183
7.1 Rejection Sampling 183
7.2 The Metropolis–Hastings Filter 185
7.3 An Overview of High-Accuracy Samplers 187
7.4 Markov Chains in Discrete Time 191
7.5 Analysis of MALA for a Cold Start 197
7.6 Analysis of MALA for a Warm Start 200
Bibliographical Notes 204
Exercises 205

8 The Proximal Sampler 208
8.1 Introduction to the Proximal Sampler 209
8.2 Convergence under Strong Log-Concavity 211
8.3 Simultaneous Flow and Time Reversal 213
8.4 Convergence under Log-Concavity 215
8.5 Convergence under Functional Inequalities 216
8.6 Implementations of the RGO and Applications 218
Bibliographical Notes 220
Exercises 220

9 Lower Bounds for Sampling 223
9.1 A Brief Discussion of Query Complexity 223

9.2 Query Complexity in One Dimension 224
9.3 Query Complexity in Constant Dimension 231
9.4 Query Complexity for Gaussians 233
Bibliographical Notes 233
Exercises 234

10 Structured Sampling 236
10.1 Coordinate Langevin 236
10.2 Mirror Langevin 236
10.3 Proximal Langevin 246
10.4 Stochastic Gradient Langevin 246
Bibliographical Notes 246
Exercises 247

11 Non-Log-Concave Sampling 249
11.1 Approximate First-Order Stationarity via Fisher Information 249
11.2 Fisher Information Bound 251
11.3 Applications of the Fisher Information Bound 253
Bibliographical Notes 255
Exercises 255

References 257
Index 269

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 elegant mathematics and canonical theory that it seemed a shame not to have available an exposition which is accessible to 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. Due to space limitations, some of these additional topics appear as supplements on my website (chewisinho.github.io).

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., Bubeck (2015); Nesterov (2018); Chewi (2025).

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 Nemirovsky and Yudin (1983, §1) for a detailed discussion. We 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 (**Bayesian statistics and uncertainty quantification.**) 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 easily 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 $\vartheta_1, \dots, \vartheta_n$ whose distributions are approximately $p_{\vartheta|X}(\cdot \mid X)$, and the expectation can be approximated to arbitrary accuracy via the Monte Carlo averages $m^{-1}\sum_{i=1}^m \varphi(\vartheta_i)$.

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

---

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

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 (**Model aggregation, randomized algorithms, etc.**) Consider the following regression model: we observe i.i.d. pairs $\{(X_i, Y_i)\}_{i\in[n]}$, assumed to be drawn according to $Y_i = f_\vartheta(X_i) + \xi_i$ for some parameter $\vartheta \in \Theta$ and noise variables ξ_i, $i \in [n]$. Given a "prior" p_ϑ over Θ, the *exponentially weighted aggregate* (EWA) estimator with parameter $\beta > 0$ is the following weighted average:

$$\widehat{f} := \int f_\theta \, \widehat{p}_{\beta,\vartheta}(\mathrm{d}\theta)\,, \quad \text{where} \quad \widehat{p}_{\beta,\vartheta}(\mathrm{d}\theta) \propto \exp\Bigl(-\beta \sum_{i=1}^n (Y_i - f_\theta(X_i))^2\Bigr)\, p_\vartheta(\mathrm{d}\theta)\,.$$

Note the similarity with the Bayesian setting; indeed, $\widehat{p}_{\beta,\vartheta}$ can be interpreted as a posterior. The estimator $\widehat{f}$, however, can be regarded from a purely frequentist lens, as an average over "models" $\theta \in \Theta$ which downweights those which incur large empirical risk. This procedure is known generally as *model aggregation*, and the risk bound developed for the EWA estimator in Dalalyan (2017b) is closely related to *PAC–Bayes theory* (Alquier, 2024). Similar weighted averages appear in the field of online learning, under various names such as *learning with expert advice* or *multiplicative weights*. Computation of these estimators can again be handled via sampling methods.

4 (**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 should benefit from a healthy understanding of the analysis of sampling algorithms, and vice versa.

5 (**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 often refer to V as the *potential energy*.

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 0.0.1 (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), $i \in [n]$, 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 $\{X_i,\ i \in [n]\}$ to be deterministic and fixed, and we posit that the outcomes $\{Y_i,\ i \in [n]\}$ 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

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

and by independence, $p_{Y_1,\dots,Y_n \mid \vartheta}(y_1, \dots, y_n \mid \theta) = \prod_{i=1}^n p_{Y_i \mid \vartheta}(y_i \mid \theta)$. A Bayes rule computation yields the posterior $p_{\vartheta \mid Y_1,\dots,Y_n} \propto \exp(-V)$, where

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

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 with 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 $(X_t)_{t \geqslant 0}$ to the stochastic differential equation (SDE)

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

With a pure gradient flow $\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t$, 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 X_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 $(X_t)_{t \geqslant 0}$, the Langevin diffusion is a noisy gradient flow, but at the level of measures $(\pi_t)_{t\geqslant 0}$, it is *precisely* a gradient flow! This remarkable connection was introduced in the seminal work of Jordan, Kinderlehrer, and Otto (Jordan et al., 1998).

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 sometimes 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., Nickl and Wang, 2024). However, some of the results also apply to restricted classes of non-log-concave measures, and in Chapter 11 we shall 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

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

Since the Brownian increment $B_{(n+1)h} - B_{nh}$ 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 $\mathsf{d}(\mu_{Nh},\pi) \leqslant \varepsilon$ 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 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 log-concave 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 $\|\mu - \pi\|_{\mathrm{TV}} \leqslant \varepsilon$, uniformly over the following class of potentials: V is twice continuously differentiable, satisfying the conditions $0 < \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$.

What This Book Does Not Contain

Here, we point out a few egregious exclusions from our selection of topics. 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 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 $a \leqslant Cb$ for a universal constant $C > 0$. Similarly, $a \gtrsim b$ or $a = \Omega(b)$ mean that $a \geqslant cb$ for a universal constant $c > 0$, and $a \asymp b$ or $a = \Theta(b)$ mean that both $a \lesssim b$ and $a \gtrsim b$. We write $a = \widetilde{O}(b)$ to mean that $a = O(b \log^{O(1)} b)$, i.e., we suppress polylogarithmic factors, and we similarly use the notation $\widetilde{\Omega}$ and $\widetilde{\Theta}$. If a theorem holds under the condition "$a \ll b$", it means that it holds provided that a is smaller than b by a sufficiently small universal constant, and "$a \ll_{\log} b$" has a similar meaning except that polylogarithmic factors are also ignored.

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 $(\partial_1 f, \dots, \partial_d f)$, 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 $\Delta f = \operatorname{tr} \nabla^2 f = \sum_{i=1}^d \partial_i^2 f$.

Finally, we sometimes use t_- to denote $\lfloor t/h \rfloor\, h$, the largest multiple of the step size h which is smaller than t. This is not to be confused with the negative part of t.

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, Nima Anari, 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, Yunbum Kook, Daniel Lacker, 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, Atsushi Nitanda, Philippe Rigollet, Lionel Riou-Durand, Adil Salim, Ruoqi Shen, Vishwak Srinivasan, Austin J. Stromme, Taiji Suzuki, Molei Tao, Kevin Tian, Jure Vogrinc, Thuy-Duong Vuong, Guillaume Wang, Lihan Wang, Andre Wibisono, Ashia Wilson, Anru R. Zhang, Matthew S. Zhang, and Konstantinos Zygalakis.

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

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):

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,. \tag{1.0.1}$$

We begin with a brief primer on 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 do 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 the spirit of rapidly progressing to the study of sampling, the goal of this section is to familiarize the reader with the basic language of stochastic calculus, without delving into technical details. The mechanics of how to perform computations can be picked up through derivations in subsequent chapters. See Steele (2001); Le Gall (2016) for mathematical expositions. We treat further topics in stochastic calculus in Chapter 3.

1.1.1 The Itô Integral

In 1828, Robert Brown observed the curious jittery motion of particles of pollen in water. Subsequent study by mathematicians and physicists clarified the nature of the so-called "Brownian motion" as arising from collision with surrounding molecules. The mathematical theory for this phenomenon was initiated by the mathematician Louis Bachelier, as well as the physicist Albert Einstein in one of his breakthrough 1905 papers.

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.

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

1 $B_0 = 0$.
2 (Independence of increments) For all positive integers k and all $0 < t_1 < \cdots < t_k$, the random variables $(B_{t_1}, B_{t_2} - B_{t_1}, \dots, B_{t_k} - B_{t_{k-1}})$ are mutually independent.
3 (Law of the increments) For all $0 \leqslant s < t < \infty$,

$$B_t - B_s \sim \mathsf{normal}(0,\ (t-s)\, I_d)\,.$$

4 (Continuity of the paths) Almost surely, $t \mapsto B_t$ is continuous.

Our immediate goal is to develop the language of *stochastic calculus*, and in particular to define integrals involving Brownian motion: given a stochastic process $(\eta_t)_{t\geqslant 0}$ in $\mathbb{R}$, 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 main technical difficulty in defining the stochastic integral is the irregularity of Brownian motion: by definition, $B_t \sim \mathsf{normal}(0,t)$, so that $|B_t|$ is typically of size $\asymp \sqrt{t}$ for small $t > 0$. In particular, this means that Brownian motion is not differentiable at 0, or indeed, anywhere. Nevertheless, the stochastic integral can be meaningfully defined, and we summarize the main steps.

The standard measure-theoretic formalism works over a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ which is *complete*, *filtered*, and *right-continuous*, meaning that there is an increasing family $(\mathscr{F}_t)_{t\geqslant 0}$ 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 $s \geqslant 0$, 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 $t \geqslant 0$. Although we make use of this formalism in this section to define the Itô integral, it is not important for the remainder of the book.

Defining the Itô integral at a single time T.

The procedure for defining the Itô integral is to work our way up, starting with simple processes and then extending the definition to richer ones. The simplest type of process is a piecewise constant one: $(\eta_t)_{t\geqslant 0}$ is of the form

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

---

[1] Existence was established by Norbert Wiener, and hence the process is often called the *Wiener process*.

for some $0 \leqslant t_0 < t_1 < \cdots < t_k$, 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

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

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

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

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

Indeed, we deduce that $(\mathcal{I}_{[0,t]}(\eta))_{t\geqslant 0}$ is a martingale from two facts: (1) H_i is $\mathscr{F}_{t_i}$-measurable for each i, and (2) $(B_t)_{t\geqslant 0}$ is a martingale.

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

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

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

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 $L^2(\mathbb{P})$. Through this isometry, we can extend the definition of the stochastic integral to the closure of the elementary processes.

The closure turns out to consist of processes $(\eta_t)_{t\geqslant 0}$ that are *progressive*[2] and satisfies the integrability condition

$$\|\eta\|_{L^2(\mathbb{P}_T)}^2 = \mathbb{E}\int_0^T \eta_t^2 \, \mathrm{d}t < \infty \,. \tag{1.1.7}$$

This means that $(\eta_t)_{t\geqslant 0}$ can be approximated by elementary processes $\{(\eta_t^{(k)})_{t\geqslant 0} : k \in \mathbb{N}\}$ 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.

---

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

The key insight, due to Kiyosi Itô, is to go back to the approximating sequence $\{(\eta_t^{(k)})_{t\geqslant 0} : k \in \mathbb{N}\}$. For each k, the Itô integral is defined as an entire process $t \mapsto \mathcal{I}_{[0,t]}(\eta^{(k)})$ via (1.1.3), and moreover this process is a continuous martingale. One can then apply powerful results on martingale convergence, leading to the following theorem.

Theorem 1.1.8 (Definition of the Itô integral; Itô isometry)**.** *Suppose that* $(\eta_t)_{t\geqslant 0}$ *is progressive and satisfies* $\mathbb{E}\int_0^T \eta_t^2 \, \mathrm{d}t < \infty$. *Then, there exists a continuous martingale, denoted* $(\int_0^t \eta_s \, \mathrm{d}B_s)_{t\geqslant 0}$, *which is adapted to* $(\mathscr{F}_t)_{t\geqslant 0}$ *and satisfies*

$$\mathbb{E}\Big[\Big|\int_0^t \eta_s \, \mathrm{d}B_s\Big|^2\Big] = \mathbb{E}\int_0^t \eta_s^2 \, \mathrm{d}s\,, \qquad \text{for all } t \in [0,T]\,. \tag{1.1.9}$$

The formula (1.1.9) *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 typically taken, namely to expand the class of allowable integrands to progressive processes η with

$$\int_0^T \eta_s^2 \, \mathrm{d}s < \infty \qquad \text{almost surely}\,. \tag{1.1.10}$$

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). Localization is also needed to state the most useful version of Itô's formula in Theorem 1.1.19. However, this is not the focus of the book, and in what follows we usually brush over such localization arguments. For now, we simply introduce the definitions in order to convey that the idea is not too complicated.

Definition 1.1.11. A **stopping time** τ is a random variable such that for each $t \geqslant 0$, the event $\{\tau \leqslant t\}$ is $\mathscr{F}_t$-measurable.

Definition 1.1.12. 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 $n \in \mathbb{N}$, $(\eta_t \, \mathbb{1}\{t \leqslant \tau_n\})_{t\geqslant 0}$ 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 1.1.13. *If η is a progressive process satisfying the condition (1.1.10), then the sequence $(\tau_n)_{n\in\mathbb{N}}$ defined by*

$$\tau_n := \inf\Bigl\{t \geqslant 0 \Bigm| \int_0^t \eta_s^2 \, \mathrm{d}s \geqslant n\Bigr\} \wedge T$$

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

The idea now is straightforward, even if the execution is not: for each progressive process η satisfying (1.1.10), 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 to the stopped process $(\eta_{t\wedge\tau_n})_{t\geqslant 0}$, which gives us a continuous martingale

$$t \mapsto \int_0^t \eta_s \, \mathbb{1}\{s \leqslant \tau_n\} \, \mathrm{d}B_s \,. \tag{1.1.14}$$

Then, we can define the Itô integral of η to be a limit of the processes (1.1.14). The details are technical, and omitted.

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

Definition 1.1.15. A process $(M_t)_{t\geqslant 0}$ is a **local martingale** if it is adapted to the filtration $(\mathscr{F}_t)_{t\geqslant 0}$ 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. $(\mathscr{F}_t)_{t\geqslant 0}$.

Proposition 1.1.16. *If η is a progressive process satisfying (1.1.10), then the Itô integral $t \mapsto \int_0^t \eta_s \, \mathrm{d}B_s$ is a continuous local martingale.*

As an example, consider $\eta_t = \exp(B_t^2)$ for $t \geqslant 0$. For large $T \geqslant 0$, this does not satisfy the condition (1.1.7), but we can still define the Itô integral $\int_0^T \eta_t \, \mathrm{d}B_t$ using localizing sequences. In general, the key advantage of localization is that we do not have to check conditions such as (1.1.7) *a priori* in order to make sense of our computations.

Looking forward.

The construction of the Itô integral may seem rather abstract, and 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. Although counterintuitive, our answer to the heavy amount of abstraction is *more abstraction*. In the next subsection, we develop the single most important computation rule in stochastic calculus (along with the Itô isometry (1.1.9)), called Itô's formula, after which we 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 the details of what we have developed thus far are not needed 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 1.1.17. A stochastic process $(X_t)_{t\geqslant 0}$ is an **Itô process** if it is of the form

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

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

Implicit in the above definition is that the process should be well-defined: the coefficients $(b_t)_{t\geqslant 0}$ and $(\sigma_t)_{t\geqslant 0}$ should be progressive processes for which the integrals exist. In other words, $\int_0^t \|b_s\| \,\mathrm{d}s < \infty$ and $\int_0^t \|\sigma_s\|_{\mathrm{HS}}^2 \,\mathrm{d}s < \infty$ almost surely, where we write $\|M\|_{\mathrm{HS}}$ for the Hilbert–Schmidt (or Frobenius) norm of M, given by

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

Also, the random variable X_0 should be $\mathscr{F}_0$-measurable, and then the process $(X_t)_{t\geqslant 0}$ is also progressive.

We refer to $(b_t)_{t\geqslant 0}$ as the **drift coefficient** and $(\sigma_t)_{t\geqslant 0}$ as the **diffusion coefficient**. When the drift coefficient is zero, then $(X_t)_{t\geqslant 0}$ is simply an Itô integral, and thus a continuous local martingale (Proposition 1.1.16). Otherwise, for a non-zero drift coefficient, the process $(X_t)_{t\geqslant 0}$ is no longer necessarily a local martingale. As a shorthand, we often write the Itô process in differential form:

$$\mathrm{d}X_t = b_t \,\mathrm{d}t + \sigma_t \,\mathrm{d}B_t\,. \tag{1.1.18}$$

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 main engine of stochastic calculus.

Although the notation (1.1.18) 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 $\xi \sim \mathsf{normal}(0, I_N)$. Note that the $\sqrt{h}$ scaling comes from the fact that the Brownian increment $B_{t+h} - B_t$ has the $\mathsf{normal}(0, hI_N)$ 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:

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

This expression suggests that $(f(X_t))_{t\geqslant 0}$ 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 1.1.19 (Itô's formula)**.** *Let* $(X_t)_{t\geqslant 0}$ *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,* $(f(X_t))_{t\geqslant 0}$ *is also an Itô process which satisfies, for* $t \geqslant 0$*:*

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

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,

$$\mathbb{E}\, f(X_t) - \mathbb{E}\, f(X_0) = \int_0^t \mathbb{E}\bigl[\langle \nabla f(X_s), b_s\rangle + \frac{1}{2}\,\langle \nabla^2 f(X_s), \sigma_s\sigma_s^{\mathsf{T}}\rangle\bigr]\,\mathrm{d}s\,, \tag{1.1.20}$$

or in differential form,

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

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

$$\begin{aligned} f(t, X_t) - f(0, X_0) = \int_0^t \bigl\{\partial_s f(s, X_s) + \langle \nabla f(s, X_s), b_s\rangle + \frac{1}{2}\,\langle \nabla^2 f(s, X_s), \sigma_s\sigma_s^{\mathsf{T}}\rangle\bigr\}\,\mathrm{d}s \\ + \int_0^t \langle \sigma_s^{\mathsf{T}}\nabla f(s, X_s), \mathrm{d}B_s\rangle\,. \end{aligned}$$

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 $\sigma : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d\times N}$. Consider the stochastic differential equation (SDE)

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}B_t\,. \tag{1.1.21}$$

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

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

In this section, we state the basic existence and uniqueness result for this SDE. The result states that if the coefficients b, σ are Lipschitz in space uniformly in time, then the SDE admits a unique solution. Uniqueness means 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 $t \geqslant 0$ and $x, y \in \mathbb{R}^d$,*

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

Then, for any complete filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geqslant 0}, \mathbb{P})$ and $x \in \mathbb{R}^d$, there exists a unique solution $(X_t)_{t\in[0,T]}$ for the SDE (1.1.21) with $X_0 = x$. Moreover, the solution $(X_t)_{t\in[0,T]}$ is a Markov process.

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 exist 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 1.2). For our purposes, the assumption of Lipschitz coefficients suffices.

1.2 Markov Semigroup Theory

More thorough treatments of Markov semigroup theory can be found in Bakry et al. (2014); van Handel (2016). 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 $(X_t)_{t\geqslant 0}$ 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 $(X_t)_{t\geqslant 0}$, its associated **Markov semigroup** $(P_t)_{t\geqslant 0}$ 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 (Semigroup property)**.** *The Markov semigroup $(P_t)_{t\geqslant 0}$ satisfies $P_0 = \mathrm{id}$ and $P_s P_t = P_t P_s = P_{s+t}$ for all $s, t \geqslant 0$.*

The starting point of developing a calculus is 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 $(P_t)_{t\geqslant 0}$ 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. 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, but see Bakry et al. (2014, Chapter 3) for details.

Example 1.2.4. Let us compute the generator of the Langevin diffusion (1.0.1). In fact, the following computation is simply a consequence of Itô's formula (Theorem 1.1.19), but it does not hurt to derive this result from scratch. We approximate

$$X_t = X_0 - \int_0^t \nabla V(X_s)\,\mathrm{d}s + \sqrt{2}\,B_t = X_0 - t\,\nabla V(X_0) + \sqrt{2}\,B_t + o(t)\,.$$

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

$$\mathbb{E}\, f(X_t) = \mathbb{E}[f(X_0) + \langle \nabla f(X_0), -t\,\nabla V(X_0) + \sqrt{2}\,B_t\rangle + \langle \nabla^2 f(X_0)\,B_t, B_t\rangle] + o(t)\,.$$

Since B_t is mean zero and independent of X_0, with $\mathbb{E}[B_t B_t^{\mathsf{T}}] = t I_d$,

$$\begin{aligned}\mathbb{E}[f(X_t) \mid X_0 = x] &= f(x) - t\,\langle \nabla f(x), \nabla V(x)\rangle + t\,\operatorname{tr} \nabla^2 f(x) + o(t)\\ &= f(x) - t\,\langle \nabla f(x), \nabla V(x)\rangle + t\,\Delta f(x) + o(t)\,.\end{aligned}$$

Hence,

$$\mathscr{L} f(x) = \lim_{t\searrow 0} \frac{\mathbb{E}[f(X_t) \mid X_0 = x] - f(x)}{t} = \Delta f(x) - \langle \nabla V(x), \nabla f(x)\rangle\,.$$

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* $t \geqslant 0$, *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 $(P_t)_{t\geqslant 0}$.

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 backward 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* $t \geqslant 0$,

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

Here is another illuminating way to express these equations. For $u_t \coloneqq P_t f$, and $\pi_t = P_t^*\pi_0$,

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

The terms "backward" and "forward" can evoke confusion, so we will not use them. Instead, we will henceforth 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 $t \geqslant 0$.

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

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

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

Corollary 1.2.9. *The stationary distribution of the Langevin diffusion* (1.0.1) *with potential V is* $\pi \propto \exp(-V)$.

1.2.2 Reversibility and the Spectrum

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

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

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 $\int f\, P_t f\, \mathrm{d}\pi \geqslant 0$ 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,

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

and integrating this yields $\int (P_t f)^2\, \mathrm{d}\pi \leqslant \int P_t(f^2)\, \mathrm{d}\pi = \int f^2\, \mathrm{d}\pi$, 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 (Non-negativity of carré du champ)**.** *For any function f, $\Gamma(f,f) \geqslant 0$.*

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

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

and sending $t \searrow 0$ yields $\mathscr{L}(f^2) \geqslant 2f\,\mathscr{L}f$. (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. It also implies $\int f\,(-\mathscr{L})f\,\mathrm{d}\pi \geqslant 0$ for all f.

Corollary 1.2.15. *For a reversible Markov semigroup, $-\mathscr{L} \geqslant 0$.*

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

$$\partial_t \rho_t = \mathscr{L} \rho_t \,. \tag{1.2.16}$$

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

Example 1.2.17. Returning to the fundamental example of the Langevin diffusion, a computation shows that

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

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

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

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.

We can reformulate Theorem 1.2.14 in this case in an illuminating way as

$$-\mathscr{L} = \nabla^{*,\pi}\,\nabla \tag{1.2.18}$$

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

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 $(P_t)_{t\geqslant 0}$ 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 $\mathscr{L} \leqslant 0$. If we have a *spectral gap*

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

then we would expect that

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

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

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

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

The Poincaré constant C_{PI} corresponds to the inverse of the spectral gap. Under the assumption of a Poincaré inequality, we can establish convergence in $L^2(\pi)$ via Markov semigroup calculus:

$$\begin{aligned}\partial_t \|P_t f\|^2_{L^2(\pi)} &= \partial_t \int (P_t f)^2 \,\mathrm{d}\pi = 2\int P_t f\, \mathscr{L} P_t f \,\mathrm{d}\pi = -2\mathscr{E}(P_t f, P_t f)\\ &\leqslant -\frac{2}{C_{\mathsf{PI}}}\, \|P_t f\|^2_{L^2(\pi)}\,.\end{aligned}$$

Such differential inequalities are encountered repeatedly throughout the book and are handled with Grönwall's inequality.

Lemma 1.2.20 (Grönwall's lemma, differential form)**.** *Let $T > 0$ and let $g : [0,T] \to \mathbb{R}$ be differentiable. Suppose that there is a constant $c \in \mathbb{R}$ such that*

$$g'(t) \leqslant c g(t)\,, \qquad \forall t \in [0,T]\,.$$

Then,

$$g(t) \leqslant g(0)\exp(ct)\,, \qquad \forall t \in [0,T]\,.$$

Proof Differentiate the function $t \mapsto \exp(-ct)\, g(t)$ and apply the assumption.

$$\partial_t[\exp(-ct)\, g(t)] = \exp(-ct)\,[-cg(t) + g'(t)] \leqslant 0\,. \qquad \square$$

Hence, the differential inequality above proves the implication from the first to the second statement below. The converse implication is left as Exercise 1.6.

Theorem 1.2.21 (PI and variance decay)**.** *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 $t \geqslant 0$,*

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

In particular, we can apply this result to the semigroup corresponding to the Langevin diffusion (1.0.1) 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 (1.2.16) 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**

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

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

Theorem 1.2.22 (PI and χ^2 decay)**.** *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* $t \geqslant 0$,

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

Example 1.2.23. For the Langevin diffusion, the Poincaré inequality reads

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

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

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

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

$$\partial_t \mathsf{KL}(\pi_t \,\|\, \pi) = \partial_t \int \rho_t \log \rho_t \, \mathrm{d}\pi = \int (\log \rho_t + 1)\, \mathscr{L}\rho_t \, \mathrm{d}\pi = -\mathscr{E}(\rho_t, \log \rho_t)\,. \tag{1.2.24}$$

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

Definition 1.2.25. 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,

$$\mathsf{KL}(\rho\pi \,\|\, \pi) \leqslant \frac{C_{\mathsf{LSI}}}{2}\,\mathscr{E}(\rho, \log\rho)\,.$$

Theorem 1.2.26 (LSI and KL decay)**.** *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 $t \geqslant 0$,

$$\mathsf{KL}(\pi_t \,\|\, \pi) \leqslant \exp\Bigl(-\frac{2t}{C_{\mathsf{LSI}}}\Bigr)\,\mathsf{KL}(\pi_0 \,\|\, \pi)\,.$$

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 PI with constant $C_{\mathsf{PI}} \leqslant C_{\mathsf{LSI}}$ (Exercise 1.7).

Example 1.2.27. For the Langevin diffusion, the LSI reads

$$\frac{2}{C_{\mathsf{LSI}}}\,\mathsf{KL}(\mu \,\|\, \pi) \leqslant \mathbb{E}_\pi\Bigl\langle \nabla\frac{\mu}{\pi}, \nabla\log\frac{\mu}{\pi}\Bigr\rangle = \mathbb{E}_\mu\Bigl[\Bigl\|\nabla\log\frac{\mu}{\pi}\Bigr\|^2\Bigr] = 4\,\mathbb{E}_\pi\Bigl[\Bigl\|\nabla\sqrt{\frac{\mu}{\pi}}\Bigr\|^2\Bigr]\,.$$

The right-hand side of the above expression is important; it is known as the (relative) **Fisher information** $\mathsf{FI}(\mu \,\|\, \pi) := \mathbb{E}_\mu[\|\nabla\log(\mu/\pi)\|^2]$. 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

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

where for a function $g : \mathbb{R}^d \to \mathbb{R}_+$ we define $\operatorname{ent}_\pi(g) := \mathbb{E}_\pi(g\log g) - \mathbb{E}_\pi g \log\mathbb{E}_\pi g$. 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 1.2.28. The **iterated carré du champ** is the operator Γ_2 defined via

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

Recalling that $\Gamma(f, g) = \frac{1}{2}\,\{\mathscr{L}(fg) - f\,\mathscr{L}g - g\,\mathscr{L}f\}$, 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 1.2.29. The Markov semigroup is said to satisfy the **Bakry–Émery criterion** with constant $\alpha > 0$ if for all functions f,

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

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

We will prove the following theorem in Chapter 2.

Theorem 1.2.30 (Bakry–Émery)**.** *Consider a diffusion Markov semigroup. Assume that the curvature-dimension condition* $\mathsf{CD}(\alpha, \infty)$ *holds. Then, a log-Sobolev inequality holds with constant* $C_{\mathsf{LSI}} \leqslant 1/\alpha$*.*

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 1.7). For the Langevin diffusion, this condition amounts to the following theorem.

Theorem 1.2.31 (CD condition for Langevin)**.** *For the Langevin diffusion* (1.0.1)*, the curvature-dimension condition* $\mathsf{CD}(\alpha, \infty)$ *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 Villani (2003) is a standard reference for optimal transport; see also Ambrosio et al. (2008); Villani (2009b); Santambrogio (2015); Chewi et al. (2025) 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

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

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 (Existence)**.** *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 Gaspard 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 Leonid 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 focus on the case $c(x, y) := \|x - y\|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$ for concreteness.

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

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

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 is actually 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 $\int f(x)\,\gamma(\mathrm{d}x,\mathrm{d}y) = \int f(x)\,\mu(\mathrm{d}x)$. Doing the same for the constraint on the second marginal of γ, we can write the Kantorovich problem as an *unconstrained* min-max problem

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

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:

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

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

$$\mathcal{D}(\mu,\nu) := \Big\{(f,g) \in L^1(\mu) \times L^1(\nu) \;\Big|\; f(x) + g(y) \leqslant \frac{\|x-y\|^2}{2} \text{ for } \mu\text{-a.e. } x, \nu\text{-a.e. } y \in \mathbb{R}^d\Big\}.$$

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

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

Since inf sup $\geqslant$ 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 $\frac{1}{2}\int \|x-y\|^2\,\gamma^\star(\mathrm{d}x,\mathrm{d}y) = \int f^\star\,\mathrm{d}\mu + \int g^\star\,\mathrm{d}\nu$, then the primal and dual values coincide and $\gamma^\star$, $f^\star$, and $g^\star$ are all optimal.

By carefully studying the dual problem, we can 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 outline a proof 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 $(x_1, y_1), \dots, (x_n, y_n)$ 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

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

   Equivalently,

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

   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 Gangbo and McCann (1996, 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 $(x_1, y_1), \dots, (x_n, y_n)$, 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 $x_1, \dots, x_n \in \mathbb{R}^d$, and let σ be

a permutation of $[n]$. Then, from convexity,

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

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

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

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, pick any $(x_0, y_0) \in S$ and consider

$$\varphi(x) := \sup\Bigl\{\sum_{i=0}^n \langle y_i, x_{i+1} - x_i\rangle \Bigm| n \in \mathbb{N}^+,\ (x_1,y_1),\dots,(x_n,y_n) \in S,\ x_{n+1} = x\Bigr\}\,.$$

This characterization is due to Rockafellar (1966).

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 $x, y \in \mathbb{R}^d$,

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

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

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

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 $\partial\varphi$,

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

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

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.

Our discussion thus far applies to 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 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,

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

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

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

On the other hand, by the definition of $\bar\varphi^*$, we have $\bar\varphi(x) + \bar\varphi^*(y) \geqslant \langle x,y\rangle$ 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. See Exercise 1.10.

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 $\gamma_{2,3}$). 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 $\gamma_{1,2}^\star$ be optimal for (μ_1, μ_2), and let $\gamma_{2,3}^\star$ be optimal for (μ_2, μ_3). Let γ be obtained by gluing $\gamma_{1,2}^\star$ and $\gamma_{2,3}^\star$, and let $\gamma_{1,3} \in \mathcal{C}(\mu_1, \mu_3)$ denote the $(1,3)$-marginal of γ. Then,

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

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 $L^2(\gamma)$,

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

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 Ambrosio et al. (2008).

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 individual particle trajectories, and the *Eulerian* perspective which tracks the evolution of the aggregate 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, $(v_t)_{t\geqslant 0}$ 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 (Continuity equation)**.** *Let $(v_t)_{t\geqslant 0}$ be a family of vector fields and suppose that the process $(X_t)_{t\geqslant 0}$ 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 $\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0$. □

***Remark* 1.3.19.** We have not justified that the curve $(\mu_t)_{t\geqslant 0}$ is regular enough to make sense of the PDE (1.3.18) in the classical sense. To be rigorous, we should interpret (1.3.18) in the "weak" sense, that is, in duality with test functions: for all smooth and compactly supported functions φ, $\partial_t \int \varphi \, \mathrm{d}\mu_t = \int \langle \nabla\varphi, v_t\rangle \, \mathrm{d}\mu_t$. For pedagogical reasons, in the subsequent discussion, we simply *assume* that (1.3.18) holds in the classical sense.

Here, Theorem 1.3.17 shows that a family of vector fields and an initial condition μ_0 give rise to a curve of measures. The punchline is that the converse holds: *every* nice curve of measures $(\mu_t)_{t\geqslant 0}$ can be interpreted as the fluid flow along a family of vector fields, i.e., we can find vector fields $(v_t)_{t\geqslant 0}$ 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 $(v_t)_{t\geqslant 0}$ 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 1.3.20 (Curves of measures as fluid flows)**.** *Let $(\mu_t)_{t\geqslant 0}$ be an absolutely continuous curve of measures.*

1. *For any family of vector fields $(\tilde v_t)_{t\geqslant 0}$ such that the continuity equation (1.3.18) holds, we have $|\dot\mu|(t) \leqslant \|\tilde v_t\|_{L^2(\mu_t)}$ for all t.*
2. *Conversely, there exists a unique choice of vector fields $(v_t)_{t\geqslant 0}$ such that the continuity equation (1.3.18) holds and $\|v_t\|_{L^2(\mu_t)} \leqslant |\dot\mu|(t)$ 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

$$v_t = \lim_{\delta \searrow 0} \frac{T_{\mu_t \to \mu_{t+\delta}} - \mathrm{id}}{\delta} \qquad \text{in } L^2(\mu_t)\,, \tag{1.3.21}$$

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

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

However, $F_{t,t+\delta} - \mathrm{id} = \delta \tilde{v}_t + o(\delta)$, so letting $\delta \searrow 0$ we obtain $|\dot{\mu}|(t) \leqslant \|\tilde{v}_t\|_{L^2(\mu_t)}$. (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 $(v_t)_{t\geqslant 0}$ satisfying the continuity equation and such that $\|v_t\|_{L^2(\mu_t)} \leqslant |\dot{\mu}|(t)$. 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 $(v_t)_{t\geqslant 0}$ is uniquely determined.
3. Gradient vector fields are optimal. Here, we show that if the continuity equation holds for the family of vector fields $(v_t)_{t\geqslant 0}$ 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 lie in S.

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 $(\tilde{v}_t)_{t\geqslant 0}$ furnishes a possibly suboptimal transport map. To fix this, the flow map for the optimal $(v_t)_{t\geqslant 0}$ 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,

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

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

4 Existence of optimal vector fields. Finally, one can show for instance that vector fields defined via limits of transport maps as in (1.3.21) indeed satisfy the continuity equation and are gradient vector fields, and are therefore optimal. 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

$$\mathsf{d}(p,q) := \inf\Big\{\int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)}\, \mathrm{d}t \;\Big|\; \gamma : [0,1] \to \mathcal{M},\ \gamma(0) = p,\ \gamma(1) = q\Big\}. \tag{1.3.22}$$

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

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

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 treat it as one for the purpose of developing calculation rules.

If the continuity equation $\partial_t\mu_t + \mathrm{div}(\mu_t v_t) = 0$ 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 1.3.20.

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

$$W_2(\mu_0, \mu_1) = \inf\Big\{\int_0^1 \|v_t\|_{L^2(\mu_t)}\,\mathrm{d}t \;\Big|\; \partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0\Big\}\,. \tag{1.3.24}$$

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,* $(\mu_t)_{t\in[0,1]}$ *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, $\int_0^1 \|v_t\|_{L^2(\mu_t)}\,\mathrm{d}t \geqslant \int_0^1 |\dot\mu|(t)\,\mathrm{d}t$. For a partition $0 \leqslant t_0 < t_1 < \cdots < t_k \leqslant 1$,

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

As the size of the partition tends to zero, we obtain $W_2(\mu_0, \mu_1) \leqslant \int_0^1 |\dot\mu|(t)\,\mathrm{d}t$. 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 a process $(X_t)_{t\geqslant 0}$ 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

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

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 1.3.25. 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 $\mu_t := \operatorname{law}(X_t)$. Then, the curve $(\mu_t)_{t\in[0,1]}$ 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 1.3.26. 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 $t \in [0,1]$,

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

where d is the induced Riemannian distance (1.3.22).

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

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

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,

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

1.4 The Langevin SDE as a Wasserstein Gradient Flow

We are now ready to interpret the Langevin diffusion (1.0.1) 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 $(\mu_t)_{t\geqslant 0}$ 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 (Wasserstein gradient)**.** *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 $(\mu_t)_{t\geqslant 0}$ 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 be rewritten as

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

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

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

and therefore

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

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

$$\partial_t\mu_t = \operatorname{div}\big(\mu_t\nabla\log\frac{\mu_t}{\pi}\big)\,.$$

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 $(\pi_t)_{t\geqslant 0}$ 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 Otto (2001), and it is often known as **Otto calculus**; the interpretation of the Langevin diffusion in this context was put forth in the seminal work Jordan et al. (1998). The paper Otto (2001) 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 Ambrosio et al. (2008).

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 $\mathcal{F} := \mathsf{KL}(\cdot \,\|\, \pi)$ 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 $(\mu_t)_{t\geqslant 0}$ 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 $(x_t)_{t\geqslant 0}$, 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 $\mathcal{E}(\mu) := \int V\,\mathrm{d}\mu$ for the energy and $\mathcal{H}(\mu) := \int \mu \log \mu$ 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 $\alpha\,\|T-\mathrm{id}\|^2_{\mu_0}$, so that $\mathcal{E}$ is α-strongly convex.

The entropy is slightly trickier. Write $T_t := (1-t)\,\mathrm{id} + t\,T$. 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,

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

Already from the fact that $-\log\det$ is convex on the space of positive semidefinite matrices, we can see that $\partial_t^2 \mathcal{H}(\mu_t) \geqslant 0$. A more careful computation (Exercise 1.15) based on the derivatives of $-\log\det$ shows that

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

We have obtained the following result.

Theorem 1.4.4 (Convexity of the KL divergence)**.** *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.

First, we describe consequences of strong convexity for gradient flows in general. Let $(\mu_t)_{t\geqslant 0}$, $(\nu_t)_{t\geqslant 0}$ denote gradient flows for a functional $\mathcal{F}$ which is minimized at $\mu_\star$. More generally, the discussion here applies to gradient flows over Riemannian manifolds, with the appropriate change of definitions. We refer to Chewi (2025) for a reference on the corresponding results for Euclidean gradient flows.

We begin with a preliminary fact.

Theorem 1.4.5 (Derivative of W_2)**.** *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 Villani (2009b, Theorem 23.9). □

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

First, suppose that $\mathcal{F}$ is α-convex along Wasserstein geodesics; we allow $\alpha < 0$. Recall from Definition 1.3.26 that this is equivalent to the first-order condition

$$\mathcal{F}(\nu) \geqslant \mathcal{F}(\mu) + \langle \nabla_{W_2}\mathcal{F}(\mu), T_{\mu\to\nu} - \mathrm{id}\rangle_\mu + \frac{\alpha}{2}\, W_2^2(\mu,\nu) \qquad \text{for all } \mu,\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\,. \tag{1.4.6}$$

We now justify the following facts.

- The gradient flow is contractive:[3] for all $t \geqslant 0$,

  $$W_2(\mu_t,\nu_t) \leqslant \exp(-\alpha t)\, W_2(\mu_0,\nu_0)\,.$$

  See Exercise 1.17. In fact, contractivity is equivalent to α-convexity. When $\nu_0 = \mu_\star$, then $\nu_t = \mu_\star$ for all $t \geqslant 0$, so this also yields convergence to the minimizer.

---

[3] “Contractive” is a misnomer when $\alpha < 0$.

- The gradient flow converges in function value: for all $t \geqslant 0$,

  $$\mathcal{F}(\mu_t) - \inf \mathcal{F} \leqslant \frac{\alpha}{2\,(\exp(\alpha t) - 1)}\, W_2^2(\mu_0, \mu_\star)\,. \tag{1.4.7}$$

  In particular, for $\alpha = 0$, it yields

  $$\mathcal{F}(\mu_t) - \inf \mathcal{F} \leqslant \frac{1}{2t}\, W_2^2(\mu_0, \mu_\star)\,.$$

  To justify this, we use Theorem 1.4.5 and (1.4.6) to obtain

  $$\partial_t W_2^2(\mu_t, \mu_\star) = 2\,\langle \nabla_{W_2}\mathcal{F}(\mu_t), T_{\mu_t \to \mu_\star} - \mathrm{id}\rangle_{\mu_t} \leqslant -(\mathcal{F}(\mu_t) - \inf \mathcal{F}) - \alpha\, W_2^2(\mu_t, \mu_\star)\,.$$

  By differentiating $t \mapsto \exp(\alpha t)\, W_2^2(\mu_t, \mu_\star)$, one readily obtains (1.4.7).

  Convexity is not necessary for the gradient flow to converge. It is always true that

  $$\partial_t (\mathcal{F}(\mu_t) - \inf \mathcal{F}) = \langle \nabla_{W_2}\mathcal{F}(\mu_t), \underbrace{-\nabla_{W_2}\mathcal{F}(\mu_t)}_{\text{tangent vector of the curve}} \rangle_{\mu_t} = -\|\nabla_{W_2}\mathcal{F}(\mu_t)\|_{\mu_t}^2\,. \tag{1.4.8}$$

Thus, $t \mapsto \mathcal{F}(\mu_t)$ is monotonically decreasing, and lower bounds on the gradient norm yield rates of convergence. One such condition, known as a *gradient domination* condition, or a **Polyak–Lojasiewicz (PL) inequality**, reads

$$\|\nabla_{W_2}\mathcal{F}(\mu)\|_\mu^2 \geqslant 2\alpha\,(\mathcal{F}(\mu) - \inf \mathcal{F}) \qquad \text{for all } \mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\,. \tag{1.4.9}$$

- The functional $\mathcal{F}$ satisfies (1.4.9) if and only if

  $$\mathcal{F}(\mu_t) - \inf \mathcal{F} \leqslant \exp(-2\alpha t)\,(\mathcal{F}(\mu_0) - \inf \mathcal{F})$$

  for all gradient flows $(\mu_t)_{t \geqslant 0}$ for $\mathcal{F}$. This is a consequence of Grönwall's lemma (Lemma 1.2.20).
- The condition (1.4.9) is implied by strong convexity: starting from (1.4.6), we take $\nu = \mu_\star$, yielding

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

  where the last line uses Young's inequality and $\|T_{\mu \to \mu^\star} - \mathrm{id}\|_\mu = W_2(\mu, \mu^\star)$. However, (1.4.9) can hold even if $\mathcal{F}$ is not strongly convex.
- To convert convergence in function value to closeness to the minimizer, we note that (1.4.9) implies the following *quadratic growth* inequality for $\mathcal{F}$ around the minimizer:

  $$\mathcal{F}(\nu) - \inf \mathcal{F} \geqslant \frac{\alpha}{2}\, W_2^2(\nu, \mu^\star) \qquad \text{for all } \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\,. \tag{1.4.10}$$

  It is easy to show that strong convexity implies (1.4.10); simply take $\mu = \mu^\star$ in (1.4.6), so that $\nabla_{W_2}\mathcal{F}(\mu^\star) = 0$. The proof of the stronger assertion—that (1.4.9) implies (1.4.10)—is sketched as Exercise 1.16. In the case where $\mathcal{F}$ is not uniquely minimized, the right-hand side should be replaced by $\frac{\alpha}{2} \inf_{\mu_\star \in \arg\min \mathcal{F}} W_2^2(\nu, \mu_\star)$.

  Finally, when $\mathcal{F}$ satisfies neither strong convexity nor the PL condition, (1.4.8) still implies convergence of the gradient norm:

$$\inf_{t \in [0,T]} \|\nabla_{W_2}\mathcal{F}(\mu_t)\|_{\mu_t} \leqslant \sqrt{\frac{\mathcal{F}(\mu_0) - \inf \mathcal{F}}{T}}\,.$$

In some cases, this can be used as the basis of a soft argument that $\mu_t \to \mu_\star$.

We now specialize these results to $\mathcal{F} = \mathsf{KL}(\cdot \,\|\, \pi)$ with $\mu_\star = \pi \propto \exp(-V)$. Let $(\mu_t)_{t\geqslant 0}$, $(\nu_t)_{t\geqslant 0}$ denote the marginal laws for two copies of the Langevin diffusion (1.0.1). Then, α-convexity of V implies the bounds

$$W_2(\mu_t, \nu_t) \leqslant \exp(-\alpha t)\, W_2(\mu_0, \nu_0) \qquad \text{and} \qquad \mathsf{KL}(\mu_t \,\|\, \pi) \leqslant \frac{\alpha}{2\,(\exp(\alpha t) - 1)}\, W_2^2(\mu_0, \pi)\,.$$

In this setting, the first inequality is implied by the following theorem (see Exercise 1.17).

Theorem 1.4.11 (Contractivity of the Langevin diffusion)**.** *Suppose that $\nabla^2 V \succeq \alpha I_d$ for some $\alpha \in \mathbb{R}$. If $(X_t)_{t\geqslant 0}$ and $(X'_t)_{t\geqslant 0}$ denote two copies of the Langevin diffusion* (1.0.1) *with potential V and driven by the same Brownian motion, then we have the almost sure contraction*

$$\|X_t - X'_t\| \leqslant \exp(-\alpha t)\, \|X_0 - X'_0\|\,.$$

On the other hand, in this setting, the second inequality above is not sharp. The following is true:

$$\mathsf{KL}(\mu_t \,\|\, \pi) \leqslant \frac{\alpha}{2\,(\exp(2\alpha t) - 1)}\, W_2^2(\mu_0, \pi)\,. \tag{1.4.12}$$

A proof of this inequality is sketched in Exercise 1.18 (see also Exercise 2.17, Exercise 3.4, and Exercise 4.7). In the case $\alpha = 0$, it yields the convergence rate

$$\mathsf{KL}(\mu_t \,\|\, \pi) \leqslant \frac{1}{4t}\, W_2^2(\mu_0, \pi)\,. \tag{1.4.13}$$

The PL inequality (1.4.9) specializes to

$$\mathsf{KL}(\mu \,\|\, \pi) \leqslant \frac{1}{2\alpha}\, \Big\|\nabla \log \frac{\mu}{\pi}\Big\|_\mu^2 = \frac{1}{2\alpha}\, \mathsf{FI}(\mu \,\|\, \pi) \qquad \text{for all } \mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\,. \tag{1.4.14}$$

This is precisely the log-Sobolev inequality (see Example 1.2.27); we have therefore recovered the Bakry–Émery theorem (Theorem 1.2.30) that $\mathsf{CD}(\alpha, \infty)$ implies LSI, as well as Theorem 1.2.26 which asserted that LSI yields exponentially fast decay in the KL divergence.

For the Langevin diffusion, the quadratic growth inequality (1.4.10) reads

$$\mathsf{KL}(\mu \,\|\, \pi) \geqslant \frac{\alpha}{2}\, W_2^2(\mu, \pi) \qquad \text{for all } \mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)\,. \tag{1.4.15}$$

This is known as **Talagrand's T_2 inequality** and it is an example of a *transport inequality*. Such inequalities have been closely studied in relation to the concentration of measure phenomenon. See Chapter 2 and van Handel (2016, Chapter 4) for more details. The fact that LSI implies the T_2 inequality, which is a consequence of Exercise 1.16, is known as the **Otto–Villani theorem** (Otto and Villani, 2000).

1.5 Overview of the Convergence Results

1.5.1 Convergence Results and Initialization

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 α-strongly convex, if and only if: for all $\mu_0, \nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, if $(\mu_t)_{t\geqslant 0}$, $(\nu_t)_{t\geqslant 0}$ are Langevin diffusions started at μ_0 and ν_0 respectively, then $W_2^2(\mu_t, \nu_t) \leqslant \exp(-2\alpha t)\, W_2^2(\mu_0, \nu_0)$.

Moreover, in this case, $\mathsf{KL}(\mu_t \,\|\, \pi) \leqslant \alpha\, W_2^2(\mu_0, \pi)/(2\,(\exp(2\alpha t) - 1))$, which then reduces to $\mathsf{KL}(\mu_t \,\|\, \pi) \leqslant W_2^2(\mu_0, \pi)/(4t)$ for $\alpha = 0$.

- The target π satisfies the log-Sobolev inequality (LSI) with constant $1/\alpha$ if and only if for all $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, along the Langevin dynamics $(\mu_t)_{t\geqslant 0}$ started at μ_0 it holds that $\mathsf{KL}(\mu_t \,\|\, \pi) \leqslant \exp(-2\alpha t)\, \mathsf{KL}(\mu_0 \,\|\, \pi)$. 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 $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, along the Langevin dynamics $(\mu_t)_{t\geqslant 0}$ started at μ_0 it holds that $\chi^2(\mu_0 \,\|\, \pi) \leqslant \exp(-2\alpha t)\, \chi^2(\mu_0 \,\|\, \pi)$. 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}-PI. We also present convergence results which hold in Rényi divergences in Section 2.2.5.

In order to interpret these results, it is helpful to compare the sizes of these quantities to each other and at initialization.

Lemma 1.5.1 (Comparison between divergences)**.** *For any* $\mu \ll \pi$,

$$2\,\|\mu - \pi\|_{\mathrm{TV}}^2 \leqslant \mathsf{KL}(\mu \,\|\, \pi) \leqslant \log(1 + \chi^2(\mu \,\|\, \pi)) \leqslant \chi^2(\mu \,\|\, \pi)\,,$$

where $\|\cdot\|_{\mathrm{TV}}$ *is the total variation distance, introduced in Definition 1.5.4.*

Proof The first inequality is known as **Pinsker's inequality**. The second inequality is a special case of Exercise 2.9, and the last inequaltiy follows from $\log(1+x) \leqslant x$ for $x \geqslant 0$. □

In the strongly log-concave case, the following initialization bound holds.

Lemma 1.5.2 (Chi-squared divergence at initialization)**.** *Let* $\pi \propto \exp(-V)$ *be such that* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ *and* $\kappa := \beta/\alpha$*. Then, if* $x_\star$ *denotes the mode of* π *and* $\mu_0 := \mathsf{normal}(x_\star, \beta^{-1} I_d)$,

$$\log(1 + \chi^2(\mu_0 \,\|\, \pi)) \leqslant \frac{d}{2} \log \kappa\,.$$

By Lemma 1.5.1, it implies that $\mathsf{KL}(\mu_0 \,\|\, \pi) \leqslant \frac{d}{2} \log \kappa$. Recall that under a T_2 transport inequality with constant α^{-1} (which is implied by α^{-1}-LSI) we have $\mathsf{KL} \geqslant \frac{\alpha}{2}\, W_2^2$, so this further implies $W_2^2(\mu_0, \pi) \leqslant (d/\alpha) \log \kappa$. This inequality can be sharpened slightly.

Lemma 1.5.3 (Wasserstein distance at initialization)**.** *Let* $\pi \propto \exp(-V)$ *be such that* $0 \preceq \alpha I_d \preceq \nabla^2 V$*. Let* $x_\star$ *denote the mode of* π*. Then, for any* μ_0 *with mean* m *and covariance* Σ,

$$W_2^2(\mu_0, \pi) \leqslant \frac{d}{\alpha} + \operatorname{tr}\Sigma + \|m - x_\star\|\,\Bigl(\|m - x_\star\| + 2\sqrt{\frac{d}{\alpha}}\Bigr)\,,$$

Proof Let $X_0 \sim \mu_0$ and $Z \sim \pi$ be indepedent. Then,

$$\begin{aligned}
W_2^2(\mu_0, \pi) &\leqslant \mathbb{E}[\|X_0 - Z\|^2] = \mathbb{E}[\|X_0 - m + m - x_\star + x_\star - Z\|^2] \\
&= \mathbb{E}[\|X_0 - m\|^2 + \|m - x_\star\|^2 + \|Z - x_\star\|^2 - 2\,\langle m - x_\star, Z - x_\star\rangle] \\
&\leqslant \mathbb{E}[\|Z - x_\star\|^2] + \operatorname{tr}\Sigma + \|m - x_\star\| \left(\|m - x_\star\| + 2\sqrt{\mathbb{E}[\|Z - x_\star\|^2]}\right) \\
&\leqslant \frac{d}{\alpha} + \operatorname{tr}\Sigma + \|m - x_\star\| \left(\|m - x_\star\| + 2\sqrt{\frac{d}{\alpha}}\right),
\end{aligned}$$

where the bound on $\mathbb{E}[\|Z - x_\star\|^2]$ follows from Lemma 4.0.1. □

In particular, for $\mu_0 = \delta_{x_\star}$, it yields $W_2^2(\delta_{x_\star}, \pi) \leqslant d/\alpha$.

Therefore, at initialization, we typically have $W_2^2(\mu_0, \pi)$, $\mathsf{KL}(\mu_0 \,\|\, \pi) = \widetilde{O}(d)$ (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.

When we study discretization in Chapter 4, the W_2 contraction under strong log-concavity is the easiest to turn into a sampling guarantee. 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.

For more general initialization results, see Chewi et al. (2024a, Appendix A).

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

$$\begin{aligned}
\|\mu - \nu\|_{\mathrm{TV}} &:= \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| = \sup_{f : \mathcal{X} \to [0,1]} \left|\int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu\right| \\
&= \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \mathbb{1}\{x \neq y\}\,\gamma(\mathrm{d}x, \mathrm{d}y) = \frac{1}{2} \int \left|\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right| \mathrm{d}\lambda\,,
\end{aligned}$$

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 1.10) and an information divergence.

The family of information divergences can be further expanded as follows.

Definition 1.5.5. 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 relative to ν is

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

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,

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

For example, the TV distance corresponds to $f(x) = \frac{1}{2}\,|x-1|$, the KL divergence corresponds to $f(x) = x\log x$, 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.

In Section 2.2.5, we will introduce the closely related family of Rényi divergences.

We conclude by stating a few key facts (without complete proofs) about f-divergences. The first is the data-processing inequality. Below, given a Markov kernel P and a mixing measure μ, we write $\mu P := \int P(x, \cdot)\,\mu(\mathrm{d}x)$ for the mixture distribution.

Theorem 1.5.6 (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*

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

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,

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

To prove joint convexity, let $\lambda \in (0,1)$ and $Z \sim \mathsf{Bernoulli}(\lambda)$. Let $\mu_0, \mu_1, \nu_0, \nu_1 \in \mathcal{P}(\mathcal{X})$. Under $\boldsymbol{\mu}$, let X be drawn from μ_Z, and under $\boldsymbol{\nu}$, let X be drawn from ν_Z. Then,

$$\mathcal{D}_f\big(\mathrm{law}_{\boldsymbol{\mu}}(X,Z) \,\big\|\, \mathrm{law}_{\boldsymbol{\nu}}(X,Z)\big) = (1-\lambda)\,\mathcal{D}_f(\mu_0 \,\|\, \nu_0) + \lambda\,\mathcal{D}_f(\mu_1 \,\|\, \nu_1)\,.$$

Also, $\mathcal{D}_f(\mathrm{law}_{\boldsymbol{\mu}}(X,Z) \,\|\, \mathrm{law}_{\boldsymbol{\nu}}(X,Z)) \geqslant \mathcal{D}_f(\mathrm{law}_{\boldsymbol{\mu}}(X) \,\|\, \mathrm{law}_{\boldsymbol{\nu}}(X))$ by the data-processing inequality, where we note that

$$\mathcal{D}_f(\mathrm{law}_\mu(X) \,\|\, \mathrm{law}_\nu(X)) = \mathcal{D}_f\big((1-\lambda)\,\mu_0 + \lambda\,\mu_1 \,\big\|\, (1-\lambda)\,\nu_0 + \lambda\,\nu_1\big)\,. \qquad \square$$

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

Theorem 1.5.7 (Donsker–Varadhan variational principle)**.** *Suppose that $\mu, \nu \in \mathcal{P}(\mathcal{X})$, where $\mathcal{X}$ is a Polish space. Then,*

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

The theorem asserts that the functionals $\mu \mapsto \mathsf{KL}(\mu \,\|\, \nu)$ and $g \mapsto \log \mathbb{E}_\nu \exp g$ are convex conjugates of each other. See Dembo and Zeitouni (2010, Lemma 6.2.13) or Rassoul-Agha and Seppäläinen (2015, Theorem 5.4) for careful proofs, or see the remark after Lemma 2.3.5.

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

Lemma 1.5.8 (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 Steele (2001); Pavliotis (2014); Le Gall (2016). For Markov semigroup theory, see Bakry et al. (2014); van Handel (2016). For optimal transport, the core theory is developed in Villani (2003); Santambrogio (2015); Chewi et al. (2025), and for a rigorous development of Otto calculus see Ambrosio et al. (2008); Villani (2009b).

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 $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geqslant 0}, \mathbb{P})$ 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.

In accordance with most of the literature, we refer to the last conclusion of Theorem 1.3.8 as Brenier's theorem, after Brenier (1991). However, see also the earlier work of Cuesta and Matrán (1989). An elegant direct proof of strong duality can be found in Santambrogio (2015, Section 1.6.3).

The perspective of the Langevin diffusion as a Wasserstein gradient flow was introduced in Jordan et al. (1998); the application of Otto calculus to functional inequalities was given in Otto and Villani (2000); and the calculation rules for Otto calculus were set out in Otto (2001). These three papers are seminal and are worth reading carefully. An alternative (but related) approach to functional inequalities via optimal transport is given in Cordero-Erausquin (2002). The use of PL inequalities in optimizaton was popularized by Karimi et al. (2016). The formal proof of the Otto–Villani theorem in Exercise 1.16 was made rigorous via entropic interpolations in Gentil et al. (2020); see Blanchet and Bolte (2018) for a generalization.

The Efron–Stein inequality in Exercise 1.1 is just one example of the use of martingales to derive concentration inequalities; see Boucheron et al. (2013); van Handel (2016) for more on this topic.

The upper bound (1.E.2) in Exercise 1.9 is surprisingly sharp: it holds that

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

see Carrillo and Vaes (2021, Lemma 3.5).

The proof of the dynamical formulation of dual optimal transport in Exercise 1.11 is carried out rigorously in Villani (2003, Section 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 (Evans, 2010, Section 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 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 Villani (2003, Section 5.4).

The proof of the sharp KL convergence bound in Exercise 1.18 was carried out for $\alpha = 0$ in Otto and Villani (2001), and for $\alpha > 0$ in Liang et al. (2024).

Exercises

A Primer on Stochastic Calculus

⊵ **Exercise 1.1** (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:

$$\operatorname{var} f(X_1,\dots,X_n) \leqslant \mathbb{E}\sum_{k=1}^n \operatorname{var}\bigl(f(X_1,\dots,X_n) \bigm| X_{-k}\bigr)\,,$$

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

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

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., Boucheron et al. (2013).

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

⊵ **Exercise 1.2** (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 1.3** (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 $(B_t)_{t \geqslant 0}$. 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 $(\eta_t)_{t \geqslant 0}$ 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 1.4** (Generator for general SDEs)
Compute the generator for a general time-homogeneous SDE $\mathrm{d}X_t = b(X_t) \, \mathrm{d}t + \sigma(X_t) \, \mathrm{d}B_t$.

⊵ **Exercise 1.5** (Basic properties of the Markov semigroup)
Let $(P_t)_{t \geqslant 0}$ be a Markov semigroup with carré du champ Γ.

1. Prove that if $\phi : \mathbb{R} \to \mathbb{R}$ is convex, then $P_t \phi(f) \geqslant \phi(P_t f)$ and $\mathscr{L} \phi(f) \geqslant \phi'(f) \, \mathscr{L} f$ whenever the expressions are well-defined.
2. If $(X_t)_{t \geqslant 0}$ denotes the Markov process associated with the semigroup, assumed to have continuous sample paths, 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, $(f(X_t))_{t \geqslant 0}$ 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 $\Gamma(f, g) \leqslant \sqrt{\Gamma(f, f) \, \Gamma(g, g)}$.

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

⊵ **Exercise 1.6** (Functional inequalities and exponential decay)
Prove the equivalence between the Poincaré inequality and exponential decay of variance (Theorem 1.2.22), and the equivalence between the log-Sobolev inequality and exponential decay of entropy (Theorem 1.2.26).

⊵ **Exercise 1.7** (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

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

⊵ **Exercise 1.8** (Mixing of the Ornstein–Uhlenbeck process)
Consider the Ornstein–Uhlenbeck process $(X_t)_{t \geqslant 0}$ introduced in Exercise 1.3. 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

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

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, show that

$$\mathscr{E}(P_t f, P_t f) \leqslant \exp(-2t)\, \mathscr{E}(f, f)\,.$$

Explain why this implies a Poincaré inequality for the standard Gaussian distribution.

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

In Chapter 2, we will generalize these calculations to prove Theorem 1.2.30.

The Geometry of Optimal Transport

⊵ **Exercise 1.9** (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

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

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 $W_2(\nu_0, \nu_1) \geqslant W_2(\mu_0, \mu_1)$.

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

⊵ **Exercise 1.10** (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 if optimal dual potentials (f, g) exist, they can be taken to be **c-conjugates**, i.e., $f = g^c$ and $g = f^c$ where

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

and that

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

Under general conditions, equality holds; see Villani (2009b, 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 $p \geqslant 1$, we can define the **p-Wasserstein distance**

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

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

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

⊵ Exercise 1.11 (Dynamical formulations of optimal transport)

The formula (1.3.24) 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:

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

Although the problems (1.3.24) and (1.E.5) both identify geodesics in the Wasserstein space, the latter problem has more favorable properties. Namely, the minimizing curves in (1.3.24) 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* $p_t \coloneqq \mu_t v_t$ and write

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

which is now a strictly convex problem in the variables (μ, p). This convenient reformulation is known as the **Benamou–Brenier formula** (Benamou and Brenier, 1999).

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

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

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**

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

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

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

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 $[0,1]\times\mathbb{R}^d\to\mathbb{R}$, $(t,x)\mapsto u_t(x)$,

$$\begin{aligned}\int u_1\,\mathrm{d}\mu_1 - \int u_0\,\mathrm{d}\mu_0 &= \int_0^1\Bigl(\partial_t\int u_t\,\mathrm{d}\mu_t\Bigr)\,\mathrm{d}t = \int_0^1\Bigl(\int(\partial_t u_t\,\mu_t + u_t\,\partial_t\mu_t)\Bigr)\,\mathrm{d}t\\ &= \int_0^1\Bigl(\int\bigl(\partial_t u_t + \bigl\langle\nabla u_t, \frac{p_t}{\mu_t}\bigr\rangle\bigr)\,\mathrm{d}\mu_t\Bigr)\,\mathrm{d}t\,.\end{aligned}$$

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

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

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

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

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 $(\mu_t)_{t\in[0,1]}$ be a Wasserstein geodesic and let $(v_t)_{t\in[0,1]}$ 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 1.12** (Wasserstein space has non-negative curvature)
Let $(\mu_t)_{t\in[0,1]}$ denote a Wasserstein geodesic. By finding an appropriate coupling, prove that for all $t \in [0,1]$ and all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

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

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 Ambrosio et al. (2008, Section 7.3).

The Langevin SDE as a Wasserstein Gradient Flow

⊵ **Exercise 1.13** (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 $\partial_t \mu_t = \operatorname{div}(\mu_t \nabla \log(\mu_t/\pi))$, with $\pi \propto \exp(-V)$, then we can interpret this as a fluid flow: let $X_0 \sim \mu_0$, and $\dot X_t = -\nabla \log(\mu_t/\pi)(X_t)$, so that $X_t \sim \mu_t$. Compute $\mathbb{E}\,\varphi(X_\delta)$ 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 Sections 1.2 and 1.4, but it is illuminating to directly connect the Langevin diffusion to the Wasserstein gradient flow.

⊵ **Exercise 1.14** (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 $(\pi_t)_{t\geqslant 0}$ 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 1.15** (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 $\partial_t^2|_{t=0} \mathcal{F}(\mu_t) \leqslant \beta\, \|T - \mathrm{id}\|_{\mu_0}^2$.

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.

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

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

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 $(x_t)_{t\geqslant 0}$, $(y_t)_{t\geqslant 0}$ be two gradient flows for V. Show that $\|x_t - y_t\|^2 \leqslant \exp(-2\alpha t)\,\|x_0 - y_0\|^2$.
2 Next, prove Theorem 1.4.11. *Hint*: Apply Itô's formula (Theorem 1.1.19) to $f(x, x') := \|x - x'\|^2$.
3 Let $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$ be an α-convex functional, and let $(\mu_t)_{t\geqslant 0}$, $(\nu_t)_{t\geqslant 0}$ be gradient flows for $\mathcal{F}$. Prove that

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

using the following steps. First, compute the derivative of $t \mapsto W_2^2(\mu_t, \nu_t)$ using Theorem 1.4.5. Next, apply the strong convexity inequality (1.4.6) to obtain two inequalities $\mathcal{F}(\mu_t) \geqslant \mathcal{F}(\nu_t) + \cdots$ and $\mathcal{F}(\nu_t) \geqslant \mathcal{F}(\mu_t) + \cdots$. Adding these two inequalities, deduce that $\partial_t W_2^2(\mu_t, \nu_t) \leqslant -2\alpha\,W_2^2(\mu_t, \nu_t)$.

⊵ **Exercise 1.18** (Sharp KL convergence)
Let $(\mu_t)_{t\geqslant 0}$ denote the marginal law of the Langevin diffusion (1.0.1) with $\nabla^2 V \succeq \alpha I_d \succeq 0$.

1 Either prove via direct calculation that $\mathsf{FI}(\mu_t \,\|\, \pi) \leqslant \exp(-2\alpha t)\,\mathsf{FI}(\mu_0 \,\|\, \pi)$, or refer to the proof of Theorem 1.2.30 in Chapter 2.
2 Consider the Lyapunov function $\mathcal{L}_t := A_t\,\mathsf{FI}(\mu_t \,\|\, \pi) + 2B_t\,\mathsf{KL}(\mu_t \,\|\, \pi) + W_2^2(\mu_t, \pi)$. Choose A_t, B_t carefully to ensure that $\dot{\mathcal{L}}_t \leqslant -\alpha\mathcal{L}_t$, and thereby deduce the bounds

$$\mathsf{FI}(\mu_t \,\|\, \pi) \leqslant \frac{\alpha^2\,W_2^2(\mu_0, \pi)}{\exp(2\alpha t)\,(1 - \exp(-\alpha t))^2}\,, \qquad \mathsf{KL}(\mu_t \,\|\, \pi) \leqslant \frac{\alpha\,W_2^2(\mu_0, \pi)}{2\,(\exp(2\alpha t) - 1)}\,.$$

Overview of the Convergence Results

⊵ **Exercise 1.19** (Divergences at initialization)
Prove Lemma 1.5.2. In fact, prove the following stronger fact: $\log\sup(\mu_0/\pi) \leqslant \frac{d}{2}\log\kappa$.

⊵ **Exercise 1.20** (First moment of an exponential distribution)
Consider the probability measure $\pi \propto \exp(-\|\cdot\|)$ over $\mathbb{R}^d$. Show that the first moment behaves as $\int \|\cdot\|\,\mathrm{d}\pi = d$. This yields a natural example of a log-concave distribution for which the Wasserstein distance at initialization is expected to scale as d, rather than $\sqrt{d}$.

⊵ **Exercise 1.21** (Sharpness of the initialization bounds)
How sharp are Lemma 1.5.2 and Lemma 1.5.3? Consider the case where π is a Gaussian.

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

Let $C_c^\infty(\mathbb{R}^d)$ denote the set of compactly supported and smooth functions $\mathbb{R}^d \to \mathbb{R}$. The main inequalities that we study in this chapter are the following:

- the **Poincaré inequality** (**PI**), as specialized to the Langevin diffusion (see Example 1.2.23):

$$\operatorname{var}_\pi(f) \leqslant C_{\mathsf{PI}}\, \mathbb{E}_\pi[\|\nabla f\|^2]\,, \qquad \text{for all } f \in C_c^\infty(\mathbb{R}^d)\,.$$

- the **log-Sobolev inequality** (**LSI**), as specialized to the Langevin diffusion (see Example 1.2.27):

$$\operatorname{ent}_\pi(f^2) \leqslant 2C_{\mathsf{LSI}}\, \mathbb{E}_\pi[\|\nabla f\|^2]\,, \qquad \text{for all } f \in C_c^\infty(\mathbb{R}^d)\,.$$

- and **Talagrand's T_2 inequality**

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

In addition, using the W_1 metric introduced in Exercise 1.10, we consider

- **Talagrand's T_1 inequality**

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

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 new 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 compactly supported and smooth functions, once established they can be extended to a wider class of functions by density arguments; see Bakry et al. (2014, Chapter 3). 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 1.2.30) 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 $W_1 \leqslant W_2$, 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 1.7.
- In general, PI and T_1 are incomparable.

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 $c(x,y) \geqslant \delta \|x-y\|^2$ 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 $K \subseteq \mathbb{R}^d$,*

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

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

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

Proof sketch Fix $\lambda \in \mathbb{R}$. Using the dual formulation given in Exercise 1.10,

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

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 and the compact support of f, one can justify that the infimum is attained at a point h with $\|h\| = O(\varepsilon)$. Then,

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

Hence,

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

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 $f \in C_{\mathrm{c}}^\infty(\mathbb{R}^d)$ 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

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

On the other hand, the linearization (1.E.1) of the KL divergence in Exercise 1.7 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 $\Gamma_2(f,f) \geqslant \alpha\,\Gamma(f,f)$ (denoted $\mathsf{CD}(\alpha,\infty)$). 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

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

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

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

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 $\mathsf{CD}(0,\infty)$.

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 $\nabla^2 V \geq \alpha I_d \succ 0$, then the curvature-dimension condition $\mathsf{CD}(\alpha, \infty)$ holds, which was asserted as Theorem 1.2.31.

As a first illustration of the use of curvature, let us generalize the calculation in Exercise 1.8 from the Ornstein–Uhlenbeck diffusion to the Langevin diffusion. From the computation $\partial_t \operatorname{var}_\pi P_t f = -2\mathscr{E}(P_t f, P_t f) = -2\int \Gamma(P_t f, P_t f)\,\mathrm{d}\pi$ and $\operatorname{var}_\pi P_t f \to 0$ as $t \to \infty$ by ergodicity, we obtain the identity $\operatorname{var}_\pi f = 2\int_0^\infty \mathscr{E}(P_t f, P_t f)\,\mathrm{d}t$. For the Ornstein–Uhlenbeck semigroup, we have the explicit identity $\nabla P_t f = \exp(-t)\,P_t \nabla f$, which leads to $\mathscr{E}(P_t f, P_t f) \leqslant \exp(-2t)\,\mathscr{E}(f, f)$ and hence $\operatorname{var}_\pi f \leqslant \mathscr{E}(f, f)$, which is a Poincaré inequality for the standard Gaussian measure.

To extend this idea to more general diffusions, we show that the curvature-dimension condition is equivalent to a gradient bound.

Theorem 2.2.6 (Gradient bound)**.** *The following are equivalent.*

1 *The curvature-dimension condition* $\mathsf{CD}(\alpha, \infty)$ *holds.*

2 *For all* f *and* $t \geqslant 0$,

$$\Gamma(P_t f, P_t f) \leqslant \exp(-2\alpha t)\,P_t \Gamma(f, f)\,.$$

---

[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.

Proof $\underline{(1) \implies (2)}$: We differentiate the following quantity:

$$\partial_s [P_s \Gamma(P_{t-s} f, P_{t-s} f)] = P_s \bigl(\mathscr{L}\Gamma(P_{t-s} f, P_{t-s} f) - 2\Gamma(P_{t-s} f, \mathscr{L} P_{t-s} f)\bigr)\,.$$

Applying the definition of the iterated carré du champ and $\mathsf{CD}(\alpha, \infty)$ yields

$$\partial_s [P_s \Gamma(P_{t-s} f, P_{t-s} f)] = 2 P_s \Gamma_2(P_{t-s} f, P_{t-s} f) \geqslant 2\alpha\, P_s \Gamma(P_{t-s} f, P_{t-s} f)\,.$$

Integrating this from $s = 0$ to $s = t$ yields $P_t \Gamma(f, f) \geqslant \exp(2\alpha t)\, \Gamma(P_t f, P_t f)$.

$\underline{(2) \implies (1)}$: A Taylor expansion of the inequality around $t = 0$ yields

$$\Gamma(f, f) + 2t\, \Gamma(f, \mathscr{L} f) + o(t) \leqslant \Gamma(f, f) + t\,\{-2\alpha\, \Gamma(f, f) + \mathscr{L}\Gamma(f, f)\} + o(t)\,,$$

which clearly recovers the $\mathsf{CD}(\alpha, \infty)$ condition $\Gamma_2 \geqslant \alpha \Gamma$. □

Substituting this into $\operatorname{var}_\pi f = 2 \int_0^\infty \mathscr{E}(P_t f, P_t f)\, \mathrm{d}t$ now shows that as soon as the Markov semigroup satisfies $\mathsf{CD}(\alpha, \infty)$, the Poincaré inequality $\mathsf{PI}(\alpha^{-1})$ holds.

2.2.2 The Brascamp–Lieb Inequality

In this section, we prove a far-reaching refinement of the preceding argument and thereby establish the Brascamp–Lieb inequality, 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]

Lemma 2.2.7 (Dualizing the Poincaré inequality)**.** *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 $A : \mathbb{R}^d \to \mathsf{PD}(d)$ is a matrix-valued function mapping into the space of symmetric positive definite matrices such that for all smooth $u : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}_\pi[(\mathscr{L} u)^2] \geqslant \mathbb{E}_\pi \langle \nabla u, A\, \nabla u \rangle\,. \tag{2.2.8}$$

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

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

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

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

□

---

[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 $(-\mathscr{L})^{-1}$.

The point is that the condition (2.2.8) 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 2.2.9 (Brascamp–Lieb inequality)**.** *Let* $\pi \propto \exp(-V)$*, where* V *is strictly convex on* $\mathbb{R}^d$ *and twice continuously differentiable. Then, for all* $f : \mathbb{R}^d \to \mathbb{R}$*,*

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

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

***Remark* 2.2.10.** Lemma 2.2.7 shows that $\mathsf{PI}(\alpha^{-1})$ is equivalent to the *integrated* version of the $\mathsf{CD}(\alpha,\infty)$ condition: $\int \Gamma_2(f,f)\,\mathrm{d}\pi \geqslant \alpha \int \Gamma(f,f)\,\mathrm{d}\pi$ for all f.

In Exercise 2.3, we give another proof of Theorem 2.2.9 by linearizing a transport inequality. First, we introduce the transport cost.

Definition 2.2.11. 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

$$\mathcal{D}_V(\mu,\nu) := \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \int D_V(x,y)\,\gamma(\mathrm{d}x,\mathrm{d}y)\,.$$

The Bregman transport cost will also play a key role in Section 10.2, in which we will prove the following transport inequality (see also Exercise 2.3).

Theorem 2.2.12 (Bregman transport inequality)**.** *Suppose that* $V : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable. Then, for* $\pi \propto \exp(-V)$ *and all* $\mu \in \mathcal{P}(\mathbb{R}^d)$*,*

$$\mathcal{D}_V(\mu,\pi) \leqslant \mathsf{KL}(\mu \,\|\, \pi)\,.$$

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. When V is strongly convex, $\nabla^2 V \succeq \alpha I_d \succ 0$, then $D_V(x,y) \geqslant \frac{\alpha}{2}\,\|x-y\|^2$, so the Bregman transport inequality implies $\mathsf{T}_2(\alpha^{-1})$.

2.2.3 Proof of the Bakry–Émery Theorem

In this section, we prove the Bakry–Émery theorem (Theorem 1.2.30), i.e., that $\mathsf{CD}(\alpha, \infty)$ implies $\mathsf{LSI}(\alpha^{-1})$. We introduce the notion of a diffusion semigroup.

Definition 2.2.13. The Markov semigroup $(P_t)_{t \geqslant 0}$ 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:

$$\Gamma(\phi \circ f, g) = \phi'(f)\, \Gamma(f, g)\,. \tag{2.2.14}$$

More generally, for functions $f_1, \dots, f_k$ and $\Psi : \mathbb{R}^k \to \mathbb{R}$,

$$\Gamma\big(\Psi(f_1, \dots, f_k), g\big) = \sum_{i=1}^k (\partial_i \Psi)(f_1, \dots, f_k)\, \Gamma(f_i, g)\,.$$

Equivalently, the chain rule can be stated for the generator (Exercise 2.8):

$$\mathscr{L}\phi(f) = \phi'(f)\, \mathscr{L}f + \phi''(f)\, \Gamma(f, f)\,. \tag{2.2.15}$$

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 1.2.30) 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 $\phi(x) := x \log x$, 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

$$\begin{aligned} \operatorname{ent}_\pi^\phi f &:= \int \phi(f)\, \mathrm{d}\pi - \phi\Big(\int f\, \mathrm{d}\pi\Big) = -\int_0^\infty \Big(\partial_t \int \phi(P_t f)\, \mathrm{d}\pi\Big)\, \mathrm{d}t \\ &= \int_0^\infty \mathscr{E}(\phi' \circ P_t f, P_t f)\, \mathrm{d}t\,. \end{aligned}$$

We now specialize our calculations to the entropy function $\phi(x) = x \log x$ and use reversibility of the semigroup.

$$\begin{aligned} \mathscr{E}(\phi' \circ P_t f, P_t f) &= \int (\log P_t f)\, (-\mathscr{L}) P_t f\, \mathrm{d}\pi = \int (\log P_t f)\, P_t(-\mathscr{L} f)\, \mathrm{d}\pi \\ &= \int P_t \log P_t f\, (-\mathscr{L}) f\, \mathrm{d}\pi = \int \Gamma(P_t \log P_t f, f)\, \mathrm{d}\pi \\ &\leqslant \sqrt{\int \frac{\Gamma(f, f)}{f}\, \mathrm{d}\pi \int f\, \Gamma(P_t \log P_t f, P_t \log P_t f)\, \mathrm{d}\pi} \end{aligned}$$

where the last line uses the Cauchy–Schwarz inequality (Exercise 1.5). By the chain rule for the carré du champ, we have

$$\Gamma(\log f, f) = \frac{\Gamma(f,f)}{f}\,.$$

By applying the gradient bound (Theorem 2.2.6) and the chain rule,

$$\begin{aligned}\mathscr{E}(\log P_t f, P_t f) &\leqslant \exp(-\alpha t)\sqrt{\int \Gamma(\log f, f)\,\mathrm{d}\pi \int f\, P_t \Gamma(\log P_t f, \log P_t f)\,\mathrm{d}\pi}\\ &= \exp(-\alpha t)\sqrt{\int \Gamma(\log f, f)\,\mathrm{d}\pi \int P_t f\, \Gamma(\log P_t f, \log P_t f)\,\mathrm{d}\pi}\\ &= \exp(-\alpha t)\sqrt{\int \Gamma(\log f, f)\,\mathrm{d}\pi \int \Gamma(\log P_t f, P_t f)\,\mathrm{d}\pi}\end{aligned}$$

which is rearranged to yield

$$\mathscr{E}(\log P_t f, P_t f) \leqslant \exp(-2\alpha t)\,\mathscr{E}(\log f, f)\,. \tag{2.2.16}$$

This shows that under $\mathsf{CD}(\alpha,\infty)$, the Fisher information (introduced in Example 1.2.27) decays exponentially fast. Another proof of this fact is given as Exercise 3.5.

Substituting this into the representation above,

$$\operatorname{ent}_\pi f = \int_0^\infty \mathscr{E}(\log P_t f, P_t f)\,\mathrm{d}t \leqslant \mathscr{E}(\log f, f)\int_0^\infty \exp(-2\alpha t)\,\mathrm{d}t \leqslant \frac{1}{2\alpha}\,\mathscr{E}(\log f, f)\,,$$

which is the log-Sobolev inequality. □

2.2.4 Local Inequalities

The curvature-dimension condition is a *pointwise* inequality that does not explicitly involve the stationary distribution π. Interestingly, this means that the Poincaré and log-Sobolev inequalities can be strengthened to local inequalities, which in turn turn out to be equivalent to the curvature-dimension condition. To illustrate this, we state local versions of the Poincaré inequality, along with an interesting reverse inequality.

Theorem 2.2.17 (Local Poincaré inequality). *Let $(P_t)_{t\geqslant 0}$ be a diffusion semigroup. The following are equivalent.*

1 *The curvature-dimension condition* $\mathsf{CD}(\alpha,\infty)$ *holds.*
2 *For all f and $t \geqslant 0$,*

$$P_t(f^2) - (P_t f)^2 \leqslant \frac{1-\exp(-2\alpha t)}{\alpha}\, P_t \Gamma(f,f)\,. \tag{2.2.18}$$

3 *For all f and $t \geqslant 0$,*

$$P_t(f^2) - (P_t f)^2 \geqslant \frac{\exp(2\alpha t)-1}{\alpha}\, \Gamma(P_t f, P_t f)\,. \tag{2.2.19}$$

Consider the local Poincaré inequality (2.2.18). When evaluated at a point x, and keeping in mind that $P_t f(x) = \mathbb{E}_{\delta_x P_t} f$ where $\delta_x P_t$ is the law of the Markov process started at δ_x, it can be interpreted as follows: *for all x and all $t \geqslant 0$, the measure $\delta_x P_t$ satisfies a Poincaré inequality with constant* $(1 - \exp(-2\alpha t))/\alpha$. If $\alpha > 0$, then as $t \to \infty$ it recovers the Poincaré inequality for the stationary measure with constant $1/\alpha$, but the local inequality is considerably stronger: for example, it is meaningful even for $\alpha \leqslant 0$. When $\alpha = 0$, the constant should be interpreted in terms of its limiting value, i.e., $\lim_{\alpha \to 0}(1 - \exp(-2\alpha t))/\alpha = 2t$. Similarly, the inequality (2.2.19) can be understood as a local *reverse* Poincaré inequality.

To give the flavor for the arguments, let ϕ be a convex function. The proofs begin with the following differentiation:

$$\partial_s P_s \phi(P_{t-s} f) = P_s\big(\mathscr{L}\phi(P_{t-s} f) - \phi'(P_{t-s} f)\, \mathscr{L} P_{t-s} f\big) = P_s\big(\phi''(P_{t-s} f)\, \Gamma(P_{t-s} f, P_{t-s} f)\big)\,,$$

where we applied the diffusion chain rule (2.2.15). It yields

$$P_t \phi(f) - \phi(P_t f) = \int_0^t P_s\big(\phi''(P_{t-s} f)\, \Gamma(P_{t-s} f, P_{t-s} f)\big)\, \mathrm{d}s\,.$$

By Jensen's inequality, the left-hand side is always non-negative, but the right-hand side quantifies this effect. Indeed, Markov semigroup calculations are designed to extract useful information from the lack of commutativity of the semigroup P_t and the application of the non-linear function ϕ. If we apply this to $\phi : x \mapsto x^2$, we obtain

$$P_t(f^2) - (P_t f)^2 = 2 \int_0^t P_s \Gamma(P_{t-s} f, P_{t-s} f)\, \mathrm{d}s\,. \tag{2.2.20}$$

Under $\mathsf{CD}(\alpha, \infty)$, we have $P_s \Gamma(P_{t-s} f, P_{t-s} f) \leqslant \exp(-2\alpha\,(t-s))\, P_t \Gamma(f, f)$ from Theorem 2.2.6, and substituting this into (2.2.20) yields the local Poincaré inequality (2.2.18). The other implications in Theorem 2.2.17 are left as Exercise 2.7.

A salient feature of these arguments is that by their pointwise nature—i.e., they are not integrated w.r.t. π—we never invoke integration by parts. They are also perfectly valid even when the stationary measure is infinite, e.g., for standard Brownian motion.

Finally, we also record the local log-Sobolev inequality (a proof of the implication 1 ⇒ 3 for the Langevin diffusion is given in Exercise 4.4).

Theorem 2.2.21 (Local log-Sobolev inequality)**.** *Let $(P_t)_{t \geqslant 0}$ be a diffusion semigroup. The following are equivalent.*

1. *The curvature-dimension condition* $\mathsf{CD}(\alpha, \infty)$ *holds.*
2. *For all f and $t \geqslant 0$,* $\sqrt{\Gamma(P_t f, P_t f)} \leqslant \exp(-\alpha t)\, P_t \sqrt{\Gamma(f, f)}$.
3. *For all $f \geqslant 0$ and $t \geqslant 0$,*

$$P_t(f \log f) - P_t f \log P_t f \leqslant \frac{1 - \exp(-2\alpha t)}{2\alpha}\, P_t \frac{\Gamma(f, f)}{f}\,.$$

4. *For all $f \geqslant 0$ and $t \geqslant 0$,*

$$P_t(f \log f) - P_t f \log P_t f \geqslant \frac{\exp(2\alpha t) - 1}{2\alpha}\, \frac{\Gamma(P_t f, P_t f)}{P_t f}\,. \tag{2.2.22}$$

2.2.5 Convergence in Rényi Divergence

One curiosity is that the log-Sobolev inequality directly implies a Poincaré inequality (Exercise 1.7), 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 (Mironov, 2017).

Definition 2.2.23. For $q > 1$, the **Rényi divergence** of order q between μ and π is defined by

$$\mathcal{R}_q(\mu \,\|\, \pi) := \frac{1}{q-1} \log \int \Bigl(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Bigr)^q \,\mathrm{d}\pi \tag{2.2.24}$$

if $\mu \ll \pi$, and $\mathcal{R}_q(\mu \,\|\, \pi) := +\infty$ otherwise.

Rényi divergences are monotonic in the order: if $1 < q \leqslant q' < \infty$, then $\mathcal{R}_q \leqslant \mathcal{R}_{q'}$ (this follows from Jensen's inequality, see Exercise 2.9). Some notable special cases include:

- As $q \searrow 1$, $\mathcal{R}_q \searrow \mathsf{KL}$; hence, we set $\mathcal{R}_1 = \mathsf{KL}$.
- For $q = 2$, we have the identity $\mathcal{R}_2 = \log(1 + \chi^2)$.
- As $q \nearrow \infty$, $\mathcal{R}_q \nearrow \mathcal{R}_\infty$, where $\mathcal{R}_\infty(\mu \,\|\, \pi) := \log \|\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\|_{L^\infty(\pi)}$.

Remarkably, Vempala and Wibisono (2019) show that the Poincaré and log-Sobolev inequalities imply convergence of the Langevin diffusion in every Rényi divergence.

Theorem 2.2.25 (Rényi convergence, Vempala and Wibisono (2019))**.** *Let $(P_t)_{t\geqslant 0}$ be a reversible diffusion Markov semigroup, and let $(\pi_t)_{t\geqslant 0}$ 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 $q \geqslant 1$,

$$\mathcal{R}_q(\pi_t \,\|\, \pi) \leqslant \exp\Bigl(-\frac{2t}{qC_{\mathsf{LSI}}}\Bigr)\, \mathcal{R}_q(\pi_0 \,\|\, \pi)\,.$$

2 Suppose that a Poincaré inequality holds with constant C_{PI}. Then, for all $q \geqslant 2$,

$$\mathcal{R}_q(\pi_t \,\|\, \pi) \leqslant \begin{cases} \mathcal{R}_q(\pi_0 \,\|\, \pi) - \dfrac{2t}{qC_{\mathsf{PI}}}\,, & \text{if } \mathcal{R}_q(\pi_t \,\|\, \pi) \geqslant 1\,, \\ \exp\Bigl(-\dfrac{2t}{qC_{\mathsf{PI}}}\Bigr)\, \mathcal{R}_q(\pi_0 \,\|\, \pi)\,, & \text{if } \mathcal{R}_q(\pi_0 \,\|\, \pi) \leqslant 1\,. \end{cases}$$

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,

$$\begin{aligned} \partial_t \mathcal{R}_q(\pi_t \,\|\, \pi) &= \frac{1}{q-1}\, \frac{\partial_t \int \rho_t^q \,\mathrm{d}\pi}{\int \rho_t^q \,\mathrm{d}\pi} = \frac{q}{q-1}\, \frac{\int \rho_t^{q-1}\, \mathscr{L}\rho_t \,\mathrm{d}\pi}{\int \rho_t^q \,\mathrm{d}\pi} = -\frac{q}{q-1}\, \frac{\int \Gamma(\rho_t^{q-1}, \rho_t)\,\mathrm{d}\pi}{\int \rho_t^q \,\mathrm{d}\pi} \\ &= -\frac{4}{q}\, \frac{\mathscr{E}(\rho_t^{q/2}, \rho_t^{q/2})}{\int \rho_t^q \,\mathrm{d}\pi}\,. \end{aligned}$$

Log-Sobolev case. The log-Sobolev inequality reads (due to the chain rule) as

$$2C_{\mathsf{LSI}}\,\mathscr{E}(f,f) \geqslant \operatorname{ent}_\pi(f^2)\,.$$

Applying this to $f = \rho^{q/2}$, we obtain

$$\begin{aligned}
2C_{\mathsf{LSI}}\,\mathscr{E}(\rho^{q/2},\rho^{q/2}) &\geqslant q\int \rho^q \log\rho\,\mathrm{d}\pi - \Bigl(\int \rho^q\,\mathrm{d}\pi\Bigr)\log\Bigl(\int \rho^q\,\mathrm{d}\pi\Bigr) \\
&= q\,\partial_q \int \rho^q\,\mathrm{d}\pi - \Bigl(\int \rho^q\,\mathrm{d}\pi\Bigr)\log\Bigl(\int \rho^q\,\mathrm{d}\pi\Bigr)
\end{aligned}$$

and hence

$$\begin{aligned}
\frac{4}{q}\,\frac{\mathscr{E}(\rho^{q/2},\rho^{q/2})}{\int \rho^q\,\mathrm{d}\pi} &\geqslant \frac{2}{C_{\mathsf{LSI}}}\,\partial_q \log\int \rho^q\,\mathrm{d}\pi - \frac{2}{qC_{\mathsf{LSI}}}\log\int \rho^q\,\mathrm{d}\pi \\
&= \frac{2}{C_{\mathsf{LSI}}}\,\partial_q[(q-1)\,\mathcal{R}_q(\rho\pi \,\|\, \pi)] - \frac{2\,(q-1)}{qC_{\mathsf{LSI}}}\,\mathcal{R}_q(\rho\pi \,\|\, \pi) \\
&= \frac{2}{C_{\mathsf{LSI}}}\,\mathcal{R}_q(\rho\pi \,\|\, \pi) + \frac{2\,(q-1)}{C_{\mathsf{LSI}}}\,\underbrace{\partial_q\mathcal{R}_q(\rho\pi \,\|\, \pi)}_{\geqslant 0} - \frac{2\,(q-1)}{qC_{\mathsf{LSI}}}\,\mathcal{R}_q(\rho\pi \,\|\, \pi) \\
&\geqslant \frac{2}{qC_{\mathsf{LSI}}}\,\mathcal{R}_q(\rho\pi \,\|\, \pi)
\end{aligned}$$

where we used the fact that the Rényi divergence is monotonic in the order.

Poincaré case. Next, applying a Poincaré inequality to $f = \rho^{q/2}$,

$$\begin{aligned}
C_{\mathsf{PI}}\,\mathscr{E}(\rho^{q/2},\rho^{q/2}) \geqslant \operatorname{var}_\pi(\rho^{q/2}) &= \int \rho^q\,\mathrm{d}\pi - \Bigl(\int \rho^{q/2}\,\mathrm{d}\pi\Bigr)^2 \\
&= \Bigl(\int \rho^q\,\mathrm{d}\pi\Bigr)\Bigl[1 - \frac{\exp((q-2)\,\mathcal{R}_{q/2}(\rho\pi \,\|\, \pi))}{\exp((q-1)\,\mathcal{R}_q(\rho\pi \,\|\, \pi))}\Bigr] \\
&\geqslant \Bigl(\int \rho^q\,\mathrm{d}\pi\Bigr)\bigl\{1 - \exp\bigl(-\mathcal{R}_q(\rho\pi \,\|\, \pi)\bigr)\bigr\}
\end{aligned}$$

where we used the monotonicity of the Rényi divergence in the order. Hence,

$$\begin{aligned}
\frac{4}{q}\,\frac{\mathscr{E}(\rho^{q/2},\rho^{q/2})}{\int \rho^q\,\mathrm{d}\pi} &\geqslant \frac{4}{qC_{\mathsf{PI}}}\,\bigl\{1 - \exp\bigl(-\mathcal{R}_q(\rho\pi \,\|\, \pi)\bigr)\bigr\} \\
&\geqslant \frac{2}{qC_{\mathsf{PI}}}\begin{cases} 1\,, & \text{if } \mathcal{R}_q(\rho\pi \,\|\, \pi) \geqslant 1\,, \\ \mathcal{R}_q(\rho\pi \,\|\, \pi)\,, & \text{if } \mathcal{R}_q(\rho\pi \,\|\, \pi) \leqslant 1\,. \end{cases}
\end{aligned}$$

□

To interpret this theorem, the Poincaré result states that after an initial waiting period of time $O(qC_{\mathsf{PI}}\,\mathcal{R}_q(\pi_0 \,\|\, \pi))$, 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 $\log(1 + \chi^2)$, which is substantially stronger.

Another approach to Rényi convergence under $\mathsf{CD}(\alpha,\infty)$ is based on Harnack inequalities, see Exercise 2.17.

2.2.6 Beyond Strong Log-Concavity

Most of the results thus far have assumed $\mathsf{CD}(\alpha, \infty)$ for $\alpha > 0$, which in the case of the Langevin diffusion is equivalent to strong log-concavity of π. What can we assert about the validity of functional inequalities more generally?

First, consider the case $\alpha = 0$: π is log-concave but not uniformly so. For example, π could be the Laplace distribution with density $\pi(x) = \exp(-\frac{1}{2}|x|)$ on $\mathbb{R}$. In this case, since π only has subexponential tails, we cannot hope for a log-Sobolev inequality due to Theorem 2.4.3 below.[4] The Brascamp–Lieb inequality (Theorem 2.2.9) holds as a useful substitute for the Poincaré inequality, but a deep conjecture of Kannan, Lovász, and Simonovits (Kannan et al., 1995) goes much further: they 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 $C_{\mathsf{PI}} \geqslant 1$, 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 2.4.22), see Guédon and Milman (2011); Eldan (2013); Lee and Vempala (2017); Chen (2021); Klartag and Lehec (2022), culminating in the current state-of-the-art result described next.

Theorem 2.2.26 (Klartag (2023))**.** *If π is an isotropic log-concave measure over $\mathbb{R}^d$, then it satisfies a Poincaré inequality with $C_{\mathsf{PI}} \lesssim \log d$.*

In the non-convex case, the Poincaré and log-Sobolev inequalities can still hold, but with constants that scale exponentially with the amount of non-convexity. The most favorable case is when the potential V has a benign landscape. In this setting, there is an explicit expression for the low-temperature limit of the Poincaré and log-Sobolev constants, which shows that they scale as β^{-1} (where β is the inverse temperature).

Theorem 2.2.27 (Chewi and Stromme (2024))**.** *Assume that V is C^2, admits a unique minimizer $x_\star$, and that there exists $C > 0$ such that $\Delta V \leqslant C\,(1 + \|\nabla V\|^2)$. Furthermore, assume that V satisfies a PL inequality with constant $C_{\mathsf{PL}} < \infty$, that is,*

$$V - \inf V \leqslant \frac{C_{\mathsf{PL}}}{2}\,\|\nabla V\|^2\,.$$

For $\beta > 0$, let π_β be the probability measure with density $\pi_\beta \propto \exp(-\beta V)$. Then,

$$\lim_{\beta\to\infty} \beta C_{\mathsf{PI}}(\pi_\beta) = \frac{1}{\lambda_{\min}(\nabla^2 V(x_\star))}\,, \qquad \lim_{\beta\to\infty} \beta C_{\mathsf{LSI}}(\pi_\beta) = C_{\mathsf{PL}}\,.$$

Note that the second formula connects the PL constant of V to the asymptotic limit of the LSI constant of π_β, and that the LSI is the analogue of the PL condition for the KL divergence over the Wasserstein space (see Section 1.4.2). This is somewhat analogous to Theorem 1.4.4, which connects the strong convexity of V to strong convexity of the KL divergence over the Wasserstein space, but the connection between the PL conditions is only true in the low-temperature limit. Interestingly, the asymptotic limit of the Poincaré constant only depends on the local behavior of V via the curvature at the minimizer, although Theorem 2.2.27 does impose global assumptions (that V admits a unique minimizer and satisfies the PL condition).

---

[4] However, if we further assume that π has support with diameter R, then a log-Sobolev inequality holds for π with constant $O(R^2)$; see Exercise 3.9.

When V violates the unique minimizer assumption, the functional inequality constants generally scale exponentially. A special case of the **Eyring–Kramers formula** (Eyring, 1935; Kramers, 1940) asserts that if V has two separated minima x_0, x_1, then the Poincaré constant for π can be expected to scale as $\exp(|V(x_1) - V(x_0)|)$, that is, exponential in the size of the “energy barrier”. See Menz and Schlichting (2014) for a precise statement.

The next section is devoted to techniques for establishing functional inequalities in various settings.

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 $0 < c \leqslant \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \leqslant C < \infty$. 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, Holley and Stroock (1987))**.** *Suppose that π satisfies either a Poincaré or log-Sobolev inequality. Then, if μ satisfies $0 < c \leqslant \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \leqslant C < \infty$, then μ also satisfies the corresponding functional inequality with constant*

$$C_{\mathsf{PI}}(\mu) \leqslant \frac{C}{c}\, C_{\mathsf{PI}}(\pi) \qquad \text{or} \qquad C_{\mathsf{LSI}}(\mu) \leqslant \frac{C}{c}\, C_{\mathsf{LSI}}(\pi)$$

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,

$$\begin{aligned}\operatorname{var}_\mu f = \inf_{m\in\mathbb{R}} \mathbb{E}_\mu[|f - m|^2] &\leqslant C \inf_{m\in\mathbb{R}} \mathbb{E}_\pi[|f - m|^2] = C \operatorname{var}_\pi f \\ &\leqslant C\, C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\pi\, \Gamma(f,f) \leqslant \frac{C}{c}\, C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\mu\, \Gamma(f,f)\,.\end{aligned}$$

The proof for the log-Sobolev inequality is similar once we have the variational principle

$$\operatorname{ent}_\nu f = \inf_{t>0} \mathbb{E}_\nu \underbrace{\Bigl[f \log \frac{f}{t} - f + t\Bigr]}_{\geqslant 0}$$

for any $f : \mathbb{R}^d \to \mathbb{R}_+$, which we leave as Exercise 2.12. □

One can also state a perturbation principle for the T_2 inequality, but it is more involved and the constants are less precise, see Bakry et al. (2014, Proposition 9.6.3).

Proposition 2.3.2 (Bounded perturbations for T_2). *Suppose that* π *satisfies a* T_2 *inequality. If* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *satisfies* $0 < C^{-1} \leqslant \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \leqslant C < \infty$, *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 Lipschitz 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 (Lipschitz mapping). *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*

$$C_{\mathsf{PI}}(\mu) \leqslant L^2\, C_{\mathsf{PI}}(\pi) \qquad \text{or} \qquad C_{\mathsf{LSI}}(\mu) \leqslant L^2\, C_{\mathsf{LSI}}(\pi)$$

respectively.

Proof Assume for simplicity that T is continuously differentiable, so that $\|\nabla T\|_{\mathrm{op}} \leqslant L$. Then, for $f : \mathbb{R}^d \to \mathbb{R}$, by applying the Poincaré inequality for π,

$$\begin{aligned}\operatorname{var}_\mu f = \operatorname{var}_\pi(f \circ T) &\leqslant C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\pi[\|\nabla(f \circ T)\|^2] \leqslant C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\pi[\|\nabla T\|_{\mathrm{op}}^2\, \|\nabla f \circ T\|^2] \\ &\leqslant C_{\mathsf{PI}}(\pi)\, L^2\, \mathbb{E}_\pi[\|\nabla f \circ T\|^2] = C_{\mathsf{PI}}(\pi)\, L^2\, \mathbb{E}_\mu[\|\nabla f\|^2]\,.\end{aligned}$$

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 Lipschitz Perturbations

Whereas Proposition 2.3.1 implies that the LSI for strongly log-concave measures is robust against bounded perturbations, the following result shows that it is also robust against Lipschitz perturbations.

Proposition 2.3.4 (Lipschitz perturbations). *Let* $\pi \propto \exp(-V - W)$, *where* V *is* α*-strongly convex* $(\alpha > 0)$ *and* W *is* L*-Lipschitz. Then,*

$$C_{\mathsf{LSI}}(\pi) \leqslant \frac{1}{\alpha} \exp\Bigl(\frac{4L}{\sqrt{\alpha}} + \frac{L^2}{\alpha}\Bigr)\,.$$

This result is established via Proposition 2.3.3 and the construction of a suitable Lipschitz transport map from the Gaussian; see Exercise 3.9.

2.3.4 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 2.3.5 (Variational principle for entropy)**.** *For* $f : \mathbb{R}^d \to \mathbb{R}_{>0}$,

$$\operatorname{ent}_\pi f = \sup\{\mathbb{E}_\pi[fg] \mid g : \mathbb{R}^d \to \mathbb{R} \text{ such that } \mathbb{E}_\pi \exp g \leqslant 1\}\,.$$

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,

$$\operatorname{ent}_\pi f = \mathbb{E}_\pi\bigl[f \log \frac{f}{\mathbb{E}_\pi f}\bigr] = \underbrace{\mathbb{E}_\mu\Bigl[f \exp(-g) \log \frac{f \exp(-g)}{\mathbb{E}_\mu[f \exp(-g)]}\Bigr]}_{=\operatorname{ent}_\mu(f \exp(-g)) \geqslant 0} + \mathbb{E}_\pi[fg]\,.$$

Equality holds if $g = \log(f/\mathbb{E}_\pi f)$. □

***Remark* 2.3.6.** The variational principle above is essentially a reformulation of the Donsker–Varadhan variational principle (Theorem 1.5.7).

Lemma 2.3.7 (Subadditivity of variance and entropy)**.** *If* $X_1, \dots, X_n$ *are independent random variables, then*

$$\operatorname{var} f(X_1, \dots, X_n) \leqslant \mathbb{E} \sum_{i=1}^n \operatorname{var}\bigl(f(X_1, \dots, X_n) \bigm| X_{-i}\bigr)\,,$$

$$\operatorname{ent} f(X_1, \dots, X_n) \leqslant \mathbb{E} \sum_{i=1}^n \operatorname{ent}\bigl(f(X_1, \dots, X_n) \bigm| X_{-i}\bigr)\,.$$

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.,*

$$\operatorname{var}\bigl(f(X_1, \dots, X_n) \bigm| X_{-i} = x_{-i}\bigr) = \operatorname{var} f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)\,,$$
$$\operatorname{ent}\bigl(f(X_1, \dots, X_n) \bigm| X_{-i} = x_{-i}\bigr) = \operatorname{ent} f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)\,.$$

Proof The subadditivity of the variance was established in Exercise 1.1, so we turn towards the entropy. Let $Z := f(X_1, \dots, X_n)$ and

$$\Delta_i = \log \mathbb{E}[Z \mid X_1, \dots, X_i] - \log \mathbb{E}[Z \mid X_1, \dots, X_{i-1}]\,,$$

so that

$$\operatorname{ent} Z = \mathbb{E}[Z\,(\log Z - \log \mathbb{E}\, Z)] = \mathbb{E}\Bigl[Z \sum_{i=1}^n \Delta_i\Bigr]\,.$$

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

$$\mathbb{E}[Z\Delta_i] = \mathbb{E}\,\mathbb{E}[Z\Delta_i \mid X_{-i}] \leqslant \mathbb{E}\,\mathrm{ent}(Z \mid X_{-i})\,. \qquad \square$$

Proposition 2.3.8 (Functional inequalities and convolutions)**.** *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*

$$C_{\mathsf{PI}}(\pi) \leqslant C_{\mathsf{PI}}(\pi_1) + C_{\mathsf{PI}}(\pi_2) \qquad \text{or} \qquad C_{\mathsf{LSI}}(\pi) \leqslant C_{\mathsf{LSI}}(\pi_1) + C_{\mathsf{LSI}}(\pi_2)$$

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 (Lemma 2.3.7),

$$\begin{aligned} \mathrm{var}_\pi f = \mathrm{var}\, f(X+Y) &\leqslant \mathbb{E}\,\mathrm{var}\big(f(X+Y) \bigm| Y\big) + \mathbb{E}\,\mathrm{var}\big(f(X+Y) \bigm| X\big) \\ &\leqslant \{C_{\mathsf{PI}}(\pi_1) + C_{\mathsf{PI}}(\pi_2)\}\, \mathbb{E}[\|\nabla f(X+Y)\|^2]\,, \end{aligned}$$

and a similar argument holds for the entropy. $\square$

2.3.5 Mixtures

Suppose that π is a mixture, $\pi = \mu P := \int P_x\, \mu(\mathrm{d}x)$, 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 2.3.9 (PI for mixtures, Bardet et al. (2018))**.** *Let* μP *be a mixture and assume that each* P_x *satisfies a Poincaré inequality with constant* $C_{\mathsf{PI}}(P)$*. Also, assume that*

$$C_{\chi^2} := \sup_{x, x' \in \mathrm{supp}(\mu)} \chi^2(P_x \,\|\, P_{x'}) < \infty\,. \tag{2.3.10}$$

Then, μP *satisfies a Poincaré inequality with constant*

$$C_{\mathsf{PI}}(\mu P) \leqslant \Big(1 + \frac{C_{\chi^2}}{2}\Big)\, C_{\mathsf{PI}}(P)\,.$$

Proof Let $f : \mathbb{R}^d \to \mathbb{R}$, and let $X, X' \overset{\text{i.i.d.}}{\sim} \mu$. By the law of total 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,

$$\begin{aligned}\operatorname{var} \mathbb{E}_{P_X} f &= \frac{1}{2}\,\mathbb{E}[|\mathbb{E}_{P_X} f - \mathbb{E}_{P_{X'}} f|^2] = \frac{1}{2}\,\mathbb{E}\Bigl[\Bigl|\int f\,\bigl(\frac{\mathrm{d}P_X}{\mathrm{d}P_{X'}} - 1\bigr)\,\mathrm{d}P_{X'}\Bigr|^2\Bigr] \\ &\leqslant \frac{1}{2}\,\mathbb{E}[(\operatorname{var}_{P_{X'}} f)\,\chi^2(P_X \,\|\, P_{X'})] \leqslant \frac{C_{\chi^2}}{2}\,\mathbb{E}\operatorname{var}_{P_X} f\,.\end{aligned}$$

Hence,

$$\operatorname{var}_{\mu P} f \leqslant \bigl(1 + \frac{C_{\chi^2}}{2}\bigr)\,\mathbb{E}\operatorname{var}_{P_X} f \leqslant \bigl(1 + \frac{C_{\chi^2}}{2}\bigr)\,C_{\mathsf{PI}}(P)\,\underbrace{\mathbb{E}\,\mathbb{E}_{P_X}\Gamma(f,f)}_{=\mathbb{E}_{\mu P}\Gamma(f,f)}\,. \qquad \square$$

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 $f : \mathbb{R}^d \to \mathbb{R}$,

$$\operatorname{ent}_\pi(f^2) \leqslant 2C\,\mathbb{E}_\pi\,\Gamma(f,f) + D\,\mathbb{E}_\pi[f^2]\,. \tag{2.3.11}$$

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 result shows how this is achieved.

Lemma 2.3.12 (Tightening a defective LSI, Wang (2024))**.** *Suppose that π satisfies the defective log-Sobolev inequality* (2.3.11)*, together with a Poincaré inequality. Then, π satisfies an log-Sobolev inequality with constant*

$$C_{\mathsf{LSI}} \leqslant C + \frac{C_{\mathsf{PI}} D}{2}\,.$$

We also need one change of measure lemma.

Lemma 2.3.13 (Chen et al. (2021a))**.** *Suppose that μ, ν are probability measures and f is a positive function. Then,*

$$\mathbb{E}_\mu(f)\log\frac{\mathbb{E}_\mu(f)}{\mathbb{E}_\nu(f)} \leqslant \operatorname{ent}_\mu(f) + \mathbb{E}_\mu(f)\log\bigl(1 + \chi^2(\mu \,\|\, \nu)\bigr)\,.$$

Proof By rescaling, we may assume $\mathbb{E}_\mu f = 1$. Recall the Donsker–Varadhan variational principle (Theorem 1.5.7), which states

$$\mathsf{KL}(\eta \,\|\, \eta') = \sup\bigl\{\mathbb{E}_\eta\, g - \log \mathbb{E}_{\eta'} \exp g \bigm| g : \mathcal{X} \to \mathbb{R} \text{ is bounded and measurable}\bigr\}\,.$$

If we take $\eta = f\mu$, $\eta' = \nu$, and $g = \log(f/\mathbb{E}_\nu f)$, then

$$\mathbb{E}_\mu\bigl[f\log\frac{f}{\mathbb{E}_\nu f}\bigr] = \mathbb{E}_\eta \log\frac{f}{\mathbb{E}_\nu f} \leqslant \mathsf{KL}(\eta \,\|\, \nu) + \underbrace{\log \mathbb{E}_\nu \frac{f}{\mathbb{E}_\nu f}}_{=0} = \mathbb{E}_\mu\bigl[f\log\bigl(f\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\bigr)\bigr]\,.$$

By subtracting $\mathbb{E}_\mu(f\log f)$ from both sides, we obtain

$$\log\frac{1}{\mathbb{E}_\nu f} = \mathbb{E}_\mu\bigl[f\log\frac{1}{\mathbb{E}_\nu f}\bigr] \leqslant \mathbb{E}_\mu\bigl[f\log\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\bigr]\,.$$

Next, applying the Donsker–Varadhan variational principle a second time with $\eta = f\mu$, $\eta' = \mu$, and $g = \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ yields

$$\mathbb{E}_\mu\Bigl[f \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Bigr] = \mathbb{E}_\eta \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \leqslant \mathsf{KL}(\eta \,\|\, \mu) + \log \mathbb{E}_\mu \frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \operatorname{ent}_\mu f + \log\bigl(1 + \chi^2(\mu \,\|\, \nu)\bigr)\,,$$

which is what we wanted to show. □

We can now prove the log-Sobolev inequality for mixtures.

Proposition 2.3.14 (LSI for mixtures, Chen et al. (2021a))**.** *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*

$$C_{\chi^2} := \sup_{x,x' \in \operatorname{supp}(\mu)} \chi^2(P_x \,\|\, P_{x'}) < \infty\,.$$

Then, μP satisfies a log-Sobolev inequality with constant

$$C_{\mathsf{LSI}}(\mu P) \leqslant C_{\mathsf{LSI}}(P)\,\Bigl\{2 + \frac{1}{2}\,\Bigl(1 + \frac{C_{\chi^2}}{2}\Bigr) \log(1 + C_{\chi^2})\Bigr\}\,.$$

Proof We begin, as in the proof of Proposition 2.3.9, with a decomposition of the entropy:

$$\operatorname{ent}_{\mu P}(f^2) = \mathbb{E} \operatorname{ent}_{P_X}(f^2) + \operatorname{ent} \mathbb{E}_{P_X}(f^2)\,.$$

As before, it is the second term that is difficult to control.

Applying Lemma 2.3.13,

$$\begin{aligned} \operatorname{ent} \mathbb{E}_{P_X}(f^2) = \mathbb{E}\Bigl[\mathbb{E}_{P_X}(f^2) \log \frac{\mathbb{E}_{P_X}(f^2)}{\mathbb{E}_{\mu P}(f^2)}\Bigr] &\leqslant \mathbb{E}\bigl[\operatorname{ent}_{P_X}(f^2) + \mathbb{E}_{P_X}(f^2) \log\bigl(1 + \chi^2(P_X \,\|\, \mu P)\bigr)\bigr] \\ &\leqslant \mathbb{E} \operatorname{ent}_{P_X}(f^2) + \log(1 + C_{\chi^2})\, \mathbb{E}_{\mu P}(f^2)\,. \end{aligned}$$

Hence,

$$\begin{aligned} \operatorname{ent}_{\mu P}(f^2) &\leqslant 2\, \mathbb{E} \operatorname{ent}_{P_X}(f^2) + \log(1 + C_{\chi^2})\, \mathbb{E}_{\mu P}(f^2) \\ &\leqslant 4 C_{\mathsf{LSI}}(P)\, \mathbb{E}_{\mu P} \Gamma(f, f) + \log(1 + C_{\chi^2})\, \mathbb{E}_{\mu P}(f^2)\,, \end{aligned}$$

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 2.3.9 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

$$\chi^2(P_x \,\|\, P_{x'}) = \exp \frac{\|x - x'\|^2}{\sigma^2} - 1 \leqslant \exp \frac{4R^2}{\sigma^2} - 1$$

for $x, x' \in \mathsf{B}(0,R)$, we deduce that μP satisfies a log-Sobolev inequality with constant

$$C_{\mathsf{LSI}}(\mu P) \leqslant 2 \left(\sigma^2 + R^2 \exp \frac{4R^2}{\sigma^2}\right).$$

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. Another approach to the log-Sobolev inequality for Gaussian mixtures is given in Exercise 3.9.

We extend the results of this section in Exercise 2.14.

2.3.6 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 2.3.7). Indeed, if $f : \mathbb{R}^{Nd} \to \mathbb{R}$ and if $X_i \sim \pi_i$ are independent for $i \in [N]$,

$$\begin{aligned} \operatorname{var}_\pi f = \operatorname{var} f(X_1, \dots, X_N) &\leqslant \mathbb{E} \sum_{i=1}^N \operatorname{var}\bigl(f(X_1, \dots, X_N) \bigm| X_{-i}\bigr) \\ &\leqslant \max_{i \in [N]} C_{\mathsf{PI}}(\pi_i)\, \mathbb{E} \sum_{i=1}^N \mathbb{E}[\|\nabla_i f(X_1, \dots, X_N)\|^2 \mid X_{-i}] \\ &= \max_{i \in [N]} C_{\mathsf{PI}}(\pi_i)\, \mathbb{E}_\pi[\|\nabla f\|^2]\,. \end{aligned}$$

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). See the supplementary material for a proof.

Theorem 2.3.17 (Marton's tensorization, Marton (1996)). *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

$$\inf_{\gamma_i \in \mathcal{C}(\pi_i, \nu_i)} \varphi\Big(\int c_i \, \mathrm{d}\gamma_i\Big) \leqslant 2\sigma^2 \,\mathsf{KL}(\nu_i \,\|\, \pi_i)\,, \qquad \forall \nu_i \in \mathcal{P}(\mathcal{X}_i)\,,\ \forall i \in [N]\,.$$

Then, it holds that

$$\inf_{\gamma \in \mathcal{C}(\pi, \nu)} \sum_{i=1}^{N} \varphi\Big(\int c_i(x_i, y_i)\, \gamma(\mathrm{d}x_{1:N}, \mathrm{d}y_{1:N})\Big) \leqslant 2\sigma^2 \,\mathsf{KL}(\nu \,\|\, \pi)\,, \qquad \forall \nu \in \mathcal{P}(\mathcal{X})\,.$$

We can apply Marton's tensorization (Theorem 2.3.17) with the convex function $\varphi(x) := x$ and cost functions $c_i := \mathsf{d}_i^2$, with d_i a lower semicontinuous metric on $\mathcal{X}_i$. This immediately yields the following corollary.

Corollary 2.3.18 (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 $\mathsf{d}(x_{1:N}, y_{1:N})^2 := \sum_{i=1}^N \mathsf{d}(x_i, y_i)^2$ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_N$.*

2.4 Concentration of Measure and Isoperimetry

Arguably, the most prominent application of functional inequalities has been to the concentration of measure phenomenon, which is an indispensable tool in high-dimensional probability and statistics. Since it is not the focus of this book, we only mention a few key results, focusing on the connection with *isoperimetric inequalities* due to their relevance to Chapter 7. See, however, the supplementary material to this book for more details.

Since many of the arguments hold on a general Polish space (that is, a complete separable metric space) $(\mathcal{X}, \mathsf{d})$ 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

$$A^\varepsilon := \{x \in \mathcal{X} \mid \mathsf{d}(x, A) < \varepsilon\}\,.$$

The **concentration function** $\alpha_\pi : \mathbb{R}_+ \to [0, 1]$ is defined via

$$\alpha_\pi(\varepsilon) = \sup\Bigl\{\pi\bigl((A^\varepsilon)^\mathrm{c}\bigr) \Bigm| A \subseteq \mathcal{X} \text{ is a Borel subset with } \pi(A) \geqslant \frac{1}{2}\Bigr\}\,.$$

Typically, we have $\alpha_\pi(\varepsilon) \leqslant C_0 \exp(-\varepsilon/C_1)$ or $\alpha_\pi(\varepsilon) \leqslant C_0 \exp(-\varepsilon^2/C_1)$ 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. We first 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 more useful for 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 $\mathbb{P}\{X \leqslant m\} \wedge \mathbb{P}\{X \geqslant m\} \geqslant \frac{1}{2}$.

Theorem 2.4.2 (Blow-up and Lipschitz functions)**.** *Suppose that* $(\mathcal{X}, \mathsf{d}, \pi)$ *has concentration function* α_π. *Then, for any* 1*-Lipschitz function* $f : \mathcal{X} \to \mathbb{R}$ *and* $\varepsilon \geqslant 0$,

$$\pi\{f \geqslant \operatorname{med} f + \varepsilon\} \leqslant \alpha_\pi(\varepsilon)\,.$$

Conversely, suppose that for all 1*-Lipschitz functions* $f : \mathcal{X} \to \mathbb{R}$, *it holds that*

$$\pi\{f \geqslant \operatorname{med} f + \varepsilon\} \leqslant \beta(\varepsilon)\,.$$

Then, the concentration function α_π *of* $(\mathcal{X}, \mathsf{d}, \pi)$ *satisfies* $\alpha_\pi \leqslant \beta$.

Proof ($\Longrightarrow$) Consider the set $A := \{f \leqslant \operatorname{med} f\}$. 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 $\mathsf{d}(x, y) < \varepsilon$, so $f(x) - \operatorname{med} f = f(y) - \operatorname{med} f + f(x) - f(y) \leqslant \mathsf{d}(x, y) < \varepsilon$. Hence,

$$\pi\{f - \operatorname{med} f < \varepsilon\} \geqslant \pi(A^\varepsilon) \geqslant 1 - \alpha(\varepsilon)\,.$$

($\Longleftarrow$) The function $f := \mathsf{d}(\cdot, A)$ is 1-Lipschitz, and if $\pi(A) \geqslant \frac{1}{2}$ then 0 is a median of f. Thus, it holds that

$$\pi(A^\varepsilon) = \pi\{f - \operatorname{med} f < \varepsilon\} \geqslant 1 - \beta(\varepsilon)\,. \qquad \square$$

More broadly, it is a general principle that many statements about sets have an equivalent reformulation in terms of functions.

Whether by bounding the concentration function, or by establishing concentration inequalities for Lipschitz functions directly, by now there is a wealth of techniques for proving results of the following type. We remark that while the following theorem asserts concentration around the mean, whereas Theorem 2.4.2 considers concentration around the median, the difference is merely superficial since the mean and median are essentially equivalent (see, e.g., Milman, 2009).

Theorem 2.4.3 (Functional inequalities and concentration)**.** *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 $t \geqslant 0$,

$$\pi\{f - \mathbb{E}_\pi f \geqslant t\} \leqslant 3\exp\Bigl(-\frac{t}{\sqrt{C_{\mathsf{PI}}}}\Bigr)\,.$$

2 If π satisfies Talagrand's T_1 inequality with constant C_{T_1}, then for all $t \geqslant 0$,

$$\pi\{f - \mathbb{E}_\pi f \geqslant t\} \leqslant \exp\Bigl(-\frac{t^2}{2C_{\mathsf{T}_1}}\Bigr)\,.$$

In particular, recalling that $C_{\mathsf{T}_1} \leqslant C_{\mathsf{T}_2} \leqslant C_{\mathsf{LSI}}$, the second statement furnishes a concentration inequality under a log-Sobolev inequality. Thus, a Poincaré inequality implies subexponential concentration, and a log-Sobolev inequality implies sub-Gaussian concentration. Moreover, due to tensorization results (Section 2.3.6), the constants involved are dimension-free for many situations of interest. Consequently, Theorem 2.4.3 is immensely useful for high-dimensional applications. Here is a basic illustration.

Example 2.4.4. Suppose that γ is the standard Gaussian measure on $\mathbb{R}^d$. From the Bakry–Émery theorem (Theorem 1.2.30), 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 $\operatorname{var}\|Z\| \leqslant 1$, i.e., $\sqrt{d-1} \leqslant \mathbb{E}\|Z\| \leqslant \sqrt{d}$.

The concentration result above now shows that the standard Gaussian "lives" on a thin spherical shell of radius $\sqrt{d}$ and width $O(1)$.

2.4.2 *Classical Isoperimetry Results*

We now 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.

The connection between concentration and isoperimetry began with the work of Lévy, who found the isoperimetric inequality on the sphere.

Theorem 2.4.5 (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 $\sigma_d(A) \notin \{0, 1\}$. Let C be a spherical cap with the same measure as A. Then, for all $\varepsilon > 0$,*

$$\sigma_d(A^\varepsilon) \geqslant \sigma_d(C^\varepsilon)\,.$$

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 $\mathsf{B}(x_0, r)$ for some $x_0 \in \mathbb{S}^d$ and $r > 0$, where the balls are defined w.r.t. d.

Note that Theorem 2.4.5 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, $C = \mathsf{B}(x_0, \frac{\pi}{2})$, and so

$$\sigma_d(A^\varepsilon) \geqslant \sigma_d(C^\varepsilon) = \sigma_d\Bigl(\mathsf{B}\bigl(x_0, \frac{\pi}{2} + \varepsilon\bigr)\Bigr)\,.$$

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 Exercise 2.18.

Theorem 2.4.6 (Concentration on the sphere)**.** *Let σ_d be the uniform measure on the unit sphere $\mathbb{S}^d$ in dimension $d \geqslant 2$, equipped with the geodesic distance d. Then,*

$$\alpha_{\sigma_d}(\varepsilon) \leqslant \exp\bigl(-\frac{(d-1)\,\varepsilon^2}{2}\bigr)\,, \qquad \textit{for all } \varepsilon > 0\,.$$

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

$$H_{x_0,t} \coloneqq \{x \in \mathbb{R}^d \mid \langle x, x_0\rangle \leqslant t\}\,.$$

Theorem 2.4.7 (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 $\varepsilon > 0$,*

$$\gamma_d(A^\varepsilon) \geqslant \gamma_d(H_{x_0,t}^\varepsilon)\,.$$

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 $H_{e,t}^\varepsilon = H_{e,t+\varepsilon}$, then

$$\gamma_d(A^\varepsilon) \geqslant \Phi\bigl(\Phi^{-1}\bigl(\gamma_d(A)\bigr) + \varepsilon\bigr)\,. \tag{2.4.8}$$

In particular, if $\gamma_d(A) = \frac{1}{2}$, then $\Phi^{-1}(\frac{1}{2}) = 0$, so

$$\alpha_{\gamma_d}(\varepsilon) \leqslant \Phi(-\varepsilon) \leqslant \frac{1}{2}\exp\bigl(-\frac{\varepsilon^2}{2}\bigr)\,.$$

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 2.4.9. Given a non-empty Borel set A and a measure π on a Polish space $(\mathcal{X}, \mathsf{d})$, the **Minkowski content** of A under π is

$$\pi^+(A) := \liminf_{\varepsilon \searrow 0} \frac{\pi(A^\varepsilon) - \pi(A)}{\varepsilon} \,.$$

Definition 2.4.10. For a measure π on a Polish space $(\mathcal{X}, \mathsf{d})$, 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 (2.4.8) yields

$$\gamma_d(A^\varepsilon) \geqslant \gamma_d(A) + \phi\big(\Phi^{-1}\big(\gamma_d(A)\big)\big)\, \varepsilon + o(\varepsilon) \,,$$

where $\phi = \Phi'$ is the Gaussian density. Hence, we can identify the isoperimetric profile of the standard Gaussian as

$$\mathcal{I}_{\gamma_d}(p) = \phi\big(\Phi^{-1}(p)\big) \,. \tag{2.4.11}$$

Actually, (2.4.11) is *equivalent* to the Gaussian isoperimetric inequality (2.4.8). Here, (2.4.8) is called the *integral* form of the inequality, whereas (2.4.11) is called the *differential* form. The following theorem shows how to convert between the two forms.

Theorem 2.4.12 (Bobkov and Houdré (1997))**.** *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 $\pi(A) \notin \{0,1\}$,

$$\pi(A^\varepsilon) \geqslant F\big(F^{-1}\big(\pi(A)\big) + \varepsilon\big) \,.$$

2 For all Borel A with $\pi(A) \notin \{0,1\}$,

$$\pi^+(A) \geqslant \mathcal{I}\big(\pi(A)\big) \,.$$

Proof sketch Let $\bar\omega(p,\varepsilon) := F(F^{-1}(p) + \varepsilon)$. Then, one can show that $\bar\omega$ satisfies the semigroup property $\bar\omega(\bar\omega(p,\varepsilon),\varepsilon') = \bar\omega(p, \varepsilon + \varepsilon')$. Using this, one shows that to prove $\pi(A^\varepsilon) \geqslant \omega(\pi(A))$ it suffices to consider $\varepsilon \searrow 0$. A Taylor expansion yields

$$\begin{aligned} \pi(A^\varepsilon) &\geqslant \pi(A) + \pi^+(A)\, \varepsilon + o(\varepsilon) \,, \\ \bar\omega(\pi(A), \varepsilon) &= \pi(A) + \mathcal{I}(\pi(A))\, \varepsilon + o(\varepsilon) \,, \end{aligned}$$

from which we deduce that $\pi^+(A) \geqslant \mathcal{I}(\pi(A))$ for all A if and only if $\pi(A^\varepsilon) \geqslant \bar\omega(\pi(A), \varepsilon)$ for all A and all $\varepsilon > 0$. ☐

2.4.3 *Cheeger Isoperimetry*

We now consider a class of probability measures which is characterized by a lower bound on the isoperimetric profile.

Definition 2.4.13. A probability measure π satisfies a **Cheeger isoperimetric inequality** with constant $\mathsf{Ch} > 0$ if for all Borel sets $A \subseteq \mathcal{X}$,

$$\pi^+(A) \geqslant \frac{1}{\mathsf{Ch}}\, \pi(A)\, \pi(A^c)\,. \tag{2.4.14}$$

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 (2.4.14) is the differential form of the inequality. By applying Theorem 2.4.12, one shows that the inequality (2.4.14) implies, for any $\varepsilon \in [0, \mathsf{Ch}]$,

$$\pi(A^\varepsilon) - \pi(A) \geqslant \frac{\varepsilon}{2\,\mathsf{Ch}}\, \pi(A)\, \pi(A^c)\,. \tag{2.4.15}$$

For all $\varepsilon > 0$, Theorem 2.4.12 also implies that

$$\alpha_\pi(\varepsilon) \leqslant \exp\bigl(-\frac{\varepsilon}{\mathsf{Ch}}\bigr)\,,$$

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 2.4.16 (Coarea inequality)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be Lipschitz. Then,*

$$\int \|\nabla f\|\, \mathrm{d}\pi \geqslant \int_{-\infty}^{\infty} \pi^+\{f > t\}\, \mathrm{d}t\,.$$

***Remark* 2.4.17.** On a general metric space $(\mathcal{X}, \mathsf{d})$, we define

$$\|\nabla f\|(x) := \limsup_{y \in \mathcal{X},\, \mathsf{d}(x,y) \searrow 0} \frac{|f(x) - f(y)|}{\mathsf{d}(x,y)}\,.$$

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 $f \geqslant 0$. Let $f_\varepsilon(x) := \sup_{\mathsf{d}(x,\cdot)<\varepsilon} f$ and $A_t := \{f > t\}$. We can check that

$A_t^\varepsilon = \{f_\varepsilon > t\}$, and for $g \geqslant 0$ 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 2.4.18 (Functional form of Cheeger isoperimetry)**.** *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*

$$\mathbb{E}_\pi |f - \mathbb{E}_\pi f| \leqslant 2 \, \mathsf{Ch} \, \mathbb{E}_\pi \|\nabla f\| \, . \tag{2.4.19}$$

Proof sketch (2) $\Longrightarrow$ (1): Apply (2.4.19) 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) $\Longrightarrow$ (2): Let $A_t := \{f > t\}$. Applying the coarea inequality and the Cheeger isoperimetric inequality,

$$\begin{aligned}
2 \, \mathsf{Ch} \, \mathbb{E}_\pi \|\nabla f\| &\geqslant 2 \, \mathsf{Ch} \int_{-\infty}^{\infty} \pi^+\{f > t\} \, \mathrm{d}t \\
&\geqslant 2 \int_{-\infty}^{\infty} \pi(A_t) \, \pi(A_t^{\mathsf{c}}) \, \mathrm{d}t = \int_{-\infty}^{\infty} \mathbb{E}_\pi |\mathbb{1}_{A_t} - \pi(A_t)| \, \mathrm{d}t \\
&\geqslant \sup_{\|g\|_{L^\infty(\pi)} \leqslant 1} \int_{-\infty}^{\infty} \Bigl(\int g \, \{\mathbb{1}_{A_t} - \pi(A_t)\} \, \mathrm{d}\pi \Bigr) \, \mathrm{d}t \\
&= \sup_{\|g\|_{L^\infty(\pi)} \leqslant 1} \int_{-\infty}^{\infty} \Bigl(\int \{g - \mathbb{E}_\pi g\} \, \mathbb{1}_{A_t} \, \mathrm{d}\pi \Bigr) \, \mathrm{d}t = \sup_{\|g\|_{L^\infty(\pi)} \leqslant 1} \int \{g - \mathbb{E}_\pi g\} \, f \, \mathrm{d}\pi \\
&= \mathbb{E}_\pi |f - \mathbb{E}_\pi f| \, .
\end{aligned}$$

□

2.4.4 L^p–L^q Poincaré Inequalities

In this section, we work on Euclidean space for simplicity.

The inequality (2.4.19) can be considered an "L^1 variant" of the Poincaré inequality. More generally, we can define the following family of inequalities.

Definition 2.4.20 (L^p–L^q Poincaré inequality)**.** For $p, q \in [1, \infty]$ with $q \geqslant p$, the L^p–L^q **Poincaré inequality** asserts that for all smooth $f : \mathbb{R}^d \to \mathbb{R}$,

$$\|f - \mathbb{E}_\pi f\|_{L^p(\pi)} \leqslant C_{p,q} \, \bigl\| \|\nabla f\| \bigr\|_{L^q(\pi)} \, .$$

In this new notation, the usual Poincaré inequality is an L^2–L^2 Poincaré inequality with $C_{2,2} = \sqrt{C_{\mathsf{PI}}}$, whereas the inequality (2.4.19) is an L^1–L^1 Poincaré inequality.

These inequalities form a hierarchy via Hölder's inequality.

Proposition 2.4.21 (Milman (2009)). *Suppose $p, q, \hat p, \hat q \in [1,\infty]$ are such that $p \leqslant \hat p$ and $q \leqslant \hat q$, 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 $\operatorname{med}_\pi f = 0$, which we can arrange by adding a constant. Define the function $g := (\operatorname{sgn} f)\,|f|^{\hat p/p}$, which still satisfies $\operatorname{med}_\pi g = 0$. Throughout the proof, we use the fact that for any function h, the $L^p(\pi)$ norms of $h - \mathbb{E}_\pi h$ and $h - \operatorname{med}_\pi h$ are equivalent up to a universal constant (Milman, 2009). Applying the L^p–L^q Poincaré inequality to g together with Hölder's inequality,

$$\begin{aligned}\|f - \operatorname{med}_\pi f\|_{L^{\hat p}(\pi)}^{\hat p/p} &= \|g - \operatorname{med}_\pi g\|_{L^p(\pi)} \lesssim \|g - \mathbb{E}_\pi g\|_{L^p(\pi)} \leqslant C_{p,q}\, \big\|\, \|\nabla g\|\, \big\|_{L^q(\pi)} \\ &= \frac{\hat p}{p}\, C_{p,q}\, \big\|\, |f|^{\hat p/p - 1}\, \|\nabla f\|\, \big\|_{L^q(\pi)} \\ &= \frac{\hat p}{p}\, C_{p,q}\, \|f - \operatorname{med}_\pi f\|_{L^{\hat p}(\pi)}^{\hat p/p-1}\, \big\|\, \|\nabla f\|\, \big\|_{L^{\hat q}(\pi)}\,,\end{aligned}$$

where we leave it to the reader to check that the exponents work out correctly. If we rearrange this inequality, then

$$\|f - \mathbb{E}_\pi f\|_{L^{\hat p}(\pi)} \lesssim \|f - \operatorname{med}_\pi f\|_{L^{\hat p}(\pi)} \lesssim \frac{\hat p}{p}\, C_{p,q}\, \big\|\, \|\nabla f\|\, \big\|_{L^{\hat q}(\pi)}\,. \qquad \square$$

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 2.4.18 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 $1 \leqslant p \leqslant q < \infty$, 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 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 2.4.22 (Reversing the hierarchy, Milman (2009)). *Let π 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 1.2.30), π satisfies a Poincaré inequality with $C_{2,2}^2 = C_{\mathsf{PI}} \leqslant 1/\alpha$. By reversing the hierarchy, we see that this implies a Cheeger isoperimetric inequality.

Corollary 2.4.23. *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 2.4.24. *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,

$$\mathcal{I}_\pi(p) \geqslant \mathcal{I}_\pi\big(\frac{1}{2}\big)\, p\,. \tag{2.4.25}$$

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 $(P_t)_{t\geqslant 0}$ with stationary distribution π. We start with a semigroup calculation.

Proposition 2.4.26. *Assume that the Markov semigroup $(P_t)_{t\geqslant 0}$ is reversible and satisfies the curvature-dimension condition* $\mathsf{CD}(0,\infty)$. *For all $t > 0$ and $p \in [2,\infty]$,*

$$\big\| \sqrt{\Gamma(P_t f, P_t f)} \big\|_{L^p(\pi)} \leqslant \frac{1}{\sqrt{2t}}\, \|f\|_{L^p(\pi)} \tag{2.4.27}$$

and

$$\|f - P_t f\|_{L^1(\pi)} \leqslant \sqrt{2t}\, \big\| \sqrt{\Gamma(f,f)} \big\|_{L^1(\pi)}\,. \tag{2.4.28}$$

Proof The local reverse Poincaré inequality (2.2.19) yields

$$\sqrt{\Gamma(P_t f, P_t f)} \leqslant \frac{1}{\sqrt{2t}} \sqrt{P_t(f^2)}$$

so that

$$\big\| \sqrt{\Gamma(P_t f, P_t f)} \big\|_{L^p(\pi)} \leqslant \frac{1}{\sqrt{2t}} \{\mathbb{E}_\pi P_t(|f|^p)\}^{1/p} \leqslant \frac{1}{\sqrt{2t}} \|f\|_{L^p(\pi)}\,.$$

The last inequality is the dual of the $p = \infty$ case. Indeed, for g with $\|g\|_{L^\infty(\pi)} \leqslant 1$,

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

and hence, by the Cauchy–Schwarz inequality for the carré du champ (Exercise 1.5),

$$\begin{aligned}\int (f - P_t f)\, g\,\mathrm{d}\pi &= \int_0^t \Big(\int \Gamma(f, P_s g)\,\mathrm{d}\pi\Big)\,\mathrm{d}s \leqslant \int_0^t \Big(\int \sqrt{\Gamma(f,f)\,\Gamma(P_s g, P_s g)}\,\mathrm{d}\pi\Big)\,\mathrm{d}s \\ &\leqslant \big\| \sqrt{\Gamma(f,f)} \big\|_{L^1(\pi)} \int_0^t \big\| \sqrt{\Gamma(P_s g, P_s g)} \big\|_{L^\infty(\pi)}\,\mathrm{d}s \\ &\leqslant \big\| \sqrt{\Gamma(f,f)} \big\|_{L^1(\pi)} \|g\|_{L^\infty(\pi)} \int_0^t \frac{1}{\sqrt{2s}}\,\mathrm{d}s \leqslant \sqrt{2t}\, \big\| \sqrt{\Gamma(f,f)} \big\|_{L^1(\pi)}\,. \end{aligned}$$ □

We are now ready to prove Milman's theorem.

Proof of Milman's theorem, Theorem 2.4.22 By approximating the indicator function $\mathbb{1}_A$ with a smooth function and applying the inequality (2.4.28), we can justify the bound

$$\sqrt{2t}\,\pi^+(A) \geqslant \|\mathbb{1}_A - P_t\mathbb{1}_A\|_{L^1(\pi)}\,.$$

Next, a calculation shows that

$$\begin{aligned}\mathbb{E}_\pi|\mathbb{1}_A - P_t\mathbb{1}_A| &= 2\left\{\pi(A)\,\pi(A^{\mathrm{c}}) - \mathbb{E}\big[\big(\mathbb{1}_A - \pi(A)\big)\,\big(P_t\mathbb{1}_A - \pi(A)\big)\big]\right\}\\ &\geqslant 2\,\{\pi(A)\,\pi(A^{\mathrm{c}}) - \underbrace{\|\mathbb{1}_A - \pi(A)\|_{L^\infty(\pi)}}_{\leqslant 1}\,\|P_t\mathbb{1}_A - \pi(A)\|_{L^1(\pi)}\}\,.\end{aligned}$$

From the L^1–L^∞ Poincaré inequality and (2.4.27),

$$\|P_t\mathbb{1}_A - \pi(A)\|_{L^1(\pi)} \leqslant C_{1,\infty}\,\big\|\,\|\nabla P_t\mathbb{1}_A\|\,\big\|_{L^\infty(\pi)} \leqslant \frac{C_{1,\infty}}{\sqrt{2t}}\,\|\mathbb{1}_A\|_{L^\infty(\pi)} \leqslant \frac{C_{1,\infty}}{\sqrt{2t}}\,.$$

Hence, we have

$$\sqrt{2t}\,\pi^+(A) \geqslant 2\left\{\pi(A)\,\pi(A^{\mathrm{c}}) - \frac{C_{1,\infty}}{\sqrt{2t}}\right\}.$$

Now choose $\sqrt{2t} = 2C_{1,\infty}/(\pi(A)\,\pi(A^{\mathrm{c}}))$ to obtain

$$\pi^+(A) \geqslant \frac{1}{2C_{1,\infty}}\,\pi(A)^2\,\pi(A^{\mathrm{c}})^2\,.$$

This inequality is not fully satisfactory, but if we take $\pi(A) = \frac{1}{2}$ then we deduce from this that $\mathcal{I}_\pi(\frac{1}{2}) \geqslant 1/(32C_{1,\infty})$, and from (2.4.25) we conclude. □

2.4.5 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 $\mathcal{I}_{\gamma_d}(p) \sim p\sqrt{2\log(1/p)}$ 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 (1997).

Theorem 2.4.29 (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* $f : \mathbb{R}^d \to [0,1]$,

$$\sqrt{\alpha}\,\mathcal{I}_{\gamma_d}(\mathbb{E}_\pi f) \leqslant \mathbb{E}_\pi\sqrt{\alpha\,\mathcal{I}_{\gamma_d}(f)^2 + \Gamma(f,f)}\,. \tag{2.4.30}$$

As this formulation suggests, Theorem 2.4.29 has a proof via Markov semigroup theory, for which we refer readers to Bakry et al. (2014, Section 8.5.2). By converting the functional inequality back into an isoperimetric statement, one can deduce the following comparison theorem.

Theorem 2.4.31 (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,*

$$\mathcal{I}_\pi \geqslant \sqrt{\alpha}\,\mathcal{I}_{\gamma_d}\,.$$

We explore these results further in the exercises.

2.5 Riemannian Manifolds

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. 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.

2.5.1 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 $(p_t)_{t\geqslant 0}$ such that the velocity $\dot p_t$ of the curve equals $-\nabla f(p_t)$ for all $t \geqslant 0$.

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 $p \in \mathcal{M}$.[5] 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

---

[5] Geometers would write that X is a *section* of the tangent bundle.

$Xf : \mathcal{M} \to \mathbb{R}$, 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

$$Z\langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z Y \rangle \,. \tag{2.5.1}$$

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 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

$$\mathsf{d}(p, q) = \inf\Bigl\{ \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)}\, \mathrm{d}t \Bigm| \gamma(0) = p,\ \gamma(1) = q \Bigr\} \,.$$

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[6] 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 (2.5.1), 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 $\nabla^2 f(p_t)[\dot{p}_t, \dot{p}_t]$. 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.[7]

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. 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. On Euclidean space, it vanishes: $\mathsf{Riem} = 0$. Also, the Riemann curvature tensor is fully determined by the **sectional**

---

[6] 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 $\operatorname{tr} A = \sum_{i=1}^d \langle e_i, A e_i \rangle_p$.

[7] 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.

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 $(B_t)_{t\geqslant 0}$ 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; see Hsu (2002).

More generally, a stochastic process $(Z_t)_{t\geqslant 0}$ 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 1.10 that the optimal transport problem can be posed with other costs; in particular, we take the cost to be $c(x, y) = \mathsf{d}(x, y)^2$, 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 $\mathsf{d}^2/2$-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[8]), most of the facts familiar to us from the Euclidean setting continue to hold when generalized appropriately. We refer to Villani (2009b) for more details.

2.5.2 Geometry of Markov Semigroups

We now indicate how Markov semigroups 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,

$$\begin{aligned}\Gamma_2(f,f) &= \frac{1}{2}\,\{\mathscr{L}(\|\nabla f\|^2) - 2\,\langle \nabla f, \nabla \mathscr{L} f\rangle\}\\ &= \frac{1}{2}\,\{\Delta(\|f\|^2) - 2\,\langle \nabla f, \nabla \Delta f\rangle\} + \langle \nabla f, \nabla^2 V\,\nabla f\rangle\,.\end{aligned}$$

Unlike in Section 2.2.1, however, we now have to apply the Bochner identity

$$\frac{1}{2}\,\Delta(\|\nabla f\|^2) = \langle \nabla f, \nabla \Delta f\rangle + \|\nabla^2 f\|_{\mathrm{HS}}^2 + \mathsf{Ric}(\nabla f, \nabla f) \tag{2.5.2}$$

which shows that

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

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 $\mathsf{Ric} + \nabla^2 V$. If $\mathsf{Ric} + \nabla^2 V \succeq \alpha$, in the sense that $\mathsf{Ric}(X,X) + \langle X, \nabla^2 V\, X\rangle \geqslant \alpha\,\|X\|^2$ for any vector field X on $\mathcal{M}$, then the curvature-dimension condition $\Gamma_2 \geqslant \alpha\Gamma$ holds. Since the proof of the Bakry–Émery theorem (Theorem 1.2.30) only relied on the $\mathsf{CD}(\alpha, \infty)$ 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

$$\|\nabla^2 f\|_{\mathrm{HS}}^2 \geqslant \frac{1}{d}\,(\operatorname{tr} \nabla^2 f)^2 = \frac{1}{d}\,(\Delta f)^2\,.$$

---

[8] 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.

This observation motivates the following definition.

Definition 2.5.3. A Markov semigroup is said to satisfy the **curvature-dimension condition** with curvature lower bound α and dimension bound d, denoted $\mathsf{CD}(\alpha, d)$, if for all functions f,

$$\Gamma_2(f, f) \geqslant \alpha\,\Gamma(f, f) + \frac{1}{d}\,(\mathscr{L} f)^2\,. \tag{2.5.4}$$

As the name suggests, the following theorem holds.

Theorem 2.5.5. *Let $\mathcal{M}$ be a complete Riemannian manifold with volume measure $\mathfrak{m}$, and let $\alpha > 0$, $d \geqslant 1$. Consider the Markov semigroup associated with standard Brownian motion on $\mathcal{M}$. Then, the following two statements are equivalent.*

1. $\mathsf{CD}(\alpha, d)$ *holds.*
2. $\mathsf{Ric} \succeq \alpha$ *and* $\dim \mathcal{M} \leqslant d$.

As an example, one can show that the unit sphere $\mathbb{S}^d$ satisfies $\mathsf{Ric} = d - 1$, so that the $\mathsf{CD}(d-1, d)$ condition holds. Then, using the Bakry–Émery theorem (Theorem 1.2.30), or by using Markov semigroup calculus to prove that the curvature-dimension condition implies Bobkov's functional form of the Gaussian isoperimetric inequality (Theorem 2.4.29; see Bakry et al., 2014, Corollary 8.5.4), one can now deduce results such as concentration on the sphere (Theorem 2.4.6).

Besides providing an abstract framework for deriving functional inequalities, it is worth noting that the condition (2.5.4) no longer makes any mention of the ambient space except through the Markov semigroup $(P_t)_{t \geqslant 0}$ 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 survey the literature here, we show one illustrative example to give the flavor of the results. First, one shows that the $\mathsf{CD}(\alpha, d)$ condition implies a Sobolev inequality.

Theorem 2.5.6 (Sobolev inequality)**.** *Consider a diffusion Markov semigroup satisfying the* $\mathsf{CD}(\alpha, d)$ *condition for some $\alpha > 0$ and $d > 2$. Then, for all $p \in [1, \frac{2d}{d-2}]$ and all functions f,*

$$\frac{1}{p-2}\Big\{\Big(\int |f|^p \,\mathrm{d}\pi\Big)^{2/p} - \int f^2 \,\mathrm{d}\pi\Big\} \leqslant \frac{d-1}{\alpha d} \int \Gamma(f, f)\,\mathrm{d}\pi\,. \tag{2.5.7}$$

From this Sobolev inequality, one can then deduce a diameter bound for the Markov semigroup. Here, the diameter is defined as follows:

$$\operatorname{diam}\big((P_t)_{t\geqslant 0}\big) \coloneqq \sup\Big\{\pi\text{-}\operatorname*{ess\,sup}_{x,y\in\mathcal{X}} |f(x) - f(y)| \;\Big|\; \|\Gamma(f, f)\|_{L^\infty(\pi)} \leqslant 1\Big\}\,.$$

Theorem 2.5.8 (Diameter bound)**.** *Suppose that the Markov semigroup $(P_t)_{t\geqslant 0}$ satisfies the Sobolev inequality (2.5.7). Then,*

$$\operatorname{diam}\big((P_t)_{t\geqslant 0}\big) \leqslant \pi\sqrt{\frac{d-1}{\alpha}}\,.$$

The diameter bound is sharp, as it is attained by the sphere, and together with Theorem 2.5.6 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.

2.5.3 Displacement Convexity of Entropy

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 $\mathsf{Ric} + \nabla^2 V \ge \alpha$.

Theorem 2.5.9. *Let $\mathcal{M}$ be a Riemannian manifold and let $\pi \propto \exp(-V)$ be a density w.r.t. the volume measure. Then, the following are equivalent:*

1. *It holds that $\mathsf{Ric} + \nabla^2 V \ge \alpha$.*
2. *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 such that for all $t \in [0, 1]$,*

$$\mathsf{KL}(\mu_t \,\|\, \pi) \leqslant (1-t)\,\mathsf{KL}(\mu_0 \,\|\, \pi) + t\,\mathsf{KL}(\mu_1 \,\|\, \pi) - \frac{\alpha\, t\,(1-t)}{2}\, W_2^2(\mu_0, \mu_1)\,. \tag{2.5.10}$$

A key observation is that the condition (2.5.10) still makes sense when we replace $\mathcal{M}$ with a space that admits a notion of geodesics, i.e., there is no need for a smooth structure *a priori*. Similarly, there are ways to to formulate the general $\mathsf{CD}(\alpha, d)$ condition via displacement convexity, but they are considerably more complicated, and we omit them for simplicity. This observation is the starting place for the influential **Lott–Sturm–Villani theory of synthetic Ricci curvature lower bounds** (Sturm, 2006a,b; Lott and Villani, 2009). See the supplement for further context and details.

2.6 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

$$\operatorname{ent}_\pi(f^2) \leqslant 2C\,\mathscr{E}(f, f) \qquad \text{for all } f\,. \tag{2.6.1}$$

On the other hand, recall from Theorem 1.2.26 (which still holds for discrete state spaces) that the exponential decay of the KL divergence is equivalent to the inequality

$$\operatorname{ent}_\pi(f) \leqslant \frac{C}{2}\,\mathscr{E}(f, \log f) \qquad \text{for all } f \geqslant 0\,. \tag{2.6.2}$$

When the carré du champ satisfies a chain rule, then (2.6.1) and (2.6.2) are equivalent, but in general the first inequality (2.6.1) is strictly stronger.

Lemma 2.6.3. *The inequality* (2.6.1) *implies inequality* (2.6.2).

See Exercise 2.23. The first inequality (2.6.1) is often simply called the *log Sobolev inequality*, whereas the second inequality (2.6.2) 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.

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 2.5, 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 (Ollivier, 2007, 2009) introduced the following notion of curvature.

Definition 2.6.4. 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 $x, y \in \mathcal{X}$,

$$W_1\big(P(x,\cdot), P(y,\cdot)\big) \leqslant (1-\kappa)\,\mathsf{d}(x, y)\,. \tag{2.6.5}$$

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 $\mathsf{Ric} \geq \alpha$, 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

$$W_1\big(P(x,\cdot), P(y,\cdot)\big) \leqslant \Big(1 - \frac{\alpha\varepsilon^2}{2\,(d+2)} + O(\varepsilon^3)\Big)\,\mathsf{d}(x, y)\,.$$

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 Ollivier (2007, 2009). 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 (Exercise 2.25).

Theorem 2.6.6 (Ollivier (2009))**.** *Suppose that P is a reversible Markov kernel on a metric space $(\mathcal{X}, \mathsf{d})$, 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) (Ollivier and Villani, 2012; Gozlan et al., 2014; Léonard, 2017); using ideas from Bakry–Émery theory (Klartag et al., 2016; Fathi and Shu, 2018); and defining a modified W_2 distance for which the Markov process becomes a gradient flow of the KL divergence (Maas, 2011; Erbar and Maas, 2012; Mielke, 2013).

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 Bakry et al. (2014); van Handel (2016) 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 notoriously difficult to study. See Villani (2003, Chapter 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 Cordero-Erausquin (2017). One might wonder whether a "log-Sobolev" version of the Brascamp–Lieb inequality holds, but the answer is unfortunately negative (Bobkov and Ledoux, 2000).

The general idea used in the proof of Lemma 2.2.7 is attributed to Hörmander (2007). In the proof of Lemma 2.2.7, we assumed the solvability of the Poisson equation; this can be avoided via a density argument, see Cordero-Erausquin et al. (2004); Barthe and Cordero-Erausquin (2013). The proof of the dimensional Brascamp–Lieb inequality in Exercise 2.4 is taken from Bolley et al. (2018), and the bound on $\operatorname{var}_\pi V$ therein was used in Chewi (2021) to show that the entropic barrier is an optimal self-concordant barrier. Also, Exercise 2.5 is from Helffer and Sjöstrand (1994) and Exercise 2.6 is from Klartag (2023). Finally, we caution the reader that the Brascamp–Lieb inequality in Theorem 2.2.9 should not be confused with another family of inequalities, which are unfortunately also known as Brascamp–Lieb inequalities and described in, e.g., Villani (2003, Section 6.3).

The convergence in Rényi divergence of the Langevin diffusion was obtained earlier, under stronger assumptions in Cao et al. (2019). 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 **Latala–Oleszkiewicz inequalities** (**LOI**) (Latala and Oleszkiewicz, 2000). The convergence of the Langevin diffusion under an LO inequality is given in Chewi et al. (2024a). One can also consider variants of Sobolev inequalities (Chafaï, 2004).

For further works investigating functional inequalities in the low temperature regime, see Kinoshita and Suzuki (2022); Li and Erdogdu (2023); Chen and Sridharan (2025); Gong et al. (2025).

The Prékopa–Leindler inequality given in Exercise 2.11 can be used to deduce other functional inequalities, such as the log-Sobolev inequality and the Bregman transport inequality; see Bobkov and

Ledoux (2000); Gentil (2008). It was subsequently generalized to Riemannian manifolds in Cordero-Erausquin et al. (2001).

The tightening lemma (Lemma 2.3.12) is usually proven via the Rothaus lemma (Rothaus, 1985); the sharper form stated in the lemma is due to Wang (2024). Exercise 2.14 essentially contains the main results of Chen et al. (2021a) (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 2.14). Earlier works on functional inequalities for mixtures include Chafaï and Malrieu (2010); Zimmermann (2013, 2016); Bardet et al. (2018); Schlichting (2019).

There are many treatments on concentration of measure, e.g., Ledoux (2001); Boucheron et al. (2013); Bakry et al. (2014); van Handel (2016); Vershynin (2018). The monographs Bobkov and Houdré (1997); Ledoux (2001); Bakry et al. (2014) are excellent sources to learn about isoperimetry. The exposition of the functional form of Cheeger's inequality (Theorem 2.4.18) as well as Milman's theorem (Theorem 2.4.22) were inspired by the treatment in Alonso-Gutiérrez and Bastero (2015). 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 2.4.18 and Proposition 2.4.21, 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 Bakry et al. (2014, Chapter 8). Also, more refined "two-level" isoperimetric inequalities have been pioneered in Talagrand (1991), which has applications in its own right.

From Exercise 2.15, the Harnack inequality was first established in Wang (1997), and the log-Harnack inequality proof is from Bobkov et al. (2001). These inequalities are variants of parabolic Harnack inequalities from the PDE literature, and they are notable for holding with dimension-free constants. Moreover, whereas parabolic Harnack inequalities typically compare the semigroup at two different times, the inequalities from Exercise 2.15 avoids this by considering the commutation of the Markov semigroup with strictly convex functions such as the power function $(\cdot)^p$. There is a huge literature on these inequalities which we cannot survey here, but we refer to F.-Y. Wang's monographs for details (Wang, 2013, 2014). The convexity principle is taken from Altschuler and Chewi (2024b), which also contains further discussion. Further proofs of the inequalities in Exercise 2.17 are given as Exercise 3.4 and Exercise 4.7.

The Gaussian isoperimetric inequality is due to Sudakov and Cirel'son (1974) and Borell (1975). It has since been extended and refined in various ways, e.g., in the context of noise stability (Borell, 1985; Isaksson and Mossel, 2012; Eldan, 2015; Mossel and Neeman, 2015; Kindler et al., 2018).

Section 2.5 draws upon many resources, which we list here: do Carmo (1992) for Riemannian geometry; Hsu (2002) for diffusions on manifolds; Villani (2009b) for optimal transport on manifolds, including synthetic Ricci curvature lower bounds; and Ledoux (2000); Bakry et al. (2014) for the geometry of Markov semigroups. The curvature-dimension condition and its equivalences have been explored in a vast number of works, e.g., Sturm (2006a,b); Lott and Villani (2009); Wang (2011).

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, Theorem 1.3.8). In this exercise, we linearize the Monge–Ampère equation to gain insight into the infinitesimal behavior of optimal transport.

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**:

$$-\mathscr{L}u = f\,, \qquad \text{where} \qquad \mathscr{L}u := \Delta u - \Big\langle \nabla \log\frac{1}{\mu}, \nabla u\Big\rangle\,.$$

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 1.7, deduce that the $\mathsf{T}_2(C)$ inequality implies

$$C\int f^2\,\mathrm{d}\mu \geqslant \int f\,(-\mathscr{L})^{-1} f\,\mathrm{d}\mu\,.$$

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 the formula (2.2.5) for the iterated carré du champ.

⊵ 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 2.2.12). Then, by applying Proposition 2.1.1, give another proof of the Brascamp–Lieb inequality (Theorem 2.2.9).

⊵ **Exercise 2.4** (Dimensional improvement of the Brascamp–Lieb inequality)
In finite-dimensional space, one can improve upon the Brascamp–Lieb inequality (Theorem 2.2.9) 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

$$\mathbb{E}_\pi[f^2] \leqslant \mathbb{E}_\pi[(f + \mathscr{L} u)^2] + \mathbb{E}_\pi \langle \nabla f, (\nabla^2 V)^{-1} \nabla f \rangle - \mathbb{E}_\pi[\|\nabla^2 u\|_{\mathrm{HS}}^2] \, .$$

2 Prove that $\mathbb{E}_\pi[\|\nabla^2 u\|_{\mathrm{HS}}^2] \geqslant d^{-1} (\mathbb{E}_\pi \Delta u)^2$, and that

$$\mathbb{E}_\pi \Delta u = \operatorname{cov}_\pi(f, V) - \mathbb{E}_\pi[(f + \mathscr{L} u) \, V] \, .$$

3 Choose u to solve $-\mathscr{L} u = f + \lambda \, (V - \mathbb{E}_\pi V)$ for some $\lambda \geqslant 0$ and substitute this into the previous parts. Optimize over λ and prove that

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

4 In particular, deduce that

$$\operatorname{var}_\pi V \leqslant \frac{d \, \mathbb{E}_\pi \langle \nabla V, (\nabla^2 V)^{-1} \nabla V \rangle}{d + \mathbb{E}_\pi \langle \nabla V, (\nabla^2 V)^{-1} \nabla V \rangle} \leqslant d \, .$$

⊵ **Exercise 2.5** (Helffer–Sjöstrand identity)
Let $\mathscr{L}$ denote the generator corresponding to $\pi \propto \exp(-V)$, $\nabla^2 V \succ 0$. Heuristically derive the following *identity*: for all $f, g : \mathbb{R}^d \to \mathbb{R}$,

$$\operatorname{cov}_\pi(f, g) = \mathbb{E}_\pi \langle \nabla f, (-\mathscr{L} + \nabla^2 V)^{-1} \nabla g \rangle \, .$$

Note that since $-\mathscr{L} \succcurlyeq 0$, this implies the Brascamp–Lieb inequality (Theorem 2.2.9)!

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 (1.2.18) and (2.2.4).

⊵ **Exercise 2.6** (Improved Lichnerowicz inequality)
Suppose that π is α-strongly log-concave. The goal of this exercise is to prove the following improved Lichnerowicz inequality:

$$C_{\mathsf{PI}}(\pi) \leqslant \sqrt{\frac{\|\operatorname{cov}_\pi\|_{\mathrm{op}}}{\alpha}} \, .$$

Note that this interpolates between the Bakry–Émery result $C_{\mathsf{PI}}(\pi) \leqslant 1/\alpha$ and the KLS conjecture $C_{\mathsf{PI}}(\pi) \lesssim \|\operatorname{cov}_\pi\|_{\mathrm{op}}$.

1 Let u be an eigenfunction corresponding to the spectral gap of $-\mathscr{L}$. Let $\lambda := 1/C_{\mathsf{PI}}(\pi)$. By writing $\lambda^2 = \int (\mathscr{L} u)^2 \, \mathrm{d}\pi$, bringing in the iterated carré du champ (as in the proof of Lemma 2.2.7), and applying strong log-concavity and the Poincaré inequality, show that $\|\int \nabla u \, \mathrm{d}\pi\|^2 \geqslant \alpha$.

2 Next, write $\|\int \nabla u \, \mathrm{d}\pi\| = \sup_{\theta \in \mathbb{R}^d,\, \|\theta\|=1} \int \langle \nabla u, \theta \rangle \, \mathrm{d}\pi$. Use the integration by parts formula and Cauchy–Schwarz to show that $\|\int \nabla u \, \mathrm{d}\pi\|^2 \leqslant \lambda^2 \, \|\mathrm{cov}_\pi\|_{\mathrm{op}}$. Conclude the proof.

⊵ **Exercise 2.7** (Local Poincaré inequality)

In Theorem 2.2.17, prove that $\mathsf{CD}(\alpha, \infty)$ implies (2.2.19), and that a Taylor expansion of either (2.2.18) or (2.2.19) implies back $\mathsf{CD}(\alpha, \infty)$.

Hint: To establish (2.2.19), use (2.2.20).

⊵ **Exercise 2.8** (Diffusion chain rule)

Verify the equivalence between (2.2.14) and (2.2.15).

⊵ **Exercise 2.9** (Monotonicity of Rényi divergences in the order)

Prove that for $1 \leqslant q \leqslant q' \leqslant \infty$, $\mathcal{R}_q \leqslant \mathcal{R}_{q'}$.

⊵ **Exercise 2.10** (Hypercontractivity)

Let $(P_t)_{t \geqslant 0}$ be a Markov semigroup with stationary distribution π, and let $\alpha \geqslant 0$. Assume that the semigroup either satisfies the diffusion chain rule, or that it is reversible. Show that the log-Sobolev inequality in the form

$$\mathrm{ent}_\pi(f^2) \leqslant 2 C_{\mathsf{LSI}} \, \mathscr{E}(f, f)$$

for all f is equivalent to the following **hypercontractivity** statement: for all functions f, $t \geqslant 0$, and $p \geqslant 1$, if we set $p(t) := 1 + (p-1)\exp(2t/C_{\mathsf{LSI}})$, then

$$\|P_t f\|_{L^{p(t)}(\pi)} \leqslant \|f\|_{L^p(\pi)} \, .$$

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 $t \mapsto \log \|P_t f\|_{L^{p(t)}(\pi)}$. In the case where the semigroup is assumed to be reversible (but does not satisfy the diffusion chain rule), one needs to show the following inequality: for any non-negative function h and any $q > 1$, it holds that $\mathscr{E}(h^{q-1}, h) \geqslant \frac{4\,(q-1)}{q^2} \, \mathscr{E}(h^{q/2}, h^{q/2})$. Show that this follows from the numerical inequality $(b^{q-1} - a^{q-1})\,(b-a) \geqslant \frac{4\,(q-1)}{q^2} \, (b^{q/2} - a^{q/2})^2$ for all $0 \leqslant a \leqslant b$.

⊵ **Exercise 2.11** (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 $x, y \in \mathbb{R}^d$,

$$\log h\big((1-t)\,x + t\,y\big) \geqslant (1-t)\log f(x) + t \log g(y) - \frac{\alpha}{2}\, t\,(1-t)\,\|x-y\|^2 \, .$$

Prove that

$$\log \int h \, \mathrm{d}\pi \geqslant (1-t) \log \int f \, \mathrm{d}\pi + t \log \int g \, \mathrm{d}\pi \, . \tag{2.E.1}$$

Hint: Let μ_0, μ_1 achieve equality in the Donsker–Varadhan variational principle (Theorem 1.5.7), so that

$$\begin{aligned}\log \mathbb{E}_\pi f &= \mathbb{E}_{\mu_0} \log f - \mathsf{KL}(\mu_0 \mathbin{\|} \pi)\,,\\ \log \mathbb{E}_\pi g &= \mathbb{E}_{\mu_1} \log g - \mathsf{KL}(\mu_1 \mathbin{\|} \pi)\,.\end{aligned}$$

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 $\log \mathbb{E}_\pi h$.

2 The inequality (2.E.1) continues to hold when π is replaced by Lebesgue measure, if we set $\alpha = 0$ in the assumption.[9] 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 $f := \pi(x_0, \cdot)$, $g := \pi(x_1, \cdot)$, and $h := \pi((1-t)\,x_0 + t\,x_1, \cdot)$.

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 $(x, y) \mapsto \log \pi(x, y) + \frac{1}{2}\,\langle (x, y), \Sigma^{-1}\,(x, y)\rangle$ is concave, where

$$\Sigma = \begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2}\\ \Sigma_{2,1} & \Sigma_{2,2}\end{bmatrix} \succ 0\,.$$

Prove that for the marginal π^1 of π on $\mathbb{R}^{d_1+d_2}$, $x \mapsto \log \pi^1(x) + \frac{1}{2}\,\langle x, (\Sigma_{1,1})^{-1}\,x\rangle$ 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 $A, B \subseteq \mathbb{R}^d$,

$$\operatorname{vol}\big((1-t)\,A + t\,B\big) \geqslant \operatorname{vol}(A)^{1-t}\operatorname{vol}(B)^t\,. \tag{2.E.2}$$

By scaling A and B and choosing t, show that (2.E.2) can be upgraded to

$$\operatorname{vol}(A+B)^{1/d} \geqslant \operatorname{vol}(A)^{1/d} + \operatorname{vol}(B)^{1/d}\,.$$

Operations Preserving Functional Inequalities

⊵ Exercise 2.12 (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 2.2.11). 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 2.13 (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 $x \in \mathbb{R}$, $\mu((-\infty, T(x)]) = \pi((-\infty, x])$.

1 Differentiate this relation to obtain a formula for $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(T(x))$.

---

[9] 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 $R \to \infty$.

2. Substitute this into the KL divergence $\mathsf{KL}(\mu \,\|\, \pi) = \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(T(x))\, \pi(\mathrm{d}x)$ and use the inequality $t - 1 - \log t \geqslant 0$ 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 2.14** (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 2.3.9 except that we replace the assumption (2.3.10) with the weaker condition
$$C_{\chi^2,2} \coloneqq \sqrt{\mathbb{E}[\chi^2(P_X \,\|\, P_{X'})^2]} < \infty\,, \tag{2.E.3}$$
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
$$\begin{aligned} \operatorname{var} \mathbb{E}_{P_X} f &\leqslant \mathbb{E} \min\{(\operatorname{var}_{P_X} f)\, \chi^2(\mu P \,\|\, P_X), (\operatorname{var}_{\mu P} f)\, \chi^2(P_X \,\|\, \mu P)\} \\ &\leqslant \mathbb{E} \sqrt{(\operatorname{var}_{P_X} f)\, \chi^2(\mu P \,\|\, P_X)\, (\operatorname{var}_{\mu P} f)\, \chi^2(P_X \,\|\, \mu P)}\,. \end{aligned}$$
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
$$\mathbb{E}\big[\mathbb{E}_{P_X}(f^2) \log\big(1 + \chi^2(P_X \,\|\, P_{X'})\big)\big] \leqslant \mathbb{E}_{\mu P}(f^2) \log(1 + C_{\chi^2})\,,$$
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 2.3.5) to prove that for a suitable constant $C > 0$ (depending on $C_{\chi^2,2}$),
$$2\, \mathbb{E}\big[\mathbb{E}_{P_X}(f^2)\, \big\{\log\big(1 + \chi^2(P_X \,\|\, P_{X'})\big) - C\big\}\big] \leqslant \operatorname{ent} \mathbb{E}_{P_X}(f^2)\,.$$
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 $\mathsf{B}(0, R)$, we assume that μ has sub-Gaussian tails:
$$\iint \exp \frac{\|x - x'\|^2}{\sigma_{\mathsf{sG}}^2}\, \mu(\mathrm{d}x)\, \mu(\mathrm{d}x') \leqslant C_{\mathsf{sG}}\,.$$
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.

The next group of exercises introduces Harnack and reverse transport inequalities.

⊵ **Exercise 2.15** (Harnack and reverse transport inequalities)

Let P be a Markov kernel over a space $\mathcal{X}$, and let $x, y \in \mathcal{X}$.

1 Use Donsker–Varadhan duality to show that

$$Pf(x) \leqslant C_{\log}(x, y) + \log P(\exp f)(y) \qquad \text{for all bounded } f : \mathcal{X} \to \mathbb{R}, \tag{2.E.4}$$

if and only if

$$\mathsf{KL}(\delta_x P \,\|\, \delta_y P) \leqslant C_{\log}(x, y)\,. \tag{2.E.5}$$

2 Similarly, let $p, q \in (1, \infty)$ be a pair of Hölder conjugate exponents, and use duality to show that

$$Pf(x) \leqslant C_p(x, y)\, P(f^p)(y)^{1/p} \qquad \text{for all bounded } f : \mathcal{X} \to \mathbb{R}_+, \tag{2.E.6}$$

if and only if

$$\mathcal{R}_q(\delta_x P \,\|\, \delta_y P) \leqslant \frac{q}{q-1} \log C_p(x, y)\,. \tag{2.E.7}$$

We refer to (2.E.4) as a **log-Harnack inequality**, (2.E.6) as a **Harnack inequality**, and (2.E.5) and (2.E.7) as **reverse transport inequalities**.

3 Show that if (2.E.6) holds for every $p > 1$ with $C_p = \exp(\frac{C}{p-1})$, then (2.E.4) holds with $C_{\log} \leqslant C$. *Hint*: Let $p \to \infty$.

⊵ **Exercise 2.16** (Convexity principle)
Let P be a Markov kernel over $\mathcal{X}$.

1 Suppose that $\mathsf{KL}(\delta_x P \,\|\, \delta_y P) \leqslant \rho(x, y)$ for all $x, y \in \mathcal{X}$. Using the joint convexity of the KL divergence (Theorem 1.5.6), prove that

$$\mathsf{KL}(\mu P \,\|\, \nu P) \leqslant \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \rho(x, y)\, \gamma(\mathrm{d}x, \mathrm{d}y)\,.$$

2 Suppose that $\mathcal{R}_q(\delta_x P \,\|\, \delta_y P) \leqslant \rho(x, y)$ for all $x, y \in \mathcal{X}$. Although $\mathcal{R}_q$ is not jointly convex for $q > 1$, it is a monotonic transformation of an f-divergence. Use this to give an upper bound on $\mathcal{R}_q(\mu P \,\|\, \nu P)$ in terms of an optimal transport problem between μ and ν.

⊵ **Exercise 2.17** (Harnack inequalities via semigroup arguments)
Let $(P_t)_{t\geqslant 0}$ be a Markov semigroup over $\mathbb{R}^d$, satisfying the $\mathsf{CD}(\alpha, \infty)$ condition with $\alpha \in \mathbb{R}$.

1 Let $x, y \in \mathbb{R}^d$ and let $(x_s)_{s\in[0,t]}$ be a curve joining x to y. Differentiate $P_s(\log P_{t-s} f)(x_s)$ w.r.t. s and apply the diffusion chain rule and the gradient bound (Theorem 2.2.6). By optimizing over the choice of curves $(x_s)_{s\in[0,t]}$, prove the log-Harnack inequality

$$P_t(\log f)(x) \leqslant \log P_t f(y) + \frac{\alpha\, \|x - y\|^2}{2\,(\exp(2\alpha t) - 1)}\,.$$

By Exercise 2.15 and Exercise 2.16, it implies that for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathsf{KL}(\mu P_t \,\|\, \nu P_t) \leqslant \frac{\alpha\, W_2^2(\mu, \nu)}{2\,(\exp(2\alpha t) - 1)}\,. \tag{2.E.8}$$

Note that this strengthens (1.4.12) in that it allows the second argument to be νP_t for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$.

2 Find an analogous Markov semigroup argument to prove that

$$(P_t f(x))^p \leqslant P_t(f^p)(y) \exp\Bigl(\frac{\alpha p \, \|x-y\|^2}{2\,(p-1)\,(\exp(2\alpha t)-1)}\Bigr)\,, \tag{2.E.9}$$

and hence

$$\mathcal{R}_q(\delta_x P_t \,\|\, \delta_y P_t) \leqslant \frac{\alpha q \, \|x-y\|^2}{2\,(\exp(2\alpha t)-1)}\,. \tag{2.E.10}$$

3 Here, we consider an alternative argument. Starting from the local log-Sobolev inequality in the form (2.2.22), establish (2.E.9) by differentiating $s \mapsto \frac{p}{\xi_s} \log P_t(f^{\xi_s})(x_s)$, where $\xi_s := 1 + (p-1)\,s$. Argue that this reasoning can be reversed, so that (2.E.9) implies back the local log-Sobolev inequality, and hence $\mathsf{CD}(\alpha, \infty)$.

Concentration of Measure and Isoperimetry

⊵ **Exercise 2.18** (Isoperimetry on the sphere)

Prove Theorem 2.4.6 from the spherical isoperimetric inequality (Theorem 2.4.5). To do so, use the fact that the measure of $\mathsf{B}(x_0, r)$ is

$$\sigma_d\bigl(\mathsf{B}(x_0, r)\bigr) = \frac{\int_0^r (\sin\theta)^{d-1}\,\mathrm{d}\theta}{\int_0^\pi (\sin\theta)^{d-1}\,\mathrm{d}\theta}\,,$$

or prove this fact yourself. It is also acceptable to establish a weaker bound of the form $\alpha_{\sigma_d}(\varepsilon) \leqslant C\exp(-cd\varepsilon^2)$ for universal constants $c, C > 0$.

⊵ **Exercise 2.19** (Gaussian isoperimetry)

Consider the Gaussian isoperimetric inequality in Theorem 2.4.7.

1 In the spirit of Theorem 2.4.18, show that the functional inequality (2.4.30) is equivalent to an isoperimetric statement. Consequently, deduce the comparison theorem (Theorem 2.4.31) from Theorem 2.4.29.

2 Show that the functional form of the Gaussian isoperimetric inequality in (2.4.30) is preserved (up to constants) under Lipschitz mappings (in other words, prove the analogue of Proposition 2.3.3 for (2.4.30)).

⊵ **Exercise 2.20** (Gaussian isoperimetry implies LSI)

In a similar spirit to Exercise 1.7, show that the Gaussian isoperimetric inequality (2.4.30) implies the log-Sobolev inequality.

⊵ **Exercise 2.21** (Isoperimetry and optimal transport in dimension one)

Let $\mu, \pi \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R})$, and let T denote the optimal transport map from π to μ.

1 Show that $T = F_\mu^{-1} \circ F_\pi$, where F_μ, F_π denote the CDFs of μ and π respectively.

2 Use the Monge–Ampère equation (Exercise 2.1) to deduce that T is L-Lipschitz if and only if $\mu \circ F_\mu^{-1} \geqslant L^{-1}\,\pi \circ F_\pi^{-1}$, i.e., the isoperimetric profile of μ is lower bounded by the isoperimetric profile of π.

3 When $\pi = \mathsf{normal}(0, I_d)$, the existence of a Lipschitz transport map from π to μ implies that μ satisfies a Gaussian isoperimetric inequality, by Exercise 2.19. Use the preceding part to deduce the converse of this statement holds in dimension one. What happens when π is the Laplace distribution with density $\pi(x) = \frac{1}{2}\exp(-|x|)$?

Riemannian Manifolds

⊵ **Exercise 2.22** (Lichnerowicz inequality)
Under the $\mathsf{CD}(\alpha, d)$ condition (2.5.4), the spectral gap estimate for $-\mathscr{L}$ can be sharpened to $\lambda_{\min}(-\mathscr{L}) \geqslant \alpha d/(d-1)$, 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 $\mathsf{CD}(\alpha, d)$ and deduce that $\lambda \geqslant \alpha d/(d-1)$.

Discrete Space and Time

⊵ **Exercise 2.23** (LSI implies MLSI)
Prove Lemma 2.6.3.

Hint: Prove that $4\,(\sqrt{a} - \sqrt{b})^2 = (\int_a^b t^{-1/2}\, \mathrm{d}t)^2 \leqslant (\log a - \log b)\,(a-b)$ for all $a, b > 0$.

⊵ **Exercise 2.24** (Coarse Ricci curvature on graphs)
Show that if (2.6.5) holds for all $x, y \in \mathcal{X}$, then $W_1(\mu P, \nu P) \leqslant (1-\kappa)\, W_1(\mu, \nu)$ for all $\mu, \nu \in \mathcal{P}_1(\mathcal{X})$. Furthermore, show that when d is the shortest path metric on a graph, it suffices to check (2.6.5) for neighbors, that is, for pairs (x, y) with $\mathsf{d}(x, y) = 1$.

⊵ **Exercise 2.25** (Coarse Ricci curvature implies PI)
Suppose that $\mathcal{X}$ has bounded diameter and that P satisfies (2.6.5). Let $f : \mathcal{X} \to \mathbb{R}$ be Lipschitz and satisfy $\int f\, \mathrm{d}\pi = 0$, where π is the stationary distribution for P. Prove that $\limsup_{n\to\infty} (\operatorname{var}_\pi P^n f)^{1/n} \leqslant (1-\kappa)^2$, and deduce that the operator norm of P restricted to the subspace $\mathcal{V} \subseteq L^2(\pi)$ consisting of Lipschitz functions with zero mean is at most $1-\kappa$. Next, using the fact that $\mathcal{V}$ is dense in $L^2(\pi)$, establish Theorem 2.6.6. (See Section 7.4 for background on discrete-time chains.)

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 1.1.19).

Finite variation vs. quadratic variation.

As a first step toward understanding the difficulties we faced when constructing the Itô integral, we recall the classical condition under which it is possible to integrate a continuous process $(\eta_t)_{t\in[0,T]}$ against another continuous process $(A_t)_{t\in[0,T]}$ to form the integral $\int_0^T \eta_t \, \mathrm{d}A_t$. This condition states that 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

$$\operatorname{mesh}(t_i : i = 0, 1, \dots, n) := \max_{i\in[n]} |t_i - t_{i-1}|\,,$$

then it holds that

$$\lim_{\operatorname{mesh}(t_i \,:\, i=0,1,\dots,n)\searrow 0} \sum_{i=1}^n |A_{t_i} - A_{t_{i-1}}| < \infty\,.$$

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 $\int_{[0,T]} \eta_t \, \mathrm{d}A_t := \int_{[0,T]} \eta_t \, \mu_A(\mathrm{d}t)$. 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 difficulty, 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 $B_{t_i} - B_{t_{i-1}} \sim \mathsf{normal}(0, T/n)$, we expect (heuristically) that

$$\liminf_{n\to\infty} \sum_{i=1}^{n} \underbrace{|B_{t_i} - B_{t_{i-1}}|}_{\asymp \sqrt{T/n}} \gtrsim \liminf_{n\to\infty} n \cdot \sqrt{\frac{T}{n}} = \infty \,.$$

On the other hand, if we change the definition slightly, then we expect (heuristically) that

$$\limsup_{n\to\infty} \sum_{i=1}^{n} \underbrace{|B_{t_i} - B_{t_{i-1}}|^2}_{\asymp T/n} \lesssim \limsup_{n\to\infty} n \cdot \frac{T}{n} < \infty \,.$$

We say that Brownian motion has finite **quadratic variation**. We will show in fact that the above limit is well-defined and non-trivial 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.

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. The process $[M, M]$ of Theorem 3.1.1 is called the **quadratic variation** of M.

Note that Theorem 3.1.1 provides another interpretation of the quadratic variation: when we compose the martingale M with the convex function $(\cdot)^2$, then M^2 is no longer a martingale.[2] The quadratic variation is the process that we can subtract from M^2 in order to make it a martingale.

---

[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.

[2] It is, however, a submartingale.

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,

$$\begin{aligned}\mathbb{E}\Big[\Big|\sum_{i=1}^{n}\{(B_{t_i}-B_{t_{i-1}})^2-(t_i-t_{i-1})\}\Big|^2\Big] &= \sum_{i=1}^{n}\mathbb{E}[|(B_{t_i}-B_{t_{i-1}})^2-(t_i-t_{i-1})|^2] \\ &\leqslant \sum_{i=1}^{n}\mathbb{E}[(B_{t_i}-B_{t_{i-1}})^4] = 3\sum_{i=1}^{n}(t_i-t_{i-1})^2 \\ &\leqslant 3\,\mathrm{mesh}(t_i : i=0,1,\dots,n)\underbrace{\sum_{i=1}^{n}(t_i-t_{i-1})}_{=T} \to 0\,.\end{aligned}$$

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 1.5).

Semimartingales.

We often consider solutions to SDEs with non-zero drift coefficients, which therefore are not continuous local martingales. To accommodate this addition, we consider the following definition.

Definition 3.1.3. A process $(X_t)_{t\in[0,T]}$ is called a **continuous semimartingale** if we can write $X = A + M$, where A is an adapted 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,

$$\sum_{i=1}^{n}(\Delta_{t_i}-\Delta_{t_{i-1}})^2 \leqslant \mathrm{mesh}(t_i : i=0,1,\dots,n)\underbrace{\sum_{i=1}^{n}|\Delta_{t_i}-\Delta_{t_{i-1}}|}_{\text{bounded as } n\to\infty} \to 0$$

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

$$\begin{aligned}
&\Big|\sum_{i=1}^n (X_{t_i} - X_{t_{i-1}})^2 - \sum_{i=1}^n (M_{t_i} - M_{t_{i-1}})^2\Big| \\
&\quad = \Big|\sum_{i=1}^n (A_{t_i} - A_{t_{i-1}})^2 + 2\sum_{i=1}^n (A_{t_i} - A_{t_{i-1}})\,(M_{t_i} - M_{t_{i-1}})\Big| \\
&\quad \leqslant \sum_{i=1}^n (A_{t_i} - A_{t_{i-1}})^2 + 2\sqrt{\Big(\sum_{i=1}^n (A_{t_i} - A_{t_{i-1}})^2\Big)\Big(\sum_{i=1}^n (M_{t_i} - M_{t_{i-1}})^2\Big)}\,,
\end{aligned}$$

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:

$$[X,Y] := \frac{1}{2}\,([X+Y, X+Y] - [X,X] - [Y,Y])\,.$$

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

$$[X,Y]_t = \int_0^t \langle \sigma_s^X, \sigma_s^Y\rangle\,\mathrm{d}s\,, \qquad \textit{for } t \in [0,T]\,.$$

Itô's formula revisited.

Finally, we conclude this section by revisiting Itô's formula (Theorem 1.1.19) 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*

$$f(X_t) = f(X_0) + \sum_{i=1}^d \int_0^t \partial_i f(X_s)\,\mathrm{d}X_s^i + \frac{1}{2}\sum_{i,j=1}^d \int_0^t \partial_i\partial_j f(X_s)\,\mathrm{d}[X^i, X^j]_s\,.$$

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 1.1.19. 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 $(\Omega, \mathscr{F}, \mathbb{P})$. 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 $\mathscr{F}$ 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 $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}$. 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 characterizing 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*

$$h \in \mathcal{H} := \Big\{h \in C([0,T]) \;\Big|\; h(0) = 0,\; \int_0^T \|\dot h(t)\|^2 \, \mathrm{d}t < \infty\Big\}\,. \tag{3.2.2}$$

If this holds, then

$$\frac{\mathrm{d}(T_h)_\# \mathbf{W}}{\mathrm{d}\mathbf{W}}(\omega) = \exp\Big(\int_0^T \langle \dot h(t), \mathrm{d}\omega_t\rangle - \frac{1}{2}\int_0^T \|\dot h(t)\|^2 \, \mathrm{d}t\Big)\,. \tag{3.2.3}$$

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 (3.2.3), 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

$$\frac{\mathrm{d}(T_h)_\# \gamma}{\mathrm{d}\gamma}(x) = \exp\Big(\langle h, x\rangle - \frac{1}{2}\|h\|^2\Big)\,.$$

This bears a striking resemblance to (3.2.3). 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 (3.2.3) behaves as if

$$\mathrm{d}\mathbf{W}(\omega) \text{ “}\propto\text{” } \exp\Big(-\frac{1}{2}\|\omega\|_{\mathcal{H}}^2\Big)\,\mathrm{d}\omega\,. \tag{3.2.4}$$

The charming part about (3.2.4) 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, (3.2.4) 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 $\|x\|_{\mathcal{H}} := \sup\{|\ell(\omega)| \mid \ell \in \mathcal{B}^*,\ \|\ell\|_{L^2(\mu)} \leqslant 1\}$ 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

$$\mathrm{d}X_t = b_t\,\mathrm{d}t + \mathrm{d}B_t \iff X_t = \int_0^t b_s\,\mathrm{d}s + B_t\,, \tag{3.2.5}$$

where $(b_t)_{t\geqslant 0}$ is an adapted process. If $(b_t)_{t\geqslant 0}$ is in fact deterministic, then the law of X_t is a centered Gaussian with covariance $(\int_0^t b_s\,\mathrm{d}s)^{\otimes 2} + tI_d$, and is therefore fully explicit. 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 (3.2.5):

$$X_{n+1} := F(X_n) + \xi_n\,, \qquad n = 0,1,2,\dots\,,$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is a deterministic map and $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. Gaussian variables. Again, due to the non-linear mapping F, the law of X_n is hard to describe exactly, yet once we *condition* on X_n, the law of X_{n+1} is an explicit Gaussian. This observation makes it straightforward to write down an explicit and simple expression for the joint law of $(X_0, X_1, \dots, X_N)$. See Exercise 3.3 for further elaboration on this discrete-time approach.

Returning to (3.2.5), 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

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{W}} = \exp\Bigl(\int_0^T \langle b_s, \mathrm{d}B_s\rangle - \frac{1}{2}\int_0^T \|b_s\|^2\,\mathrm{d}s\Bigr)\,. \tag{3.2.6}$$

Note that this exactly mirrors Theorem 3.2.1, except that now we allow for adapted processes $(b_t)_{t\geqslant 0}$.

Let us now discuss how we would establish (3.2.6) carefully. Actually, we will proceed in the opposite order from our informal discussion: we will first *define* **P** via the formula (3.2.6), and then investigate the effect of the change of measure from **W** to **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 **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 **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 (3.2.6) defines a valid probability measure **P**. In other words, we need the **W**-expectation of the right-hand side of (3.2.6) to equal 1. Actually, for any $t \in [0,T]$, let us write $\mathbf{W}_t$ to be the restriction of **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

$$t \mapsto \frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{W}_t} \overset{?}{=} \exp\Bigl(\int_0^t \langle b_s, \mathrm{d}B_s\rangle - \frac{1}{2}\int_0^t \|b_s\|^2\,\mathrm{d}s\Bigr)$$

to have constant **W**-expectation, equal to 1 for all $t \in [0,T]$. This would follow if we knew that this defined a **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 **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,

$$\mathcal{E}(M) := \exp\bigl(M - \frac{1}{2}\,[M,M]\bigr) \tag{3.2.7}$$

is called the **exponential martingale** associated with M. Is it actually a martingale? Applying Itô's formula in the form (3.1.6),

$$\mathrm{d}\mathcal{E}(M)_t = \mathcal{E}(M)_t\,\Bigl(\mathrm{d}M_t - \frac{1}{2}\,\mathrm{d}[M,M]_t + \frac{1}{2}\,\mathrm{d}[M,M]_t\Bigr) = \mathcal{E}(M)_t\,\mathrm{d}M_t\,,$$

so $\mathcal{E}(M)$ is a stochastic integral. From Proposition 1.1.16, this tells us that $\mathcal{E}(M)$ is a continuous *local* **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[3] 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 **P** defined via (3.2.6) is a probability measure on path space. We now claim that under **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 (3.2.6) and characteristic functions; we leave it as Exercise 3.2. It leads to the following theorem.

[3] We borrow the terminology from Steele (2001).

Theorem 3.2.8 (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* $[M,M] = \int_0^\cdot \|b_s\|^2 \,\mathrm{d}s$ *denote the quadratic variation. Define the exponential martingale*

$$\mathcal{E}(M) := \exp\bigl(M - \frac{1}{2}\,[M,M]\bigr)\,.$$

Assume that $\mathcal{E}(M)$ *is a* $\mathbf{W}$*-martingale and define the measure* $\mathbf{P}$ *on path space via*

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{W}} = \mathcal{E}(M)_T\,.$$

Then, under $\mathbf{P}$*,*

$$t \mapsto \tilde B_t := B_t - [B,M]_t = B_t - \int_0^t b_s\,\mathrm{d}s \qquad \text{is a standard Brownian motion}\,.$$

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 $\mathbb{E}^{\mathbf{W}}\,\mathcal{E}(M)_T \leqslant 1$. 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 (3.2.6) 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**, which reads $\mathbb{E}^{\mathbf{W}} \exp(\frac{1}{2}\,[M,M]_T) < \infty$. 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 can often use the technique of localization, mentioned in Section 1.1.1, together with lower semicontinuity, to avoid these conditions altogether. In this book, we omit discussion of these technical arguments.

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 Radon–Nikodym 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 $(h_t(X_t))_{t\in[0,T]}$,

$$\mathrm{d}h_t(X_t) = (\partial_t h_t + \mathscr{L}_t h_t)(X_t)\,\mathrm{d}t + \langle \nabla h_t(X_t), \sigma_t(X_t)\,\mathrm{d}B_t\rangle\,, \tag{3.3.2}$$

and we deduce that this is a martingale if and only if

$$\partial_t h_t + \mathscr{L}_t h_t = 0\,,$$

where $\mathscr{L}_t f := \frac{1}{2}\,\langle \sigma_t\sigma_t^{\mathsf{T}}, \nabla^2 f\rangle + \langle b_t, \nabla f\rangle$ 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 $h_t^{\leftarrow} := h_{T-t}$, 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 $M := \int_0^{\cdot} \langle \Delta, \mathrm{d}B\rangle$, then Girsanov's theorem (Theorem 3.2.8) yields

$$\mathrm{d}\big(\log \frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{Q}_t}\big) = \langle \Delta_t, \mathrm{d}B_t\rangle - \frac{1}{2}\,\|\Delta_t\|^2\,\mathrm{d}t\,.$$

On the other hand, Itô's formula and (3.3.2) yield

$$\begin{aligned}\mathrm{d}\big(\log h_t(X_t)\big) &= \frac{1}{h_t(X_t)}\,\mathrm{d}h_t(X_t) - \frac{1}{2\,h_t(X_t)^2}\,\mathrm{d}[h_\cdot(X), h_\cdot(X)]_t \\ &= \frac{1}{h_t(X_t)}\,\langle \nabla h_t(X_t), \sigma_t(X_t)\,\mathrm{d}B_t\rangle - \frac{1}{2\,h_t(X_t)^2}\,\|\sigma_t(X_t)^{\mathsf{T}}\,\nabla h_t(X_t)\|^2\,\mathrm{d}t\end{aligned}$$

and we quickly deduce that

$$\Delta_t = \sigma_t(X_t)^{\mathsf{T}}\,\nabla \log h_t(X_t)\,.$$

Finally, we have obtained

$$\begin{aligned}\mathrm{d}X_t &= \big(b_t(X_t) + \sigma_t(X_t)\,\Delta_t\big)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}\tilde{B}_t \\ &= \big(b_t(X_t) + \sigma_t(X_t)\,\sigma_t(X_t)^{\mathsf{T}}\,\nabla \log h_t(X_t)\big)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}\tilde{B}_t\,.\end{aligned}$$

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 $\pi_t := \operatorname{law}^{\mathbf{Q}}(X_t)$,

$$\begin{aligned}
\mathbb{E}^{\mathbf{Q}}[\eta_t\, f(X_T)] &= \mathbb{E}^{\mathbf{Q}}\big[\eta_t\, \mathbb{E}^{\mathbf{Q}}[f(X_T) \mid X_t]\big] = \mathbb{E}^{\mathbf{Q}}\Big[\eta_t \int f(x_T)\, P_{t,T}(X_t, \mathrm{d}x_T)\Big] \\
&= \mathbb{E}^{\mathbf{Q}}\Big[\eta_t \int f(x_T)\, \frac{\mathrm{d}P_{t,T}(X_t, \cdot)}{\mathrm{d}\pi_T}(x_T)\, \pi_T(\mathrm{d}x_T)\Big] \\
&= \int f(x_T)\, \mathbb{E}^{\mathbf{Q}}\Big[\eta_t\, \frac{\mathrm{d}P_{t,T}(X_t, \cdot)}{\mathrm{d}\pi_T}(x_T)\Big]\, \pi_T(\mathrm{d}x_T)\,.
\end{aligned}$$

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

$$\frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}\mathbf{Q}_t} = h_t^{x_T}(X_t) \qquad \text{for all } 0 \leqslant t < T\,.$$

We are now in the setting of Doob's transform. In particular, under $\mathbf{P}$,

$$\mathrm{d}X_t = \big(b_t(X_t) + \sigma_t(X_t)\,\sigma_t(X_t)^{\mathsf{T}}\, \nabla \log h_t^{x_T}(X_t)\big)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}\tilde{B}_t\,, \qquad 0 \leqslant t < T\,. \tag{3.3.3}$$

Example 3.3.4 (Brownian bridge)**.** Suppose that $B = X$ is a standard Brownian motion under $\mathbf{Q}$. Then, $P_{s,t} = \mathsf{normal}(0, (t-s)\, I_d)$ 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

$$h_t^{x_T}(x) \propto \exp\Big(-\frac{\|x - x_T\|^2}{2\,(T-t)}\Big) \qquad \Longrightarrow \qquad \nabla \log h_t^{x_T}(x) = -\frac{x - x_T}{T-t}\,.$$

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 $\pi_t := \operatorname{law}(X_t)$ is the marginal law of X at time t. We will construct another SDE $X^{\leftarrow}$ such that $\operatorname{law}(X_t^{\leftarrow}) = \pi_{T-t}$ 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

$$\partial_t \pi_t = \frac{1}{2} \langle \sigma_t \sigma_t^\mathsf{T}, \nabla^2 \pi_t \rangle - \operatorname{div}(\pi_t b_t)\,.$$

Therefore, the time reversal $\pi_t^\leftarrow := \pi_{T-t}$ satisfies

$$\partial_t \pi_t^\leftarrow = -\frac{1}{2} \langle \sigma_{T-t} \sigma_{T-t}^\mathsf{T}, \nabla^2 \pi_t^\leftarrow \rangle + \operatorname{div}(\pi_t^\leftarrow b_{T-t})\,.$$

Next, we note that

$$\langle \sigma_{T-t} \sigma_{T-t}^\mathsf{T}, \nabla^2 \pi_t^\leftarrow \rangle = \operatorname{div}(\sigma_{T-t} \sigma_{T-t}^\mathsf{T}\, \nabla \pi_t^\leftarrow) = \operatorname{div}\bigl(\pi_t^\leftarrow\, (\sigma_{T-t} \sigma_{T-t}^\mathsf{T}\, \nabla \log \pi_t^\leftarrow)\bigr)\,,$$

and hence we can write

$$\partial_t \pi_t^\leftarrow = \frac{1}{2} \langle \sigma_{T-t} \sigma_{T-t}^\mathsf{T}, \nabla^2 \pi_t^\leftarrow \rangle + \operatorname{div}\bigl(\pi_t^\leftarrow\, (b_{T-t} - \sigma_{T-t} \sigma_{T-t}^\mathsf{T} \nabla \log \pi_t^\leftarrow)\bigr)\,. \tag{3.3.5}$$

We can therefore read off the SDE

$$\mathrm{d}X_t^\leftarrow = \{-b_{T-t}(X_t^\leftarrow) + \sigma_{T-t} \sigma_{T-t}^\mathsf{T}\, \nabla \log \pi_t^\leftarrow(X_t^\leftarrow)\}\, \mathrm{d}t + \sigma_{T-t}\, \mathrm{d}B_t\,. \tag{3.3.6}$$

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 $\operatorname{law}^{\mathbf{P}}(X_t) = \pi_{t|T}(\cdot \mid x_T)$. Therefore, from (3.3.3) and writing $\pi_{t|T} := \pi_{t|T}(\cdot \mid x_T)$ to lighten the notation, we deduce that

$$\partial_t \pi_{t|T} = \frac{1}{2} \langle \sigma_t \sigma_t^\mathsf{T}, \nabla^2 \pi_{t|T} \rangle - \operatorname{div}\Bigl(\pi_{t|T}\, \bigl(b_t + \sigma_t \sigma_t^\mathsf{T}\, \nabla \log \frac{\mathrm{d}P_{t,T}}{\mathrm{d}\pi_T}\bigr)\Bigr)\,.$$

The time reversal $\pi_{t|T}^\leftarrow := \pi_{T-t|T}$ satisfies

$$\partial_t \pi_{t|T}^\leftarrow = -\frac{1}{2} \langle \sigma_{T-t} \sigma_{T-t}^\mathsf{T}, \nabla^2 \pi_{t|T}^\leftarrow \rangle + \operatorname{div}\Bigl(\pi_{t|T}^\leftarrow\, \bigl(b_{T-t} + \sigma_{T-t} \sigma_{T-t}^\mathsf{T}\, \nabla \log \frac{\mathrm{d}P_{T-t,T}}{\mathrm{d}\pi_T}\bigr)\Bigr)\,. \tag{3.3.7}$$

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

$$\nabla \log \frac{\pi_{T-t|T}(\cdot \mid x_T)}{\pi_{T-t}} = \nabla \log \pi_{t|T}^\leftarrow - \nabla \log \pi_t^\leftarrow\,.$$

Substituting this into (3.3.7) and applying the logarithmic derivative trick,

$$\partial_t \pi_{t|T}^\leftarrow = \frac{1}{2} \langle \sigma_{T-t} \sigma_{T-t}^\mathsf{T}, \nabla^2 \pi_{t|T}^\leftarrow \rangle + \operatorname{div}\bigl(\pi_{t|T}^\leftarrow\, (b_{T-t} - \sigma_{T-t} \sigma_{T-t}^\mathsf{T}\, \nabla \log \pi_t^\leftarrow)\bigr)\,. \tag{3.3.8}$$

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 $X^{\leftarrow}_t \sim \pi^{\leftarrow}_{t|T}$.

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 Föllmer (1985). 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, $\gamma := \mathsf{normal}(0, I_d)$. 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

$$\mathrm{d}X_t = \nabla \log P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\,\mathrm{d}t + \mathrm{d}\tilde{B}_t\,, \qquad X_0 = 0\,,$$

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

$$\mathsf{KL}(\mathbf{P} \,\|\, \mathbf{Q}) = \mathbb{E}^{\mathbf{P}} \log \frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_1) = \mathbb{E}_\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\gamma} = \mathsf{KL}(\mu \,\|\, \gamma)\,. \tag{3.4.1}$$

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 1.5.6) we have $\mathsf{KL}(\hat{\mathbf{P}} \,\|\, \mathbf{Q}) \geqslant \mathsf{KL}(\mu \,\|\, \gamma)$. 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 3.2.8), this entropy optimality property becomes

$$\mathsf{KL}(\mathbf{P} \,\|\, \mathbf{Q}) = \frac{1}{2}\,\mathbb{E}^{\mathbf{P}} \int_0^1 \Big\|\nabla \log P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\Big\|^2\,\mathrm{d}t \leqslant \frac{1}{2}\,\mathbb{E}^{\hat{\mathbf{P}}} \int_0^1 \|\hat{b}_t(X_t)\|^2\,\mathrm{d}t = \mathsf{KL}(\hat{\mathbf{P}} \,\|\, \mathbf{Q})\,. \tag{3.4.2}$$

Thus, 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 (2013), although he attributes the idea earlier, e.g., to Borell (2000). Here, we demonstrate its power to establish the Gaussian log-Sobolev and T_2 inequalities, and refer to Lehec (2013) and subsequent literature for further applications.

Transport inequality.

Under **P**, we know that $X_1 \sim \mu$ and $\tilde{B}_1 \sim \gamma$. Hence,

$$\begin{aligned} W_2^2(\mu,\gamma) \leqslant \mathbb{E}^{\mathbf{P}}[\|X_1 - \tilde{B}_1\|^2] &= \mathbb{E}^{\mathbf{P}}\Big[\Big\|\int_0^1 \nabla \log P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\,\mathrm{d}t\Big\|^2\Big] \\ &\leqslant \mathbb{E}^{\mathbf{P}}\int_0^1 \Big\|\nabla \log P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\Big\|^2\,\mathrm{d}t = 2\,\mathsf{KL}(\mu \,\|\, \gamma)\,, \end{aligned}$$

where we used (3.4.1) and (3.4.2) in the last line.

Log-Sobolev inequality.

Let $h_t \coloneqq P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}$ and recall from the construction of the Doob transform that $(h_t(X_t))_{t\in[0,1]}$ is a **Q**-martingale. We claim that $(\nabla \log h_t(X_t))_{t\in[0,1]}$ is a **P**-martingale. To prove this, let $s \leqslant t$ and let $A_s \in \mathscr{F}_s$. Then,

$$\begin{aligned} \mathbb{E}^{\mathbf{P}_t}[\nabla \log h_t(X_t)\,\mathbb{1}_{A_s}] = \mathbb{E}^{\mathbf{P}_t}\Big[\frac{\nabla h_t(X_t)}{h_t(X_t)}\,\mathbb{1}_{A_s}\Big] &= \mathbb{E}^{\mathbf{Q}_t}[\nabla h_t(X_t)\,\mathbb{1}_{A_s}] = \mathbb{E}^{\mathbf{Q}_t}[\nabla P_{1-t}h_1(X_t)\,\mathbb{1}_{A_s}] \\ &= \mathbb{E}^{\mathbf{Q}_t}[P_{1-t}\nabla h_1(X_t)\,\mathbb{1}_{A_s}] = \mathbb{E}^{\mathbf{Q}_s}\big[\mathbb{E}[P_{1-t}\nabla h_1(X_t) \mid \mathscr{F}_s]\,\mathbb{1}_{A_s}\big] \\ &= \mathbb{E}^{\mathbf{Q}_s}[P_{1-s}\nabla h_1(X_s)\,\mathbb{1}_{A_s}]\,. \end{aligned}$$

Applying this equality for $s = t$ yields $\mathbb{E}^{\mathbf{P}_t}[\nabla \log h_t(X_t)\,\mathbb{1}_{A_s}] = \mathbb{E}^{\mathbf{P}_s}[\nabla \log h_s(X_s)\,\mathbb{1}_{A_s}]$, or

$$\mathbb{E}^{\mathbf{P}}[\nabla \log h_t(X_t) \mid \mathscr{F}_s] = \nabla \log h_s(X_s)\,.$$

In particular, $t \mapsto \|\nabla \log h_t(X_t)\|^2$ is a **P**-submartingale.

Now, using the equality for the KL divergence and the submartingale property,

$$\mathsf{KL}(\mu \,\|\, \gamma) = \frac{1}{2}\,\mathbb{E}^{\mathbf{P}}\int_0^1 \Big\|\nabla \log P_{1-t}\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_t)\Big\|^2\,\mathrm{d}t \leqslant \frac{1}{2}\,\mathbb{E}^{\mathbf{P}}\Big[\Big\|\nabla \log\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}(X_1)\Big\|^2\Big] = \frac{1}{2}\,\mathsf{FI}(\mu \,\|\, \gamma)\,.$$

3.4.2 Connection to Stochastic Localization

In this section, we relate the Föllmer process to Eldan's **stochastic localization** scheme (introduced in Eldan (2013)), 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 Klartag and Putterman (2021); El Alaoui et al. (2022) 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 (Kannan et al., 1995), but instead seeks to produce a *random* measure-valued process $(p_t)_{t\geqslant 0}$. This process is such that $p_0 = \mu$, $p_\infty = \delta_X$ for some random variable X, and $(p_t)_{t\geqslant 0}$ is a martingale. The last property implies that $X \sim \mu$, and indeed, $\mathbb{E}\,p_t = \mu$ for all $t \geqslant 0$. 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 $\theta_t \coloneqq tX + B_t$, where as usual $(B_t)_{t\geqslant 0}$ 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 $(p_t)_{t\geqslant 0}$ 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 $(\breve{B}_t)_{t\geqslant 0}$ 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 $p_t = \operatorname{law}(X \mid \breve{\theta}_t) = \rho_{0|t^{-1}}(\cdot \mid \theta_t/t)$. Thus, we have the explicit expression

$$p_t(x) \propto \mu(x) \exp\Bigl(-\frac{t\,\|x - \theta_t/t\|^2}{2}\Bigr)\,.$$

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 Klartag and Putterman (2021). Then, a calculation starting from

$$\mathrm{d}\log p_t(x) = -\mathrm{d}\Bigl(\frac{t\,\|x - \theta_t/t\|^2}{2}\Bigr) - \mathrm{d}\log \int \exp\Bigl(-\frac{t\,\|y - \theta_t/t\|^2}{2}\Bigr)\,\mu(\mathrm{d}y)$$

eventually yields (Exercise 3.6)

$$\mathrm{d}p_t(x) = p_t(x)\,\langle x - a_t, \mathrm{d}B_t\rangle\,, \tag{3.4.3}$$

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 3.7 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:

$$\underset{\mathbf{P}\in\mathcal{P}(C([0,1]))}{\text{minimize}} \quad \mathsf{KL}(\mathbf{P} \,\|\, \mathbf{W}^{\mu,\varepsilon}) \qquad \text{such that} \qquad (X_0)_\#\mathbf{P} = \mu\,,\ (X_1)_\#\mathbf{P} = \nu\,.$$

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

$$\mathsf{KL}(\mathbf{P} \,\|\, \mathbf{W}^{\mu,\varepsilon}) = \mathsf{KL}\big(\mathrm{law}_{\mathbf{P}}(X_0, X_1) \,\big\|\, \mathrm{law}_{\mathbf{W}^{\mu,\varepsilon}}(X_0, X_1)\big) + \mathbb{E}^{\mathbf{P}}\,\mathsf{KL}(\mathbf{P}^{|X_0,X_1} \,\|\, \mathbf{W}^{\mu,\varepsilon|X_0,X_1})\,,$$

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 $\gamma := \mathrm{law}_{\mathbf{P}}(X_0, X_1)$,

$$\begin{aligned}\mathsf{KL}(\gamma \,\|\, \eta) &= \frac{1}{2\varepsilon}\int \|x - y\|^2\,\gamma(\mathrm{d}x, \mathrm{d}y) + \int \log\frac{\gamma(x, y)}{\mu(x)}\,\gamma(\mathrm{d}x, \mathrm{d}y) + \text{const.}\\ &= \frac{1}{2\varepsilon}\int \|x - y\|^2\,\gamma(\mathrm{d}x, \mathrm{d}y) + \mathsf{KL}(\gamma \,\|\, \mu \otimes \nu) + \text{const.}\end{aligned}$$

where we used the fact that $\int \log\nu(y)\,\gamma(\mathrm{d}x, \mathrm{d}y) = \int \log\nu\,\mathrm{d}\nu$ 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

$$\mathcal{T}_{c,\varepsilon}(\mu, \nu) := \inf_{\gamma\in\mathcal{C}(\mu,\nu)}\Big[\int c(x, y)\,\gamma(\mathrm{d}x, \mathrm{d}y) + \varepsilon\,\mathsf{KL}(\gamma \,\|\, \mu \otimes \nu)\Big]\,.$$

Entropic optimal transport was introduced to speed up computation of optimal transport costs in Cuturi (2013); 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*

$$\sup_{(f,g)\in L^1(\mu)\times L^1(\nu)} \Big\{\int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu - \varepsilon \iint \exp\Big(\frac{f \oplus g - c}{\varepsilon}\Big)\,\mathrm{d}(\mu\otimes\nu) + \varepsilon\Big\}$$

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 in (1.3.7), we have replaced the "hard" constraint of $f \oplus g \leqslant c$ with the "soft" constraint of adding the Lagrangian 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 $\varphi_\varepsilon \to \varphi$ and $\psi_\varepsilon \to \psi$ of the entropic potentials to their unregularized counterparts as $\varepsilon \searrow 0$ (see Nutz and Wiesel, 2022). The condition that γ_ε has marginals μ and ν yields the coupled equations

$$\begin{aligned}\varphi_\varepsilon(x) &= \varepsilon \log \int \exp\Big(\frac{\langle x, y\rangle - \psi_\varepsilon(y)}{\varepsilon}\Big)\,\nu(\mathrm{d}y)\,,\\ \psi_\varepsilon(y) &= \varepsilon \log \int \exp\Big(\frac{\langle x, y\rangle - \varphi_\varepsilon(x)}{\varepsilon}\Big)\,\mu(\mathrm{d}x)\,.\end{aligned} \tag{3.5.4}$$

From these expressions, one can prove the following lemma (see Exercise 3.10).

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 (Chewi and Pooladian (2023))**.** *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 Chewi and Pooladian, 2023):

Corollary 3.5.7 (Caffarelli's contraction theorem, Caffarelli (2000))**.** *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*

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

Proof Integration by parts and $\mathbb{E}_\pi \nabla V = 0$ yield

$$\mathbb{E}_\pi \nabla f = \int \nabla f\,\mathrm{d}\pi = \int \left(f\,\nabla\log\frac{1}{\pi}\right)\mathrm{d}\pi = \mathbb{E}_\pi[(f - \mathbb{E}_\pi f)\,\nabla V]\,.$$

Therefore,

$$\begin{aligned}\langle \mathbb{E}_\pi \nabla f, (\mathbb{E}_\pi \nabla^2 V)^{-1}\,\mathbb{E}_\pi \nabla f\rangle &= \mathbb{E}_\pi\big[(f - \mathbb{E}_\pi f)\,\langle \nabla V, (\mathbb{E}_\pi \nabla^2 V)^{-1}\,\mathbb{E}_\pi \nabla f\rangle\big] \\ &\leqslant \sqrt{(\operatorname{var}_\pi f)\,\mathbb{E}_\pi\big\langle \mathbb{E}_\pi \nabla f, (\mathbb{E}_\pi \nabla^2 V)^{-1}\,(\nabla V)^{\otimes 2}\,(\mathbb{E}_\pi \nabla^2 V)^{-1}\,\mathbb{E}_\pi \nabla f\big\rangle}\,.\end{aligned}$$

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,*

$$(\mathbb{E}_\pi \nabla^2 V)^{-1} \preceq \operatorname{cov}_\pi \preceq \mathbb{E}_\pi[(\nabla^2 V)^{-1}]\,,$$

where the second bound holds if π *is strictly log-concave.*

Proof The lower and upper bounds follow respectively from the Cramér–Rao inequality (Theorem 3.5.8) and the Brascamp–Lieb inequality (Theorem 2.2.9) by taking test functions $f = \langle e, \cdot \rangle$ for unit vectors $e \in \mathbb{R}^d$. □

Proof of Theorem 3.5.6 In the proof, we write $\mu \propto \exp(-V)$ and $\nu \propto \exp(-W)$. For any point $x \in \mathbb{R}^d$, 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,

$$\nabla^2 \psi_\varepsilon(y) = \varepsilon^{-1} \operatorname{cov}_{\gamma_\varepsilon^{X|Y=y}} \succeq \varepsilon^{-1} \left(\mathbb{E}_{\gamma_\varepsilon^{X|Y=y}}[\varepsilon^{-1} \nabla^2 \varphi_\varepsilon + \nabla^2 V]\right)^{-1} \succeq \left(\mathbb{E}_{\gamma_\varepsilon^{X|Y=y}}[\nabla^2 \varphi_\varepsilon + \varepsilon\beta I]\right)^{-1}.$$

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

$$\lambda_{\max}\big(\nabla^2 \varphi_\varepsilon(x)\big) \leqslant \big((L_\varepsilon + \varepsilon\beta)^{-1} + \varepsilon\alpha\big)^{-1}$$

and hence

$$L_\varepsilon \leqslant \big((L_\varepsilon + \varepsilon\beta)^{-1} + \varepsilon\alpha\big)^{-1}.$$

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 (Caffarelli, 2000).

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 2.4.29 by first establishing it for Gaussians and appealing to Caffarelli contraction. Another approach to constructing Lipschitz transport maps is explored in Exercise 3.8.

Bibliographical Notes

The discussion of quadratic variation and Girsanov's theorem is heavily inspired by the treatment in Le Gall (2016). 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., Folland (1999). See Steele (2001, 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 Steele (2001) is warmly recommended.

The proof of Exercise 3.4 was first given in Arnaudon et al. (2006). See Altschuler and Chewi (2024b) for further discussion.

The time reversal of SDEs was formulated rigorously under a finite entropy condition in Föllmer (1985). See Cattiaux et al. (2023) for an extended discussion. The proof of Exercise 3.5 is taken from Conforti et al. (2025).

Stochastic localization was introduced by R. Eldan in Eldan (2013), and he and other researchers have subsequently used the Föllmer process and stochastic localization to great effect for various applications; see, e.g., Eldan (2015, 2016); Eldan et al. (2017); Lee and Vempala (2017); Eldan (2018); Eldan and Lee (2018); Klartag (2018); Lee and Vempala (2018); Eldan (2020); Eldan et al. (2020); Klartag and Putterman (2021); Eldan et al. (2022); Eldan and Shamir (2022); Eldan (2023); Klartag (2023). Algorithmic applications were developed in El Alaoui et al. (2022); Montanari (2023) and are closely related to the proximal sampler in Chapter 8. These ideas can be placed in the more general framework of localization schemes (Chen and Eldan, 2022), and can be interpreted in terms of renormalization and the Polchinski flow (Bauerschmidt et al., 2024). See also Kook and Vempala (2025); Shi et al. (2025) for recent surveys.

The Kim–Milman map from Exercise 3.8 was introduced in Kim and Milman (2012). It has been used to establish functional inequalities in subsequent works, e.g., Mikulincer and Shenfeld (2023); Brigati and Pedrotti (2024); Chewi et al. (2024b); Fathi et al. (2024). See Exercise 3.9 and the closely related approach of the Brownian transport map (Mikulincer and Shenfeld, 2024). In Exercise 3.9, it is shown that a log-concave distribution over a set of diameter R has log-Sobolev constant $O(R^2)$; if in addition the distribution is assumed to be isotropic, then the estimate improves to $O(R)$, which is tight (Lee and Vempala, 2018).

See Chewi (2024, Chapter 4) for background on entropic optimal transport, and Chen et al. (2021b) for an introduction to the Schrödinger bridge problem. An entropic optimal transport approach to Caffarelli's contraction theorem was first given in Fathi et al. (2020), and the proof in Section 3.5.2 was generalized in Gozlan and Sylvestre (2025). Entropic optimal transport was also used to prove other functional inequalities in Ledoux (2018); Gentil et al. (2020). In the Gaussian case, closed-form expressions were computed for the entropic optimal transport cost in Janati et al. (2020); Mallasto et al. (2022), generalizing Exercise 3.11, and for the Schrödinger bridge in Bunne et al. (2023).

Exercises

Change of Measure in Path Space

⊵ **Exercise 3.1** (Sobolev embedding)

Let $h \in \mathcal{H}$, where $\mathcal{H}$ is the Cameron–Martin space defined in (3.2.2). Prove that h is $1/2$-Hölder continuous. *Hint*: Use the fundamental theorem of calculus and Cauchy–Schwarz.

This is quite suggestive, since Brownian paths are s-Hölder continuous for every $s < 1/2$.

⊵ **Exercise 3.2** (Proof of Girsanov's theorem)
Let $0 = t_0 < t_1 < \cdots < t_n \leqslant T$ and $\theta_1, \dots, \theta_n \in \mathbb{R}^d$. Let b, B, and $\tilde B$ be as in Theorem 3.2.8. Using the formula for the Radon–Nikodym derivative of $\mathbf{P}$ w.r.t. $\mathbf{W}$, compute $\mathbb{E}^{\mathbf{P}} \exp(\mathbf{i} \sum_{i=1}^n \langle \theta_i, \tilde B_{t_i} - \tilde B_{t_{i-1}} \rangle)$, where $\mathbf{i} := \sqrt{-1}$, and deduce that $\tilde B$ is a $\mathbf{P}$-Brownian motion.

⊵ **Exercise 3.3** (Girsanov's theorem in discrete time)
To gain intuition for Girsanov's theorem, consider the following discrete approximations to two SDEs:

$$\begin{aligned} X_{n+1} &:= X_n - h\, b_n^X(X_n) + \sqrt{h}\, \xi_n\,, \\ Y_{n+1} &:= Y_n - h\, b_n^Y(Y_n) + \sqrt{h}\, \xi_n\,, \end{aligned}$$

where $X_0 = Y_0 = x$ and $(\xi_n)_{n \in \mathbb{N}}$ is a sequence of i.i.d. $\mathsf{normal}(0, I_d)$ variables. Let $\mathbf{P}$, $\mathbf{Q}$ denote the joint distributions of $(X_0, X_1, \dots, X_N)$ and $(Y_0, Y_1, \dots, Y_N)$ respectively.

1 Write down an expression for the Radon–Nikodym derivative $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}$. Let $h \searrow 0$ and $N \to \infty$ such that $Nh \to T$, and compare the result with Girsanov's theorem (Theorem 3.2.8).
2 Write down an expression for $\mathsf{KL}(\mathbf{P} \parallel \mathbf{Q})$. Ensure that it makes sense (i.e., it should be manifestly non-negative).

⊵ **Exercise 3.4** (Harnack inequalities via shifted Girsanov)
In this exercise, we give another proof of the Harnack inequalities established in Exercise 2.17. Let V be α-convex for some $\alpha \in \mathbb{R}$, and let $(P_t)_{t \geqslant 0}$ denote the Langevin semigroup with potential V.

1 Let $\mathbf{P}'$ be the path measure under which $(B'_t)_{t \geqslant 0}$ is a standard Brownian motion, and consider the coupled system of SDEs

$$\begin{aligned} \mathrm{d}X_t &= \{-\nabla V(X_t) + \eta_t\,(Y_t - X_t)\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_t\,, & X_0 &= x\,, \\ \mathrm{d}Y_t &= -\nabla V(Y_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_t\,, & Y_0 &= y\,, \end{aligned}$$

where $(\eta_t)_{t \in [0,T]}$ is a deterministic non-negative process to be chosen later. Assume that we can choose $(\eta_t)_{t \in [0,T]}$ such that $X_T = Y_T$. Finally, let $\mathbf{P}$ be the path measure under which $(B_t)_{t \geqslant 0}$ is a standard Brownian motion, where

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,,$$

that is, $\mathrm{d}B_t = \mathrm{d}B'_t + \frac{\eta_t}{\sqrt{2}}\,(X_t - Y_t)\,\mathrm{d}t$. Use the data-processing inequality (Theorem 1.5.6) and Girsanov's theorem (Theorem 3.2.8) to justify the inequality

$$\mathsf{KL}(\delta_y P_T \parallel \delta_x P_T) \leqslant \mathsf{KL}(\mathbf{P}' \parallel \mathbf{P}) = \frac{1}{4}\,\mathbb{E}^{\mathbf{P}'} \int_0^T \eta_t^2\, \|X_t - Y_t\|^2\,\mathrm{d}t\,.$$

2 Use Itô's formula to show that

$$\|X_t - Y_t\|^2 \leqslant \exp\Bigl(-2\alpha t - 2\int_0^t \eta_s\,\mathrm{d}s\Bigr)\,\|x - y\|^2\,.$$

In particular, $X_T = Y_T$ provided $\int_0^T \eta_t\,\mathrm{d}t = \infty$. Substitute this into the inequality above and optimize over the choice of $(\eta_t)_{t \in [0,T]}$ in order to establish (2.E.8).
3 Can you extend the argument above in order to establish (2.E.10) as well? *Hint*: Use the fact that the exponential martingale (3.2.7) is a supermartingale.

Doob's Transform

⊵ **Exercise 3.5** (Fisher information bound via time reversal)
Let $(X_t)_{t\geqslant 0}$ denote the Langevin diffusion, let $\mu_t := \operatorname{law}(X_t)$, and let $(X_t^{\leftarrow})_{t\in[0,T]}$ denote the time reversal on $[0,T]$.

1 Show that

$$\mathrm{d}X_t^{\leftarrow} = \{\nabla V(X_t^{\leftarrow}) + 2\,\nabla\log\mu_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,.$$

2 Using Itô's formula and the Fokker–Planck equation, show that

$$\mathrm{d}\nabla\log\frac{\mu_{T-t}}{\pi}(X_t^{\leftarrow}) = \nabla^2 V(X_t^{\leftarrow})\,\nabla\log\frac{\mu_{T-t}}{\pi}(X_t^{\leftarrow})\,\mathrm{d}t + \sqrt{2}\,\nabla^2\log\frac{\mu_{T-t}}{\pi}(X_t^{\leftarrow})\,\mathrm{d}B_t\,.$$

Note that while we would naïvely expect Itô's formula to produce a third derivative, there is a miraculous cancellation, due to the time reversal, which removes this term!

3 Using Itô's formula again, show that if $\nabla^2 V \succeq \alpha I_d$ for some $\alpha\in\mathbb{R}$, then

$$\mathsf{FI}(\mu_T \,\|\, \pi) \leqslant \exp(-2\alpha T)\,\mathsf{FI}(\mu_0 \,\|\, \pi)\,.$$

Note that this provides a stochastic calculus proof of (2.2.16).

Föllmer Drift

⊵ **Exercise 3.6** (Derivation of stochastic localization)
Derive the evolution equation (3.4.3) for stochastic localization.

⊵ **Exercise 3.7** (Föllmer and stochastic localization)
Let $\mathsf{BB}_t := (1-t)\,B_{t/(1-t)}$. By solving the SDE that we derived for Brownian bridge in Example 3.3.4 (with $x_1 = 0$), show that BB is indeed a Brownian bridge. Then, verify that $F_t = (1-t)\,\theta_{t/(1-t)}$.

⊵ **Exercise 3.8** (Kim–Milman map)
In this exercise and the next, we construct a Lipschitz transport map from the standard Gaussian $\gamma := \mathsf{normal}(0, I_d)$ to μ, under appropriate conditions on μ. By Proposition 2.3.3, this implies functional inequalities for μ.

1 For $t \geqslant 0$, let μ_t denote the law of the Ornstein–Uhlenbeck process at time t, started at $\mu_0 := \mu$. Let $F_t : \mathbb{R}^d \to \mathbb{R}^d$ denote the flow map at time t, corresponding to the following ODE:

$$\partial_t F_t(x) = -\nabla\log\frac{\mu_t}{\gamma}(F_t(x))\,, \qquad F_0 = \mathrm{id}\,.$$

Argue that $(F_t)_\# \mu = \mu_t$. *Hint*: Recall the ODE interpretation of the Wasserstein gradient flow, e.g., Exercise 1.13.

2 Let $T_t := F_t^{-1}$, so that $(T_t)_\# \mu_t = \mu$. Show that T_t is Lipschitz with

$$\|T_t\|_{\mathrm{Lip}} \leqslant \exp\int_0^t \lambda_{\max}\bigl(\nabla^2\log\frac{\mu_s}{\gamma}\bigr)\,\mathrm{d}s\,.$$

Thus, we expect that $(T_\infty)_\#\gamma = \mu$, with Lipschitz constant bounded by $\exp\int_0^\infty \lambda_{\max}(\nabla^2\log(\mu_t/\gamma))\,\mathrm{d}t$. (This requires justification, but we accept it as a fact here.)

3 Establish the identity

$$\nabla^2 \log \frac{\mu_t}{\gamma}(y) = \frac{1}{\exp(2t) - 1} \Big(\frac{\operatorname{cov}_{\nu_{1-\exp(-2t),\, \exp(-t)\, y}}}{1 - \exp(-2t)} - I_d\Big)\,,$$

where for each $\tau > 0$ and $y \in \mathbb{R}^d$ we define the probability measure $\nu_{\tau,y}$ via

$$\nu_{\tau,y}(\mathrm{d}x) \propto \exp\Big(-\frac{\|x-y\|^2}{2\tau} + \frac{\|x\|^2}{2}\Big)\, \mu(\mathrm{d}x)\,.$$

4 Suppose that μ is α-strongly log-concave, $\alpha > 1$. (We may assume that the strong log-concavity parameter is sufficiently large, since we can rescale the measure.) Use the identity from the preceding part, together with Exercise 2.11, to compute the integral, and conclude that there exists a Lipschitz transport map from γ to μ with constant at most $\alpha^{-1/2}$.

⊵ **Exercise 3.9** (Kim–Milman map II)

Here, we obtain further conditions for the existence of a Lipschitz transport map from γ to μ.

1 Show that if $\operatorname{diam}\operatorname{supp}\mu \leqslant R$, then

$$\nabla^2 \log \frac{\mu_t}{\gamma} \preceq \exp(-2t) \Big(\frac{R^2}{(1-\exp(-2t))^2} - \frac{1}{1-\exp(-2t)}\Big)\, I_d\,. \tag{3.E.1}$$

2 Suppose that $\operatorname{diam}\operatorname{supp}\mu \leqslant R$ and $\nabla^2 \log(1/\mu) \succeq \alpha I_d$, where $\alpha \leqslant 0$. Using a similar estimate as the preceding exercise, write down another bound for $\nabla^2 \log(\mu_t/\gamma)$ which makes use of the assumption $\nabla^2 \log(1/\mu) \succeq \alpha I_d$. By taking the minimum of this estimate and (3.E.1), show that there exists a Lipschitz transport map from γ to μ with constant at most $R \exp((1-\alpha R^2)/2)$. In particular, conclude that if $\alpha = 0$, then μ satisfies a log-Sobolev inequality with constant $\mathrm{e}R^2$.

3 Next, suppose that $\mu = \rho * \gamma$, where $\operatorname{diam}\operatorname{supp}\rho \leqslant R$. Prove that

$$\nabla^2 \log \frac{\mu_t}{\gamma} \preceq R^2 \exp(-2t)\, I_d\,,$$

and deduce that there is a Lipschitz transport map from γ to μ with constant $\exp(R^2/2)$. Compare this with Example 2.3.15.

4 Next, suppose that $\mu \propto \exp(-V - W)$, where V is α-strongly convex ($\alpha > 1$), and W is L-Lipschitz. Argue that

$$\|\operatorname{cov}_{\nu_{\tau,y}}\|_{\mathrm{op}} \leqslant \Big(\sqrt{\|\operatorname{cov}_{\tilde\nu_{\tau,y}}\|_{\mathrm{op}}} + W_2(\nu_{\tau,y}, \tilde\nu_{\tau,y})\Big)^2\,,$$

where

$$\tilde\nu_{\tau,y}(x) \propto \exp\Big(-\frac{\|x-y\|^2}{2\tau} + \frac{\|x\|^2}{2} - V(x)\Big)\,,$$

i.e., $\tilde\nu_{t,y}$ omits the Lipschitz perturbation W. Use Corollary 3.5.9 to control the first term, and the log-Sobolev and T_2 inequalities to control the second term, in order to deduce that $\|\operatorname{cov}_{\nu_{\tau,y}}\|_{\mathrm{op}} \leqslant (\sqrt{1/\alpha_\tau} + L/\alpha_\tau)^2$, where $\alpha_\tau := \alpha - 1 + 1/\tau$. Use this to prove that there exists a Lipschitz transport map from γ to μ with constant $\alpha^{-1/2} \exp(2L/\sqrt{\alpha} + L^2/(2\alpha))$, and thereby deduce Proposition 2.3.4.

Schrödinger Bridge

⊵ **Exercise 3.10** (Entropic potentials)
Derive (3.5.4) and use it to prove Lemma 3.5.5 for the entropic potentials.

⊵ **Exercise 3.11** (Entropic optimal transport between Gaussians)
Compute the entropic optimal transport cost between $\mathsf{normal}(0, 1)$ and $\mathsf{normal}(0, \alpha)$ by making a suitable ansatz for the entropic Brenier potentials.

⊵ **Exercise 3.12** (Poincaré constant and log-smoothness)
Suppose that $\pi \propto \exp(-V)$, where π satisfies a Poincaré inequality with constant $1/\alpha$, and $\nabla^2 V \preceq \beta I_d$. Show that $\alpha \leqslant \beta$.

Hint: Use the Poincaré inequality to prove an upper bound on cov_π, and combine it with the lower bound of Corollary 3.5.9.

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

$$\hat{X}_{(n+1)h} := \hat{X}_{nh} - h\,\nabla V(\hat{X}_{nh}) + \sqrt{2}\,(B_{(n+1)h} - B_{nh})\,. \tag{LMC}$$

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 also include "optimization boxes" which provide background and context for the corresponding results in optimization theory. Finally, κ always denotes the ratio β/α.

Before proceeding, we state the following fundamental lemma, which will be used repeatedly.

Lemma 4.0.1 (Basic lemma)**.** *Let* $\pi \propto \exp(-V)$.

1 *If* $\nabla^2 V \succeq \alpha I_d \succ 0$ *and* V *is minimized at* $x_\star$, *then* $\mathbb{E}_\pi[\|\cdot - x_\star\|^2] \leqslant d/\alpha$.
2 *If* $\nabla^2 V \preceq \beta I_d$, *then* $\mathbb{E}_\pi[\|\nabla V\|^2] \leqslant \beta d$.

Proof We use the fact that for the generator $\mathscr{L} = \Delta - \langle \nabla V, \nabla \cdot \rangle$ of the Langevin diffusion, $\mathbb{E}_\pi \mathscr{L} f = 0$ for all test functions $f : \mathbb{R}^d \to \mathbb{R}$.

1 Take $f = \frac{1}{2}\|\cdot - x_\star\|^2$. By strong convexity,

$$0 = \mathbb{E}_\pi \mathscr{L} f = d - \mathbb{E}_\pi \langle \nabla V, \cdot - x_\star \rangle \leqslant d - \alpha\, \mathbb{E}_\pi[\|\cdot - x_\star\|^2]\,.$$

2 Take $f = V$. Since $\nabla^2 V \preceq \beta I_d$, then $\Delta V \leqslant \beta d$, whence

$$0 = \mathbb{E}_\pi \mathscr{L} V = \mathbb{E}_\pi[\Delta V - \|\nabla V\|^2] \leqslant \beta d - \mathbb{E}_\pi[\|\nabla V\|^2]\,. \qquad \square$$

A stronger sub-Gaussian concentration bound for ∇V is given as Lemma 6.2.7.

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 $x_{n+1} = x_n - h\,\nabla V(x_n)$. 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,

$$\begin{aligned}\|x_{n+1} - y\|^2 &= \|x_n - h\,\nabla V(x_n) - y\|^2 \\ &= \|x_n - y\|^2 - 2h\,\langle \nabla V(x_n), x_n - y\rangle + h^2\,\|\nabla V(x_n)\|^2 \\ &\leqslant (1 - \alpha h)\,\|x_n - y\|^2 - 2h\,\{V(x_n) - V(y)\} + h^2\,\|\nabla V(x_n)\|^2\,.\end{aligned}$$

Using the smoothness of V,

$$V(x_{n+1}) - V(x_n) \leqslant \langle \nabla V(x_n), x_{n+1} - x_n\rangle + \frac{\beta}{2}\,\|x_{n+1} - x_n\|^2 = -h\,\big(1 - \frac{\beta h}{2}\big)\,\|\nabla V(x_n)\|^2\,.$$

For any $h \leqslant 1/\beta$, it yields $\|\nabla V(x_n)\|^2 \leqslant \frac{2}{h}\,\{V(x_n) - V(x_{n+1})\}$, and hence

$$\|x_{n+1} - y\|^2 \leqslant (1 - \alpha h)\,\|x_n - y\|^2 - 2h\,\{V(x_{n+1}) - V(y)\}\,.$$

In particular, for $y = x_\star$, it yields $\|x_{n+1} - x_\star\|^2 \leqslant (1 - \alpha h)\,\|x_n - x_\star\|^2$. Choosing $h = 1/\beta$, one obtains $\|x_N - x_\star\| \leqslant \varepsilon$ in $N = O(\kappa \log(\|x_0 - x_\star\|/\varepsilon))$ iterations, where $\kappa := \beta/\alpha$.

We now consider the corresponding result for LMC.

Theorem 4.1.2 (Coupling analysis of LMC)**.** *For $n \in \mathbb{N}$, let $\hat\mu_{nh}$ denote the law of the n-th iterate of LMC with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, provided $h \ll 1/\beta\kappa$, for all $N \in \mathbb{N}$,*

$$W_2(\hat\mu_{Nh}, \pi) \leqslant \exp\Big(-\frac{\alpha N h}{2}\Big)\,W_2(\mu_0, \pi) + O\Big(\frac{\beta d^{1/2} h^{1/2}}{\alpha}\Big)\,. \tag{4.1.3}$$

If we set $h \asymp \varepsilon^2/\beta\kappa d$, then for any $\varepsilon \in [0, \sqrt{d}]$ we obtain $\sqrt{\alpha}\,W_2(\hat\mu_{Nh}, \pi) \leqslant \varepsilon$ after

$$N = O\Big(\frac{\kappa^2 d}{\varepsilon^2}\log\frac{\alpha W_2^2(\mu_0, \pi)}{\varepsilon^2}\Big) \qquad \textit{iterations}\,.$$

***Remark* 4.1.4.** We pause to make a few comments about the assumptions and result.

1 Typically we assume that we have access to the mode $x_\star$ of π, since the complexity of finding the minimizer of V via convex optimization is typically less than the complexity of sampling. In this case, we can initialize at $\mu_0 = \delta_{x_\star}$, which satisfies $\alpha W_2^2(\mu_0, \pi) \leqslant d$ by Lemma 4.0.1.

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 $s_\lambda : x \mapsto \lambda x$, then information divergences such as KL satisfy $\mathsf{KL}((s_\lambda)_\# \mu \,\|\, (s_\lambda)_\# \pi) = \mathsf{KL}(\mu \,\|\, \pi)$, by the data-processing inequality (Theorem 1.5.6). 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 $\sqrt{\alpha}\, W_2(\cdot, \pi) \leqslant \sqrt{2\, \mathsf{KL}(\cdot \,\|\, \pi)}$. 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 $\widetilde{O}(\kappa d/\varepsilon^2)$.

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.

Proof of Theorem 4.1.2 **1. One-step discretization bound.** Suppose that the Langevin diffusion and the LMC algorithm are both initialized at the same measure $\hat{\mu}_0$. We will first bound the discretization error $W_2^2(\hat{\mu}_h, \pi_h)$ in one step.

We couple the two processes by taking $\hat{X}_0 = X_0$ and using the same Brownian motion:

$$\begin{aligned}\hat{X}_h &= X_0 - h\, \nabla V(X_0) + \sqrt{2}\, B_h\,,\\ X_h &= X_0 - \int_0^h \nabla V(X_t)\, \mathrm{d}t + \sqrt{2}\, B_h\,.\end{aligned}$$

Then,

$$\begin{aligned}W_2^2(\hat{\mu}_h, \pi_h) &\leqslant \mathbb{E}[\|\hat{X}_h - X_h\|^2] \leqslant \mathbb{E}\Big[\Big\|\int_0^h \nabla V(X_t)\, \mathrm{d}t - h\, \nabla V(X_0)\Big\|^2\Big]\\ &\leqslant h \int_0^h \mathbb{E}[\|\nabla V(X_t) - \nabla V(X_0)\|^2]\, \mathrm{d}t\,.\end{aligned}$$

Therefore, we just have to bound the movement $\|X_t - X_0\| = \|-\int_0^t \nabla V(X_s)\, \mathrm{d}s + \sqrt{2}\, B_t\|$ of the Langevin diffusion in time t. Roughly, we expect $\|\int_0^t \nabla V(X_s)\, \mathrm{d}s\| = O(\sqrt{d}\, t)$ 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 4.1.8. For $t \leqslant 1/3\beta$, it yields the bound

$$\mathbb{E}[\|X_t - X_0\|^2] \leqslant 8t^2\, \mathbb{E}[\|\nabla V(X_0)\|^2] + 8dt$$

and hence

$$W_2^2(\hat{\mu}_h, \pi_h) \leqslant \beta^2 h \int_0^h \mathbb{E}[\|X_t - X_0\|^2]\, \mathrm{d}t \leqslant 3\beta^2 h^4\, \mathbb{E}[\|\nabla V(X_0)\|^2] + 4\beta^2 d h^3\,.$$

2. Multi-step discretization bound. We produce a coupling of $\hat{\mu}_{(n+1)h}$ and π as follows. First, let

$\hat{X}_{nh} \sim \hat{\mu}_{nh}$ and $X_{nh} \sim \pi$ be optimally coupled. Using the same Brownian motion, we set

$$\begin{aligned} \hat{X}_{(n+1)h} &:= \hat{X}_{nh} - h\,\nabla V(\hat{X}_{nh}) + \sqrt{2}\,(B_{(n+1)h} - B_{nh})\,, \\ X_t &:= X_{nh} - \int_{nh}^{t} \nabla V(X_s)\,\mathrm{d}s + \sqrt{2}\,(B_t - B_{nh})\,, \qquad \text{for } t \in [nh, (n+1)h]\,. \end{aligned}$$

Clearly $\hat{X}_{(n+1)h} \sim \hat{\mu}_{(n+1)h}$; also, since π is stationary for the Langevin diffusion, then $X_{(n+1)h} \sim \pi$. We also introduce an auxiliary process: let

$$\bar{X}_t := \hat{X}_{nh} - \int_{nh}^{t} \nabla V(\bar{X}_s)\,\mathrm{d}s + \sqrt{2}\,(B_t - B_{nh}) \qquad \text{for } t \in [nh, (n+1)h]$$

denote the Langevin diffusion started at $\hat{X}_{kh}$. We bound

$$\begin{aligned} W_2(\hat{\mu}_{(n+1)h}, \pi) &\leqslant \sqrt{\mathbb{E}[\|\hat{X}_{(n+1)h} - X_{(n+1)h}\|^2]} \\ &\leqslant \sqrt{\mathbb{E}[\|\bar{X}_{(n+1)h} - X_{(n+1)h}\|^2]} + \sqrt{\mathbb{E}[\|\hat{X}_{(n+1)h} - \bar{X}_{(n+1)h}\|^2]}\,. \end{aligned}$$

Now we examine the two terms. In the first term, both $\bar{X}$ and X 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 1.4.11):

$$\mathbb{E}[\|\bar{X}_{(n+1)h} - X_{(n+1)h}\|^2] \leqslant \exp(-2\alpha h)\,\mathbb{E}[\|\hat{X}_{nh} - X_{kh}\|^2] = \exp(-2\alpha h)\,W_2^2(\hat{\mu}_{nh}, \pi)\,.$$

For the second term, $\hat{X}$ is the LMC algorithm and $\bar{X}$ is the continuous-time Langevin diffusion, both initialized at the same distribution $\hat{\mu}_{nh}$. Hence, we can apply our one-step discretization bound from before and deduce that

$$\begin{aligned} \mathbb{E}[\|\hat{X}_{(n+1)h} - \bar{X}_{(n+1)h}\|^2] &\leqslant 3\beta^2 h^4\,\mathbb{E}[\|\nabla V(\hat{X}_{nh})\|^2] + 4\beta^2 dh^3 \\ &\lesssim \beta^4 h^4\,\mathbb{E}[\|\hat{X}_{nh} - X_{nh}\|^2] + \beta^2 h^4\,\mathbb{E}_\pi[\|\nabla V\|^2] + \beta^2 dh^3 \\ &\leqslant \beta^4 h^4\,W_2^2(\hat{\mu}_{nh}, \pi) + \beta^3 dh^4 + \beta^2 dh^3\,, \end{aligned}$$

where we used the basic lemma (Lemma 4.0.1). Note that the second term can be dropped since we are assuming $h \ll 1/\beta$.

Combining everything and using $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ for $x, y \geqslant 0$,

$$W_2(\hat{\mu}_{(n+1)h}, \pi) \leqslant \exp(-\alpha h)\,W_2(\hat{\mu}_{nh}, \pi) + O\big(\beta^2 h^2\,W_2(\hat{\mu}_{nh}, \pi) + \beta d^{1/2} h^{3/2}\big)\,.$$

Provided $h \ll 1/\beta\kappa$, we can ensure that $\exp(-\alpha h) + O(\beta^2 h^2) \leqslant \exp(-\alpha h/2)$, therefore absorbing the extra Wasserstein error term. It yields

$$W_2(\hat{\mu}_{(n+1)h}, \pi) \leqslant \exp\big(-\frac{\alpha h}{2}\big)\,W_2(\hat{\mu}_{nh}, \pi) + O(\beta d^{1/2} h^{3/2})\,.$$

After iterating this recursion, it implies

$$W_2(\hat{\mu}_{Nh}, \pi) \leqslant \exp\Big(-\frac{\alpha N h}{2}\Big)\,W_2(\hat{\mu}_0, \pi) + O\Big(\frac{\beta d^{1/2} h^{1/2}}{\alpha}\Big)\,.$$

This finishes the proof. □

***Remark* 4.1.5.** By inspecting the proof, one can see that that the following stronger inequality holds. Let us denote by $\hat{P}^{\mathsf{LMC}}$ 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 $x, y \in \mathbb{R}^d$,

$$W_2\big(\hat{P}^{\mathsf{LMC}}(x,\cdot),\, P(y,\cdot)\big) \leqslant \exp\big(-\frac{\alpha h}{2}\big)\,\|x-y\| + O\big(\beta h^2\,\|\nabla V(y)\| + \beta d^{1/2} h^{3/2}\big)\,. \tag{4.1.6}$$

If we square this inequality and apply Young's inequality, it implies

$$\begin{aligned} W_2^2\big(\hat{P}^{\mathsf{LMC}}(x,\cdot),\, P(y,\cdot)\big) &\leqslant \exp(-\alpha h)\,\|x-y\|^2 + O\big(\beta^2 h^4\,\|\nabla V(y)\|^2 + \beta^2 d h^3\big) \\ &\qquad + O\big(\|x-y\|\,(\beta h^2\,\|\nabla V(y)\| + \beta d^{1/2} h^{3/2})\big) \\ &\leqslant \exp\big(-\frac{\alpha h}{2}\big)\,\|x-y\|^2 + O\Big(\frac{\beta^2 h^3\,\|\nabla V(y)\|^2}{\alpha} + \frac{\beta^2 d h^2}{\alpha}\Big)\,. \end{aligned} \tag{4.1.7}$$

Iterating either (4.1.6) or (4.1.7), together with a coupling argument, implies 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 4.1.8 (Langevin movement bound)**.** *Let $(X_t)_{t\geqslant 0}$ denote the Langevin diffusion and let $(\pi_t)_{t\geqslant 0}$ denote its law. Assume that ∇V is β-Lipschitz. Then, provided that $t \leqslant 1/3\beta$,*

$$\mathbb{E}[\|X_t - X_0\|^2] \leqslant 8t^2\,\mathbb{E}[\|\nabla V(X_0)\|^2] + 8dt\,.$$

Proof By definition,

$$\mathbb{E}[\|X_t - X_0\|^2] = \mathbb{E}\Big[\Big\|-\int_0^t \nabla V(X_s)\,\mathrm{d}s + \sqrt{2}\,B_t\Big\|^2\Big] \leqslant 2t\int_0^t \mathbb{E}[\|\nabla V(X_s)\|^2]\,\mathrm{d}s + 4\,\mathbb{E}[\|B_t\|^2]\,.$$

Using the β-Lipschitzness of ∇V, $\|\nabla V(X_s)\| \leqslant \|\nabla V(X_0)\| + \beta\,\|X_s - X_0\|$. Thus,

$$\mathbb{E}[\|X_t - X_0\|^2] \leqslant 4\beta^2 t\int_0^t \mathbb{E}[\|X_s - X_0\|^2]\,\mathrm{d}s + 4t^2\,\mathbb{E}[\|\nabla V(X_0)\|^2] + 4dt\,.$$

Applying Grönwall's inequality (Lemma 1.2.20), it implies

$$\mathbb{E}[\|X_t - X_0\|^2] \leqslant \{4t^2\,\mathbb{E}[\|\nabla V(X_0)\|^2] + 4dt\}\exp(4\beta^2 t^2)\,.$$

Finally, use the assumption $t \leqslant 1/3\beta$ 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,

$$V(x_{n+1}) - V(x_n) \leqslant -\frac{h}{2}\,\|\nabla V(x_n)\|^2\,,$$

provided that $h \leqslant 1/\beta$ (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–Lojasiewicz (PL) inequality**

$$\|\nabla V(x)\|^2 \geqslant 2\alpha\,\{V(x) - V(x_\star)\} \qquad \text{for all } x \in \mathbb{R}^d\,.$$

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

$$\begin{aligned} V(x_{n+1}) - V(x_\star) &= V(x_{n+1}) - V(x_n) + V(x_n) - V(x_\star) \\ &\leqslant -\frac{h}{2}\,\|\nabla V(x_n)\|^2 + V(x_n) - V(x_\star) \leqslant (1-\alpha h)\,\{V(x_n) - V(x_\star)\}\,. \end{aligned}$$

Setting $h = 1/\beta$, we can achieve $V(x_N) - V(x_\star) \leqslant \varepsilon^2$ in $O(\kappa \log(V(x_0) - V(x_\star)/\varepsilon^2))$ iterations under PL and smoothness.

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 Vempala and Wibisono (2019) (slightly refined using a lemma from Chewi et al., 2024a). 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: we set

$$\hat{X}_t := \hat{X}_{t_-} - (t - t_-)\,\nabla V(\hat{X}_{t_-}) + \sqrt{2}\,(B_t - B_{t_-})\,, \tag{4.2.2}$$

where we have introduced the notation $t_- := \lfloor t/h \rfloor\, h$.

Proposition 4.2.3. *Let $(\hat{\mu}_t)_{t\geqslant 0}$ be the law of the interpolated process (4.2.2). Then,*

$$\partial_t \hat{\mu}_t = \operatorname{div}\Bigl[\hat{\mu}_t\,\Bigl(\nabla \log \frac{\hat{\mu}_t}{\pi} + \mathbb{E}[\nabla V(\hat{X}_{t_-}) - \nabla V(\hat{X}_t) \mid \hat{X}_t = \cdot]\Bigr)\Bigr]\,.$$

Proof Let $\hat{\mu}_{t|t_-}$ denote the law of $\hat{X}_t$ conditioned on $\hat{X}_{t_-}$ at time t_-. Then, $(\hat{\mu}_{t|t_-})_{t\in[t_-,t_-+h]}$ satisfies the Fokker–Planck equation

$$\partial_t \hat{\mu}_{t|t_-}(\cdot \mid \hat{X}_{t_-}) = \Delta \hat{\mu}_{t|t_-}(\cdot \mid \hat{X}_{t_-}) + \operatorname{div}\bigl(\hat{\mu}_{t|t_-}(\cdot \mid \hat{X}_{t_-})\,\nabla V(\hat{X}_{t_-})\bigr)\,.$$

Next, we take the expectation of the above equation; since $\mathbb{E}\,\hat{\mu}_{t|t_-}(\cdot \mid \hat{X}_{t_-}) = \hat{\mu}_t$,

$$\partial_t \hat{\mu}_t = \Delta \hat{\mu}_t + \operatorname{div} \mathbb{E}[\hat{\mu}_{t|t_-}(\cdot \mid \hat{X}_{t_-})\,\nabla V(\hat{X}_{t_-})]\,.$$

Note that $\hat\mu_{t|t_-}(x_t \mid x_{t_-})\,\hat\mu_{t_-}(\mathrm{d}x_{t_-}) = \hat\mu_t(x_t)\,\hat\mu_{t_-|t}(\mathrm{d}x_{t_-} \mid x_t)$. So,

$$\begin{aligned}\mathbb{E}[\hat\mu_{t|t_-}(x_t \mid \hat X_{t_-})\,\nabla V(\hat X_{t_-})] &= \int \hat\mu_{t|t_-}(x_t \mid x_{t_-})\,\nabla V(x_{t_-})\,\hat\mu_{t_-}(\mathrm{d}x_{t_-})\\ &= \hat\mu_t(x_t)\int \nabla V(x_{t_-})\,\hat\mu_{t_-|t}(\mathrm{d}x_{t_-} \mid x_t)\\ &= \hat\mu_t(x_t)\,\mathbb{E}[\nabla V(\hat X_{t_-}) \mid \hat X_t = x_t]\,.\end{aligned}$$

Therefore,

$$\begin{aligned}\partial_t\hat\mu_t &= \Delta\hat\mu_t + \operatorname{div}\bigl(\hat\mu_t\,\mathbb{E}[\nabla V(\hat X_{t_-}) \mid \hat X_t = \cdot]\bigr)\\ &= \operatorname{div}\Bigl(\hat\mu_t\,\nabla\log\frac{\hat\mu_t}{\pi}\Bigr) + \operatorname{div}\bigl(\hat\mu_t\,\mathbb{E}[\nabla V(\hat X_{t_-}) - \nabla V(\hat X_t) \mid \hat X_t = \cdot]\bigr)\,. \qquad \square\end{aligned}$$

Corollary 4.2.4. *Along the law $(\hat\mu_t)_{t\geqslant 0}$ of the interpolated process (4.2.2),*

$$\partial_t\,\mathsf{KL}(\hat\mu_t \,\|\, \pi) \leqslant -\frac{3}{4}\,\mathsf{FI}(\hat\mu_t \,\|\, \pi) + \mathbb{E}[\|\nabla V(\hat X_t) - \nabla V(\hat X_{t_-})\|^2]\,.$$

Recall that the Fisher information is $\mathsf{FI}(\hat\mu \,\|\, \pi) \coloneqq \mathbb{E}_{\hat\mu}[\|\nabla\log(\hat\mu/\pi)\|^2]$ if $\hat\mu$ has a smooth density with respect to π.

Proof Using Proposition 4.2.3,

$$\begin{aligned}\partial_t\,\mathsf{KL}(\hat\mu_t \,\|\, \pi) &= -\,\mathbb{E}_{\hat\mu_t}\Bigl\langle\nabla\log\frac{\hat\mu_t}{\pi},\,\nabla\log\frac{\hat\mu_t}{\pi} + \mathbb{E}[\nabla V(\hat X_{t_-}) - \nabla V(\hat X_t) \mid \hat X_t = \cdot]\Bigr\rangle\\ &= -\,\mathsf{FI}(\hat\mu_t \,\|\, \pi) + \mathbb{E}_{\hat\mu_t}\Bigl\langle\nabla\log\frac{\hat\mu_t}{\pi},\,\mathbb{E}[\nabla V(\hat X_t) - \nabla V(\hat X_{t_-}) \mid \hat X_t = \cdot]\Bigr\rangle\,.\end{aligned}$$

Using Young's inequality,

$$\begin{aligned}&\mathbb{E}_{\hat\mu_t}\Bigl\langle\nabla\log\frac{\hat\mu_t}{\pi},\,\mathbb{E}[\nabla V(\hat X_t) - \nabla V(\hat X_{t_-}) \mid \hat X_t = \cdot]\Bigr\rangle\\ &\qquad\leqslant \frac{1}{4}\,\mathsf{FI}(\hat\mu_t \,\|\, \pi) + \mathbb{E}\bigl[\|\mathbb{E}[\nabla V(\hat X_t) - \nabla V(\hat X_{t_-}) \mid \hat X_t]\|^2\bigr]\\ &\qquad\leqslant \frac{1}{4}\,\mathsf{FI}(\hat\mu_t \,\|\, \pi) + \mathbb{E}[\|\nabla V(\hat X_t) - \nabla V(\hat X_{t_-})\|^2]\,. \qquad \square\end{aligned}$$

In the convergence proof for LMC, we will use the following lemma. It can be seen as an extension of the second part of the basic lemma (Lemma 4.0.1).

Lemma 4.2.5 (Chewi et al. (2024a, Lemma 16)). *Suppose that $\pi \propto \exp(-V)$ where $\nabla^2 V \preceq \beta I_d$. Then, for any probability measure $\mu \ll \pi$,*

$$\mathbb{E}_\mu[\|\nabla V\|^2] \leqslant \mathsf{FI}(\mu \,\|\, \pi) + 2\beta d\,.$$

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 $\Delta V \leqslant \beta d$. Thus, using the fundamental integration by parts identity (Theorem 1.2.14),

$$\begin{aligned}\mathbb{E}_\mu[\|\nabla V\|^2] = \mathbb{E}_\mu[\Delta V - \mathscr{L} V] &\leqslant \beta d + \int (-\mathscr{L} V)\,\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\pi = \beta d + \int \Bigl\langle \nabla V, \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Bigr\rangle\,\mathrm{d}\pi \\ &= \beta d + \int \Bigl\langle \nabla V, \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Bigr\rangle\,\mathrm{d}\mu \\ &\leqslant \beta d + \frac{1}{2}\,\mathbb{E}_\mu[\|\nabla V\|^2] + \frac{1}{2}\,\mathsf{FI}(\mu \,\|\, \pi)\,.\end{aligned}$$

Rearranging the inequality yields the result. ☐

Finally, recall that an LSI implies $\mathsf{KL}(\cdot \,\|\, \pi) \leqslant \frac{C_{\mathsf{LSI}}}{2}\,\mathsf{FI}(\cdot \,\|\, \pi)$. In the following theorem, we note that $\kappa := \beta/\alpha \geqslant 1$ as a consequence of Exercise 3.12.

Theorem 4.2.6 (Vempala and Wibisono (2019))**.** *For $n \in \mathbb{N}$, let $\hat\mu_{nh}$ denote the law of the n-th iterate of* LMC *with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies $C_{\mathsf{LSI}}(\pi) \leqslant 1/\alpha$ and that ∇V is β-Lipschitz. Then, for all $h \leqslant 1/4\beta$, for all $N \in \mathbb{N}$,*

$$\mathsf{KL}(\mu_{Nh} \,\|\, \pi) \leqslant \exp(-\alpha N h)\,\mathsf{KL}(\mu_0 \,\|\, \pi) + O\Bigl(\frac{\beta^2 d h}{\alpha}\Bigr)\,.$$

In particular, if $\kappa := \beta/\alpha$, for all $\varepsilon \in [0, \kappa\sqrt{d}]$ and for $h \asymp \varepsilon^2/\beta\kappa d$, we obtain the guarantee $\sqrt{\mathsf{KL}(\hat\mu_{Nh} \,\|\, \pi)} \leqslant \varepsilon$ after

$$N = O\Bigl(\frac{\kappa^2 d}{\varepsilon^2}\log\frac{\mathsf{KL}(\hat\mu_0 \,\|\, \pi)}{\varepsilon^2}\Bigr) \qquad \textit{iterations}\,.$$

Proof In light of Corollary 4.2.4, we focus our attention on the discretization error term

$$\begin{aligned}\mathbb{E}[\|\nabla V(\hat X_t) - \nabla V(\hat X_{t_-})\|^2] &\leqslant \beta^2\,\mathbb{E}[\|\hat X_t - \hat X_{t_-}\|^2] \\ &= \beta^2 h^2\,\mathbb{E}[\|\nabla V(\hat X_{t_-})\|^2] + 2\beta^2\,\mathbb{E}[\|B_t - B_{t_-}\|^2]\,.\end{aligned}$$

In order to apply Lemma 4.2.5, it is more convenient to have $\mathbb{E}[\|\nabla V(\hat X_t)\|^2]$ instead of $\mathbb{E}[\|\nabla V(\hat X_{t_-})\|^2]$. So, we use

$$\mathbb{E}[\|\nabla V(\hat X_{t_-})\|^2] \leqslant 2\,\mathbb{E}[\|\nabla V(\hat X_t)\|^2] + 2\,\mathbb{E}[\|\nabla V(\hat X_t) - \nabla V(\hat X_{t_-})\|^2]\,.$$

If $h \leqslant 1/2\beta$, we can combine this inequality with the previous one and rearrange to obtain

$$\begin{aligned}\mathbb{E}[\|\nabla V(\hat X_t) - \nabla V(\hat X_{t_-})\|^2] &\leqslant 4\beta^2 h^2\,\mathbb{E}[\|\nabla V(\hat X_t)\|^2] + 4\beta^2\,\mathbb{E}[\|B_t - B_{t_-}\|^2] \\ &= 4\beta^2 h^2\,\mathbb{E}[\|\nabla V(\hat X_t)\|^2] + 4\beta^2 d h\,.\end{aligned}$$

For the first term, we apply Lemma 4.2.5, yielding for $h \leqslant 1/4\beta$

$$4\beta^2 h^2\,\mathbb{E}[\|\nabla V(\hat X_t)\|^2] \leqslant 4\beta^2 h^2\,\mathsf{FI}(\hat\mu_t \,\|\, \pi) + 8\beta^3 d h^2 \leqslant \frac{1}{4}\,\mathsf{FI}(\hat\mu_t \,\|\, \pi) + 2\beta^2 d h\,.$$

Combining with our differential inequality from Corollary 4.2.4 and LSI,

$$\partial_t\,\mathsf{KL}(\hat\mu_t \,\|\, \pi) \leqslant -\frac{1}{2}\,\mathsf{FI}(\hat\mu_t \,\|\, \pi) + 6\beta^2 d h \leqslant -\alpha\,\mathsf{KL}(\hat\mu_t \,\|\, \pi) + 6\beta^2 d h\,.$$

This implies that

$$\partial_t[\exp(\alpha t)\,\mathsf{KL}(\hat\mu_t \,\|\, \pi)] \leqslant 6\beta^2 d h\exp(\alpha t)\,,$$

and upon integration,

$$\mathsf{KL}(\hat{\mu}_{Nh} \parallel \pi) \leqslant \exp(-\alpha N h)\, \mathsf{KL}(\hat{\mu}_0 \parallel \pi) + O\Big(\frac{\beta^2 d h}{\alpha}\Big)\,. \qquad \square$$

Recall from Theorem 2.2.25 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 above proof 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 Durmus et al. (2019), which is inspired by convex optimization. This proof yields the state-of-the-art dependence of LMC on the condition number κ of the target π, and in Chapter 10, we will also use it to analyze a "mirror" variant of LMC.

Let the target be $\pi = \exp(-V)$ (for this proof, we assume that V is normalized so that $\int \exp(-V) = 1$; this just simplifies the notation but changes neither the algorithm nor the analysis). We now view sampling as the composite optimization problem of minimizing the objective

$$\mathsf{KL}(\mu \parallel \pi) := \underbrace{\int V \, \mathrm{d}\mu}_{=:\mathcal{E}(\mu)} + \underbrace{\int \mu \log \mu}_{=:\mathcal{H}(\mu)}\,,$$

where the two terms are the *energy* and the (negative) *entropy*. Accordingly, we break up the iterates of LMC into the steps

$$\hat{X}_{nh}^+ := \hat{X}_{nh} - h\, \nabla V(\hat{X}_{nh})\,,$$
$$\hat{X}_{(n+1)h} := \hat{X}_{nh}^+ + \sqrt{2}\, (B_{(n+1)h} - B_{nh})\,.$$

The first step is simply a deterministic gradient descent update on the function V. If we write $\hat{\mu}_{nh}^+$ for the law of $\hat{X}_{nh}^+$, then in the space of measures one can show that $\hat{\mu}_{nh}^+$ is obtained from $\hat{\mu}_{nh}$ 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 Wibisono (2018) 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 (Exercise 4.8)! 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—it requires care to design one that is asymptotically unbiased. Recall from Exercise 1.15 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 (see Optimization Box 8.0.1), but unfortunately the "backward" step for the entropy cannot be easily implemented.

---

[1] Technically this is only true if the step size h is chosen so that $h\, \|\nabla^2 V\|_{\mathrm{op}} \leqslant 1$. 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.

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,

$$\|x_{n+1} - y\|^2 \leqslant (1 - \alpha h)\, \|x_n - y\|^2 - 2h\, \{V(x_n) - V(y)\} + h^2\, \|\nabla V(x_n)\|^2 \,. \tag{4.3.2}$$

Moreover, for $h \leqslant 1/\beta$,

$$\|x_{n+1} - y\|^2 \leqslant (1 - \alpha h)\, \|x_n - y\|^2 - 2h\, \{V(x_{n+1}) - V(y)\} \,. \tag{4.3.3}$$

Inspired by Ambrosio et al. (2008), 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 $V(x_n) - V(x_\star) \geqslant 0$ to conclude that $\|x_{n+1} - x_\star\|^2 \leqslant (1 - \alpha h)\, \|x_n - x_\star\|^2$, 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

$$V(x_N) - V(x_\star) \leqslant \frac{1}{N} \sum_{n=0}^{N-1} \{V(x_{n+1}) - V(x_\star)\} \leqslant \frac{\beta\, \|x_0 - x_\star\|^2}{2N} \,,$$

where the first inequality follows from the descent lemma.

Smooth case.

In analogy, we aim to prove the following key lemma.

Lemma 4.3.4 (EVI with error for LMC)**.** *Let $\pi = \exp(-V)$ be the target and assume that $0 \preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Let $(\hat\mu_{nh})_{n\in\mathbb{N}}$ be the marginal laws of LMC with step size $h \leqslant 1/\beta$. Then,*

$$2h\, \mathsf{KL}(\hat\mu_{(n+1)h} \,\|\, \pi) \leqslant (1 - \alpha h)\, W_2^2(\hat\mu_{nh}, \pi) - W_2^2(\hat\mu_{(n+1)h}, \pi) + 2\beta d h^2 \,. \tag{4.3.5}$$

From this, we deduce the following results.

Theorem 4.3.6 (Durmus et al. (2019)). *Suppose that* $\pi = \exp(-V)$ *is the target distribution and that* V *satisfies* $0 \preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. *Let* $(\hat\mu_{nh})_{n\in\mathbb{N}}$ *denote the marginal laws of* LMC.

1 *(Weakly convex case) Suppose that* $\alpha = 0$. *For any* $\varepsilon \in [0, \sqrt{d}]$, *if we take step size* $h \asymp \varepsilon^2/\beta d$, *then for the mixture distribution* $\bar\mu_{Nh} := N^{-1}\sum_{n=1}^N \hat\mu_{nh}$ *it holds that* $\sqrt{\mathsf{KL}(\bar\mu_{Nh} \,\|\, \pi)} \leqslant \varepsilon$ *after*

$$N = O\Big(\frac{\beta d\, W_2^2(\hat\mu_0, \pi)}{\varepsilon^4}\Big) \qquad \textit{iterations}\,.$$

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* $h \asymp \varepsilon^2/\beta d$ *we obtain* $\sqrt{\alpha}\, W_2(\hat\mu_{Nh}, \pi) \leqslant \varepsilon$ *and* $\sqrt{\mathsf{KL}(\bar\mu_{Nh,2Nh} \,\|\, \pi)} \leqslant \varepsilon$ *after*

$$N = O\Big(\frac{\kappa d}{\varepsilon^2}\log\frac{\alpha W_2^2(\hat\mu_0, \pi)}{\varepsilon^2}\Big) \qquad \textit{iterations}\,,$$

where $\bar\mu_{Nh,2Nh} := N^{-1}\sum_{n=N+1}^{2N} \hat\mu_{nh}$.

Proof We use Lemma 4.3.4.

1 By summing the inequality (4.3.5) and using the convexity of the KL divergence,

$$\mathsf{KL}(\bar\mu_{Nh} \,\|\, \pi) \leqslant \frac{1}{N}\sum_{n=1}^N \mathsf{KL}(\hat\mu_{nh} \,\|\, \pi) \leqslant \frac{W_2^2(\hat\mu_0, \pi)}{2Nh} + \beta d h\,.$$

The result follows from our choice of h and N.

2 First, we prove the W_2 guarantee. Using the fact that $\mathsf{KL}(\hat\mu_{(n+1)h} \,\|\, \pi) \geqslant 0$ and iterating the inequality (4.3.5) we obtain

$$\begin{aligned} W_2^2(\hat\mu_{Nh}, \pi) &\leqslant (1-\alpha h)^N\, W_2^2(\hat\mu_0, \pi) + 2\beta d h^2 \sum_{n=0}^{N-1} (1-\alpha h)^n \\ &\leqslant \exp(-\alpha N h)\, W_2^2(\hat\mu_0, \pi) + O(\kappa d h)\,. \end{aligned}$$

With our choice of h and N, we obtain $\sqrt{\alpha}\, W_2(\hat\mu_{Nh}, \pi) \leqslant \varepsilon$.

Next, forget about the previous N iterations of LMC and consider $\hat\mu_{Nh}$ to be the new initialization to LMC. Applying the weakly convex result yields the KL guarantee $\sqrt{\mathsf{KL}(\bar\mu_{Nh,2Nh} \,\|\, \pi)} \leqslant \varepsilon$. □

We next turn toward 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. First, let $Z \sim \pi$ be optimally coupled to $\hat X_{nh}$. Then, $\mathcal{E}(\hat\mu_{nh}^+) - \mathcal{E}(\pi) = \mathbb{E}[V(\hat X_{nh}^+) - V(Z)]$. However, since $\hat X_{nh}^+$ is obtained from $\hat X_{nh}$ via a gradient descent step on V, we can apply the EVI (4.3.3) to argue that

$$\begin{aligned} \mathcal{E}(\hat\mu_{nh}^+) - \mathcal{E}(\pi) &\leqslant \frac{1}{2h}\,\mathbb{E}[(1-\alpha h)\,\|\hat X_{nh} - Z\|^2 - \|\hat X_{nh}^+ - Z\|^2] \\ &\leqslant \frac{1}{2h}\,\{(1-\alpha h)\, W_2^2(\hat\mu_{nh}, \pi) - W_2^2(\hat\mu_{nh}^+, \pi)\}\,. \end{aligned} \qquad (4.3.7)$$

2. The flow step does not substantially increase the energy. Next, using the β-smoothness of V,

$$\begin{aligned}\mathcal{E}(\hat\mu_{(n+1)h}) - \mathcal{E}(\hat\mu^+_{nh}) &= \mathbb{E}[V(\hat X_{(n+1)h}) - V(\hat X^+_{nh})] \\ &\leqslant \mathbb{E}\Bigl[\langle \nabla V(\hat X^+_{nh}), \hat X_{(n+1)h} - \hat X^+_{nh}\rangle + \frac{\beta}{2}\,\|\hat X_{(n+1)} - \hat X^+_{nh}\|^2\Bigr] \\ &= \mathbb{E}[\sqrt{2}\,\langle \nabla V(\hat X^+_{nh}), B_{(n+1)h} - B_{nh}\rangle + \beta\,\|B_{(n+1)h} - B_{nh}\|^2] \\ &= \beta d h\,. \end{aligned} \tag{4.3.8}$$

3. The flow step dissipates the entropy. Let $(Q_t)_{t\geqslant 0}$ denote the heat semigroup, i.e., $Q_t f(x) \coloneqq \mathbb{E}\, f(x + \sqrt{2}\, B_t)$, so that $\hat\mu_{(n+1)h} = \hat\mu^+_{nh} Q_h$. Then, since the heat flow is the Wasserstein gradient flow of $\mathcal{H}$, and the Wasserstein gradient of $\mathcal{H}$ is $\nabla_{W_2}\mathcal{H}(\mu) = \nabla \log \mu$, one can show that

$$\partial_t W_2^2(\hat\mu^+_{nh} Q_t, \pi) \leqslant 2\,\mathbb{E}\langle \nabla \log \hat\mu^+_{nh} Q_t(\hat X^+_{nh+t}), Z - \hat X^+_{nh+t}\rangle$$

where $\hat X^+_{nh+t} \sim \hat\mu^+_{nh} Q_t$ and $Z \sim \pi$ are optimally coupled. This follows from the formula for the gradient of the squared Wasserstein distance (Theorem 1.4.5); it may be justified more rigorously using, e.g., Ambrosio et al. (2008, Theorem 10.2.2).

On the other hand, we showed that $\mathcal{H}$ is geodesically convex in (1.4.3), so

$$\mathcal{H}(\pi) - \mathcal{H}(\hat\mu^+_{nh} Q_t) \geqslant \mathbb{E}\langle \nabla \log \hat\mu^+_{nh} Q_t(\hat X^+_{nh+t}), Z - \hat X^+_{nh+t}\rangle\,.$$

Using the fact that $t \mapsto \mathcal{H}(\hat\mu^+_{nh} Q_t)$ is decreasing (which also follows because $t \mapsto \hat\mu^+_{nh} Q_t$ is the gradient flow of $\mathcal{H}$), we then have

$$W_2^2(\hat\mu_{(n+1)h}, \pi) - W_2^2(\hat\mu^+_{nh}, \pi) \leqslant 2h\,\{\mathcal{H}(\pi) - \mathcal{H}(\hat\mu_{(n+1)h})\}\,. \tag{4.3.9}$$

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 (EVI with error for LMC, non-smooth case)**.** *Let $\pi = \exp(-V)$ be the target and assume that V is convex and L-Lipschitz. Let $(\hat\mu_{nh})_{n\in\mathbb{N}}$ denote the marginal laws of* LMC *with step size $h > 0$. Then,*

$$2h\,\mathsf{KL}(\hat\mu_{(n+1)h} \,\|\, \pi) \leqslant W_2^2(\hat\mu^+_{nh}, \pi) - W_2^2(\hat\mu^+_{(n+1)h}, \pi) + L^2 h^2\,.$$

Theorem 4.3.11 (Durmus et al. (2019))**.** *Suppose that $\pi = \exp(-V)$ is the target distribution and that V is convex and L-Lipschitz. Let $(\hat\mu_{nh})_{n\in\mathbb{N}}$ denote the marginal laws of* LMC*. For any $\varepsilon > 0$, if we take step size $h \asymp \varepsilon^2/L^2$, then for the mixture distribution $\bar\mu_{Nh} \coloneqq N^{-1}\sum_{n=1}^N \hat\mu_{nh}$ it holds that $\sqrt{\mathsf{KL}(\bar\mu_{Nh} \,\|\, \pi)} \leqslant \varepsilon$ after*

$$N = O\Bigl(\frac{L^2\, W_2^2(\hat\mu_0^+, \pi)}{\varepsilon^4}\Bigr) \qquad \textit{iterations}\,.$$

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 $\hat{X}_{(n+1)h}$. From (4.3.2),

$$2h\,\{\mathcal{E}(\hat{\mu}_{(n+1)h}) - \mathcal{E}(\pi)\} \leqslant \mathbb{E}[\|\hat{X}_{(n+1)h} - Z\|^2 - \|\hat{X}^+_{(n+1)h} - Z\|^2 + h^2\,\|\nabla V(\hat{X}_{(n+1)h})\|^2]\,.$$

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,

$$2h\,\{\mathcal{E}(\hat{\mu}_{(n+1)h}) - \mathcal{E}(\pi)\} \leqslant W_2^2(\hat{\mu}_{(n+1)h}, \pi) - W_2^2(\hat{\mu}^+_{(n+1)h}, \pi) + L^2h^2\,. \tag{4.3.12}$$

On the other hand, recall from (4.3.9) that

$$2h\,\{\mathcal{H}(\hat{\mu}_{(n+1)h}) - \mathcal{H}(\pi)\} \leqslant W_2^2(\hat{\mu}^+_{nh}, \pi) - W_2^2(\hat{\mu}_{(n+1)h}, \pi)\,.$$

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 $\mathsf{KL}(\hat{\mu}_{Nh} \,\|\, \pi_{Nh})$, where $\hat{\mu}_{Nh}$ is the law of the LMC iterate $\hat{X}_{Nh}$ and π_{Nh} is the law of the Langevin diffusion X_{Nh} initialized at $\hat{\mu}_0$. 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, namely the smoothness of V. The point, however, is that control of $\mathsf{KL}(\hat{\mu}_{Nh} \,\|\, \pi_{Nh})$ does not immediately yield quantitative convergence of $\hat{\mu}_{Nh} \to \pi$; 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 $\mathsf{KL}(\hat{\mu}_{Nh} \,\|\, \pi_{Nh})$ 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. We remark that this shortcoming can be removed in the strongly log-concave setting using a "shifted Girsanov" approach, similar to Exercise 3.4; see Altschuler and Chewi (2024c) for details.

Historically, the Girsanov method was one of the first discretization techniques utilized in the modern quantitative study of sampling (see Dalalyan and Tsybakov, 2012). The argument we present here is similar in spirit to Dalalyan and Tsybakov (2012), although we have made a few refinements.

Theorem 4.4.1 (LMC discretization bound). *Let $\pi \propto \exp(-V)$ be the target and assume that ∇V is β-Lipschitz. Let $(\hat{\mu}_{nh})_{n\in\mathbb{N}}$ denote the marginal laws of* LMC, *and let $(\pi_t)_{t\geqslant 0}$ denote the law of the Langevin diffusion initialized at $\hat{\mu}_0$. Then, for $h \lesssim 1/\beta$ and $T := Nh$,*

$$\mathsf{KL}(\hat{\mu}_T \,\|\, \pi_T) \lesssim \beta^2 h^3 \sum_{n=0}^{N-1} \mathbb{E}_{\hat{\mu}_{nh}}[\|\nabla V\|^2] + \beta^2 dhT\,.$$

Proof Let $(B_t)_{t\in[0,T]}$ be our standard Brownian motion and consider the SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad X_0 \sim \hat{\mu}_0\,.$$

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

$$\mathrm{d}X_t = -\nabla V(X_{t_-})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\tilde{B}_t\,,$$

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 $\mathrm{d}\tilde{B}_t = \mathrm{d}B_t - \mathrm{d}[B,M]_t$ where $\mathrm{d}M_t := \frac{1}{\sqrt{2}}\,\langle \nabla V(X_t) - \nabla V(X_{t_-}), \mathrm{d}B_t\rangle$, consider the exponential martingale $\mathcal{E}(M)$ associated with M.

By the data-processing inequality (Theorem 1.5.6), $\mathsf{KL}(\hat{\mu}_T \,\|\, \pi_T) \leqslant \mathsf{KL}(\mathbf{P}_T \,\|\, \mathbf{W}_T)$, so it suffices to bound the latter. By Girsanov's theorem (Theorem 3.2.8),[2]

$$\begin{aligned}\mathsf{KL}(\mathbf{P}_T \,\|\, \mathbf{W}_T) &= \mathbb{E}^{\mathbf{P}_T} \log \frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{W}_T} = \mathbb{E}^{\mathbf{P}_T} \log \mathcal{E}(M)_T \\ &= \mathbb{E}^{\mathbf{P}_T}\Big[\frac{1}{\sqrt{2}}\int_0^T \langle \nabla V(X_t) - \nabla V(X_{t_-}), \mathrm{d}B_t\rangle - \frac{1}{4}\int_0^T \|\nabla V(X_t) - \nabla V(X_{t_-})\|^2\,\mathrm{d}t\Big]\,.\end{aligned}$$

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

$$\begin{aligned}\mathsf{KL}(\mathbf{P}_T \,\|\, \mathbf{W}_T) &= \frac{1}{4}\int_0^T \mathbb{E}^{\mathbf{P}_T}[\|\nabla V(X_t) - \nabla V(X_{t_-})\|^2]\,\mathrm{d}t \leqslant \frac{\beta^2}{4}\int_0^T \mathbb{E}^{\mathbf{P}_T}[\|X_t - X_{t_-}\|^2]\,\mathrm{d}t \\ &= \frac{\beta^2}{4}\int_0^T \{(t-t_-)^2\,\mathbb{E}^{\mathbf{P}_T}[\|\nabla V(X_{t_-})\|^2] + 2d\,(t-t_-)\}\,\mathrm{d}t \\ &\leqslant \frac{\beta^2 h^3}{12}\sum_{n=0}^{N-1} \mathbb{E}^{\mathbf{P}_T}[\|\nabla V(X_{nh})\|^2] + \frac{\beta^2 d h^2 N}{4}\,.\end{aligned}$$

□

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 $\hat{\mu}_{Nh}$ 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) and the fact that the total variation distance satisfies the triangle inequality.

Corollary 4.4.2. *Let $\pi \propto \exp(-V)$ be the target, assume that $C_{\mathsf{LSI}}(\pi) \leqslant 1/\alpha$ and that ∇V is β-Lipschitz, and let $\kappa := \beta/\alpha$. Let $(\hat{\mu}_{nh})_{n\in\mathbb{N}}$ denote the marginal laws of* LMC. *Then, for all sufficiently small $\varepsilon > 0$ and for an appropriate choice of h, we obtain the guarantee $\|\hat{\mu}_{Nh} - \pi\|_{\mathrm{TV}} \leqslant \varepsilon$ provided that*

$$N = \Theta\Big(\frac{\kappa^2 d}{\varepsilon^2}\log^2\frac{\mathsf{KL}(\hat{\mu}_0 \,\|\, \pi)}{\varepsilon^2}\Big) \qquad \textit{iterations}\,.$$

---

[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.

For the Pinsker approach, we can leverage the fact that we can bound $\|\hat{\mu}_{Nh} - \pi_{Nh}\|_{\mathrm{TV}}^2$ either by $\mathsf{KL}(\hat{\mu}_{Nh} \parallel \pi_{Nh})$ or $\mathsf{KL}(\pi_{Nh} \parallel \hat{\mu}_{Nh})$, depending on which is easier. The latter KL divergence is controlled in exactly the same way as in Theorem 4.4.1, except that in the resulting bound we replace $\mathbb{E}_{\hat{\mu}_{nh}}[\|\nabla V\|^2]$ with $\mathbb{E}_{\pi_{nh}}[\|\nabla V\|^2]$.

Proof of Corollary 4.4.2 The LSI implies exponential convergence of the Langevin diffusion to its target in KL divergence (Theorem 1.2.26 and Theorem 1.2.30), and we obtain $\sqrt{\mathsf{KL}(\pi_{Nh} \parallel \pi)} \leqslant \varepsilon/\sqrt{2}$ with the choice $T = Nh \asymp \alpha^{-1} \log(\mathsf{KL}(\hat{\mu}_0 \parallel \pi)/\varepsilon^2)$.

Next, if we simply apply the discretization bound for $\mathsf{KL}(\hat{\mu}_{Nh} \parallel \pi_{Nh})$ in Theorem 4.4.1, it becomes tricky to bound $\mathbb{E}_{\hat{\mu}_{nh}}[\|\nabla V\|^2]$ without further assumptions. Instead, as discussed above, we use a bound for $\mathsf{KL}(\pi_{Nh} \parallel \hat{\mu}_{Nh})$. We apply Talagrand's T_2 inequality, which is implied by the LSI (Exercise 1.16), and the data-processing inequality[3] (Theorem 1.5.6):

$$\mathbb{E}_{\pi_{nh}}[\|\nabla V\|^2] \lesssim \mathbb{E}_{\pi}[\|\nabla V\|^2] + \beta^2\, W_2^2(\pi_{nh}, \pi) \lesssim \beta d + \frac{\beta^2}{\alpha}\, \mathsf{KL}(\pi_{nh} \parallel \pi) \leqslant \beta d + \frac{\beta^2}{\alpha}\, \mathsf{KL}(\hat{\mu}_0 \parallel \pi)\,.$$

Similarly to Theorem 4.4.1, we then obtain

$$\mathsf{KL}(\pi_{Nh} \parallel \hat{\mu}_{Nh}) \lesssim \beta^3 d h^2 T + \frac{\beta^4 h^2 T}{\alpha}\, \mathsf{KL}(\hat{\mu}_0 \parallel \pi) + \beta^2 d h T\,.$$

We now choose h to make this at most $\varepsilon^2/2$. To simplify the calculation, we assume that ε is sufficiently small so that the third term is dominant, as is typically the case in applications. This leads to the choice $h = \Theta(\varepsilon^2/\beta\kappa d \log(\mathsf{KL}(\hat{\mu}_0 \parallel \pi)/\varepsilon^2))$.

By the triangle inequality and Pinsker's inequality,

$$\|\hat{\mu}_{Nh} - \pi\|_{\mathrm{TV}} \leqslant \|\hat{\mu}_{Nh} - \pi_{Nh}\|_{\mathrm{TV}} + \|\pi_{Nh} - \pi\|_{\mathrm{TV}} \leqslant \sqrt{\frac{1}{2}\, \mathsf{KL}(\pi_{Nh} \parallel \hat{\mu}_{Nh})} + \sqrt{\frac{1}{2}\, \mathsf{KL}(\pi_{Nh} \parallel \pi)} \leqslant \varepsilon\,.$$

Finally, plugging in the choice of h into $Nh \asymp \alpha^{-1} \log(\mathsf{KL}(\hat{\mu}_0 \parallel \pi)/\varepsilon^2)$ yields the result. □

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 (2.2.24); recall that $\mathsf{KL} = \mathcal{R}_1$. We have a continuous-time result for the Langevin diffusion in Rényi divergence (Theorem 2.2.25), 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* (see Lemma 6.2.5). 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 Dalalyan and Tsybakov (2012), 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: Dalalyan and Tsybakov (2012); Dalalyan (2017a,b); Durmus and Moulines (2017, 2019). The Girsanov argument of Theorem 4.4.1 is also due to Dalalyan and Tsybakov (2012).

---

[3] The data-processing inequality is not available to us if we bound the KL divergence with the opposite order of arguments.

The coupling analysis of Section 4.1 will be generalized into a local error framework in Section 5.1.1. Using this framework, Exercise 5.2 will show that the rate for LMC can be improved under higher-order smoothness assumptions on V. Guarantees in KL divergence under higher-order smoothness were also obtained via the interpolation method of Section 4.2 in Mou et al. (2022).

The proof of Exercise 4.7 is taken from Altschuler and Chewi (2024b), and the examples of Exercise 4.8 and Exercise 4.9 are from Wibisono (2018).

Another notable proof technique that we have omitted from this chapter is **reflection coupling** (Eberle, 2011, 2016). 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**).

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. In particular, for $\Sigma = I_d$, show that the step size can be taken to be $h \asymp \varepsilon/d^{1/2}$ to control the bias.

⊵ **Exercise 4.2** (Second moment bounds for LMC)
Assume that $\beta I_d \succeq \nabla^2 V \succeq \alpha I_d \succ 0$ and $\nabla V(0) = 0$. Prove that if $h \leqslant 1/\beta$, then the LMC iterates initialized at δ_0 with step size $h > 0$ have uniformly bounded second moment: $\sup_{n\in\mathbb{N}} \mathbb{E}[\|\hat{X}_{nh}\|^2] \lesssim d/\alpha$. (Write a recursion for $\mathbb{E}[\|\hat{X}_{(n+1)h}\|^2]$ in terms of $\mathbb{E}[\|\hat{X}_{nh}\|^2]$. In order to prove the result for all step sizes $h \leqslant 1/\beta$, you may need to appeal to coercivity of the gradient, Lemma 5.2.3.)

⊵ **Exercise 4.3** (LMC with decaying step size)
Consider LMC initialized at $\delta_{x_\star}$, where $x_\star$ is the mode of π and π satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Show that by considering LMC with a decaying step size schedule $h_n = 1/(C\beta\kappa + c\alpha n)$, where $C, c > 0$ are appropriately chosen universal constants, one can obtain an iteration complexity which removes the logarithmic factor in Theorem 4.1.2 for all $N \gtrsim \kappa^2$.

Hint: The term $C\beta\kappa$ in the choice of step size is so that the one-step bound in Remark 4.1.5 is valid. However, the resulting computation is tedious, so here is a simpler exercise that still conveys the main point. Take the step size schedule $h_n = 1/(c\alpha n)$ for $c > 0$ sufficiently small and use the one-step bound from Remark 4.1.5 (pretending that it is still valid) to obtain an iteration complexity bound without a logarithmic term.

⊵ **Exercise 4.4** (Uniform-in-time LSI for LMC under strong log-concavity)

1 Let $(\hat{\mu}_{nh})_{n\in\mathbb{N}}$ denote the laws of the iterates of LMC, and assume that $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$, where α is allowed to be negative. Show that if h is small enough (depending on β only), and if $C_{\mathsf{LSI}}(\mu_0) \leqslant 1/\alpha_0$, then for all $n \in \mathbb{N}$,

$$C_{\mathsf{LSI}}(\hat{\mu}_{nh}) \leqslant \frac{c_1(\alpha, n)}{\alpha_0} + c_2(\alpha, n)\,,$$

for some $c_1(\alpha, n)$ and $c_2(\alpha, n)$ that you should compute. In particular, if $\alpha > 0$, then the LSI is uniform-in-time: $\sup_{n\in\mathbb{N}} C_{\mathsf{LSI}}(\hat{\mu}_{nh}) < \infty$.

Hint: Use Proposition 2.3.3 and Proposition 2.3.8.

2 Take the continuous-time limit $h \searrow 0$ as $nh \to t$ in the previous part, and deduce that the local LSI (Theorem 2.2.21) holds.

Proof via Interpolation Argument

⊵ Exercise 4.5 (More general interpolations)

Let $(X_t)_{t\geqslant 0}$ be a stochastic process with C^1 sample paths, and let $\mu_t := \operatorname{law}(X_t)$. Prove that

$$\partial_t \mu_t + \operatorname{div}(\mu_t b_t) = 0\,,$$

where $b_t := \mathbb{E}[\dot{X}_t \mid X_t = \cdot]$. Note that this Fokker–Planck equation also describes the marginal laws of the process $\dot{Y}_t = b_t(Y_t)$, which is Markovian.

More generally, for any $\gamma > 0$, show that $\partial_t \mu_t + \operatorname{div}(\mu_t b_t^\gamma) = \gamma\,\Delta\mu_t$, where $b_t^\gamma := b_t + \gamma\,\nabla \log \mu_t$. This corresponds to the SDE $\mathrm{d}Y_t = b_t^\gamma(Y_t)\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t$.

⊵ Exercise 4.6 (Fokker–Planck vs. Girsanov's theorem)

Consider the Fokker–Planck equations for two SDEs:

$$\partial_t \mu_t + \operatorname{div}(\mu_t b_t) = \gamma\,\Delta\mu_t\,, \qquad \partial_t \bar{\mu}_t + \operatorname{div}(\bar{\mu}_t \bar{b}_t) = \gamma\,\Delta\bar{\mu}_t\,.$$

By differentiating $t \mapsto \mathsf{KL}(\mu_t \,\|\, \bar{\mu}_t)$ (similarly to the calculation in Corollary 4.2.4) show that $\mathsf{KL}(\mu_T \,\|\, \bar{\mu}_T) \leqslant \frac{1}{4\gamma}\int_0^T \mathbb{E}_{\mu_t}[\|b_t - \bar{b}_t\|^2]\,\mathrm{d}t$. Compare this with what can be obtained from Girsanov's theorem (Theorem 3.2.8).

⊵ Exercise 4.7 (Harnack inequalities via Fokker–Planck)

In this exercise, we provide yet another proof of the Harnack inequalities from Exercise 2.17 and Exercise 3.4. Consider Fokker–Planck equations for the Langevin diffusion and for an auxiliary process, given as follows:

$$\partial_t \mu_t = \operatorname{div}\Bigl(\mu_t\,\nabla \log \frac{\mu_t}{\pi}\Bigr)\,, \quad \partial_t \nu_t = \operatorname{div}\Bigl(\nu_t\,\nabla \log \frac{\nu_t}{\pi}\Bigr)\,, \quad \partial_t \mu_t' = \operatorname{div}\Bigl(\mu_t'\,\bigl(\nabla \log \frac{\mu_t'}{\pi} - \eta_t\,(T_{\mu_t'\to\mu_t} - \mathrm{id})\bigr)\Bigr)$$

with $\mu_0' = \nu_0$ and $(\eta_t)_{t\geqslant 0}$ chosen so that $\mu_T' = \mu_T$. Here, $T_{\mu_t'\to\mu_t}$ is the optimal transport map from μ_t' to μ_t. Assuming that $\nabla^2 V \succeq \alpha I$, differentiate $t \mapsto \mathsf{KL}(\mu_t' \,\|\, \nu_t)$ as in the preceding exercise, and use this to again establish (2.E.8). Can you also adapt this argument to Rényi divergences?

Proof via Convex Optimization

⊵ Exercise 4.8 (Forward–flow discretization is biased)

Consider optimization of a composite objective $f+g$, where $f(x) := \frac{1}{2}\,(x-1)^2$ and $g(x) := \frac{1}{2}\,(x+1)^2$. Write down the "forward–flow" splitting scheme for minimizing $f+g$, compute the limiting point of the algorithm with step size h, and deduce that this scheme is biased.

⊵ **Exercise 4.9** (Proximal operator for the entropy)
As described further in Optimization Box 8.0.1, the sampling analogue of a “backward” discretization step on the entropy would be to solve the following optimization problem:

$$\operatorname{prox}_{h\mathcal{H}}(\nu) := \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \Bigl\{\mathcal{H}(\mu) + \frac{1}{2h}\, W_2^2(\mu, \nu)\Bigr\}\,.$$

In general, this is intractable. However, show that when $\nu = \mathsf{normal}(0, \Sigma)$, the solution μ is also Gaussian, and determine the mean and covariance of μ.

⊵ **Exercise 4.10** (LMC without a two-phase analysis)
In Theorem 4.3.6, we deduced the rate for the strongly convex case via a two-phase analysis; in particular, we throw away the first N iterates and average the following N iterates, which can be interpreted as a “burn-in” phase. Here, we show that this is unnecessary, provided that we use a different averaging scheme. Starting with (4.3.5), multiply both sides by $(1 - \alpha h)^{-(n+1)}$ and sum across the iterations. Show that a suitable average $\bar{\mu}_{Nh}$ of the first iterates achieves the KL guarantee in Theorem 4.3.6 for the strongly convex case.

Proof via Girsanov’s Theorem

⊵ **Exercise 4.11** (Convergence under PI)
Follow the proof of Corollary 4.4.2, but instead of assuming that $C_{\mathsf{LSI}}(\pi) \leqslant 1/\alpha$, assume instead that $C_{\mathsf{PI}}(\pi) \leqslant 1/\alpha$. What is the final iteration complexity if $\log \chi^2(\hat{\mu}_0 \,\|\, \pi) = \widetilde{O}(d)$? (See Lemma 1.5.2.)

⊵ **Exercise 4.12** (Diverging Girsanov bounds)
Consider the case where $\pi = \mathsf{normal}(0, I_d)$ and show that for the path measures in the proof of Theorem 4.4.1, $\mathsf{KL}(\boldsymbol{P}_T \,\|\, \boldsymbol{W}_T)$ *necessarily* grows linearly with T; thus, the diverging bound is a fundamental limitation of the Girsanov-based approach. See, however, Altschuler and Chewi (2024c) for how to circumvent this with the shifted Girsanov method. On the other hand, argue that $\mathsf{KL}(\hat{\mu}_T \,\|\, \pi_T)$ remains bounded as $T \to \infty$, so the data-processing inequality is a bit loose here.

⊵ **Exercise 4.13** (Sharper Girsanov bound)
Show that in Theorem 4.4.1, the following sharper upper bound holds:

$$\mathsf{KL}(\hat{\mu}_T \,\|\, \pi_T) \leqslant \frac{1}{4} \int_0^T \mathbb{E}\bigl[\bigl\|\mathbb{E}[\nabla V(\hat{X}_{t_-}) \mid \hat{X}_t] - \nabla V(\hat{X}_t)\bigr\|^2\bigr]\, \mathrm{d}t\,.$$

Thus, Girsanov’s theorem can also capture some of the benefits of the interpolation method. *Hint*: Similarly to Exercise 4.5, use Proposition 4.2.3 to argue that there is a Markov process which is different from the interpolation of LMC, yet has the same marginal laws as LMC. Apply Girsanov’s theorem to this process instead.

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 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 Shen and Lee (2019). The application to the Langevin diffusion was carried out in He et al. (2020) and more recently in Yu et al. (2024).

Consider the continuous-time Langevin diffusion from time nh to $(n+1)h$:

$$X_{(n+1)h} = X_{nh} - \int_{nh}^{(n+1)h} \nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,(B_{(n+1)h} - B_{nh})\,.$$

In the Euler discretization, we approximate the second term via $-h\,\nabla V(X_{nh})$. However, if we want an *unbiased* estimator of the integral, then we can introduce an auxiliary random variable $u_n \sim \mathsf{uniform}([0,1])$ and use

$$X_{(n+1)h} \approx X_{nh} - h\,\nabla V(X_{nh+u_n h}) + \sqrt{2}\,(B_{(n+1)h} - B_{nh})\,.$$

To compute this approximation, however, we need to know $X_{nh+u_n h}$. We have

$$X_{nh+u_n h} = X_{nh} - \int_{nh}^{nh+u_n h} \nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,(B_{nh+u_n h} - B_{nh})\,.$$

Note that we are in the same situation as before. In particular, if we desire, we can draw another uniform random variable u_n' and approximate the second term above via $-u_n h\,\nabla V(X_{nh+u_n u_n' h})$. 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 $X_{nh+u_n h}$ via an Euler–Maruyama step:

$$X_{nh+u_n h} \approx X_{nh} - u_n h \, \nabla V(X_{nh}) + \sqrt{2}\,(B_{nh+u_n h} - B_{nh})\,.$$

To summarize, the randomized midpoint discretization of the Langevin diffusion, which we will call RM–LMC, is the following update:

$$\begin{aligned} \hat{X}_{(n+1)h} &:= \hat{X}_{nh} - h\,\nabla V(\hat{X}^+_{nh+u_n h}) + \sqrt{2}\,(B_{(n+1)h} - B_{nh})\,, \\ \hat{X}^+_{nh+u_n h} &:= \hat{X}_{nh} - u_n h\,\nabla V(\hat{X}_{nh}) + \sqrt{2}\,(B_{nh+u_n h} - B_{nh})\,, \end{aligned} \tag{RM–LMC}$$

where $(u_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. $\mathsf{uniform}([0,1])$ random variables which are independent of $\hat{X}_0$ 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_n and ξ'_n, and set

$$\begin{aligned} B_{nh+u_n h} - B_{nh} &:= \sqrt{u_n h}\,\xi_n\,, \\ B_{(n+1)h} - B_{nh} &:= \sqrt{u_n h}\,\xi_n + \sqrt{(1-u_n)h}\,\xi'_n\,. \end{aligned}$$

We now state the complexity bound for RM–LMC.

Theorem 5.1.1 (RM–LMC). *For $n \in \mathbb{N}$, let $\hat{\mu}_{nh}$ denote the law of the n-th iterate of RM–LMC with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, provided $h \ll 1/\beta\kappa^{1/2}$, for all $N \in \mathbb{N}$,*

$$W_2^2(\hat{\mu}_{Nh}, \pi) \leqslant \exp(-\alpha N h)\, W_2^2(\hat{\mu}_0, \pi) + O\Big(\frac{\beta^2 d h^2}{\alpha} + \frac{\beta^4 d h^3}{\alpha^2}\Big)\,.$$

In particular, if we choose $h \asymp \varepsilon/\beta d^{1/2}$, then for any $\varepsilon \in [0, d^{1/2}/\kappa^{1/4}]$ we obtain the guarantee $\sqrt{\alpha}\, W_2(\hat{\mu}_{Nh}, \pi) \leqslant \varepsilon$ after

$$N = \widetilde{O}\Big(\frac{\kappa d^{1/2}}{\varepsilon} \vee \frac{\kappa^{4/3} d^{1/3}}{\varepsilon^{2/3}}\Big) \qquad \textit{iterations}\,.$$

This result is a substantial improvement over the $\widetilde{O}(\kappa d/\varepsilon^2)$ rate obtained for LMC in Theorem 4.3.6. To establish the theorem, we now introduce a powerful framework known as *local error analysis*, which streamlines the analysis of RM–LMC as well as other sophisticated discretizations of SDEs.

See the Bibliographical Notes for more recent developments.

5.1.1 Local Error Analysis

The intuition behind randomized midpoint is that the gradient evaluated at a uniformly random chosen time is an "unbiased" estimator of the stochastic integral. But what exactly does "bias" mean in this context? Suppose that the "error" incurred by a discretization method at iteration n is denoted ω_n, and for the sake of argument, suppose that $\omega_1, \dots, \omega_N$ are i.i.d. with mean m and variance σ^2. Then, the total error over N iterations, $\sum_{n=1}^N \omega_n$, has size roughly $Nm + N^{1/2}\sigma$, where the first term represents systematic bias that adds up over the iterations and the second term represents stochastic fluctuations which cancel out and hence only add up as the square root of the number of iterations. Although this model for the errors is clearly unrealistic, it motivates the development of a framework that separates out the effects of "systematic bias" versus "stochastic fluctuations", which in the sequel will be called the "weak error" and the "strong error".

The following framework, known as **local error analysis** or **mean-squared error analysis**, is classical in the literature on numerical methods for SDEs (Milstein and Tretyakov, 2021).

Lemma 5.1.2 (Local error analysis)**.** *Let* $\hat P$, P *be two Markov kernels over* $\mathbb{R}^d$ *and let* $x, y \in \mathbb{R}^d$*. Suppose that there are jointly defined random variables* $\hat X \sim \delta_x \hat P$, $X \sim \delta_x P$, $Y \sim \delta_y P$ *satisfying the following four conditions:*

- *(Contractivity)* $\|X - Y\|_{L^2} \leqslant L\,\|x - y\|$.
- *(Coupling)* $\|(X - x) - (Y - y)\|_{L^2} \leqslant \zeta\,\|x - y\|$.
- *(Weak error)* $\|\mathbb{E}\,\hat X - \mathbb{E}\,X\| \leqslant \mathscr{E}_{\text{weak}}(x)$.
- *(Strong error)* $\|\hat X - X\|_{L^2} \leqslant \mathscr{E}_{\text{strong}}(x)$.

Then, the following one-step Wasserstein bound holds:

$$W_2^2(\delta_x \hat P, \delta_y P) \leqslant L^2\,\|x - y\|^2 + 2\left(\mathscr{E}_{\text{weak}}(x) + \zeta \mathscr{E}_{\text{strong}}(x)\right)\|x - y\| + \mathscr{E}_{\text{strong}}(x)^2\,. \tag{5.1.3}$$

In particular, if the above conditions hold for all $x, y \in \mathbb{R}^d$ *for some* $L < 1$*, then for any probability measures* μ_0, ν_0,

$$W_2^2(\mu_0 \hat P^N, \nu_0 P^N) \leqslant L^N\,W_2^2(\mu_0, \nu_0) + \frac{(\bar{\mathscr{E}}_{\text{weak}} + \zeta \bar{\mathscr{E}}_{\text{strong}})^2}{L\,(1 - L)^2} + \frac{(\bar{\mathscr{E}}_{\text{strong}})^2}{1 - L}\,,$$

where we set

$$\bar{\mathscr{E}}_{\text{weak}} := \max_{n=0,1,\dots,N-1} \|\mathscr{E}_{\text{weak}}\|_{L^2(\mu_0 \hat P^n)}\,, \qquad \bar{\mathscr{E}}_{\text{strong}} := \max_{n=0,1,\dots,N-1} \|\mathscr{E}_{\text{strong}}\|_{L^2(\mu_0 \hat P^n)}\,.$$

Proof Expand out the square and apply Cauchy–Schwarz:

$$\begin{aligned}
\mathbb{E}[\|\hat X - Y\|^2] &= \mathbb{E}[\|X - Y\|^2 + 2\,\langle \hat X - X, X - Y\rangle + \|\hat X - X\|^2] \\
&\leqslant L^2\,\|x - y\|^2 + \mathbb{E}[2\,\langle \hat X - X, x - y\rangle + 2\,\langle \hat X - X, (X - x) - (Y - y)\rangle + \|\hat X - X\|^2] \\
&\leqslant L^2\,\|x - y\|^2 + 2\left(\mathscr{E}_{\text{weak}}(x) + \zeta \mathscr{E}_{\text{strong}}(x)\right)\|x - y\| + \mathscr{E}_{\text{strong}}(x)^2\,.
\end{aligned}$$

To iterate this bound, we apply Young's inequality to the middle term to obtain

$$2\left(\mathscr{E}_{\text{weak}}(x) + \zeta \mathscr{E}_{\text{strong}}(x)\right)\|x - y\| \leqslant L\,(1 - L)\,\|x - y\|^2 + \frac{1}{L\,(1 - L)}\left(\mathscr{E}_{\text{weak}}(x) + \zeta \mathscr{E}_{\text{strong}}(x)\right)^2.$$

Hence,

$$W_2^2(\delta_x \hat P, \delta_y P) \leqslant L\,\|x - y\|^2 + \frac{1}{L\,(1 - L)}\left(\mathscr{E}_{\text{weak}}(x) + \zeta \mathscr{E}_{\text{strong}}(x)\right)^2 + \mathscr{E}_{\text{strong}}(x)^2\,.$$

A coupling argument (c.f. Lemma 8.2.2) and iteration yield the result. □

In applications, P is taken to be an idealized process (e.g., the exact Langevin diffusion), whereas $\hat P$ is a numerical discretization. The lemma adopts four assumptions, two of which are solely properties of the kernel P. The remaining two assumptions, which quantify the accuracy of the numerical scheme, are purely local in that they only consider one step of the algorithm and idealized process started from the same starting point, and are therefore typically easy to compute.

Note that the effective time horizon arising from a contraction factor L is $N_{\text{eff}} := 1/(1 - L)$. So, if we disregard the term involving ζ (which is often small), the final bound shows that the error over N steps is bounded by $N_{\rm eff}\,\bar{\mathscr{E}}_{\rm weak} + N_{\rm eff}^{1/2}\,\bar{\mathscr{E}}_{\rm strong}$, in accordance with the intuition presented above.

As stated, the local error framework requires contractivity ($L < 1$) and only produces guarantees in the W_2 metric. Both of these points were addressed in Altschuler and Chewi (2024c), which established a local error framework in KL divergence. One advantage of such a framework is that the errors do not grow exponentially with the number of iterations, even in the case $L > 1$.

***Remark* 5.1.4.** It is often more convenient to have the weak and strong error terms depend on y rather than x, because the expectations are then taken under the idealized process $\{\nu_0 P^n\}_{n\geqslant 0}$. To handle this, often $\mathscr{E}_{\rm weak}, \mathscr{E}_{\rm strong}$ are $L_{\rm weak}$-, $L_{\rm strong}$-Lipschitz respectively, in which case (5.1.3) implies

$$W_2^2(\delta_x \hat{P}, \delta_y P) \leqslant L'^2\,\|x-y\|^2 + 2\,\big(\mathscr{E}_{\rm weak}(y) + \zeta\mathscr{E}_{\rm strong}(y)\big)\,\|x-y\| + 2\mathscr{E}_{\rm strong}(y)^2\,,$$

with $L'^2 = L^2 + 2L_{\rm weak} + 2\zeta L_{\rm strong} + 2L_{\rm strong}^2$. If $L' < 1$, this can be iterated to yield

$$W_2^2(\mu_0\hat{P}^N, \nu_0 P^N) \leqslant L'^N\,W_2^2(\mu_0,\nu_0) + \frac{(\bar{\mathscr{E}}_{\rm weak} + \zeta\bar{\mathscr{E}}_{\rm strong})^2}{L'\,(1-L')^2} + \frac{2\,(\bar{\mathscr{E}}_{\rm strong})^2}{1-L'}\,,$$

where now we take

$$\bar{\mathscr{E}}_{\rm weak} = \max_{n=0,1,\dots,N-1} \|\mathscr{E}_{\rm weak}\|_{L^2(\nu_0 P^n)}\,, \qquad \bar{\mathscr{E}}_{\rm strong} = \max_{n=0,1,\dots,N-1} \|\mathscr{E}_{\rm strong}\|_{L^2(\nu_0 P^n)}\,.$$

***Remark* 5.1.5.** If we do not wish to take advantage of the weak error, then we instead have the simpler recursion $W_2(\delta_x\hat{P}, \delta_y P) \leqslant L\,\|x-y\| + \mathscr{E}_{\rm strong}(x)$. This is iterated to yield

$$W_2(\mu_0\hat{P}^N, \nu_0 P^N) \leqslant L^N\,W_2(\mu_0,\nu_0) + \frac{\max_{n=0,1,\dots,N-1}\|\mathscr{E}_{\rm strong}\|_{L^2(\mu_0\hat{P}^n)}}{1-L}\,.$$

If $\mathscr{E}_{\rm strong}$ is $L_{\rm strong}$-Lipschitz, then for $L' := L + L_{\rm strong}$,

$$W_2(\mu_0\hat{P}^N, \nu_0 P^N) \leqslant L'^N\,W_2(\mu_0,\nu_0) + \frac{\max_{n=0,1,\dots,N-1}\|\mathscr{E}_{\rm strong}\|_{L^2(\nu_0 P^n)}}{1-L'}\,.$$

To conclude this section, we show that the Langevin diffusion satisfies the first two assumptions of Lemma 5.1.2.

Lemma 5.1.6 (Local error assumptions for the Langevin diffusion)**.** *Let $(X_t)_{t\geqslant 0}$, $(Y_t)_{t\geqslant 0}$ denote two copies of the Langevin diffusion with potential V and started at x and y respectively. Assume that $0 \preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, the following statements hold.*

1 $\|X_h - Y_h\|_{L^\infty} \leqslant \exp(-\alpha h)\,\|x-y\|$.
2 $\|X_h - x - (Y_h - y)\|_{L^\infty} \leqslant \beta h\,\|x-y\|$.

Proof The first statement is Exercise 1.17. For the second statement,

$$\|X_h - x - (Y_h - y)\| = \Big\|\int_0^h \{\nabla V(X_t) - \nabla V(Y_t)\}\,\mathrm{d}t\Big\| \leqslant \beta\int_0^h \|X_t - Y_t\|\,\mathrm{d}t \leqslant \beta h\,\|x-y\|\,. \qquad \square$$

5.1.2 Analysis of Randomized Midpoint

To prove Theorem 5.1.1, we compute the weak and strong errors for RM–LMC.

Lemma 5.1.7 (Local errors for RM–LMC). *Let $\hat P$, P denote the kernels for RM–LMC and for the Langevin diffusion run for time h respectively. Assume that $-\beta I_d \preceq \nabla^2 V \preceq \beta I_d$ and $h \ll 1/\beta$. There is a coupling $\hat X_h \sim \delta_x \hat P$, $X_h \sim \delta_x P$ such that the following hold.*

1. *(Weak error)* $\|\mathbb{E}\,\hat X_h - \mathbb{E}\,X_h\| \lesssim \beta^2 h^3\,\|\nabla V(x)\| + \beta^2 d^{1/2} h^{5/2}$.
2. *(Strong error)* $\|\hat X_h - X_h\|_{L^2} \lesssim \beta h^2\,\|\nabla V(x)\| + \beta d^{1/2} h^{3/2}$.

Proof We begin with the strong error.

$$\|\hat X_h - X_h\| = \Big\|\int_0^h \{\nabla V(\hat X_{uh}^+) - \nabla V(X_t)\}\,\mathrm{d}t\Big\| \leqslant \beta \int_0^h \|\hat X_{uh}^+ - X_t\|\,\mathrm{d}t\,.$$

Next,

$$\begin{aligned}\|\hat X_{uh}^+ - X_t\|_{L^2} &= \Big\|\int_0^t \nabla V(X_s)\,\mathrm{d}s - uh\,\nabla V(x) + \sqrt{2}\,(B_{uh} - B_t)\Big\|_{L^2}\\ &\lesssim h\,\|\nabla V(x)\| + \int_0^t \|\nabla V(X_s)\|_{L^2}\,\mathrm{d}s + d^{1/2}h^{1/2}\\ &\lesssim h\,\|\nabla V(x)\| + \beta\int_0^t \|X_s - x\|_{L^2}\,\mathrm{d}s + d^{1/2}h^{1/2} \lesssim h\,\|\nabla V(x)\| + d^{1/2}h^{1/2}\,,\end{aligned}$$

where we applied Lemma 4.1.8. This yields the strong error estimate.

For the weak error, note that by taking the expectation over the uniform random variable, we obtain

$$\mathbb{E}[\hat X_h \mid (B_t)_{t\in[0,h]}] = x - \int_0^h \nabla V(\hat X_t^+)\,\mathrm{d}t + \sqrt{2}\,B_h\,.$$

This leads to substantial cancellations:

$$\begin{aligned}\|\mathbb{E}\,\hat X_h - \mathbb{E}\,X_h\| &= \Big\|\mathbb{E}\int_0^h \{\nabla V(\hat X_t^+) - \nabla V(X_t)\}\,\mathrm{d}t\Big\| \leqslant \beta\int_0^h \mathbb{E}\|\hat X_t^+ - X_t\|\,\mathrm{d}t\\ &= \beta\int_0^h \mathbb{E}\Big\|\int_0^t \{\nabla V(X_s) - \nabla V(x)\}\,\mathrm{d}s\Big\|\,\mathrm{d}t \leqslant \beta^2\int_0^h\int_0^t \mathbb{E}\|X_s - x\|\,\mathrm{d}s\,\mathrm{d}t\\ &\lesssim \beta^2 h^2\,\{h\,\|\nabla V(x)\| + d^{1/2}h^{1/2}\}\,,\end{aligned}$$

where we again applied Lemma 4.1.8. □

Proof of Theorem 5.1.1 We apply Lemma 5.1.2 via Remark 5.1.4 and Lemma 5.1.6. In particular, for $h \ll 1/\beta\kappa^{1/2}$, we can take $L' = \exp(-\alpha h/2)$. By Lemma 4.0.1, we have the estimates $\bar{\mathscr{E}}_{\rm weak} \lesssim \beta^2 d^{1/2} h^{5/2}$ and $\bar{\mathscr{E}}_{\rm strong} \lesssim \beta d^{1/2} h^{3/2}$. The result follows after some algebraic simplifications.[1] □

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 (2011). 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.

[1] Check this with paper and pen.

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

$$\boldsymbol{\pi}(x, p) \propto \exp\bigl(-V(x) - \frac{1}{2}\,\|p\|^2\bigr)\,.$$

The first marginal of $\boldsymbol{\pi}$ is $\pi \propto \exp(-V)$, so a sample from $\boldsymbol{\pi}$ yields, upon projecting to the first coordinate, 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:[2]

$$\begin{aligned}\dot{x}_t &= \nabla_p H(x_t, p_t) = p_t\,,\\ \dot{p}_t &= -\nabla_x H(x_t, p_t) = -\nabla V(x_t)\,.\end{aligned}$$

Introducing the anti-symmetric 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 $(x_t, p_t)_{t \geqslant 0}$ evolve via Hamilton's equations started at $(x_0, p_0) \sim \boldsymbol{\pi}$,

$$\begin{aligned}\partial_t\big|_{t=0}\, \mathbb{E}\, f(x_t, p_t) &= \mathbb{E}\langle \nabla f(x_0, p_0), (\dot{x}_0, \dot{p}_0)\rangle = \int \langle \nabla f(x_0, p_0), (\dot{x}_0, \dot{p}_0)\rangle\, \boldsymbol{\pi}(\mathrm{d}x_0, \mathrm{d}p_0)\\ &= -\int f \operatorname{div}\Bigl(\begin{bmatrix} p \\ -\nabla V \end{bmatrix} \boldsymbol{\pi}\Bigr)\\ &= -\int f \Bigl\{\operatorname{div}\begin{bmatrix} p \\ -\nabla V \end{bmatrix} + \Bigl\langle \begin{bmatrix} p \\ -\nabla V \end{bmatrix}, \begin{bmatrix} \nabla_x \log \boldsymbol{\pi} \\ \nabla_p \log \boldsymbol{\pi} \end{bmatrix}\Bigr\rangle\Bigr\}\, \mathrm{d}\boldsymbol{\pi} = 0\,.\end{aligned}$$

Further properties of the Hamiltonian dynamics are explored in Exercise 5.5.

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 $\mathsf{KL}((F_t)_\# \boldsymbol{\mu} \,\|\, (F_t)_\# \boldsymbol{\pi}) = \mathsf{KL}(\boldsymbol{\mu} \,\|\, \boldsymbol{\pi})$ 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 $n = 0, 1, 2, \dots$:

1 *Refresh* the velocity by drawing $P'_{nT} \sim \mathsf{normal}(0, I_d)$.
2 *Integrate* Hamilton's equations: set $(X_{(n+1)T}, P_{(n+1)T}) := F_T(X_{nT}, P'_{nT})$.

---

[2] In contrast, Newton's law $\ddot{x}_t = -\nabla V(x_t)$ is a second-order ODE.

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, Chen and Vempala (2019))**.** *Assume that the target* $\pi \propto \exp(-V)$ *satisfies* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. *For* $n \in \mathbb{N}$, *let* π_{nT} *denote the law of the* n*-th iterate* X_{nT} *of ideal HMC with integration time* $T > 0$. *Then, if we set* $T = 1/2\sqrt{\beta}$, *we obtain*

$$W_2^2(\mu_{NT}, \pi) \leqslant \exp\Bigl(-\frac{N}{16\kappa}\Bigr)\, W_2^2(\mu_0, \pi)\,.$$

It is known that the convergence rate in this theorem is optimal, see Chen and Vempala (2019). 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* $(x_t, p_t)_{t \geqslant 0}$ *and* $(x_t', p_t')_{t \geqslant 0}$ *denote two solutions to Hamilton's equations of motion with* $p_0 = p_0'$ *and a potential* V *satisfying* $\|\nabla^2 V\|_{\mathrm{op}} \leqslant \beta$. *Then, for all* $t \in [0, 1/2\sqrt{\beta}]$, *it holds that*

$$\frac{1}{2}\,\|x_0 - x_0'\|^2 \leqslant \|x_t - x_t'\|^2 \leqslant 2\,\|x_0 - x_0'\|^2\,.$$

Proof First, note that $\partial_t \|p_t - p_t'\| \leqslant \|\nabla V(x_t) - \nabla V(x_t')\| \leqslant \beta\,\|x_t - x_t'\|$. It follows that $|\partial_t\,\|x_t - x_t'\|| \leqslant \|p_t - p_t'\| \leqslant \beta \int_0^t \|x_s - x_s'\|\,\mathrm{d}s$ and hence

$$\|x_t - x_t'\| \leqslant \|x_0 - x_0'\| + \beta \int_0^t \int_0^s \|x_r - x_r'\|\,\mathrm{d}r\,\mathrm{d}s\,.$$

Applying an ODE comparison lemma, one may deduce that $\|x_t - x_t'\| \leqslant \sqrt{2}\,\|x_0 - x_0'\|$. 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*

$$\beta\,\langle \nabla f(x) - \nabla f(y), x - y\rangle \geqslant \|\nabla f(x) - \nabla f(y)\|^2\,.$$

Proof See Exercise 5.6. □

We now prove the contraction result for the Hamiltonian dynamics.

Proposition 5.2.4 (Contraction of Hamilton's equations)**.** *Consider any two solutions* $(x_t, p_t)_{t\geqslant 0}$ *and* $(x'_t, p'_t)_{t\geqslant 0}$ *to Hamilton's equations of motion with* $p_0 = p'_0$ *and a potential* V *satisfying* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$*. Then, for all* $t \in [0, 1/2\sqrt{\beta}]$*, it holds that*

$$\|x_t - x'_t\|^2 \leqslant \exp\bigl(-\frac{\alpha t^2}{4}\bigr)\,\|x_0 - x'_0\|^2\,.$$

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

$$\rho_t \coloneqq \frac{\langle \nabla V(x_t) - \nabla V(x'_t),\, x_t - x'_t\rangle}{\|x_t - x'_t\|^2}\,.$$

To bound $\|p_t - p'_t\|^2$, we use $|\partial_t\|p_t - p'_t\|| \leqslant \|\nabla V(x_t) - \nabla V(x'_t)\|$. Also, by the coercivity lemma (Lemma 5.2.3),

$$\|\nabla V(x_t) - \nabla V(x'_t)\|^2 \leqslant \beta\,\langle \nabla V(x_t) - \nabla V(x'_t), x_t - x'_t\rangle = \beta\rho_t\,\|x_t - x'_t\|^2 \leqslant 2\beta\rho_t\,\|x_0 - x'_0\|^2$$

where we used Lemma 5.2.2. Hence, by the Cauchy–Schwarz inequality,

$$\|p_t - p'_t\|^2 \leqslant \Bigl|\int_0^t \bigl|\partial_s\|p_s - p'_s\|\bigr|\,\mathrm{d}s\Bigr|^2 \leqslant \Bigl|\int_0^t \sqrt{2\beta\rho_s}\,\|x_0 - x'_0\|\,\mathrm{d}s\Bigr|^2 \leqslant 2\beta t\,\|x_0 - x'_0\|^2 \int_0^t \rho_s\,\mathrm{d}s\,.$$

From this and Lemma 5.2.2, we deduce

$$\partial_t^2\|x_t - x'_t\|^2 \leqslant -\Bigl(\rho_t - 4\beta t\int_0^t \rho_s\,\mathrm{d}s\Bigr)\,\|x_0 - x'_0\|^2\,.$$

Integrating and using $\rho \geqslant 0$,

$$\begin{aligned}
\partial_t\|x_t - x'_t\|^2 &\leqslant -\Bigl(\int_0^t \rho_s\,\mathrm{d}s - 4\beta\int_0^t s\int_0^s \rho_r\,\mathrm{d}r\,\mathrm{d}s\Bigr)\,\|x_0 - x'_0\|^2\\
&\leqslant -\Bigl(\int_0^t \rho_s\,\mathrm{d}s - 2\beta t^2\int_0^t \rho_s\,\mathrm{d}s\Bigr)\,\|x_0 - x'_0\|^2\\
&= -(1 - 2\beta t^2)\,\Bigl(\int_0^t \rho_s\,\mathrm{d}s\Bigr)\,\|x_0 - x'_0\|^2 \leqslant -\frac{\alpha t}{2}\,\|x_0 - x'_0\|^2\,.
\end{aligned}$$

Integrating again then yields

$$\|x_t - x'_t\|^2 \leqslant \|x_0 - y_0\|^2 - \frac{\alpha t^2}{4}\,\|x_0 - x'_0\|^2 \leqslant \exp\bigl(-\frac{\alpha t^2}{4}\bigr)\,\|x_0 - x'_0\|^2\,. \qquad \square$$

If we choose $T = 1/2\sqrt{\beta}$, then the contraction factor is $\exp(-1/16\kappa) \leqslant 1 - 1/32\kappa$. In particular, let P denote the transition kernel of one step of ideal HMC with integration time T. We have the W_1 contraction

$$W_1\bigl(P((x,p),\cdot), P((x',p'),\cdot)\bigr) \leqslant \bigl(1 - \frac{1}{32\kappa}\bigr)\,\|(x,p) - (x',p')\|\,,$$

which, in the language of Section 2.6, says that the coarse Ricci curvature of P is bounded below by $\kappa/32$. Hence, by Theorem 2.6.6, we immediately obtain the following corollary.

Corollary 5.2.5. *Assume that the target* $\pi \propto \exp(-V)$ *satisfies* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. *Let* P *be the Markov kernel for ideal HMC with integration time* $T = {}^1\!/_{2\sqrt{\beta}}$. *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, Chen and Vempala (2019) also provide the following sampling guarantee.

Theorem 5.2.6 (Unadjusted HMC, Chen and Vempala (2019))**.** *Assume that the target* $\pi \propto \exp(-V)$ *satisfies* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ *and* $\nabla V(0) = 0$. *Then, there is a sampling algorithm based on a discretization of ideal HMC which outputs* $\hat{\mu}$ *satisfying* $\sqrt{\alpha}\, W_2(\hat{\mu}, \pi) \leqslant \varepsilon$ *using*

$$\widetilde{O}\Bigl(\frac{\kappa^{3/2} d^{1/2}}{\varepsilon}\Bigr) \qquad \text{gradient queries}\,.$$

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

$$\begin{aligned} \mathrm{d}X_t &= P_t\,\mathrm{d}t\,, \\ \mathrm{d}P_t &= -\nabla V(X_t)\,\mathrm{d}t - \gamma P_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,. \end{aligned} \tag{5.3.1}$$

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 5.3.2. Consider the following ODE system with $p_0 = 0$:

$$\begin{aligned} \dot{x}_t &= p_t\,, \\ \dot{p}_t &= -\nabla V(x_t) - \gamma p_t\,. \end{aligned}$$

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 $\gamma = 2\sqrt{\alpha}$,

$$V(x_t) - V(x_\star) \leqslant 2\exp(-\sqrt{\alpha}\, t)\,\{V(x_0) - V(x_\star)\}\,, \tag{5.3.3}$$

see Exercise 5.7. 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 discretization of this system to yield 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., Nesterov (2018, Section 2.2) for a discrete-time scheme which achieves $V(x_N) - V(x_\star) \leqslant \varepsilon^2$ 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 Nesterov (1983), and his algorithm is optimal amongst all first-order algorithms (Nemirovsky and Yudin, 1983). The continuous-time formulation of acceleration was introduced later, in Su et al. (2016).

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 incomplete.

Unlike the Langevin diffusion, the underdamped Langevin diffusion is *not* a reversible Markov process. Moreover, it is an example of non-trivial *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 (see Section 5.3.4).

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

$$\partial_t \boldsymbol{\pi}_t = \gamma \Delta_p \boldsymbol{\pi}_t + \operatorname{div}\Bigl(\boldsymbol{\pi}_t \begin{bmatrix} -p \\ \nabla V + \gamma p \end{bmatrix}\Bigr)$$

for the evolution of the law $(\boldsymbol{\pi}_t)_{t\geqslant 0}$ of the underdamped Langevin diffusion. We can write this as the continuity equation

$$\partial_t \boldsymbol{\pi}_t = \operatorname{div}\Bigl(\boldsymbol{\pi}_t \begin{bmatrix} \nabla_p \log \boldsymbol{\pi} \\ -\nabla_x \log \boldsymbol{\pi} - \gamma \nabla_p \log \boldsymbol{\pi} + \gamma \nabla_p \log \boldsymbol{\pi}_t \end{bmatrix}\Bigr)$$

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

$$\operatorname{div}\Bigl(\boldsymbol{\pi}_t \begin{bmatrix} -\nabla_p \log \boldsymbol{\pi}_t \\ \nabla_x \log \boldsymbol{\pi}_t \end{bmatrix}\Bigr) = 0\,,$$

we have the more interpretable expression

$$\partial_t \boldsymbol{\pi}_t = \operatorname{div}\Bigl(\boldsymbol{\pi}_t\, \boldsymbol{J}_\gamma \begin{bmatrix} \nabla_x \log(\boldsymbol{\pi}_t/\boldsymbol{\pi}) \\ \nabla_p \log(\boldsymbol{\pi}_t/\boldsymbol{\pi}) \end{bmatrix}\Bigr)\,, \qquad \boldsymbol{J}_\gamma := \begin{bmatrix} 0 & -1 \\ 1 & \gamma \end{bmatrix} \otimes I_d\,,$$

or

$$\partial_t \boldsymbol{\pi}_t = \operatorname{div}\bigl(\boldsymbol{\pi}_t\, \boldsymbol{J}_\gamma\, [\nabla_{W_2} \mathsf{KL}(\cdot \,\|\, \boldsymbol{\pi})](\boldsymbol{\pi}_t)\bigr)\,. \tag{5.3.4}$$

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 Cheng et al. (2018). Note that we use the same change of variables as in Exercise 5.7.

Theorem 5.3.5 (Contraction of underdamped Langevin). *Let $(X_t^0, P_t^0)_{t\geqslant 0}$ and $(X_t^1, P_t^1)_{t\geqslant 0}$ be two copies of the underdamped Langevin diffusion, driven by the same Brownian motion. Assume that the potential satisfies $0 \preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, defining the modified norm*

$$|||(x,p)|||^2 := \|x\|^2 + \Bigl\|x + \frac{2}{\gamma}\,p\Bigr\|^2$$

and for all $\gamma \geqslant \sqrt{2\beta}$, we obtain the contraction

$$|||(X_t^1, P_t^1) - (X_t^0, P_t^0)||| \leqslant \exp\Bigl(-\frac{\alpha t}{2\gamma}\Bigr)\, |||(X_0^1, P_0^1) - (X_0^0, P_0^0)|||\,.$$

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 1.1.19),

$$\begin{aligned} \mathrm{d}\bigl(\delta X_t + \frac{2}{\gamma}\,\delta P_t\bigr) &= \bigl[\delta P_t - \frac{2}{\gamma}\,\{\nabla V(X_t^1) - \nabla V(X_t^0)\} - 2\,\delta P_t\bigr]\,\mathrm{d}t \\ &= \Bigl[-\delta P_t - \frac{2}{\gamma}\,\underbrace{\Bigl(\int_0^1 \nabla^2 V\bigl((1-s)\,X_t^0 + s\,X_t^1\bigr)\,\mathrm{d}s\Bigr)}_{=:H_t}\,\delta X_t\Bigr]\,\mathrm{d}t \\ &= \Bigl[-\frac{\gamma}{2}\,\bigl(\delta X_t + \frac{2}{\gamma}\,\delta P_t\bigr) + \bigl(\frac{\gamma}{2} - \frac{2}{\gamma}\,H_t\bigr)\,\delta X_t\Bigr]\,\mathrm{d}t \end{aligned}$$

as well as

$$\mathrm{d}(\delta X_t) = \delta P_t\,\mathrm{d}t = \bigl[\frac{\gamma}{2}\,\bigl(\delta X_t + \frac{2}{\gamma}\,\delta P_t\bigr) - \frac{\gamma}{2}\,\delta X_t\bigr]\,\mathrm{d}t$$

so that

$$\begin{aligned} &\frac{1}{2}\,\partial_t\bigl\{\bigl\|\delta X_t + \frac{2}{\gamma}\,\delta P_t\bigr\|^2 + \|\delta X_t\|^2\bigr\} \\ &\quad = -\Bigl\langle \begin{bmatrix} \delta X_t + {}^2\!/\!\gamma\,\delta P_t \\ \delta X_t \end{bmatrix}, \begin{bmatrix} {}^\gamma\!/\!2\,I_d & {}^1\!/\!2\,({}^2\!/\!\gamma\,H_t - \gamma I_d) \\ {}^1\!/\!2\,({}^\gamma\!/\!2\,H_t - \gamma I_d) & {}^\gamma\!/\!2\,I_d \end{bmatrix} \begin{bmatrix} \delta X_t + ({}^2\!/\!\gamma)\,\delta P_t \\ \delta X_t \end{bmatrix} \Bigr\rangle\,. \end{aligned}$$

One checks that for $\gamma \geqslant \sqrt{2\beta}$, then the eigenvalues of the matrix above are lower bounded by ${}^\alpha\!/\!2\gamma$. □

We check that the new norm we defined is equivalent to the Euclidean norm.

Lemma 5.3.6 (Equivalence of norms)**.** *For all $x, p \in \mathbb{R}^d$ and $\gamma > 0$,*

$$\frac{1}{3}\left(\|x\|^2 + \frac{4}{\gamma^2}\|p\|^2\right) \leqslant |||(x,p)|||^2 \leqslant 3\left(\|x\|^2 + \frac{4}{\gamma^2}\|p\|^2\right).$$

Proof The upper bound follows from

$$|||(x,p)|||^2 = \left\|x + \frac{2}{\gamma}p\right\|^2 + \|x\|^2 \leqslant 2\left(\|x\|^2 + \frac{4}{\gamma^2}\|p\|^2\right) + \|x\|^2.$$

The lower bound follows from

$$\frac{4}{\gamma^2}\|p\|^2 \leqslant 2\left\|x + \frac{2}{\gamma}p\right\|^2 + 2\|x\|^2. \qquad \square$$

Consequently, the contraction result in Theorem 5.3.5 for $\gamma = \sqrt{2\beta}$ implies

$$\|X_t^1 - X_t^0\|^2 + \frac{2}{\beta}\|P_t^1 - P_t^0\|^2 \leqslant 9\exp\left(-\frac{\alpha t}{\sqrt{2\beta}}\right)\left(\|X_0^1 - X_0^0\|^2 + \frac{2}{\beta}\|P_0^1 - P_0^0\|^2\right).$$

***Remark* 5.3.7.** If we compare Theorem 5.3.5 with Optimization Box 5.3.2, 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 5.3.5 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 nh, then the rest of the SDE is a linear SDE, and can be integrated exactly. Namely, consider

$$\begin{aligned} \mathrm{d}\hat{X}_t &= \hat{P}_t\,\mathrm{d}t\,, \\ \mathrm{d}\hat{P}_t &= -\nabla V(\hat{X}_{nh})\,\mathrm{d}t - \gamma\hat{P}_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,, \end{aligned} \qquad \text{for } t \in [nh, (n+1)h]\,. \tag{ULMC}$$

Then, the solution to the SDE is given explicitly in the following lemma.

Lemma 5.3.8 (Implementation of ULMC)**.** *Conditioned on $(\hat{X}_{nh}, \hat{P}_{nh})$, the law of the next iterate $(\hat{X}_{(n+1)h}, \hat{P}_{(n+1)h})$ is explicitly given as $\mathsf{normal}(M_{(n+1)h}, \Sigma)$ where*

$$M_{(n+1)h} = \begin{bmatrix} \hat{X}_{nh} + \gamma^{-1}(1 - \exp(-\gamma h))\,\hat{P}_{nh} - \gamma^{-1}(h - \gamma^{-1}(1 - \exp(-\gamma h)))\,\nabla V(\hat{X}_{nh}) \\ \hat{P}_{nh}\exp(-\gamma h) - \gamma^{-1}(1 - \exp(-\gamma h))\,\nabla V(\hat{X}_{nh}) \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} \frac{2}{\gamma}\{h - \frac{2}{\gamma}(1 - \exp(-\gamma h)) + \frac{1}{2\gamma}(1 - \exp(-2\gamma h))\} & * \\ \frac{1}{\gamma}\{1 - 2\exp(-\gamma h) + \exp(-2\gamma h)\} & 1 - \exp(-2\gamma h) \end{bmatrix} \otimes I_d\,.$$

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 which is a refinement of Cheng et al. (2018) (and matches the guarantee obtained in Dalalyan and Riou-Durand, 2020).

Theorem 5.3.9 (Coupling analysis of ULMC)**.** *For $n \in \mathbb{N}$, let $\hat{\boldsymbol{\mu}}_{nh}$ denote the law of the n-th iterate of* ULMC *with step size $h \asymp \varepsilon/\sqrt{\beta\kappa d}$, friction parameter $\gamma = \sqrt{2\beta}$, and initialized at $\hat\mu_0 \otimes \mathsf{normal}(0, I_d)$. Also, let $\hat\mu_{nh}$ denote the law of $\hat X_{nh}$. Assume that the target $\pi \propto \exp(-V)$ satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, we obtain the guarantee $\sqrt{\alpha}\, W_2(\hat\mu_{Nh}, \pi) \leqslant \varepsilon$ after*

$$N = O\Big(\frac{\kappa^{3/2} d^{1/2}}{\varepsilon} \log \frac{\sqrt{\alpha}\, W_2(\hat\mu_0, \pi)}{\varepsilon}\Big) \qquad \textit{iterations}\,.$$

Proof We apply Remark 5.1.5, in the norm $|||\cdot|||$. Let $(\hat X_t, \hat P_t)_{t\geqslant 0}$ denote ULMC and let $(X_t, P_t)_{t\geqslant 0}$ denote the continuous-time underdamped Langevin diffusion, both driven by the same Brownian motion and started at (x, p). We want to bound the distance $\mathbb{E}[|||(\hat X_h, \hat P_h) - (X_h, P_h)|||^2]$. According to Lemma 5.3.6, it suffices to bound $\mathbb{E}[\|\hat X_h - X_h\|^2]$ and $\mathbb{E}[\|\hat P_h - P_h\|^2]$ separately.

First,

$$\mathbb{E}[\|\hat X_h - X_h\|^2] = \mathbb{E}\Big[\Big\|\int_0^h \{\hat P_t - P_t\}\,\mathrm{d}t\Big\|^2\Big] \leqslant h \int_0^h \mathbb{E}[\|\hat P_t - P_t\|^2]\,\mathrm{d}t\,.$$

Next,

$$\begin{aligned}\mathbb{E}[\|\hat P_t - P_t\|^2] &= \mathbb{E}\Big[\Big\|\int_0^t \{-\nabla V(x) + \nabla V(X_s) - \gamma\,(\hat P_s - P_s)\}\,\mathrm{d}s\Big\|^2\Big] \\ &\lesssim h \int_0^t \big(\mathbb{E}[\|\nabla V(X_s) - \nabla V(x)\|^2] + \gamma^2\, \mathbb{E}[\|\hat P_s - P_s\|^2]\big)\,\mathrm{d}s\,.\end{aligned}$$

By Grönwall's inequality, if $h \leqslant 1/\gamma = 1/\sqrt{2\beta}$, then

$$\mathbb{E}[\|\hat P_t - P_t\|^2] \lesssim h \int_0^t \mathbb{E}[\|\nabla V(X_s) - \nabla V(x)\|^2]\,\mathrm{d}s \leqslant \beta^2 h \int_0^t \mathbb{E}[\|X_s - x\|^2]\,\mathrm{d}t\,.$$

Again, we need a movement bound for the underdamped Langevin diffusion, which is done in Lemma 5.3.10. Substituting this in and assuming $h \ll 1/\beta^{1/2}$,

$$\begin{aligned}\mathbb{E}[|||(\hat X_h, \hat P_h) - (X_h, P_h)|||^2] &\lesssim \mathbb{E}[\|\hat X_h - X_h\|^2] + \frac{1}{\beta}\,\mathbb{E}[\|\hat P_h - P_h\|^2] \\ &\lesssim \beta h^4\,\|p\|^2 + \beta^{3/2} d h^5 + \beta h^6\,\|\nabla V(x)\|^2\,.\end{aligned}$$

This is a bound on $\mathscr{E}_{\mathrm{strong}}(x, p)$. We apply Remark 5.1.5 with $L = \exp(-\alpha h/\sqrt{8\beta})$ (by Theorem 5.3.5) and $L_{\mathrm{strong}} = O(\beta h^2)$, so that $L' \leqslant \exp(-\alpha h/4\sqrt{\beta})$ for $h \ll 1/\beta^{1/2}\kappa$. Then, since $\|\mathscr{E}_{\mathrm{strong}}\|_{L^2(\pi)} \lesssim \beta^{1/2} d^{1/2} h^2$, we obtain

$$\mathcal{W}_2(\hat{\boldsymbol{\mu}}_{Nh}, \boldsymbol{\pi}) \leqslant \exp\Big(-\frac{\alpha N h}{4\sqrt{\beta}}\Big)\, \mathcal{W}_2(\hat{\boldsymbol{\mu}}_0, \boldsymbol{\pi}) + O\Big(\frac{\beta d^{1/2} h}{\alpha}\Big)$$

where $\mathcal{W}_2$ denotes the 2-Wasserstein distance in the $|||\cdot|||$ norm. Choosing the step size appropriately completes the proof. □

The next lemma provides the movement bound (Exercise 5.9).

Lemma 5.3.10 (Underdamped movement bound)**.** *Let $(X_t, P_t)_{t \geqslant 0}$ denote the underdamped Langevin diffusion with potential V satisfying $-\beta I_d \preceq \nabla^2 V \preceq \beta I_d$. If $t \leqslant 1/\sqrt{\beta} \wedge 1/\gamma$, then*

$$\mathbb{E}[\|X_t - X_0\|^2] \lesssim t^2\, \mathbb{E}[\|P_0\|^2] + \gamma d t^3 + t^4\, \mathbb{E}[\|\nabla V(X_0)\|^2]\,.$$

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 Shen and Lee (2019); Yu et al. (2024), but we state a modified variant from Altschuler et al. (2025). We set, for $t \in [0, h]$,

$$\begin{aligned}
\hat X^+_{nh+t} &:= \hat X_{nh} + \frac{1-\exp(-\gamma t)}{\gamma}\, \hat P_{nh} - \frac{1}{\gamma}\,\Bigl(t - \frac{1-\exp(-\gamma t)}{\gamma}\Bigr)\, \nabla V(\hat X_{nh}) + \xi^{(1)}_{nh,t}\,,\\
\hat X_{(n+1)h} &:= \hat X_{nh} + \frac{1-\exp(-\gamma h)}{\gamma}\, \hat P_{nh} - \frac{1}{\gamma}\,\Bigl(h - \frac{1-\exp(-\gamma h)}{\gamma}\Bigr)\, \nabla V(\hat X^+_{nh+u_n h}) + \xi^{(1)}_{nh,h}\,,\\
\hat P_{(n+1)h} &:= \exp(-\gamma h)\, \hat P_{nh} - \frac{1}{\gamma}\,(1-\exp(-\gamma h))\, \nabla V(\hat X^+_{nh+v_n h}) + \xi^{(2)}_{nh,h}\,,
\end{aligned} \tag{RM–ULMC}$$

where $(u_n, v_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random pairs, independent of the Brownian motion, with the following marginal densities on $[0, 1]$: $u_n \sim q_u$ and $v_n \sim q_v$, where

$$q_u(u) = \frac{h\,(1-\exp(-\gamma\,(1-u)\,h))}{h - (1-\exp(-\gamma h))/\gamma}\,, \qquad q_v(v) = \frac{\gamma h \exp(-\gamma\,(1-v)\,h)}{1-\exp(-\gamma h)}\,.$$

Also, we define the following Brownian integrals:

$$\begin{aligned}
\xi^{(1)}_{nh,t} &:= \sqrt{\frac{2}{\gamma}} \int_{nh}^{nh+t} (1-\exp(-\gamma\,(nh+t-s)))\, \mathrm{d}B_s\,,\\
\xi^{(2)}_{nh,t} &:= \sqrt{2\gamma} \int_{nh}^{nh+t} \exp(-\gamma\,(nh+t-s))\, \mathrm{d}B_s\,.
\end{aligned}$$

Although the form of the algorithm seems complicated, it is designed to facilitate the computation of the local errors. Since the proof of the following theorem is a combination of the proof techniques introduced for the randomized midpoint discretization (Theorem 5.1.1) and the analysis of underdamped Langevin (Theorem 5.3.9), we leave it as an exercise.

Theorem 5.3.11 (RM–ULMC)**.** *For $n \in \mathbb{N}$, let $\hat\mu_{nh}$ denote the law of $\hat X_{nh}$ along RM–ULMC with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$. If we choose $h \asymp \varepsilon^{2/3}/\beta^{1/2} d^{1/3}$, then for all sufficiently small $\varepsilon > 0$ and with an appropriate initialization, we obtain the guarantee $\sqrt{\alpha}\, W_2(\hat\mu_{Nh}, \pi) \leqslant \varepsilon$ after*

$$N = \widetilde O\Bigl(\frac{\kappa d^{1/3}}{\varepsilon^{2/3}}\Bigr) \qquad \textit{iterations}\,.$$

See the Bibliographical Notes for further developments.

5.3.4 Hypocoercivity

Our study of the underdamped Langevin diffusion has produced samplers with substantially improved sampling complexity, culminating in Theorem 5.3.11, but we have not addressed the motivating question in this section. Namely, is there an *acceleration* phenomenon for sampling, in the sense of an algorithm with a square root dependence on the condition number? (See Optimization Box 5.3.2.)

Toward this question and others, such as whether we can obtain KL divergence guarantees under functional inequality assumptions (as we did in Section 4.2), we must now enter into a discussion of the hypocoercivity arguments which until this point we had deferred due to the higher level of technical sophistication.

We start by noting that the generator of the underdamped Langevin diffusion can be written (see Exercise 5.13) as $\mathscr{L} = \gamma \mathscr{L}_{\mathsf{OU}} + \mathscr{L}_{\mathsf{Ham}}$, where $\mathscr{L}_{\mathsf{OU}}$ is the generator of the Ornstein–Uhlenbeck process (Exercise 1.3) acting on the momentum coordinate,

$$\mathscr{L}_{\mathsf{OU}} f := \Delta_p f - \langle p, \nabla_p f\rangle\,,$$

and $\mathscr{L}_{\mathsf{Ham}}$ captures the Hamiltonian part of the dynamics,

$$\mathscr{L}_{\mathsf{Ham}} f := \langle p, \nabla_x f\rangle - \langle \nabla V, \nabla_p f\rangle\,.$$

As it turns out, $\mathscr{L}_{\mathsf{Ham}}$ is *anti-symmetric* in $L^2(\pi)$, i.e., $\mathscr{L}_{\mathsf{Ham}}^* = -\mathscr{L}_{\mathsf{Ham}}$ (Exercise 5.14). Therefore, if we write the Fokker–Planck equation for the underdamped Langevin diffusion in terms of the relative density $\rho_t := \frac{\mathrm{d}\pi_t}{\mathrm{d}\pi}$, we obtain the differential equation

$$\partial_t \rho_t = \mathscr{L}^* \rho_t\,, \qquad \mathscr{L}^* = \gamma \mathscr{L}_{\mathsf{OU}} - \mathscr{L}_{\mathsf{Ham}}\,. \tag{5.3.12}$$

Now suppose that we wish to prove convergence in the chi-squared divergence. The first approach to try is to simply differentiate $t \mapsto \|\rho_t - 1\|_{L^2(\pi)}^2$, as we did for the Langevin diffusion. We find that

$$\partial_t \|\rho_t - 1\|_{L^2(\pi)}^2 = 2\gamma\, \langle \rho_t - 1, \mathscr{L}_{\mathsf{OU}}\,(\rho_t - 1)\rangle_{L^2(\pi)} - 2\,\langle \rho_t - 1, \mathscr{L}_{\mathsf{Ham}}\,(\rho_t - 1)\rangle_{L^2(\pi)}\,.$$

Unfortunately, this naïve calculation cannot succeed. Due to the anti-symmetry of $\mathscr{L}_{\mathsf{Ham}}$, the second term vanishes. Moreover, the first term is not *coercive*, in the sense that the putative inequality $\langle \rho_t - 1, \mathscr{L}_{\mathsf{OU}}\,(\rho_t - 1)\rangle_{L^2(\pi)} \lesssim -\|\rho_t - 1\|_{L^2(\pi)}^2$ generally fails, because $\mathscr{L}_{\mathsf{OU}}$ has a large kernel: namely, all test functions which depend only on the position coordinate.

However, there is a certain intuition that the underdamped Langevin diffusion ought to converge anyway: the $\mathscr{L}_{\mathsf{OU}}$ operator is coercive when restricted to the momentum coordinate, and the Hamiltonian dynamics encoded in $\mathscr{L}_{\mathsf{Ham}}$ causes the position and momentum coordinates to "mix". Indeed, it is the *interaction* of the two parts of the dynamics that drives the system to equilibrium.

We now give an exposition to the theory of **hypocoercivity**, as treated in Villani's monograph (Villani, 2009a), which places the above intuition into a mathematical framework. The theory has two key ingredients. First, the "interaction" between the two parts of the dynamics is handled by studying the *commutator* of the two operators. Second, in order to take advantage of the non-zero commutator, we perform an analysis in a *modified norm* (similar in spirit to the twisted metric in the coupling argument of Theorem 5.3.5). We will establish the following theorem.

Theorem 5.3.13 (Hypocoercivity)**.** *For $t > 0$, let $f_t := \rho_t - 1 := \frac{\mathrm{d}\pi_t}{\mathrm{d}\pi} - 1$ denote the centered relative density along the underdamped Langevin diffusion with friction coefficient $\gamma = c_0\sqrt{\beta}$, $c_0 > 0$. Assume that measure $\pi \propto \exp(-V)$ on $\mathbb{R}^d$ satisfies a Poincaré inequality with constant α^{-1}, and that ∇V is β-Lipschitz. Then, there is a constant $c' > 0$ depending only on c_0 such that*

$$\begin{aligned} &\|f_t\|^2_{L^2(\boldsymbol{\pi})} + \frac{1}{\beta}\,\|\nabla_x f_t\|^2_{L^2(\boldsymbol{\pi})} + \|\nabla_p f_t\|^2_{L^2(\boldsymbol{\pi})} \\ &\quad\lesssim \exp\Bigl(-\frac{c'\alpha t}{\sqrt{\beta}}\Bigr)\,\bigl(\|f_0\|^2_{L^2(\boldsymbol{\pi})} + \frac{1}{\beta}\,\|\nabla_x f_0\|^2_{L^2(\boldsymbol{\pi})} + \|\nabla_p f_0\|^2_{L^2(\boldsymbol{\pi})}\bigr)\,. \end{aligned}$$

To better appreciate the core of the argument, we follow Villani (2009a) and switch to more abstract notation. Although we are only concerned here with the underdamped Langevin diffusion, some of the following discussion will be phrased more generally to emphasize its broader applicability. We observe that $\mathscr{L}_{\mathsf{OU}} = -\nabla_p^* \nabla_p$, which makes apparent its symmetry and non-positivity, and hence study an operator $\mathscr{L}$ of the form $\mathscr{L} = -\gamma A^* A + B$, where A and B are linear operators and B is anti-symmetric. In our application to the underdamped Langevin diffusion, this actually corresponds to the adjoint $\mathscr{L}^*$ of the generator, with

$$A := \nabla_p\,, \qquad B := -\mathscr{L}_{\mathsf{Ham}}\,, \qquad C := [A, B]\,.$$

Here, $[A, B] := AB - BA$ is the commutator. As a preliminary step, we must compute these commutators, which we leave as an important exercise (Exercise 5.14):

$$[A, A^*] = \mathrm{id}\,, \qquad C = -\nabla_x\,, \qquad [B, C] = \nabla^2 V\,\nabla_p = \nabla^2 V\,A\,. \tag{Com}$$

Here, $[A, A^*]$ is interpreted as $([\partial_{p_1}, \nabla_p^*], \dots, [\partial_{p_d}, \nabla_p^*])$. Also, C commutes with A and A^*. Finally, we let $\|\cdot\|$ denote a Hilbert norm, which is taken to be $\|\cdot\|_{L^2(\boldsymbol{\pi})}$ for underdamped Langevin.

Already, the commutator calculation proves to be interesting. Indeed, due to the anti-symmetry of B, the computation of the time derivative along the dynamics[3] induced by $\mathscr{L}$ yields $2\,\langle f, \mathscr{L} f\rangle = -2\gamma\,\|Af\|^2$, and the problem is that the quantity $\|Af\|$ is not coercive, in the sense that it does not control $\|f\|$ from above. In other words, AA^* may not be strictly positive definite. For underdamped Langevin, this corresponds to $\|Af\| = \|\nabla_p f\|_{L^2(\boldsymbol{\pi})}$ and so we are missing a term—namely, the norm of the position gradient. But this position gradient shows up precisely as the commutator C.

Motivated by this observation, the plan is to instead consider $\|f\|^2 + \|Af\|^2 + \|Cf\|^2$, which corresponds to the usual squared Sobolev norm $\|f\|^2_{L^2(\boldsymbol{\pi})} + \|\nabla f\|^2_{L^2(\boldsymbol{\pi})}$ for underdamped Langevin. Importantly, this quantity *is* coercive, as a consequence of the Poincaré inequality for $\boldsymbol{\pi}$ (which follows from the assumed Poincaré inequality for π, together with the Poincaré inequality for the standard Gaussian). In other words, rather than requiring AA^* to be positive definite, we now only require $AA^* + CC^*$ to be positive definite. In general, this process can be repeated: if $AA^* + CC^*$ is still not positive definite, we can take another commutator $C_1 = [C, B]$ and add $C_1 C_1^*$, etc., and the number of times this process must be repeated corresponds to the degeneracy of the equation. For underdamped Langevin, just the addition of CC^* is enough.

However, simply differentiating $t \mapsto \|f_t\|^2 + \|Af_t\|^2 + \|Cf_t\|^2$ is still not enough to establish convergence. The other crucial ingredient is to *change the norm*. Hence, we shall consider the following Lyapunov functional

---

[3] In the abstract setting, this refers to $t \mapsto \exp(\mathscr{L}t)\,f_0$, i.e., the semigroup with generator $\mathscr{L}$.

$$\mathcal{L}_t \coloneqq \|f_t\|^2 + a\,\|Af_t\|^2 - \frac{2b}{\sqrt{\beta}}\,\langle Af_t, Cf_t\rangle + \frac{c}{\beta}\,\|Cf_t\|^2\,.$$

We will choose $a, b, c > 0$ to be universal constants such that $ac > b^2$, which will ensure that the above Lyapunov functional is equivalent up to universal constants to the norm $\|f\|^2 + \|Af\|^2 + \beta^{-1}\,\|Cf\|^2$.

As Villani writes, the key intuition is that the effect of $\mathscr{L}_{\mathsf{OU}} = -A^*A$ is to dissipate the terms $\|f\|^2$, $\|Af\|^2$, and $\|Cf\|^2$, whereas the effect of B is to dissipate the mixed term $-\langle Af, Cf\rangle$. To see this, suppose that M is an operator which commutes with A. Then, if $\partial_t g_t = -A^*Ag_t$,

$$\partial_t \|Mg_t\|^2 = -2\,\langle Mg_t, MA^*Ag_t\rangle = -2\,\langle Mg_t, A^*AMg_t\rangle = -2\,\|AMg_t\|^2 \leqslant 0\,.$$

On the other hand, if $\partial_t g_t = Bg_t$, from anti-symmetry, the definition of C, and (Com),

$$\begin{aligned}
-\partial_t\langle Ag_t, Cg_t\rangle &= -\langle ABg_t, Cg_t\rangle - \langle Ag_t, CBg_t\rangle \\
&= -\langle ABg_t, Cg_t\rangle - \langle Ag_t, BCg_t\rangle - \langle Ag_t, [C,B]g_t\rangle \\
&= -\langle ABg_t, Cg_t\rangle + \langle BAg_t, Cg_t\rangle + \langle Ag_t, [B,C]g_t\rangle \\
&= -\|Cg_t\|^2 + \langle Ag_t, \nabla^2 V\, Ag_t\rangle\,.
\end{aligned}$$

Up to an error term, we have obtained decay in the missing "C direction"!

We now proceed to the full calculation.

Proof of Theorem 5.3.13 If we differentiate $\mathcal{L}_t$, we obtain $-\mathcal{D}(f_t)$, where we calculate

$$\mathcal{D}(f) = 2\gamma\,\|Af\|^2 + \underbrace{a\,(\cdots)}_{\mathrm{I}} + \underbrace{\frac{2b}{\sqrt{\beta}}\,(\cdots)}_{\mathrm{II}} + \underbrace{\frac{c}{\beta}\,(\cdots)}_{\mathrm{III}}\,.$$

Our aim is to lower bound $\mathcal{D}(f_t)$ by a multiple of $\mathcal{L}_t$. Following Villani's notation, we divide each of the terms I, II, and III with subscripts, I_A, I_B, etc., to signify the part of the time derivative which comes from $-A^*A$ or from B. We estimate each of these six terms. Based on the above intuition, we expect I_A, II_B, and III_A to contribute terms that help us, whereas the other terms are error terms. Throughout, we repeatedly use the commutator calculations in (Com).

We start with the first term.

$$\begin{aligned}
\mathrm{I}_A &= 2a\gamma\,\langle Af, AA^*Af\rangle = 2a\gamma\,(\langle Af, A^*A^2f\rangle + \|Af\|^2) = 2a\gamma\,(\|A^2f\|^2 + \|Af\|^2)\,, \\
\mathrm{I}_B &= -2a\,\langle Af, ABf\rangle = -2a\,\langle Af, BAf\rangle - 2a\,\langle Af, Cf\rangle = -2a\,\langle Af, Cf\rangle \\
&\geqslant -2a\,\|Af\|\,\|Cf\|\,.
\end{aligned}$$

Next, for the second term, by the above calculations,

$$\begin{aligned}
\mathrm{II}_A &= -\frac{2b\gamma}{\sqrt{\beta}}\,(\langle AA^*Af, Cf\rangle + \langle Af, CA^*Af\rangle) \\
&= -\frac{2b\gamma}{\sqrt{\beta}}\,(\langle A^*A^2f, Cf\rangle + \langle Af, Cf\rangle + \langle Af, A^*CAf\rangle) \\
&= -\frac{2b\gamma}{\sqrt{\beta}}\,(2\,\langle A^2f, ACf\rangle + \langle Af, Cf\rangle) \geqslant -\frac{2b\gamma}{\sqrt{\beta}}\,(2\,\|A^2f\|\,\|ACf\| + \|Af\|\|Cf\|) \\
\mathrm{II}_B &= \frac{2b}{\sqrt{\beta}}\,(\|Cf\|^2 - \langle Af, \nabla^2V\,Af\rangle) \geqslant \frac{2b}{\sqrt{\beta}}\,(\|Cf\|^2 - \beta\,\|Af\|^2)\,.
\end{aligned}$$

Finally, for the third term,

$$\begin{aligned}
\mathrm{III}_A &= \frac{2c\gamma}{\beta}\,\langle Cf, CA^* Af\rangle = \frac{2c\gamma}{\beta}\,\|ACf\|^2\,,\\
\mathrm{III}_B &= -\frac{2c}{\beta}\,\langle Cf, CBf\rangle = -\frac{2c}{\beta}\,\big(\langle Cf, BCf\rangle + \langle Cf, [C,B]f\rangle\big) = \frac{2c}{\beta}\,\langle Cf, \nabla^2 V\, Af\rangle\\
&\geqslant -2c\,\|Af\|\,\|Cf\|\,.
\end{aligned}$$

Let $D(f) \coloneqq (\|Af\|, \frac{1}{\sqrt{\beta}}\,\|Cf\|, \|A^2 f\|, \frac{1}{\sqrt{\beta}}\,\|ACf\|)$. Now assume that $\gamma = c_0\sqrt{\beta}$ for some constant $c_0 > 0$. We have established the inequality $\mathcal{D}(f) \geqslant \langle D(f), \Sigma\, D(f)\rangle$ where

$$\Sigma = \sqrt{\beta}\begin{bmatrix} 2c_0 + 2ac_0 - 2b & -2a - 2c & & \\ & 2b & & \\ & & 2ac_0 & -4bc_0 \\ & & & 2cc_0 \end{bmatrix}.$$

Now we want to show that we can choose constants $a, b, c > 0$ with $ac > b^2$, depending only on c_0, such that the symmetric part of Σ is positive definite. In this case, since Σ is block diagonal, this just amounts to $ac > b^2$ as well as $4b\,(c_0 + ac_0 - b) - (a+c)^2 > 0$. If we scale a, b, c down so that $\max\{a, b, c\} \ll c_0 \wedge 1$ then the second condition is clearly satisfied. With this, we have obtained

$$\partial_t \mathcal{L}_t = -\mathcal{D}(f_t) \lesssim -\sqrt{\beta}\,\|Af_t\|^2 - \frac{1}{\sqrt{\beta}}\,\|Cf_t\|^2\,.$$

Applying the Poincaré inequality for π (in the norm given by $\|(x,p)\|^2 \coloneqq \|x\|^2 + \|p\|^2/\alpha$),

$$\partial_t \mathcal{L}_t \lesssim -\frac{\alpha}{\sqrt{\beta}}\,\big(\|Af_t\|^2 + \frac{1}{\alpha}\,\|Cf_t\|^2\big) \leqslant -\frac{\alpha}{2\sqrt{\beta}}\,\big(\|f_t\|^2 + \|Af_t\|^2 + \frac{1}{\alpha}\,\|Cf_t\|^2\big) \lesssim -\frac{\alpha}{\sqrt{\beta}}\,\mathcal{L}_t\,.$$

In the last line, we used the fact that $\mathcal{L}_t$ is equivalent up to constants to the Sobolev norm $\|f\|^2 + \|Af\|^2 + \frac{1}{\beta}\,\|Cf\|^2$. □

Remarkably, there is a variant of this theorem which holds for KL convergence, which proceeds via a twisted version of the Fisher information (see Exercise 5.15 for a challenge).

Theorem 5.3.14 (Entropic hypocoercivity)**.** *In the setting of Theorem 5.3.13, suppose instead that π satisfies a log-Sobolev inequality with constant α^{-1}. Then, there is a Lyapunov function*

$$\mathcal{L}_t \coloneqq \mathsf{KL}(\pi_t \,\|\, \pi) + \mathbb{E}_{\pi_t}\Big\langle \nabla \log\frac{\pi_t}{\pi}, \Big(\begin{bmatrix} a/\beta & b/\sqrt{\beta} \\ b/\sqrt{\beta} & c\end{bmatrix} \otimes I_d\Big)\, \nabla\log\frac{\pi_t}{\pi}\Big\rangle$$

such that for a universal constant $c' > 0$ and all $t \geqslant 0$,

$$\mathcal{L}_t \leqslant \exp\Big(-\frac{c'\alpha t}{\sqrt{\beta}}\Big)\,\mathcal{L}_0\,.$$

Nevertheless, Theorem 5.3.13 and Theorem 5.3.14 only yield *unaccelerated* rates of convergence.

5.3.5 Second-order lifts and a space-time Poincaré inequality

Very recently, through different techniques (based on a *space-time Poincaré inequality*), Cao, Lu, and Wang obtained the following remarkable result.

Theorem 5.3.15 (Space-time Poincaré inequality, Cao et al. (2023))**.** *Suppose that π is* log-concave *and satisfies a Poincaré inequality with constant α^{-1}, and set $\gamma = \sqrt{\alpha}$. There is a universal constant $c > 0$ such that for all $t \geqslant 0$, for underdamped Langevin,*

$$\chi^2(\boldsymbol{\pi}_t \,\|\, \boldsymbol{\pi}) \lesssim \exp(-c\sqrt{\alpha}\,t)\,\chi^2(\boldsymbol{\pi}_0 \,\|\, \boldsymbol{\pi})\,.$$

Note that this is indeed the accelerated rate! Unfortunately, this is a "Poincaré-type" result, which leads to a worse dependence on the dimension (see the discussion in Section 1.5). It is not known how to prove a "log-Sobolev" version of this result. Moreover, the current discretization methods are lossy, and therefore unable to fully carry over the acceleration effect into discrete time. Recently, Altschuler et al. (2025) used Theorem 5.3.15 to give a guarantee that RM–ULMC can achieve ε^2 error in KL divergence in $\widetilde{O}(\kappa^{5/6} d^{1/3}/\varepsilon^{2/3})$ iterations, assuming a warm start condition $\chi^2(\boldsymbol{\mu}_0 \,\|\, \boldsymbol{\pi}) = \operatorname{poly}(d)$. The $\kappa^{5/6}$ dependence on the condition number is the current state-of-the-art progress toward acceleration for sampling; see the Bibliographical Notes for further discussion.

In the remainder of this section, we present most of the proof of Theorem 5.3.15 (omitting the proof of an important yet technical PDE result, known as the divergence lemma). We follow the second-order lifting framework of Eberle and Lörler (2024a), which we now introduce.

Definition 5.3.16 (Second-order lift)**.** Let $(P_t)_{t\geqslant 0}$ be a Markov semigroup over a space $\mathcal{X}$, with generator $\mathscr{L}$, Dirichlet form $\mathscr{E}$, and which is reversible with respect to a probability measure π. Let $(\boldsymbol{P}_t)_{t\geqslant 0}$ be a Markov semigroup over a space $\mathcal{X} \times \mathcal{P}$, with generator $\boldsymbol{\mathscr{L}}$ and stationary distribution $\boldsymbol{\pi}$. We denote by $\Pi : \mathcal{X} \times \mathcal{P} \to \mathcal{X}$ the projection map $(x, p) \mapsto x$.

We say that $(\boldsymbol{P}_t)_{t\geqslant 0}$ is a **second-order lift** of $(P_t)_{t\geqslant 0}$ if $\Pi_\# \boldsymbol{\pi} = \pi$, and for all $f, g \in \operatorname{dom}\mathscr{L}$,

$$\int (f \circ \Pi)\,\boldsymbol{\mathscr{L}}(g \circ \Pi)\,\mathrm{d}\boldsymbol{\pi} = 0\,, \qquad \int \boldsymbol{\mathscr{L}}(f \circ \Pi)\,\boldsymbol{\mathscr{L}}(g \circ \Pi)\,\mathrm{d}\boldsymbol{\pi} = \mathscr{E}(f, g)\,.$$

Proposition 5.3.17 (Underdamped Langevin as a second-order lift)**.** *The underdamped Langevin diffusion* (5.3.1) *is a second-order lift of the Langevin diffusion* (1.0.1)*.*

Proof Here, $\boldsymbol{\mathscr{L}} = \mathscr{L}_{\mathsf{Ham}} + \gamma\mathscr{L}_{\mathsf{OU}}$. Note that $\mathscr{L}_{\mathsf{OU}}$ only acts on the momentum coordinate, whereas the definition of a second-order lift only cares about functions of the position variable; moreover, both $\mathscr{L}_{\mathsf{Ham}}$ and $\mathscr{L}_{\mathsf{OU}}$ leave $\boldsymbol{\pi}$ stationary. Hence, it suffices to check that $\mathscr{L}_{\mathsf{Ham}}$ is a second-order lift of $\mathscr{L}$, where $\mathscr{L}$ is the Langevin generator.

The first condition is true because

$$\int (f \circ \Pi)\,\mathscr{L}_{\mathsf{Ham}}(g \circ \Pi)\,\mathrm{d}\boldsymbol{\pi} = \int (f \circ \Pi)\,\langle p, \nabla_x g\rangle\,\mathrm{d}\boldsymbol{\pi} = 0\,,$$

by integrating over p first. For the second condition,

$$\int \mathscr{L}_{\mathsf{Ham}}(f \circ \Pi)\,\mathscr{L}_{\mathsf{Ham}}(g \circ \Pi)\,\mathrm{d}\boldsymbol{\pi} = \int \langle p, \nabla_x f\rangle\,\langle p, \nabla_x g\rangle\,\mathrm{d}\boldsymbol{\pi} = \int \langle \nabla_x f, \nabla_x g\rangle\,\mathrm{d}\pi\,,$$

which is the Dirichlet form for the Langevin diffusion. □

The general idea here is that starting with a reversible Markov process, we can lift it to a new Markov process, which is usually non-reversible and can speed up convergence. Interestingly, we can prove a lower bound on the relaxation time of *any* second-order lift, showing that a square root speed up is the best possible.

Theorem 5.3.18 (Relaxation time lower bound for second-order lifts, Eberle and Lörler (2024a))**.** *Let* $(\boldsymbol{P}_t)_{t\geqslant 0}$ *be a second-order lift of* $(P_t)_{t\geqslant 0}$*. Define the relaxation time of* $\mathscr{L}$ *to be*

$$\mathsf{relax}(\mathscr{L}) := \inf\Bigl\{t \geqslant 0 \Bigm| \|P_t f\|_{L^2(\pi)} \leqslant \mathrm{e}^{-1}\, \|f\|_{L^2(\pi)} \text{ for all } f \in L^2(\pi) \text{ with } \int f \, \mathrm{d}\pi = 0\Bigr\}\,,$$

and similarly define the relaxation time of $\boldsymbol{\mathscr{L}}$*. Then,*

$$\mathsf{relax}(\boldsymbol{\mathscr{L}}) \gtrsim \sqrt{\mathsf{relax}(\mathscr{L})}\,.$$

See Exercise 5.16. Combined with Theorem 5.3.15, this justifies that the underdamped Langevin diffusion is an *optimal* second-order lift of the Langevin diffusion.

We now proceed to outline Theorem 5.3.15. From now on, we specialize to the setting in which $\boldsymbol{\mathscr{L}}$ is the generator of the underdamped Langevin diffusion, and write $\boldsymbol{\pi} = \pi \otimes \rho$ where $\rho := \mathsf{normal}(0, I_d)$. Let $\mathfrak{m}_T$ denote the Lebesgue measure over $[0, T]$, and for $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ write

$$\mathscr{E}_T(f) := \int_0^T \mathscr{E}(f(t,\cdot))\, \mathrm{d}t = \int_0^T \|\nabla_x f(t,\cdot)\|^2_{L^2(\pi)}\, \mathrm{d}t\,.$$

Similarly, for $f : \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, write

$$\boldsymbol{\mathscr{E}}_T(f) := \int_0^T \boldsymbol{\mathscr{E}}(f(t,\cdot))\, \mathrm{d}t = \gamma \int_0^T \|\nabla_p f(t,\cdot)\|^2_{L^2(\boldsymbol{\pi})}\, \mathrm{d}t\,.$$

For a function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we let Π denote the momentum averaging operator, $\Pi g(x) := \int g(x,p)\, \rho(\mathrm{d}p)$. We abuse notation by adding or removing arguments from functions as necessary: for instance, if $g : \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we can apply the momentum averaging operator to obtain $\Pi g(t,x) := \int g(t,x,p)\, \rho(\mathrm{d}p)$; and we can further think of Πg as a function defined on $\mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}^d$ via $\Pi g(t,x,p) := \Pi g(t,x)$, i.e., it is constant with respect to the momentum variable.

We begin with preliminary calculations.

Lemma 5.3.19 (Preliminary calculations)**.**

1 *Assume that* π *is log-concave. For functions* $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *and* $g : \mathbb{R}^d \to \mathbb{R}$ *(depending on the position coordinate only),*

$$|\langle \mathscr{L}_{\mathsf{Ham}}(f - \Pi f),\, \mathscr{L}_{\mathsf{Ham}} g\rangle_{L^2(\boldsymbol{\pi})}| \leqslant \frac{2}{\sqrt{\gamma}} \sqrt{\boldsymbol{\mathscr{E}}(f)}\, \|\mathscr{L} g\|_{L^2(\pi)}\,.$$

2 *(**Divergence lemma**) Assume that* π *satisfies a Poincaré inequality with constant* $1/\alpha$*. For any* $f \in L^2(\mathfrak{m}_T \otimes \pi)$ *with* $\int f\, \mathrm{d}(\mathfrak{m}_T \otimes \pi) = 0$*, there exist functions* g*,* h *over* $[0,T] \times \mathbb{R}^d$ *with* $g(0,\cdot) = g(T,\cdot) = h(0,\cdot) = h(T,\cdot) = 0$ *such that* $f = \partial_t g + \mathscr{L} h$*, satisfying the estimates*

$$\sqrt{\mathscr{E}_T(g)} + \|\mathscr{L} h\|_{L^2(\mathfrak{m}_T \otimes \pi)} \lesssim \|f\|_{L^2(\mathfrak{m}_T \otimes \pi)}\,, \qquad \sqrt{\mathscr{E}_T(\partial_t h)} \lesssim \Bigl(1 + \frac{1}{\sqrt{\alpha}\, T}\Bigr)\, \|f\|_{L^2(\mathfrak{m}_T \otimes \pi)}\,.$$

Proof

1. The Gaussian Poincaré inequality implies

$$\begin{aligned}|\langle \mathscr{L}_{\mathsf{Ham}}(f-\Pi f),\, \mathscr{L}_{\mathsf{Ham}} g\rangle_{L^2(\pi)}| &= |\langle f-\Pi f,\, (\mathrm{id}-\Pi)\mathscr{L}_{\mathsf{Ham}}^2 g\rangle_{L^2(\pi)}| \\ &\leqslant \|f-\Pi f\|_{L^2(\pi)}\, \|(\mathrm{id}-\Pi)\mathscr{L}_{\mathsf{Ham}}^2 g\|_{L^2(\pi)} \\ &\leqslant \sqrt{\frac{\mathscr{E}(f)}{\gamma}}\, \|(\mathrm{id}-\Pi)\mathscr{L}_{\mathsf{Ham}}^2 g\|_{L^2(\pi)}\,.\end{aligned}$$

   A computation shows that $\mathscr{L}_{\mathsf{Ham}}^2 g = \langle p, \nabla_x^2 g\, p\rangle - \langle \nabla V, \nabla_x g\rangle$, so that

$$\begin{aligned}\|(\mathrm{id}-\Pi)\mathscr{L}_{\mathsf{Ham}}^2 g\|_{L^2(\pi)}^2 &= \int \mathrm{var}_\rho(\langle p, \nabla_x^2 g\, p\rangle - \langle \nabla V, \nabla_x g\rangle)\,\mathrm{d}\pi \leqslant 4\int \|\nabla_x^2 g\, p\|^2\,\mathrm{d}\boldsymbol{\pi} \\ &= 4\int \|\nabla_x^2 g\|_{\mathrm{HS}}^2\,\mathrm{d}\pi = 4\int \{\Gamma_2(g) - \langle \nabla_x g, \nabla^2 V\,\nabla_x g\rangle\}\,\mathrm{d}\pi \leqslant 4\,\|\mathscr{L} g\|_{L^2(\pi)}^2\,.\end{aligned}$$

2. The divergence lemma is a technical result in PDE theory which we do not prove here, but we discuss some considerations. A first approach would be to solve the PDE $\overline{\mathscr{L}} u = (-\partial_t^2 + \mathscr{L})u = f$ on $[0,T]\times\mathbb{R}_+$ with Neumann boundary conditions $\partial_t u(0,\cdot) = \partial_t u(T,\cdot) = 0$, and then taking $g := -\partial_t u$, $h := u$. With these choices, g satisfies the correct boundary conditions, but h does *not*. Nevertheless, if this approach were valid, then we could establish the desired estimates using standard arguments. Recall from (2.2.5) that

$$\int \|\nabla^2 h\|_{\mathrm{HS}}^2\,\mathrm{d}\pi = \int \{\Gamma_2(h) - \langle \nabla h, \nabla^2 V\,\nabla h\rangle\}\,\mathrm{d}\pi = \int \{(\mathscr{L} h)^2 - \langle \nabla h, \nabla^2 V\,\nabla h\rangle\}\,\mathrm{d}\pi\,. \tag{5.3.20}$$

   For the operator $\overline{\mathscr{L}}$, we have a similar formula

$$\int \|\overline{\nabla}^2 u\|_{\mathrm{HS}}^2\,\mathrm{d}(\mathfrak{m}_T\otimes\pi) = \int \{(\overline{\mathscr{L}} u)^2 - \langle \nabla_x u, \nabla^2 V\,\nabla_x u\rangle\}\,\mathrm{d}(\mathfrak{m}_T\otimes\pi) + \text{non-pos. term}\,,$$

   where $\overline{\nabla} = (\partial_t, \nabla_x)$ denotes the gradient operator on $[0,T]\times\mathbb{R}^d$, and the last term, non-positive, arises due to the boundary. If we suppose, for the sake of argument, that $\nabla^2 V \ge 0$, then we have $\int \|\overline{\nabla}^2 u\|_{\mathrm{HS}}^2\,\mathrm{d}(\mathfrak{m}_T\otimes\pi) \leqslant \|f\|_{L^2(\mathfrak{m}_T\otimes\pi)}^2$, which yields various estimates. For example,

$$\mathscr{E}_T(g) = \mathscr{E}_T(\partial_t h) = \|\nabla_x \partial_t u\|_{L^2(\mathfrak{m}_T\otimes\pi)}^2\,,$$

   so these terms would be controlled by $\|\overline{\nabla}^2 u\|_{L^2(\mathfrak{m}_T\otimes\pi)}^2$, and $\|\mathscr{L} h\|_{L^2(\mathfrak{m}_T\otimes\pi)}^2$ would be controlled with the help of (5.3.20).

   So, the issue is that the choices above do not fully satisfy the boundary conditions. Observe that if we also had $h(0,\cdot) = h(T,\cdot) = 0$, it would imply that f is orthogonal to the space $\mathbb{H}_0$ of functions v satisfying the PDE $(-\partial_t^2 + \mathscr{L})v = 0$. This motivates performing an orthogonal decomposition of f into the part that belongs to $\mathbb{H}_0$, and the part that belongs to $\mathbb{H}_0^\perp$. We refer to Eberle and Lörler (2024a) for a full proof.

□

Proof of Theorem 5.3.15 Our goal is to prove the *time-averaged* decay

$$\int_0^T \|\boldsymbol{P}_t^* f_0\|_{L^2(\pi)}^2\,\mathrm{d}t \leqslant C_T \int_0^T \mathscr{E}(\boldsymbol{P}_t^* f_0, \boldsymbol{P}_t^* f_0)\,\mathrm{d}t = \gamma C_T \int_0^T \|\nabla_p \boldsymbol{P}_t^* f_0\|_{L^2(\pi)}^2\,\mathrm{d}t\,, \tag{5.3.21}$$

for a suitable choice of $T > 0$ and for all f_0 with $\int f_0 \, \mathrm{d}\boldsymbol{\pi} = 0$. In turn, this implies (Exercise 5.17)

$$\chi^2(\boldsymbol{\pi}_t \parallel \boldsymbol{\pi}) \leqslant \exp\Bigl(-\frac{2\,(t-T)}{C_T}\Bigr)\,\chi^2(\boldsymbol{\pi}_0 \parallel \boldsymbol{\pi})\,. \tag{5.3.22}$$

Write $f(t, x, p) \coloneqq \boldsymbol{P}_t^* f_0(x, p)$. We start with the decomposition

$$\|f\|^2_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} = \|\Pi f\|^2_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} + \|f - \Pi f\|^2_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})}$$

and we control each term in turn. The second term is easier, as we can apply the Poincaré inequality for the Gaussian measure ρ:

$$\|f - \Pi f\|^2_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} \leqslant \|\nabla_p f\|^2_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} = \frac{\mathscr{E}_T(f)}{\gamma}\,.$$

For the first term, we apply the divergence lemma (Lemma 5.3.19) to write $-\Pi f = \partial_t g + \mathscr{L} h$ for suitable functions g, h. We can freely integrate by parts without boundary terms due to the properties of g and h, so

$$\begin{aligned} \|\Pi f\|^2_{L^2(\mathfrak{m}_T \otimes \pi)} &= -\langle \Pi f, \partial_t g + \mathscr{L} h\rangle_{L^2(\mathfrak{m}_T \otimes \pi)} = \langle \partial_t \Pi f, g\rangle_{L^2(\mathfrak{m}_T \otimes \pi)} + \mathscr{E}_T(\Pi f, h) \\ &= \langle (\partial_t + \mathscr{L}_{\mathsf{Ham}}) \Pi f,\, g + \mathscr{L}_{\mathsf{Ham}} h\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})}\,, \end{aligned}$$

where the last line follows from the second-order lift property.

We further write $\Pi f = f + \Pi f - f$ and bound terms separately. Using the dynamic equation for f, $\partial_t f = (-\mathscr{L}_{\mathsf{Ham}} + \gamma \mathscr{L}_{\mathsf{OU}}) f$, and $\langle \mathscr{L}_{\mathsf{OU}} f, g\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} = 0$ since g does not depend on p. So,

$$\langle (\partial_t + \mathscr{L}_{\mathsf{Ham}}) f,\, g + \mathscr{L}_{\mathsf{Ham}} h\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} = \gamma\,\langle \mathscr{L}_{\mathsf{OU}} f,\, g + \mathscr{L}_{\mathsf{Ham}} h\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} = \gamma\,\langle \mathscr{L}_{\mathsf{OU}} f,\, \mathscr{L}_{\mathsf{Ham}} h\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})}$$

and

$$\begin{aligned} \langle \mathscr{L}_{\mathsf{OU}} f,\, \mathscr{L}_{\mathsf{Ham}} h\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} &= -\langle \nabla_p^* \nabla_p f,\, p \cdot \nabla_x h\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} = -\langle \nabla_p f, \nabla_x h\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} \\ &\leqslant \|\nabla_p f\|_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})}\, \|\nabla_x h\|_{L^2(\mathfrak{m}_T \otimes \pi)} = \sqrt{\frac{\mathscr{E}_T(f)}{\gamma} \int_0^T \langle h, (-\mathscr{L})\, h\rangle_{L^2(\pi)}} \\ &\leqslant \sqrt{\frac{\mathscr{E}_T(f)}{\alpha\gamma}}\, \|\mathscr{L} h\|_{L^2(\mathfrak{m}_T \otimes \pi)} \lesssim \sqrt{\frac{\mathscr{E}_T(f)}{\alpha\gamma}}\, \|\Pi f\|_{L^2(\mathfrak{m}_T \otimes \pi)}\,. \end{aligned} \tag{5.3.23}$$

For the other term, we apply Lemma 5.3.19 to obtain

$$\begin{aligned} &\langle (\partial_t + \mathscr{L}_{\mathsf{Ham}})(\Pi f - f),\, g + \mathscr{L}_{\mathsf{Ham}} h\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} \\ &\quad = \langle f - \Pi f,\, \partial_t g + \mathscr{L}_{\mathsf{Ham}} \partial_t h + \mathscr{L}_{\mathsf{Ham}} g\rangle_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} + \frac{2}{\sqrt{\gamma}} \sqrt{\mathscr{E}_T(f)}\, \|\mathscr{L} h\|_{L^2(\mathfrak{m}_T \otimes \pi)} \\ &\quad \lesssim \|f - \Pi f\|_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})}\, (\|\mathscr{L}_{\mathsf{Ham}} \partial_t h\|_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})} + \|\mathscr{L}_{\mathsf{Ham}} g\|_{L^2(\mathfrak{m}_T \otimes \boldsymbol{\pi})}) + \sqrt{\frac{\mathscr{E}_T(f)}{\gamma}}\, \|\Pi f\|_{L^2(\mathfrak{m}_T \otimes \pi)} \\ &\quad \leqslant \sqrt{\frac{\mathscr{E}_T(f)}{\gamma}}\, \bigl(\sqrt{\mathscr{E}_T(g)} + \sqrt{\mathscr{E}_T(\partial_t h)} + \|\Pi f\|_{L^2(\mathfrak{m}_T \otimes \pi)}\bigr) \lesssim \bigl(1 + \frac{1}{\sqrt{\alpha}\, T}\bigr) \sqrt{\frac{\mathscr{E}_T(f)}{\gamma}}\, \|\Pi f\|_{L^2(\mathfrak{m}_T \otimes \pi)}\,. \end{aligned}$$

We have shown that

$$\|\Pi f\|_{L^2(\mathfrak{m}_T \otimes \pi)} \lesssim \frac{1}{\sqrt{\gamma}}\, \bigl(1 + \frac{\gamma}{\sqrt{\alpha}} + \frac{1}{\sqrt{\alpha}\, T}\bigr) \sqrt{\mathscr{E}_T(f)}$$

and hence

$$\|f\|_{L^2(\mathfrak{m}_T\otimes\pi)} \lesssim \frac{1}{\sqrt{\gamma}}\,\bigl(1+\frac{\gamma}{\sqrt{\alpha}}+\frac{1}{\sqrt{\alpha}\,T}\bigr)\,\sqrt{\mathscr{E}_T(f)}\,.$$

If we choose $\gamma \asymp 1/T \asymp \sqrt{\alpha}$, this verifies (5.3.21) with $C_T \asymp 1/\sqrt{\alpha}$. $\square$

Bibliographical Notes

The local error framework is classical (Milstein and Tretyakov, 2021) and has been widely used for the analysis of sampling algorithms (Erdogdu et al., 2018; Li et al., 2019; Sanz-Serna and Zygalakis, 2021; Li et al., 2022a,b). The improved complexity for LMC in Exercise 5.2 was obtained in Li et al. (2022b), whereas the rate when $\nabla^2 V$ is Lipschitz in the operator norm goes back to earlier works.

HMC and its variants are some of the most popular algorithms employed in practice, especially the **no-U-turn sampler** (**NUTS**) (Hoffman and Gelman, 2014) which adaptively sets the integration time. In terms of complexity analysis, the paper Chen and Vempala (2019) (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 Mangoubi and Vishnoi (2018); Mangoubi and Smith (2019); Bou-Rabee et al. (2020). Discretization of Itô SDEs with multiplicative noise, as in Exercise 5.3, was considered in Erdogdu et al. (2018), in which such diffusions were used to establish some global non-convex optimization guarantees. An improved rate in KL divergence, under higher-order smoothness, was obtained in Li et al. (2025).

The underdamped Langevin diffusion has been studied quantitatively in Cheng et al. (2018); Eberle et al. (2019); Dalalyan and Riou-Durand (2020); Ma et al. (2021); Zhang et al. (2023); Kim et al. (2025); Lehec (2025).

Besides randomized midpoint, the (idealized) **shifted ODE algorithm** (Foster et al., 2021) also achieves iteration complexity of $\widetilde{O}(\kappa d^{1/3}/\varepsilon^{2/3})$ when applied to the underdamped Langevin diffusion. However, implementing the ODE requires the use of a numerical integrator, and it is currently not known how to preserve the rate when fully discretized. Moreover, the rate $\widetilde{O}(\kappa d^{1/3}/\varepsilon^{2/3})$ is an optimal rate of discretization of the underdamped Langevin diffusion on the level of trajectories (Cao et al., 2021); see the Bibliographical Notes in Chapter 9.

Recent works have sharpened the results of Theorem 5.1.1 and Theorem 5.3.11. First, Kandasamy and Nagaraj (2024) obtained KL divergence guarantees (which are stronger than W_2 guarantees) in the LSI setting, with rates $\widetilde{O}(\kappa^{3/2}d^{3/4}/\varepsilon)$ for RM–LMC and $\widetilde{O}(\kappa^{17/12}d^{5/12}/\varepsilon^{1/2})$ for RM–ULMC. Then, Altschuler and Chewi (2024c) developed a KL version of the local error framework, leading to a KL divergence rate of $\widetilde{O}(\kappa d^{1/2}/\varepsilon)$ for RM–LMC in the strongly log-concave setting, as well as results for the LSI and weakly log-concave ($\nabla^2 V \succeq 0$) settings. The KL local error framework was extended to the underdamped Langevin diffusion in Altschuler et al. (2025), leading to the KL divergence rate $\widetilde{O}(\kappa d^{1/3}/\varepsilon^{2/3})$ for RM–ULMC in the strongly log-concave setting, again with further results in the LSI and weakly log-concave settings. Combined with the space-time Poincaré inequality (Theorem 5.3.15), it yields an iteration complexity of $\widetilde{O}\bigl(({}^{\kappa^{5/6}\,(d^{1/3}+\log^{1/3}(1+\chi^2))}/{}_{\varepsilon^{2/3}})\log^{4/3}\chi^2\bigr)$, where $\chi^2 \coloneqq \chi^2(\boldsymbol{\mu}_0 \,\|\, \boldsymbol{\pi})$. Finally, building on Kandasamy and Nagaraj (2024), Srinivasan and Nagaraj (2025) established the W_2 rates $\widetilde{O}((\kappa^{4/3}d^{1/3} + \kappa d^{2/3})/\varepsilon^{2/3})$ for RM–LMC, and roughly $\widetilde{O}(\kappa^{7/6}d^{1/3}/\varepsilon^{1/3})$ for RM–ULMC in the strongly log-concave setting.

The work of Zhang (2025) uses an anticipating Girsanov theorem and obtains KL and Rényi divergence bounds for non-adapted interpolations of discretizations, such as midpoint methods. Although this technique is not able to take advantage of the smaller weak error of randomized midpoint, it is well-suited for checking the cross-regularity assumptions of Altschuler and Chewi (2024c); Altschuler et al. (2025).

We remark that these rates are somewhat incomparable, and at the time of writing, the dust has not yet settled on the determination of the "correct" rates. The extension to KL divergence is also rather non-trivial; note, for instance, that a naïve adaptation of the methods of Sections 4.2 and 4.4 fails because RM–LMC does not admit a natural adapted interpolation.

The literature on hypocoercivity is large and we do not intend to comprehensively survey it here, but we give a few pointers. Our treatment of hypocoercivity is inspired by the monograph Villani (2009a), which was the first to develop the theory in a general framework. The illuminating work of Baudoin (2017) relates the hypocoercive calculations to the Bakry–Émery calculus. The approach to hypocoercivity taken here is sometimes called "H^1 hypocoercivity" because the constructed norm is equivalent to the $H^1(\boldsymbol{\pi})$ Sobolev norm, but there is another approach based on "L^2 hypocoercivity". The latter is called the "DMS framework" after Dolbeault et al. (2009); see Roussel and Stoltz (2018) for the application to underdamped Langevin.

In the case of Gaussian targets (and small perturbations thereof), the accelerated convergence rate was established in W_2 via a coupling argument in Schuh (2024). The space-time Poincaré approach builds on the earlier work of Albritton et al. (2024), and the sharp result was first established in Cao et al. (2023). Our exposition follows the second-order lifting framework of Eberle and Lörler (2024a). Further developments include Brigati (2023); Eberle and Lörler (2024b); Brigati et al. (2025); Eberle et al. (2025); Lehec (2025); Lu (2025). The space-time approach is also applicable to piecewise deterministic Markov processes (PDMPs), such as randomized Hamiltonian Monte Carlo and the zigzag sampler (Lu and Wang, 2022b).

Exercises

Randomized Midpoint Discretization

⊵ **Exercise 5.1** (Stochastic gradient LMC)

Assume, as usual, that $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. However, suppose that when we run LMC, we only have access to stochastic gradient evaluations. Namely, for each $x \in \mathbb{R}^d$, we can obtain a stochastic gradient $\hat\nabla V(x)$ which is unbiased, $\mathbb{E}\,\hat\nabla V(x) = \nabla V(x)$, and has uniformly bounded variance:

$$\mathbb{E}[\|\hat\nabla V(x) - \nabla V(x)\|^2] \leqslant \sigma^2 d \qquad \text{for all } x \in \mathbb{R}^d\,.$$

Use the local error framework to analyze this variant of LMC.

⊵ **Exercise 5.2** (LMC under higher-order smoothness)

Assume that the potential satisfies $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and furthermore assume that $\|\nabla\Delta V\| \leqslant \zeta_0 + \zeta_1\,\|\nabla V\|$. Define the scale-invariant quantities $\kappa := \beta/\alpha$, $\bar\kappa_0 := \zeta_0/\alpha^{3/2}$, and $\bar\kappa_1 := \zeta_1/\alpha$. Use the local error framework to show that LMC achieves iteration complexity $\widetilde O((\bar\kappa_0 + (\kappa + \bar\kappa_1)\sqrt{\kappa d})/\varepsilon)$ in this case. In particular, if $\kappa, \bar\kappa_0, \bar\kappa_1 = O(1)$, then the iteration complexity becomes $\widetilde O(d^{1/2}/\varepsilon)$, which should match what was obtained for the Gaussian case in Exercise 4.1. Note that if we have a Lipschitz condition in the operator norm, $\|\nabla^2 V(x) - \nabla^2 V(y)\|_{\mathrm{op}} \leqslant \rho\,\|x - y\|$, then $\|\nabla\Delta V\| \leqslant \rho d^{1/2}$, so in this case it implies the iteration complexity $\widetilde O((\kappa^{3/2} d^{1/2} + \kappa^{1/2}\rho d/\alpha)/\varepsilon)$.

Hence, if the Hessian is Lipschitz in the operator norm, the complexity of LMC improves from $\widetilde O(d/\varepsilon^2)$ to $\widetilde O(d/\varepsilon)$, and under a more stringent derivative condition (which is satisfied by Gaussians), it further improves to $\widetilde O(d^{1/2}/\varepsilon)$.

Hint: To bound the weak error, use Itô's formula.

⊵ **Exercise 5.3** (Itô SDEs with multiplicative noise)
Consider a general Itô SDE of the form

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t\,. \tag{5.E.1}$$

We consider the Euler–Maruyama discretization

$$\hat X_{(n+1)h} = \hat X_{nh} + h\,b(\hat X_{nh}) + \sigma(\hat X_{nh})\,(B_{(n+1)h} - B_{nh})\,.$$

We adopt the following assumptions:

1 (Contractivity) $2\,\langle b(x) - b(y), x-y\rangle + \|\sigma(x) - \sigma(y)\|_{\mathrm{HS}}^2 \leqslant -2\alpha\,\|x-y\|^2$ for all $x, y \in \mathbb{R}^d$.
2 (Lipschitz coefficients) b is L-Lipschitz and σ is $\sqrt{L}$-Lipschitz (in the $\|\cdot\|_{\mathrm{HS}}$ norm).
3 (Coefficient bounds) $\|b(0)\| \leqslant B_b$ and $\|\sigma(0)\|_{\mathrm{HS}} \leqslant B_\sigma$.

Writing π for the stationary distribution for (5.E.1) and $\kappa := L/\alpha$, use the local error framework to show that for all sufficiently small $\varepsilon > 0$ and for an appropriate choice of step size, we can obtain $\sqrt{\alpha}\,W_2(\hat\mu_{Nh}, \pi) \leqslant \varepsilon$ in

$$N = O\Bigl(\frac{\kappa^4}{\varepsilon^2}\,\bigl(\frac{B_b^2}{L} + B_\sigma^2\bigr)\log\frac{\alpha W_2^2(\mu_0, \pi)}{\varepsilon^2}\Bigr) \qquad \text{iterations}\,.$$

(This exercise is rather involved and also requires establishing second moment bounds in the spirit of Exercise 4.2.)

Hamiltonian Monte Carlo

⊵ **Exercise 5.4** (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 $\mu_0 = \mathsf{normal}(0, \sigma^2 I_d)$, show that the distribution $\boldsymbol{\mu}_t$ over phase space at time t of ideal HMC is a Gaussian distribution, $\mathsf{normal}(0, \Sigma_t)$, and compute $\Sigma_t \in \mathbb{R}^{2d\times 2d}$.

⊵ **Exercise 5.5** (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, $(x_t, p_t)_{t\geqslant 0}$, 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.,

$$\{f, H\} := \langle \nabla_x f, \nabla_p H\rangle - \langle \nabla_p f, \nabla_x H\rangle = 0\,,$$

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 $t \geqslant 0$ 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. Then, show that $(x_{T-t}, -p_{T-t})_{t\in[0,T]}$ also solve Hamilton's equations. In other words, if R is the momentum 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 5.6** (Coercivity)
Prove Lemma 5.2.3.

Hint: Let $z := y - \frac{1}{\beta} \{\nabla f(y) - \nabla f(x)\}$. 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 5.7** (Nesterov's algorithm in continuous time)
Consider the continuous-time formulation of Nesterov's algorithm, as given in Optimization Box 5.3.2. Assume that V is α-strongly convex and $p_0 = 0$. Prove the rate (5.3.3).

Hint: Let $z_t := x_t + \frac{2}{\gamma} p_t$. Consider the Lyapunov functional

$$\mathcal{L}_t := V(x_t) - V(x_\star) + \frac{\alpha}{2} \|z_t - x_\star\|^2 .$$

Prove that $\dot{\mathcal{L}}_t \leqslant -\sqrt{\alpha}\, \mathcal{L}_t$. 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 5.8** (Derivation of the ULMC updates)
Solve the SDE for ULMC to prove Lemma 5.3.8.

⊵ **Exercise 5.9** (Movement bound for the underdamped Langevin diffusion)
Prove the movement bound for underdamped Langevin (Lemma 5.3.10).

⊵ **Exercise 5.10** (Uniform-in-time LSI for ULMC)
Establish the analogue of Exercise 4.4 for ULMC. You may restrict attention to the contractive case $\alpha > 0$, and you can use the contraction for ULMC in the third part of Exercise 6.9.

⊵ **Exercise 5.11** (Implementation of RM–ULMC)
Show how to implement RM–ULMC by computing $\operatorname{cov}(\xi^{(1)}_{nh,s}, \xi^{(1)}_{nh,t})$, $\operatorname{cov}(\xi^{(1)}_{nh,t}, \xi^{(2)}_{nh,h})$, and $\operatorname{cov}(\xi^{(2)}_{nh,h})$.

⊵ **Exercise 5.12** (Analysis of RM–ULMC)
In this exercise, we establish Theorem 5.3.11.

1. Prove an analogue of Lemma 5.1.6 for the underdamped Langevin diffusion (work in the twisted norm $|||\cdot|||$ throughout).
2. Show that if we take the expectation over u_n and v_n, denoted $\mathbb{E}_{u_n}$ and $\mathbb{E}_{v_n}$ respectively, then

$$\begin{aligned}\mathbb{E}_{u_n} \hat{X}_{(n+1)h} &= X_{nh} + \frac{1 - \exp(-\gamma h)}{\gamma} \hat{P}_{nh} - \int_{nh}^{(n+1)h} \frac{1 - \exp(-\gamma\,((n+1)h - t))}{\gamma} \nabla V(\hat{X}_t^+)\,\mathrm{d}t \\ &\quad + \xi^{(1)}_{nh,h} , \\ \mathbb{E}_{v_n} \hat{P}_{(n+1)h} &= \exp(-\gamma h)\, \hat{P}_{nh} - \int_{nh}^{(n+1)h} \exp(-\gamma\,((n+1)h - t))\, \nabla V(\hat{X}_t^+)\,\mathrm{d}t + \xi^{(2)}_{nh,h} .\end{aligned}$$

3. By comparing the expressions in the preceding part with the exact solution of the underdamped

Langevin diffusion, establish the following local error bounds, assuming that $-\beta I_d \preceq \nabla^2 V \preceq \beta I_d$, $h \ll {}^1\!/\!\sqrt{\beta} \wedge {}^\gamma\!/\!_\beta$, and $\gamma \lesssim \sqrt{\beta}$:

$$h^{-1}\,\|\mathbb{E}\,\hat{X}_h - \mathbb{E}\,X_h\| + \|\mathbb{E}\,\hat{P}_h - \mathbb{E}\,P_h\| \lesssim \beta^2 h^4\,\|p\| + \beta^2\gamma^{1/2}d^{1/2}h^{9/2} + \beta^2 h^5\,\|\nabla V(x)\|\,,$$
$$h^{-1}\,\|\hat{X}_h - X_h\|_{L^2} + \|\hat{P}_h - P_h\|_{L^2} \lesssim \beta h^2\,\|p\| + \beta\gamma^{1/2}d^{1/2}h^{5/2} + \beta h^3\,\|\nabla V(x)\|\,.$$

Notice that the weak error is incredibly small!

4 Use the local error framework to prove Theorem 5.3.11.

⊵ **Exercise 5.13** (Fokker–Planck equation for the underdamped Langevin diffusion)
Check the expression $\mathscr{L} = \gamma\mathscr{L}_{\mathsf{OU}} + \mathscr{L}_{\mathsf{Ham}}$ for the generator of the underdamped Langevin diffusion given in Section 5.3.4. Then, check the various calculations leading up to the Fokker–Planck equations (5.3.4) and (5.3.12).

⊵ **Exercise 5.14** (Computations with adjoints and commutators)
Prove that $\mathscr{L}_{\mathsf{Ham}}^* = -\mathscr{L}_{\mathsf{Ham}}$ and that $\mathscr{L}_{\mathsf{OU}} = -\nabla_p^*\nabla_p$, where the adjoints are taken in $L^2(\boldsymbol{\pi})$. Moreover, verify the commutator relations (Com).

⊵ **Exercise 5.15** (Entropic hypocoercivity)
Adapt the proof of Theorem 5.3.13 to prove Theorem 5.3.14.

⊵ **Exercise 5.16** (Relaxation time lower bound)
Here, we prove Theorem 5.3.18.

1 Using reversibility, show that $\mathsf{relax}(\mathscr{L}) \asymp 1/\lambda$, where λ is the spectral gap of $\mathscr{L}$.
2 Let $f \in L^2(\pi)$ be such that $\int f\,\mathrm{d}\pi = 0$, $\int f^2\,\mathrm{d}\pi = 1$, and $\mathscr{E}(f,f) = \lambda$. (Here, we assume the existence of an eigenfunction for simplicity, although the proof can avoid this.) Use the second-order lift property to show that $\|\boldsymbol{\mathscr{L}}(f\circ\Pi)\|_{L^2(\boldsymbol{\pi})} \lesssim \sqrt{\lambda}\,\|f\circ\Pi\|_{L^2(\boldsymbol{\pi})}$.
3 Suppose that $(\boldsymbol{P}_t)_{t\geqslant 0}$ satisfies the decay rate $\|\boldsymbol{P}_t f\|_{L^2(\boldsymbol{\pi})} \leqslant \exp(-\omega\,(t-t_0))\,\|f\|_{L^2(\boldsymbol{\pi})}$ for all $t \geqslant 0$ and all $f \in L^2(\boldsymbol{\pi})$ with $\int f\,\mathrm{d}\boldsymbol{\pi} = 0$. Using the fact that $(-\boldsymbol{\mathscr{L}})^{-1} = \int_0^\infty \boldsymbol{P}_t\,\mathrm{d}t$ (justify!) and by considering $\int_0^\infty \|\boldsymbol{P}_t f\|_{L^2(\boldsymbol{\pi})}\,\mathrm{d}t$, prove that $t_0 + \omega^{-1} \gtrsim 1/\sqrt{\lambda}$.
4 Finally, show that the decay in the preceding part holds with $\omega^{-1} \asymp t_0 = \mathsf{relax}(\boldsymbol{\mathscr{L}})$. Combine everything together to deduce the result.

⊵ **Exercise 5.17** (Decay under a space-time Poincaré inequality)
Prove that (5.3.21) implies (5.3.22).

⊵ **Exercise 5.18** (Space-time Poincaré inequality without log-concavity)
In the first part of Lemma 5.3.19, how is the bound affected if we do not assume log-concavity, but instead assume that $\nabla^2 V \succeq -\beta I_d$ for some $\beta > 0$? How does this affect the final convergence estimate?

⊵ **Exercise 5.19** (Randomized Hamiltonian Monte Carlo)
Consider ideal Hamiltonian Monte Carlo with randomized integration times, namely, the integration times are drawn i.i.d. with the exponential distribution with rate γ.

1 Argue that the generator is given by $\gamma\,(\Pi - \mathrm{id}) + \mathscr{L}_{\mathsf{Ham}}$.
2 Adapt the proof of Theorem 5.3.15 to randomized Hamiltonian Monte Carlo. Namely, explain how to substitute for the inequalities leading up to (5.3.23).

CHAPTER 6

Convergence in Rényi Divergence

In this chapter, we study sampling guarantees which hold in Rényi divergences. Recall from Section 2.2.5 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 $\mathcal{R}_2 = \log(1 + \chi^2)$. Rényi divergence guarantees are stronger than W_2 or KL guarantees, and they have been of interest in their own right in differential privacy Mironov (2017). The results from this chapter will also be used to be used in the subsequent Chapter 7.

In Section 2.2.5, we studied the continuous-time convergence of the Langevin diffusion in Rényi divergence under either a Poincaré inequality or a log-Sobolev inequality. We will build upon these results in order to study the discretized LMC algorithm. Then, we will consider the underdamped Langevin diffusion, which was introduced in Section 5.3.

6.1 Analysis of LMC via Interpolation Argument

In this section, we follow Chewi et al. (2024a), which generalizes the interpolation 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 2.2.25 and Corollary 4.2.4, so it is left as Exercise 6.1. Throughout this section, let $q \geqslant 2$ be fixed.

Proposition 6.1.1. *Along the law $(\hat\mu_t)_{t\geqslant 0}$ of the interpolated process (4.2.2),*

$$\partial_t \mathcal{R}_q(\hat\mu_t \,\|\, \pi) \leqslant -\frac{3}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + q\,\mathbb{E}[\psi_t(\hat X_t)\,\|\nabla V(\hat X_t) - \nabla V(\hat X_{t_-})\|^2]\,,$$

where $\rho_t := \frac{\mathrm{d}\hat\mu_t}{\mathrm{d}\pi}$ and $\psi_t := \rho_t^{q-1}/\mathbb{E}_\pi(\rho_t^q)$.

In analogy with the usual Fisher information, the quantity

$$\mathsf{FI}_q(\mu \,\|\, \pi) := \frac{4}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho^{q/2})\|^2]}{\mathbb{E}_\pi(\rho^q)}\,, \qquad \rho := \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,,$$

may be considered the "Rényi Fisher information". As in the proof of Theorem 2.2.25, under a log-Sobolev inequality, we have

$$\mathsf{FI}_q(\mu \,\|\, \pi) \geqslant \frac{2}{qC_{\mathsf{LSI}}}\,\mathcal{R}_q(\mu \,\|\, \pi)\,,$$

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

$$\mathbb{E}\,\psi_t(\hat{X}_t) = \mathbb{E}_{\hat{\mu}_t}\,\psi_t = \mathbb{E}_\pi\Big[\rho_t\,\frac{\rho_t^{q-1}}{\mathbb{E}_\pi(\rho_t^q)}\Big] = 1\,,$$

so $\psi_t(\hat{X}_t)$ acts as a change of measure. The main difficulty of the proof is that whereas we know how to control the term $\|\nabla V(\hat{X}_t) - \nabla V(\hat{X}_{t_-})\|^2$ 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 $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbb{P}} = \psi_t(\hat{X}_t)$. 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 (Chewi et al. (2024a)). *For $n \in \mathbb{N}$, let $\hat{\mu}_{nh}$ denote the law of the n-th iterate of* LMC *with step size $h > 0$. Assume that the target $\pi \propto \exp(-V)$ satisfies $C_{\mathsf{LSI}}(\pi) \leqslant 1/\alpha$ and that ∇V is β-Lipschitz. Then, for all $h \leqslant {}^1/_{192\beta\kappa q^2}$, for all $N \geqslant N_0$, it holds that*

$$\mathcal{R}_q(\hat{\mu}_{Nh} \,\|\, \pi) \leqslant \exp\Big(-\frac{\alpha(N-N_0)h}{4}\Big)\,\mathcal{R}_2(\hat{\mu}_0 \,\|\, \pi) + \widetilde{O}\Big(\frac{\beta^2 dhq}{\alpha}\Big)\,,$$

where $N_0 := \lceil \frac{2}{\alpha h}\log(q-1)\rceil$. In particular, for all $\varepsilon \in [0, \sqrt{d/q}]$, if we choose the step size $h = \widetilde{\Theta}({}^{\varepsilon^2}/_{\beta\kappa d q})$, then we obtain the guarantee $\sqrt{\mathcal{R}_q(\hat{\mu}_{Nh} \,\|\, \pi)} \leqslant \varepsilon$ after

$$N = \widetilde{O}\Big(\frac{\kappa^2 dq}{\varepsilon^2}\log \mathcal{R}_2(\hat{\mu}_0 \,\|\, \pi)\Big) \qquad \textit{iterations}\,.$$

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 $\mathbb{E}[\psi_t(\hat{X}_t)\,\|\nabla V(\hat{X}_t) - \nabla V(\hat{X}_{t_-})\|^2]$, where from the β-smoothness of V and from $h \leqslant {}^1/_{3\beta}$ we have

$$\|\nabla V(\hat{X}_t) - \nabla V(\hat{X}_{t_-})\|^2 \leqslant 9\beta^2\,(t-t_-)^2\,\|\nabla V(\hat{X}_t)\|^2 + 6\beta^2\,(t-t_-)\,\|B_t - B_{t_-}\|^2\,.$$

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 $\psi_t \hat{\mu}_t$,

$$\begin{aligned}\mathbb{E}_{\psi_t \hat{\mu}_t}[\|\nabla V\|^2] &\leqslant \mathsf{FI}(\psi_t \hat{\mu}_t \parallel \pi) + 2\beta d = \mathbb{E}_{\hat{\mu}_t}\Bigl[\psi_t \, \Bigl\|\nabla \log\bigl(\psi_t \frac{\mathrm{d}\hat{\mu}_t}{\mathrm{d}\pi}\bigr)\Bigr\|^2\Bigr] + 2\beta d \\ &= \frac{\mathbb{E}_\pi[\rho_t^q \, \|\nabla \log(\rho_t^q)\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2\beta d = \frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2\beta d\,,\end{aligned}$$

where we used the identity

$$\mathbb{E}_{\hat{\mu}_t}\Bigl[\psi_t \, \Bigl\|\nabla \log\bigl(\psi_t \, \frac{\mathrm{d}\hat{\mu}_t}{\mathrm{d}\pi}\bigr)\Bigr\|^2\Bigr] = \frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} \tag{6.1.3}$$

which follows from the chain rule from calculus.

For the second error term, we must control the term $\|B_t - B_{t_-}\|^2$ under the measure $\mathbf{P}$, where $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbb{P}} = \psi_t(\hat{X}_t)$. 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 1.5.7) to relate the expectation under $\mathbf{P}$ (denoted $\mathbf{E}$) with the expectation under $\mathbb{P}$. For any random variable ζ, it yields

$$\mathbf{E}\zeta \leqslant \mathsf{KL}(\mathbf{P} \parallel \mathbb{P}) + \log \mathbb{E} \exp \zeta\,.$$

Applying this to $\zeta := c\,(\|B_t - B_{t_-}\| - \mathbb{E}\|B_t - B_{t_-}\|)^2$ for a constant $c > 0$ to be chosen later, we obtain

$$\begin{aligned}\mathbf{E}[\|B_t - B_{t_-}\|^2] &\leqslant 2\,\mathbb{E}[\|B_t - B_{t_-}\|^2] + \frac{2}{c}\,\mathbf{E}\zeta \\ &\leqslant 2d\,(t - t_-) + \frac{2}{c}\,\bigl\{\mathsf{KL}(\mathbf{P} \parallel \mathbb{P}) + \log \mathbb{E} \exp\bigl(c\,(\|B_t - B_{t_-}\| - \mathbb{E}\|B_t - B_{t_-}\|)^2\bigr)\bigr\}\,.\end{aligned}$$

Under $\mathbb{P}$, $B_t - B_{t_-} \sim \mathsf{normal}(0, (t - t_-)\,I_d)$. Applying concentration of measure for the Gaussian distribution (see, e.g., Theorem 2.4.3), if $c \lesssim \frac{1}{t - t_-}$, then

$$\mathbb{E} \exp\bigl(c\,(\|B_t - B_{t_-}\| - \mathbb{E}\|B_t - B_{t_-}\|)^2\bigr) \leqslant 2\,.$$

In fact, it suffices to take $c = \frac{1}{8\,(t - t_-)}$. Next, using the LSI for π,

$$\begin{aligned}\mathsf{KL}(\mathbf{P} \parallel \mathbb{P}) &= \mathbb{E}_{\psi_t \hat{\mu}_t} \log \psi_t = \mathbb{E}_{\psi_t \hat{\mu}_t} \log \frac{\rho_t^{q-1}}{\mathbb{E}_{\hat{\mu}_t}(\rho_t^{q-1})} = \frac{q-1}{q}\,\mathbb{E}_{\psi_t \hat{\mu}_t} \log \frac{\rho_t^q}{\mathbb{E}_{\hat{\mu}_t}(\rho_t^{q-1})^{q/(q-1)}} \\ &= \frac{q-1}{q}\,\Bigl\{\mathbb{E}_{\psi_t \hat{\mu}_t} \log \frac{\rho_t^q}{\mathbb{E}_{\hat{\mu}_t}(\rho_t^{q-1})} - \underbrace{\frac{1}{q-1} \log \mathbb{E}_{\hat{\mu}_t}(\rho_t^{q-1})\Bigr\}}_{\geqslant 0} \\ &\leqslant \frac{q-1}{q}\,\mathsf{KL}(\psi_t \hat{\mu}_t \parallel \pi) \\ &\leqslant \frac{q-1}{2\alpha q}\,\mathbb{E}_{\psi_t \hat{\mu}_t}\Bigl[\Bigl\|\nabla \log\bigl(\psi_t \, \frac{\mathrm{d}\hat{\mu}_t}{\mathrm{d}\pi}\bigr)\Bigr\|^2\Bigr] = \frac{2\,(q-1)}{\alpha q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,,\end{aligned}$$

where we applied the identity (6.1.3). Hence,

$$\begin{aligned}
&\mathbb{E}[\psi_t(\hat{X}_t)\,\|B_t - B_{t_-}\|^2] \\
&\quad\leqslant 2d\,(t-t_-) + \frac{32h\,(q-1)}{\alpha q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + (16\log 2)\,(t-t_-) \\
&\quad\leqslant 14d\,(t-t_-) + \frac{32h}{\alpha}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\,.
\end{aligned}$$

All in all, applying Proposition 6.1.1 and collecting the error terms,

$$\begin{aligned}
\partial_t \mathcal{R}_q(\hat{\mu}_t \,\|\, \pi) \leqslant -\frac{3}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} &+ 9\beta^2 q\,(t-t_-)^2\,\Big\{\frac{4\,\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + 2\beta d\Big\} \\
&+ 6\beta^2 q\,\Big\{14d\,(t-t_-) + \frac{32h}{\alpha}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)}\Big\}\,.
\end{aligned}$$

From $h \leqslant {}^1/_{192\beta\kappa q^2}$, we can absorb some of the error terms into the decay term and apply the LSI for π (see Theorem 2.2.25), yielding

$$\begin{aligned}
\partial_t \mathcal{R}_q(\hat{\mu}_t \,\|\, \pi) &\leqslant -\frac{1}{q}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q/2})\|^2]}{\mathbb{E}_\pi(\rho_t^q)} + O(\beta^3 dh^2 q + \beta^2 dhq) \\
&\leqslant -\frac{\alpha}{2q}\,\mathcal{R}_q(\hat{\mu}_t \,\|\, \pi) + O(\beta^2 dhq)\,.
\end{aligned}$$

This implies the differential inequality

$$\partial_t\Big\{\exp\Big(\frac{\alpha\,(t-t_-)}{2q}\Big)\,\mathcal{R}_q(\hat{\mu}_t \,\|\, \pi)\Big\} \lesssim \exp\Big(\frac{\alpha\,(t-t_-)}{2q}\Big)\,\beta^2 dhq \lesssim \beta^2 dhq\,.$$

Integrating this over $t \in [nh, (n+1)h]$ yields

$$\mathcal{R}_q(\hat{\mu}_{(n+1)h} \,\|\, \pi) \leqslant \exp\Big(-\frac{\alpha h}{2q}\Big)\,\mathcal{R}_q(\hat{\mu}_{nh} \,\|\, \pi) + O(\beta^2 dh^2 q)\,.$$

Unrolling the recursion,

$$\mathcal{R}_q(\hat{\mu}_{Nh} \,\|\, \pi) \leqslant \exp\Big(-\frac{\alpha Nh}{2q}\Big)\,\mathcal{R}_q(\hat{\mu}_0 \,\|\, \pi) + O\Big(\frac{\beta^2 dhq^2}{\alpha}\Big)\,. \qquad \square$$

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 **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 1.5.7). These inequalities yield an additional error term of the form $\mathsf{FI}(\psi_t\hat{\mu}_t \,\|\, \pi)$ (for the latter, this error term appears after an application of the LSI). The magical part of the calculation is that $\mathsf{FI}(\psi_t\hat{\mu}_t \,\|\, \pi)$ 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 2.10). The proof is left as Exercise 6.1.

Proposition 6.1.4 (Hypercontractivity). *Suppose that* $C_{\mathsf{LSI}}(\pi) \leqslant 1/\alpha$. *Along the law* $(\hat\mu_t)_{t\geqslant 0}$ *of the interpolated process* (4.2.2), *if we define the parameter* $q(t) := 1 + (q_0 - 1)\exp\frac{\alpha t}{2}$, *then*

$$\partial_t\Big(\frac{1}{q(t)}\log\int \rho_t^{q(t)}\,\mathrm{d}\pi\Big) \leqslant -\frac{2\,(q(t)-1)}{q(t)^2}\,\frac{\mathbb{E}_\pi[\|\nabla(\rho_t^{q(t)/2})\|^2]}{\mathbb{E}_\pi(\rho_t^{q(t)})} + \big(q(t)-1\big)\,\mathbb{E}[\psi_t(\hat X_t)\,\|\nabla V(\hat X_t) - \nabla V(\hat X_{t_-})\|^2]\,,$$

where $\rho_t := \frac{\mathrm{d}\hat\mu_t}{\mathrm{d}\pi}$ *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 $\bar q \geqslant 3$.

Initial waiting phase. We apply hypercontractivity (Proposition 6.1.4) with $q_0 = 2$ and for $t \leqslant N_0 h$, where $N_0 := \lceil \frac{2}{\alpha h}\log(\bar q - 1)\rceil$. Note that $\bar q \leqslant q(N_0 h) \leqslant 2\bar q$. The bound on the error term from the previous proof yields

$$\partial_t\Big(\frac{1}{q(t)}\log\int \rho_t^{q(t)}\,\mathrm{d}\pi\Big) \lesssim \beta^2 dh q(t)\,.$$

Integrating this over $t \in [nh, (n+1)h]$ yields

$$\frac{1}{q((n+1)h)}\log\int \rho_{(n+1)h}^{q((n+1)h)}\,\mathrm{d}\pi - \frac{1}{q(nh)}\log\int \rho_{nh}^{q(nh)}\,\mathrm{d}\pi \lesssim \beta^2 dh^2 \bar q\,.$$

Unrolling the recursion yields

$$\frac{1}{q(N_0 h)}\log\int \rho_{N_0 h}^{q(N_0 h)}\,\mathrm{d}\pi - \frac{1}{2}\log\int \rho_0^2\,\mathrm{d}\pi \lesssim \beta^2 dh^2 \bar q N_0 \leqslant \widetilde O\Big(\frac{\beta^2 dh\bar q}{\alpha}\Big)\,.$$

Finishing the convergence analysis. Next, after shifting time indices and applying the previous proof of Theorem 6.1.2 with $q = 2$,

$$\begin{aligned}\mathcal{R}_{\bar q}(\hat\mu_{(N+N_0)h}\,\|\,\pi) &\leqslant \frac{1}{q(N_0 h) - 1}\log\int \rho_{(N+N_0)h}^{q(N_0 h)}\,\mathrm{d}\pi \leqslant \frac{3}{4}\log\int \rho_{Nh}^2\,\mathrm{d}\pi + \widetilde O\Big(\frac{\beta^2 dh\bar q}{\alpha}\Big)\\ &\leqslant \frac{3}{4}\exp\Big(-\frac{\alpha N h}{4}\Big)\,\mathcal{R}_2(\hat\mu_0\,\|\,\pi) + \widetilde O\Big(\frac{\beta^2 dh\bar q}{\alpha}\Big)\,.\end{aligned}$$

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 $\mathsf{KL}(\psi_t\hat\mu_t\,\|\,\pi)$ 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 $t \geqslant 0$ (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 2.4.3, if $\pi \propto \exp(-V)$ satisfies an LSI, then V necessarily has quadratic growth at infinity). We explore this in Exercise 6.2.

6.2 Analysis of LMC via Girsanov's Theorem

We now provide a Rényi analysis for LMC based on Girsanov's theorem; see Sections 3.2.2 and 4.4 for background. The advantage of this approach is that it is more generalizable, as it is less reliant on seemingly miraculous calculations. The results of this section will also be used for the analysis of MALA in Chapter 7.

Discretization error.

Following the proof of Theorem 4.4.1 in Section 4.4, let $\mathbf{P}_T$, $\mathbf{W}_T$ be path measures such that under $\mathbf{P}_T$, $(X_t)_{t\in[0,T]}$ is the interpolated LMC process, and under $\mathbf{W}_T$, $(X_t)_{t\in[0,T]}$ is the Langevin diffusion. Both processes are initialized at some measure $\hat{\mu}_0$. Our goal is to prove the following theorem.

Theorem 6.2.1 (LMC Rényi discretization bound)**.** *Assume that the potential V is β-smooth and that $h \lesssim 1/\beta^2 q^2 T$ where $T := Nh$. Then, for all $q \geqslant 2$,*

$$\mathcal{R}_q(\mathbf{P}_T \parallel \mathbf{W}_T) \lesssim \frac{1}{q} \max_{n=0,1,\dots,N-1} \log \mathbb{E}^{\mathbf{W}_T} \exp\{O(\beta^2 h^2 q^2 T \, \|\nabla V(X_{nh})\|^2)\} + \beta^2 dhqT\,.$$

Proof From Girsanov's theorem (Theorem 3.2.8; see also Section 4.4), we have[1]

$$\mathbb{E}^{\mathbf{W}_T}\Big[\Big(\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{W}_T}\Big)^q\Big] = \mathbb{E}^{\mathbf{W}_T} \exp\Big(\frac{q}{\sqrt{2}} \int_0^T \langle \nabla V(X_{t_-}) - \nabla V(X_t), \mathrm{d}B_t\rangle \\ - \frac{q}{4} \int_0^T \|\nabla V(X_{t_-}) - \nabla V(X_t)\|^2 \, \mathrm{d}t\Big)\,.$$

Unlike the analysis in KL divergence, the stochastic integral is now inside the exponential. The first step is to remove this term, which we accomplish as follows. First, by the Cauchy–Schwarz inequality and for any $\lambda > 0$,

$$\Big\{\mathbb{E}^{\mathbf{W}_T}\Big[\Big(\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{W}_T}\Big)^q\Big]\Big\}^2 \leqslant \mathbb{E}^{\mathbf{W}_T} \exp\Big(\sqrt{2}\, q \int_0^T \langle \nabla V(X_{t_-}) - \nabla V(X_t), \mathrm{d}B_t\rangle \\ - \lambda \int_0^T \|\nabla V(X_{t_-}) - \nabla V(X_t)\|^2 \, \mathrm{d}t\Big) \\ \times \mathbb{E}^{\mathbf{W}_T} \exp\Big(\Big(\lambda - \frac{q}{2}\Big) \int_0^T \|\nabla V(X_{t_-}) - \nabla V(X_t)\|^2 \, \mathrm{d}t\Big)\,.$$

Next, recall from Section 3.2.2 that for any martingale M, the exponential martingale $\mathcal{E}(M) := \exp(M - \frac{1}{2}[M,M])$ is a local martingale. We apply this to the martingale M given by $M_t := \sqrt{2}\, q \int_0^t \langle \nabla V(X_{s_-}) - \nabla V(X_s), \mathrm{d}B_s\rangle$ by taking $\lambda = q^2$. Then, $\mathcal{E}(M)$ is a non-negative local martingale, so it is a supermartingale, and $\mathbb{E}^{\mathbf{W}_T} \mathcal{E}(M)_T \leqslant 1$. Finally, noting that the expectation above is at least 1, we can drop the square and we obtain

$$\mathbb{E}^{\mathbf{W}_T}\Big[\Big(\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{W}_T}\Big)^q\Big] \leqslant \mathbb{E}^{\mathbf{W}_T} \exp\Big(q^2 \int_0^T \|\nabla V(X_{t_-}) - \nabla V(X_t)\|^2 \, \mathrm{d}t\Big) \\ \leqslant \mathbb{E}^{\mathbf{W}_T} \exp\Big(\beta^2 q^2 \int_0^T \|X_t - X_{t_-}\|^2 \, \mathrm{d}t\Big)\,.$$

Let us first consider a single iteration, over the time interval $[0, h]$, and suppose that the processes are started at $x \in \mathbb{R}^d$. Then,

---

[1] Again, we ignore Novikov's condition, which can be avoided via localization.

$$\begin{aligned} \|X_t - x\| &= \Big\| -\int_0^t \nabla V(X_s)\,\mathrm{d}s + \sqrt{2}\,B_t \Big\| \leqslant \int_0^t \|\nabla V(X_s)\|\,\mathrm{d}s + \sqrt{2}\,\|B_t\| \\ &\leqslant \beta \int_0^t \|X_s - x\|\,\mathrm{d}s + h\,\|\nabla V(x)\| + \sqrt{2}\,\|B_t\|\,. \end{aligned}$$

By Grönwall's inequality, if $h \leqslant 1/\beta$, it yields, for all $t \in [0,h]$,

$$\|X_t - x\| \leqslant 3h\,\|\nabla V(x)\| + 3\sqrt{2} \sup_{s\in[0,t]} \|B_s\|\,. \tag{6.2.2}$$

Therefore, for any $\eta > 0$,

$$\mathbb{E}^{\mathbf{W}_T} \exp\bigl(\eta \sup_{t\in[0,h]} \|X_t - x\|^2\bigr) \leqslant \exp(18\eta h^2\,\|\nabla V(x)\|^2)\,\mathbb{E}^{\mathbf{W}_T} \exp\bigl(36\eta \sup_{t\in[0,h]} \|B_t\|^2\bigr)\,.$$

We require a tail bound for Brownian motion (Lemma 6.2.4 below), which yields

$$\log \mathbb{E}^{\mathbf{W}_T} \exp\bigl(\eta \sup_{t\in[0,h]} \|X_t - x\|^2\bigr) \lesssim \eta h^2\,\|\nabla V(x)\|^2 + \eta d h\,, \tag{6.2.3}$$

provided that $\eta \lesssim 1/h$.

We must now iterate this bound. One approach is to condition on the process up until time $(N-1)h$ and apply (6.2.3), which yields

$$\begin{aligned} &\mathbb{E}^{\mathbf{W}_T} \exp\Bigl(\beta^2 q^2 \int_0^{Nh} \|X_t - X_{t_-}\|^2\,\mathrm{d}t\Bigr) \\ &\quad\leqslant \mathbb{E}^{\mathbf{W}_T} \exp\Bigl(\beta^2 q^2 \int_0^{(N-1)h} \|X_t - X_{t_-}\|^2\,\mathrm{d}t + O\bigl(\beta^2 h^3 q^2\,\|\nabla V(X_{(N-1)h})\|^2 + \beta^2 d h^2 q^2\bigr)\Bigr) \end{aligned}$$

but now the terms in the exponential are dependent and we cannot peel off the next term. To circumvent this issue, we instead apply Jensen's inequality (or equivalently, the AM–GM inequality):

$$\mathbb{E}^{\mathbf{W}_T} \exp\Bigl(\beta^2 q^2 \int_0^T \|X_t - X_{t_-}\|^2\,\mathrm{d}t\Bigr) \leqslant \frac{1}{T}\int_0^T \mathbb{E}^{\mathbf{W}_T} \exp(\beta^2 q^2 T\,\|X_t - X_{t_-}\|^2)\,\mathrm{d}t\,.$$

Applying (6.2.3) conditionally, it yields

$$\mathbb{E}^{\mathbf{W}_T}\Bigl[\Bigl(\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{W}_T}\Bigr)^q\Bigr] \leqslant \frac{1}{T}\int_0^T \mathbb{E}^{\mathbf{W}_T} \exp\Bigl(O\bigl(\beta^2 h^2 q^2 T\,\|\nabla V(X_{t_-})\|^2 + \beta^2 d h q^2 T\bigr)\Bigr)\,\mathrm{d}t$$

provided that $\beta^2 q^2 T \lesssim 1/h$, i.e., $h \lesssim 1/\beta^2 q^2 T$. The result follows. □

In the proof, we used the following lemma; see Exercise 6.3.

Lemma 6.2.4 (Brownian tail bound)**.** *Let $(B_t)_{t\geqslant 0}$ denote a standard Brownian motion in $\mathbb{R}^d$. Then, for $\lambda, t > 0$ such that $\lambda < \frac{1}{2t}$,*

$$\mathbb{E}\exp\Bigl(\lambda \sup_{s\in[0,t]} \|B_s\|^2\Bigr) \leqslant \Bigl(\frac{1+2\lambda t}{1-2\lambda t}\Bigr)^d\,.$$

Sampling guarantees.

We now show how to combine the continuous-time Rényi analysis (Theorem 2.2.25) with the discretization bound (Theorem 6.2.1) to obtain sampling guarantees. The first ingredient we need is a weak triangle inequality (Exercise 6.4).

Lemma 6.2.5 (Weak triangle inequality)**.** *For any $q > 1$, $\lambda \in (0,1)$, and any probability measures μ, ν, π:*

$$\mathcal{R}_q(\mu \,\|\, \pi) \leqslant \frac{q-\lambda}{q-1}\,\mathcal{R}_{q/\lambda}(\mu \,\|\, \nu) + \mathcal{R}_{(q-\lambda)/(1-\lambda)}(\nu \,\|\, \pi)\,. \tag{6.2.6}$$

In particular, for $\lambda = 1/2$,

$$\mathcal{R}_q(\mu \,\|\, \pi) \leqslant \frac{q-1/2}{q-1}\,\mathcal{R}_{2q}(\mu \,\|\, \nu) + \mathcal{R}_{2q-1}(\nu \,\|\, \pi)\,.$$

Next, in order to apply Theorem 6.2.1, we also need a concentration inequality for $\|\nabla V\|$ under π. The following result is taken from Negrea (2022); Altschuler and Chewi (2023).

Lemma 6.2.7 (Score concentration)**.** *Let $\pi \propto \exp(-V)$ and assume that ∇V is β-Lipschitz. Then, ∇V is $\sqrt{\beta}$-sub-Gaussian under π, in the sense that for any $v \in \mathbb{R}^d$,*

$$\mathbb{E}_\pi \exp\langle \nabla V, v\rangle \leqslant \exp\bigl(\frac{\beta}{2}\,\|v\|^2\bigr)\,.$$

In particular, for any $0 < \delta < 1/2$, with probability at least $1-\delta$ under π,

$$\|\nabla V\| \lesssim \sqrt{\beta d} + \sqrt{\beta \log \frac{1}{\delta}}\,.$$

Proof By β-smoothness,

$$\int \exp(-V(\cdot - v)) \geqslant \int \exp\bigl(-V + \langle \nabla V, v\rangle - \frac{\beta}{2}\,\|v\|^2\bigr)\,.$$

Using $\int \exp(-V(\cdot - v)) = \int \exp(-V)$ and rearranging,

$$\exp\bigl(\frac{\beta}{2}\,\|v\|^2\bigr) \geqslant \int \exp\langle \nabla V, v\rangle\,\frac{\exp(-V)}{\int \exp(-V)} = \mathbb{E}_\pi \exp\langle \nabla V, v\rangle\,. \qquad \square$$

We also need to transfer this concentration to different measures, which is accomplished via the following general principle.

Lemma 6.2.8 (Change of measure)**.** *Suppose that a test function ϕ satisfies*

$$\pi\{\phi \geqslant \eta\} \leqslant \psi(\eta)\,, \qquad \text{for all } \eta > 0\,.$$

Then, for any measure μ and any $q > 1$,

$$\mu\{\phi \geqslant \eta\} \leqslant \psi(\eta)^{(q-1)/q} \exp\Bigl(\frac{q-1}{q}\,\mathcal{R}_q(\mu \,\|\, \pi)\Bigr)\,.$$

Proof We note the following simple calculation:

$$\mu\{\phi \geqslant \eta\} = \int \mathbb{1}\{\phi \geqslant \eta\}\, \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\, \mathrm{d}\pi \leqslant \pi\{\phi \geqslant \eta\}^{(q-1)/q} \Big\{\int \Big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Big)^q \,\mathrm{d}\pi\Big\}^{1/q}. \qquad \square$$

We can now prove the following result.

Theorem 6.2.9. *For $n \in \mathbb{N}$, let $\hat\mu_{nh}$ denote the law of the n-th iterate of* LMC *with step size $h > 0$. Assume that ∇V is β-Lipschitz, $q \geqslant 2$, and that ε is sufficiently small.*

1. *If π satisfies a log-Sobolev inequality, $C_{\mathsf{LSI}}(\pi) \leqslant 1/\alpha$, then for an appropriate choice of h we obtain $\mathcal{R}_q(\hat\mu_{Nh} \,\|\, \pi) \leqslant \varepsilon^2$ with*

$$N = \Theta\Big(\frac{\kappa^2 d q^3}{\varepsilon^2} \log^2 \frac{\mathcal{R}_{2q-1}(\hat\mu_0 \,\|\, \pi)}{\varepsilon^2}\Big) \qquad \textit{iterations}\,.$$

2. *If π satisfies a Poincaré inequality, $C_{\mathsf{PI}}(\pi) \leqslant 1/\alpha$, then for an appropriate choice of h we obtain $\mathcal{R}_q(\hat\mu_{Nh} \,\|\, \pi) \leqslant \varepsilon^2$ with*

$$N = \Theta\Big(\frac{\kappa^2 d q^3}{\varepsilon^2}\, \big(\mathcal{R}_{2q-1}(\hat\mu_0 \,\|\, \pi)^2 + \log^2 \frac{1}{\varepsilon}\big)\Big) \qquad \textit{iterations}\,.$$

Proof Let $(\pi_t)_{t\geqslant 0}$ denote the Langevin diffusion started at $\hat\mu_0$. By the weak triangle inequality (Lemma 6.2.5),

$$\mathcal{R}_q(\hat\mu_T \,\|\, \pi) \lesssim \mathcal{R}_{2q}(\hat\mu_T \,\|\, \pi_T) + \mathcal{R}_{2q-1}(\pi_T \,\|\, \pi)\,.$$

By Theorem 2.2.25, the second term is at most ε^2 for $T \asymp (q/\alpha) \log(\mathcal{R}_{2q-1}(\hat\mu_0\|\pi)/\varepsilon^2)$ in the LSI case, and for $T \asymp (q/\alpha)\, (\mathcal{R}_{2q-1}(\hat\mu_0 \,\|\, \pi) + \log 1/\varepsilon)$ in the PI case.

Next, Lemma 6.2.7 yields, for some universal $C > 0$,

$$\pi\{\|\nabla V\| \geqslant C\sqrt{\beta}\,(\sqrt{d} + \eta)\} \leqslant \exp(-\eta^2) \qquad \text{for all } \eta > 0\,.$$

By Lemma 6.2.8, for any $t \in [0, T]$,

$$\pi_t\{\|\nabla V\| \geqslant C\sqrt{\beta}\,(\sqrt{d} + \eta)\} \leqslant \exp\Big(-\frac{2q-1}{2q}\,\big(\eta^2 - \mathcal{R}_{2q}(\pi_t \,\|\, \pi)\big)\Big)\,.$$

By the data-processing inequality (Theorem 1.5.6), $\mathcal{R}_{2q}(\pi_t \,\|\, \pi) \leqslant \mathcal{R}_{2q}(\hat\mu_0 \,\|\, \pi)$. We can conclude that with probability at least $1 - \delta$ under π_t,

$$\|\nabla V\| \lesssim \sqrt{\beta\,\big(d + \mathcal{R}_{2q}(\hat\mu_0 \,\|\, \pi)\big)} + \sqrt{\beta \log \frac{1}{\delta}}\,.$$

From Theorem 6.2.1, provided that $h \lesssim 1/\beta^2 q^2 T$ and $h \lesssim 1/\beta^{3/2} q T^{1/2}$,

$$\mathcal{R}_{2q}(\hat\mu_T \,\|\, \pi_T) \lesssim \beta^3 h^2 q T\, \big(d + \mathcal{R}_{2q}(\hat\mu_0 \,\|\, \pi)\big) + \beta^2 d h q T\,.$$

For simplicity, we use the assumption that ε is sufficiently small in order to focus on the dominant term (i.e., $\beta^2 d h q T$). The theorem follows by choosing h appropriately. $\square$

If we assume that $\mathcal{R}_{2q-1}(\hat\mu_0 \,\|\, \pi) = \widetilde{O}(d)$ (see Exercise 6.6 for justification) and that $\kappa = O(1)$, then the complexity reads $\widetilde{O}(dq^3/\varepsilon^2)$ in the LSI case and $\widetilde{O}(d^3 q^3/\varepsilon^2)$ in the PI case. In the LSI case, this is worse than the result in Theorem 6.1.2 in two ways: first, the dependence on q (because here we did not take advantage of hypercontractivity); and second, Theorem 6.2.9 does not allow for taking an unbounded number of steps of LMC (it suggests that if we run LMC for too long, then we will start to drift farther away from π, which is absurd).

As mentioned above, however, the benefit of the Girsanov approach is its flexibility, which allowed us to tackle the Poincaré case. This flexibility will be important when we study the underdamped Langevin diffusion in the next section. Finally, we remark that the sampling guarantee under a Poincaré inequality will be considerably improved via the proximal sampler in Chapter 8.

6.3 Analysis of ULMC via Girsanov's Theorem

In this section, we apply the Girsanov approach to provide Rényi divergence guarantees for ULMC, which was introduced in Section 5.3. This result provides a "warm start" for MALA, as discussed in Chapter 7.

Let $\mathbf{W}_T$ be the path measure under which B is a standard Brownian motion that drives the underdamped Langevin diffusion:

$$\begin{aligned} \mathrm{d}X_t &= P_t\,\mathrm{d}t\,, \\ \mathrm{d}P_t &= -\gamma P_t\,\mathrm{d}t - \nabla V(X_t)\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,. \end{aligned}$$

Then, let $\mathbf{P}_T$ be the path measure under which $\tilde{B}$ is a standard Brownian motion, which drives the interpolated ULMC algorithm:

$$\begin{aligned} \mathrm{d}X_t &= P_t\,\mathrm{d}t\,, \\ \mathrm{d}P_t &= -\gamma P_t\,\mathrm{d}t - \nabla V(X_{t_-})\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}\tilde{B}_t\,. \end{aligned}$$

We will establish the following discretization bound.

Theorem 6.3.1 (ULMC Rényi discretization bound)**.** *Assume that the potential V is β-smooth and that the step size satisfies $h \lesssim 1/\sqrt{\beta} \wedge 1/\gamma \wedge 1/\beta^{2/3} q^{2/3} T^{1/3}$, where $T := Nh$. Then, for all $q \geqslant 2$,*

$$\begin{aligned} \mathcal{R}_q(\boldsymbol{P}_T \,\|\, \boldsymbol{W}_T) \lesssim \frac{1}{q} \max_{n=0,1,\dots,N-1} \log \mathbb{E}^{\mathbf{W}_T} \exp\Bigl\{O\Bigl(\frac{\beta^2 q^2 T}{\gamma}\,\bigl(h^2\,\|P_{nh}\|^2 + h^4\,\|\nabla V(X_{nh})\|^2\bigr)\Bigr)\Bigr\} \\ + \beta^2 d h^3 q T\,. \end{aligned}$$

Proof Since the argument is similar to the proof of Theorem 6.2.1, we omit some of the steps for brevity. First, from Girsanov's theorem (Theorem 3.2.8), we obtain

$$\begin{aligned} \mathbb{E}^{\mathbf{W}_T}\Bigl[\Bigl(\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{W}_T}\Bigr)^q\Bigr] = \mathbb{E}^{\mathbf{W}_T} \exp\Bigl(&\frac{q}{\sqrt{2\gamma}} \int_0^T \langle \nabla V(X_{t_-}) - \nabla V(X_t), \mathrm{d}B_t\rangle \\ &- \frac{q}{4\gamma} \int_0^T \|\nabla V(X_{t_-}) - \nabla V(X_t)\|^2\,\mathrm{d}t\Bigr)\,. \end{aligned}$$

By the same martingale and AM–GM argument as before, it implies

$$\begin{aligned} \mathbb{E}^{\mathbf{W}_T}\Bigl[\Bigl(\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{W}_T}\Bigr)^q\Bigr] &\leqslant \mathbb{E}^{\mathbf{W}_T} \exp\Bigl(\frac{\beta^2 q^2}{\gamma} \int_0^T \|X_t - X_{t_-}\|^2\,\mathrm{d}t\Bigr) \\ &\leqslant \frac{1}{T} \int_0^T \mathbb{E}^{\mathbf{W}_T} \exp\bigl(\frac{\beta^2 q^2 T}{\gamma}\,\|X_t - X_{t_-}\|^2\bigr)\,\mathrm{d}t\,. \end{aligned}$$

Next, we need an almost-sure version of the movement bound in Lemma 5.3.10; we leave it for the reader to check that if $h \lesssim 1/\sqrt{\beta} \wedge 1/\gamma$, then

$$\|X_t - X_{t-}\| \lesssim h\,\|P_{t_-}\| + h^2\,\|\nabla V(X_{t_-})\| + \gamma^{1/2} h \sup_{s \in [t_-, t_- + h]} \|B_s\|\,.$$

Using the Brownian tail bound from Lemma 6.2.4, it yields

$$\begin{aligned} \exp\Bigl(\frac{\beta^2 q^2 T}{\gamma}\,\|X_t - X_{t_-}\|^2\Bigr) &\leqslant \mathbb{E}^{\mathbf{W}_T} \exp\Bigl(O\Bigl(\frac{\beta^2 q^2 T}{\gamma}\,\bigl(h^2\,\|P_{t_-}\|^2 + h^4\,\|\nabla V(X_{t_-})\|^2\bigr)\Bigr)\Bigr) \\ &\qquad \times \exp\bigl(O(\beta^2 d h^3 q^2 T)\bigr) \end{aligned}$$

provided that $h^3 \lesssim 1/\beta^2 q^2 T$. This completes the proof. □

As in the previous section, this discretization bound leads to sampling guarantees, which we leave as an exercise.

Corollary 6.3.2. *For $n \in \mathbb{N}$, let $\hat\mu_{nh}$ denote the law of $\hat X_{nh}$ in ULMC with step size $h > 0$. Assume that ∇V is β-Lipschitz and for simplicity let us assume that $\mathcal{R}_2(\hat\mu_0 \,\|\, \pi) = \widetilde{O}(d)$. The following guarantees hold with appropriate choices of h and γ.*

1 *If π satisfies a log-Sobolev inequality, $C_{\mathsf{LSI}}(\pi) \leqslant 1/\alpha$, then $\|\hat\mu_{Nh} - \pi\|_{\mathrm{TV}} \leqslant \varepsilon$ with*

$$N = \widetilde\Theta\Bigl(\frac{\kappa^{3/2} d^{1/2}}{\varepsilon}\Bigr) \qquad \textit{iterations}\,.$$

2 *If π satisfies a Poincaré inequality, $C_{\mathsf{PI}}(\pi) \leqslant 1/\alpha$, then we obtain $\mathcal{R}_{3/2}(\hat\mu_{Nh} \,\|\, \pi) \leqslant \varepsilon^2$ with*

$$N = \Theta\Bigl(\frac{\kappa^{3/2} d^2}{\varepsilon}\Bigr) \qquad \textit{iterations}\,.$$

3 *If π satisfies a Poincaré inequality, $C_{\mathsf{PI}}(\pi) \leqslant 1/\alpha$, and is in addition log-concave, then we obtain $\mathcal{R}_{3/2}(\hat\mu_{Nh} \,\|\, \pi) \leqslant \varepsilon^2$ with*

$$N = \Theta\Bigl(\frac{\kappa d^2}{\varepsilon}\Bigr) \qquad \textit{iterations}\,.$$

Despite the fact that Theorem 6.3.1 holds for all Rényi orders $q \geqslant 2$, Corollary 6.3.2 only provides guarantees in TV and $\mathcal{R}_{3/2}$. This is because the hypocoercivity results of Section 5.3.4 hold in KL and χ^2 (recalling that $\chi^2 = \exp(\mathcal{R}_2) - 1$), and we need to further weaken the Rényi order when we apply Pinsker's inequality or the weak triangle inequality. At present, it is not known how to obtain an analogue of the Rényi convergence in Theorem 2.2.25 for the underdamped Langevin diffusion.

Moreover, this means that even in the strongly log-concave case, in order to retain the expected $\widetilde{O}(d^{1/2})$ dimension dependence, Corollary 6.3.2 only furnishes a result in total variation. With a bit more work, Zhang et al. (2023) managed to obtain a result in KL divergence, but we cannot hope for better as long as we rely on the entropic hypocoercivity result (Theorem 5.3.14).

Fortunately, there is another approach: one can argue that the underdamped Langevin diffusion is *regularizing*, in the sense of the following theorem (see the Bibliographical Notes). In Exercise 6.9, we outline an alternative approach which replaces the theorem below with the discrete-time analogue for one step of ULMC.

Theorem 6.3.3 (Hypocoercive regularization). *Let $(\boldsymbol{\mu}_t)_{t\geqslant 0}$, $(\bar{\boldsymbol{\mu}}_t)_{t\geqslant 0}$ denote the marginal laws of two copies of the underdamped Langevin diffusion, started at (x,p) and $(\bar x,\bar p)$ respectively. Assume that the potential V is β-smooth, and that $t \lesssim {}^1\!/\!\sqrt{\beta} \wedge {}^1\!/\!\gamma$. Then, for all $q \geqslant 1$, it holds that*

$$\mathcal{R}_q(\boldsymbol{\mu}_t \,\|\, \bar{\boldsymbol{\mu}}_t) \lesssim \frac{q}{\gamma}\Bigl(\frac{\|x-\bar x\|^2}{t^3} + \frac{\|p-\bar p\|^2}{t}\Bigr)\,.$$

This result then allows us to obtain convergence in Rényi divergence under strong log-concavity via a coupling argument.

Theorem 6.3.4 (Rényi convergence of underdamped Langevin). *Let $(\boldsymbol{\pi}_t)_{t\geqslant 0}$ denote the law of the underdamped Langevin diffusion for a potential V satisfying $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and with friction parameter $\gamma = \sqrt{2\beta}$. Assume that the diffusion is initialized at $\pi_0 \otimes \mathsf{normal}(0, I_d)$, where π_0 satisfies*

$$\inf_{\gamma\in\mathcal{C}(\pi_0,\pi)} \int \exp\Bigl(\frac{\beta\,\|x-\bar x\|^2}{\sigma^2}\Bigr)\,\gamma(\mathrm{d}x,\mathrm{d}\bar x) \leqslant 2 \tag{6.3.5}$$

for some $\sigma > 0$. Then, for all $q \geqslant 2$, we have $\mathcal{R}_q(\boldsymbol{\pi}_t \,\|\, \boldsymbol{\pi}) \leqslant \varepsilon^2 \leqslant 1/q$ for all

$$t \gtrsim \frac{1}{\sqrt{\beta}} + \frac{\sqrt{\beta}}{\alpha}\log\frac{\sigma^2 q}{\varepsilon^2}\,.$$

Proof Write $\boldsymbol{\eta}_t^{x,p}$ for the law of the underdamped Langevin diffusion at time t, started from (x,p). Let $f_q(\cdot) := (\cdot)^q - 1$, and let $\mathcal{D}_{f_q}$ denote the associated divergence (see Section 1.5.2). We take $T_0 \asymp {}^1\!/\!\sqrt{\beta}$ and invoke the joint convexity of f-divergences (Theorem 1.5.6) as well as Theorem 6.3.3 as follows. For any coupling $\boldsymbol{\gamma}$ of $\boldsymbol{\pi}_T$ and $\boldsymbol{\pi}$,

$$\begin{aligned}
&\exp\bigl((q-1)\,\mathcal{R}_q(\boldsymbol{\pi}_{T+T_0} \,\|\, \boldsymbol{\pi})\bigr) = 1 + \mathcal{D}_{f_q}(\boldsymbol{\pi}_{T+T_0} \,\|\, \boldsymbol{\pi})\\
&\quad = 1 + \mathcal{D}_{f_q}\Bigl(\int \boldsymbol{\eta}_{T_0}^{x,p}\,\boldsymbol{\gamma}(\mathrm{d}x,\mathrm{d}p,\mathrm{d}\bar x,\mathrm{d}\bar p) \Bigm\| \int \boldsymbol{\eta}_{T_0}^{\bar x,\bar p}\,\boldsymbol{\gamma}(\mathrm{d}x,\mathrm{d}p,\mathrm{d}\bar x,\mathrm{d}\bar p)\Bigr)\\
&\quad \leqslant 1 + \int \mathcal{D}_{f_q}(\boldsymbol{\eta}_{T_0}^{x,p} \,\|\, \boldsymbol{\eta}_{T_0}^{\bar x,\bar p})\,\boldsymbol{\gamma}(\mathrm{d}x,\mathrm{d}p,\mathrm{d}\bar x,\mathrm{d}\bar p)\\
&\quad = \int \exp\bigl((q-1)\,\mathcal{R}_q(\boldsymbol{\eta}_{T_0}^{x,p} \,\|\, \boldsymbol{\eta}_{T_0}^{\bar x,\bar p})\bigr)\,\boldsymbol{\gamma}(\mathrm{d}x,\mathrm{d}p,\mathrm{d}\bar x,\mathrm{d}\bar p)\\
&\quad \leqslant \int \exp\Bigl\{O\Bigl(\frac{q\,(q-1)}{\gamma}\,\bigl(\frac{\|x-\bar x\|^2}{T_0^3} + \frac{\|p-\bar p\|^2}{T_0}\bigr)\Bigr)\Bigr\}\,\boldsymbol{\gamma}(\mathrm{d}x,\mathrm{d}p,\mathrm{d}\bar x,\mathrm{d}\bar p)\\
&\quad = \int \exp\bigl\{O\bigl(q^2\,(\beta\,\|x-\bar x\|^2 + \|p-\bar p\|^2)\bigr)\bigr\}\,\boldsymbol{\gamma}(\mathrm{d}x,\mathrm{d}p,\mathrm{d}\bar x,\mathrm{d}\bar p)\,.
\end{aligned}$$

This is a "sub-Gaussian" coupling cost between $\boldsymbol{\pi}_T$ and $\boldsymbol{\pi}$. However, recall from Theorem 5.3.5 that we can find a coupling of two copies of the underdamped Langevin diffusion which yields an almost-sure contraction in the $|||\cdot|||$ norm. Hence, if γ_0 denotes the coupling that realizes the infimum in (6.3.5),

$$\begin{aligned} \mathcal{R}_q(\boldsymbol{\pi}_{T+T_0} \,\|\, \boldsymbol{\pi}) &\lesssim \frac{1}{q} \log \int \exp\bigl\{O\bigl(\beta q^2\, |||(x,p) - (\bar{x},\bar{p})|||^2\bigr)\bigr\}\, \boldsymbol{\gamma}(\mathrm{d}x, \mathrm{d}p, \mathrm{d}\bar{x}, \mathrm{d}\bar{p}) \\ &\leqslant \frac{1}{q} \log \int \exp\Bigl\{O\Bigl(\exp\bigl(-\frac{\alpha T}{\sqrt{2\beta}}\bigr)\, \beta q^2\, \|x - \bar{x}\|^2\Bigr)\Bigr\}\, \gamma_0(\mathrm{d}x, \mathrm{d}\bar{x}) \\ &\lesssim \sigma^2 q \exp\bigl(-\frac{\alpha T}{\sqrt{2\beta}}\bigr) \log \int \exp\Bigl(\frac{\beta\, \|x-\bar{x}\|^2}{\sigma^2}\Bigr)\, \gamma_0(\mathrm{d}x, \mathrm{d}\bar{x})\,, \end{aligned}$$

where the last inequality follows from (reverse) Jensen's inequality provided that the condition $\sigma^2 q^2 \exp(-\alpha T/\sqrt{2\beta}) \lesssim 1$ holds. The desired result follows. □

We are now in a position to obtain Rényi guarantees under strong log-concavity.

Theorem 6.3.6 (Altschuler and Chewi (2024a)). *For $n \in \mathbb{N}$, let $\hat{\mu}_{nh}$ denote the law of $\hat{X}_{nh}$ in ULMC with friction parameter $\gamma = \sqrt{2\beta}$ and initialized at $\hat{\boldsymbol{\mu}}_0 = \mathsf{normal}(x_\star, \beta^{-1} I_d) \otimes \mathsf{normal}(0, I_d)$, where $x_\star = \arg\min V$. Assume that the potential V satisfies $0 < \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, for any $q \geqslant 2$ and for step size $h = \widetilde{\Theta}(\varepsilon/\sqrt{\beta\kappa d q})$, we obtain $\mathcal{R}_q(\hat{\mu}_{Nh} \,\|\, \pi) \leqslant \varepsilon^2$ after*

$$N = \widetilde{\Theta}\Bigl(\frac{\kappa^{3/2} d^{1/2} q^{1/2}}{\varepsilon}\Bigr) \qquad \textit{iterations}\,.$$

Proof sketch With this choice of initialization, we can check that (6.3.5) holds with $\sigma^2 \asymp \kappa d$ (see Theorem 2.4.3 and Lemma 4.0.1). From Theorem 6.3.4 we can run the diffusion for a total time of $T \asymp (\kappa/\sqrt{\beta}) \log(\kappa d q/\varepsilon^2)$ to obtain $\mathcal{R}_{2q-1}(\boldsymbol{\pi}_T \,\|\, \boldsymbol{\pi}) \lesssim \varepsilon^2$, where $(\boldsymbol{\pi}_t)_{t \geqslant 0}$ denotes the underdamped Langevin diffusion initialized at $\boldsymbol{\mu}_0$. Next, we apply the discretization bound of Theorem 6.3.1 to control $\mathcal{R}_{2q}(\boldsymbol{\mu}_T \,\|\, \boldsymbol{\pi}_T)$. Since $\mathcal{R}_{2q}(\boldsymbol{\mu}_0 \,\|\, \boldsymbol{\pi}) \leqslant \frac{d}{2} \log \kappa$ (Exercise 1.19), the concentration and change of measure lemmas (Lemma 6.2.7 and Lemma 6.2.8) can be used to further simplify the bound, similarly to the proof of Theorem 6.2.9, eventually yielding $\mathcal{R}_{2q}(\boldsymbol{\mu}_T \,\|\, \boldsymbol{\pi}_T) \lesssim \beta^{3/2} d h^2 q T$. Choosing the step size h appropriately and applying the weak triangle inequality (Lemma 6.2.5) concludes the proof. □

Bibliographical Notes

Rényi guarantees for LMC were first considered in Vempala and Wibisono (2019), which proved convergence of LMC to its biased stationary distribution provided that the biased limit satisfies a Poincaré or log-Sobolev inequality. 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. Mixing time bounds toward the biased stationary distribution were also later considered in Altschuler and Talwar (2023).

Motivated by applications to differential privacy, Ganesh and Talwar (2020) 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. This result was refined in Erdogdu et al. (2022) via a two-phase analysis, still relying on the adaptive composition lemma. Subsequently, building off the earlier work of Chewi et al. (2021) (which essentially contains a one-step Rényi discretization argument), it was realized in Chewi et al. (2024a) that the earlier arguments of Ganesh and Talwar (2020); Erdogdu et al. (2022) can be streamlined by replacing the adaptive composition lemma entirely with Girsanov's theorem. The proofs of Theorem 6.1.2, Exercise 6.2, and Exercise 6.6 are also from Chewi et al. (2024a). The proof of Theorem 6.2.9 given here is a further refinement of Chewi et al. (2024a) using the score concentration lemma (Lemma 6.2.7).

The use of Girsanov's theorem to obtain Rényi discretization bounds for ULMC first appeared in Zhang et al. (2023), which considered a variety of settings (including weakly smooth potentials). In the strongly log-concave case, the argument was simplified in Altschuler and Chewi (2024a).

The hypocoercive regularization result (Theorem 6.3.3) is an example of a *reverse transport inequality*, which is dual to a (dimension-free, parabolic) *Harnack inequality*. We have already encountered such Harnack inequalities for the Langevin diffusion in Exercise 2.17 and Exercise 3.4; see the Bibliographic Notes to Chapter 2. In contrast, the inequality in Theorem 6.3.3 holds for the underdamped Langevin diffusion, and is consequently more difficult to establish. It was first shown in Guillin and Wang (2012, Corollary 4.6); see also Altschuler et al. (2025, Theorem 3.2), where it is established using the method of Exercise 3.4. Actually, Theorem 6.3.4, which combines short-time regularization with long-time contraction, is not the best approach, and Altschuler et al. (2025, Theorem 3.2) provides a sharper result.

The Rényi sampling guarantee for ULMC in Theorem 6.3.6 was established first in Altschuler and Chewi (2024a) via the approach outlined in Exercise 6.9. The presentation of the arguments here is inspired by the treatment in Altschuler and Chewi (2024b); Altschuler et al. (2025).

Exercises

Analysis of LMC 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 $\hat\mu_0 = \mathsf{normal}(0, \beta^{-1} I_d)$. The goal of this exercise is to prove a Rényi discretization bound for LMC under these assumptions.

1 First, show that $\hat\mu_t$ satisfies an LSI for all $t \geqslant 0$, and write down a bound for $C_{\mathsf{LSI}}(\hat\mu_t)$ (the bound should grow linearly with t).

Hint: See Exercise 4.4.

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 $\mathsf{KL}(\mathbf{P} \,\|\, \mathbb{P}) \leqslant \frac{q-1}{q} \mathsf{KL}(\psi_t \hat\mu_t \,\|\, \pi)$ and we applied the LSI for π. This time, use $\mathsf{KL}(\mathbf{P} \,\|\, \mathbb{P}) = \mathsf{KL}(\psi_t \hat\mu_t \,\|\, \hat\mu_t)$ and apply the LSI for $\hat\mu_t$ 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 2.2.25). (Since π is log-concave, it necessarily satisfies a Poincaré inequality with some constant $1/\alpha$, see the Bibliographical Notes to Chapter 2.)

Prove that if $\varepsilon \leqslant \sqrt{1/q} \wedge \sqrt{\kappa d}$, then with an appropriate choice of step size h and with $N = \widetilde{\Theta}(\kappa^2 d^2 q^3/\varepsilon^2)$ iterations of LMC, we obtain $\sqrt{\mathcal{R}_q(\hat\mu_{Nh} \,\|\, \pi)} \leqslant \varepsilon$. (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.)

Analysis of LMC via Girsanov's Theorem

⊵ **Exercise 6.3** (Sub-Gaussian bound for Brownian motion)
Prove Lemma 6.2.4.

Hint: The reflection principle states that if $(\tilde{B}_t)_{t\geqslant 0}$ is a one-dimensional Brownian motion, then for every $\eta > 0$, $\mathbb{P}(\sup_{s\in[0,t]} \tilde{B}_s > \eta) = 2\,\mathbb{P}(\tilde{B}_t > \eta)$.

⊵ **Exercise 6.4** (Weak triangle inequality)
Prove the weak triangle inequality (Lemma 6.2.5). What happens if we take $q = 1+\varepsilon$, $\lambda = 1-\varepsilon$, and send $\varepsilon \searrow 0$?

⊵ **Exercise 6.5** (Score concentration in the non-smooth case)
Extend Lemma 6.2.7 to the case where ∇V is simply assumed to be Hölder continuous with some exponent $s \in [0,1]$:

$$\|\nabla V(x) - \nabla V(y)\| \leqslant \beta\,\|x-y\|^s\,, \qquad \text{for all } x,y\in\mathbb{R}^d\,.$$

⊵ **Exercise 6.6** (A more general initialization bound)
Suppose that V is β-smooth, but not necessarily convex. In this case, it is not realistic to assume that we have access to the minimizer of V, but we can at least assume that we have computed a stationary point: $\nabla V(0) = 0$. Show that for $\hat{\mu}_0 = \mathsf{normal}(0,(2\beta)^{-1}I_d)$,

$$\mathcal{R}_\infty(\hat{\mu}_0 \,\|\, \pi) \leqslant 2 + V(0) - \inf V + \frac{d}{2}\log(4\beta\mathfrak{m}^2)\,,$$

where $\mathfrak{m} := \mathbb{E}_\pi\|\cdot\|$ is the first moment of π. This shows that if we can compute a stationary point with suboptimality gap $\widetilde{O}(d)$, then it is reasonable to assume $\mathcal{R}_\infty(\hat{\mu}_0 \,\|\, \pi) = \widetilde{O}(d)$.

Hint: Use the decomposition

$$\sup\frac{\hat{\mu}_0}{\pi} \leqslant \sup_{x\in\mathbb{R}^d}\exp\bigl(V(x) - \beta\,\|x\|^2\bigr)\,\frac{\int\exp(-V)}{\int\exp(-V-\delta\,\|\cdot\|^2)}\,\frac{\int\exp(-V-\delta\,\|\cdot\|^2)}{(\pi/\beta)^{d/2}}$$

and bound each ratio separately.

Analysis of ULMC via Girsanov's Theorem

⊵ **Exercise 6.7** (Rényi guarantees for ULMC)
This question fills some omitted details from the proofs.

1 Fill in the missing details in the proof of Theorem 6.3.1.
2 Prove Corollary 6.3.2 by combining Theorem 6.3.1 with the hypocoercivity results from Section 5.3.4.
3 Fill in the details in the proof of Theorem 6.3.6.

⊵ **Exercise 6.8** (KL guarantees via Fisher information convergence)
In this exercise, we outline the approach of Zhang et al. (2023) for obtaining KL guarantees in the strongly log-concave case. Thus, we let $(\hat{\boldsymbol{\mu}}_{nh})_{n\in\mathbb{N}}$ denote the laws of the iterates of ULMC, we let $(\boldsymbol{\pi}_t)_{t\geqslant 0}$ denote the law of the underdamped Langevin diffusion started at $\boldsymbol{\pi}_0 = \hat{\boldsymbol{\mu}}_0$, and $T := Nh$. Also, let $(\hat{\boldsymbol{\mu}}^\circ_{nh})_{n\in\mathbb{N}}$, $(\boldsymbol{\pi}^\circ_t)_{t\geqslant 0}$, and $\boldsymbol{\pi}^\circ$ denote the corresponding probability measures expressed in the twisted coordinates of Theorem 5.3.5.

1 Prove that

$$\mathsf{KL}(\hat{\boldsymbol{\mu}}_T^\circ \,\|\, \boldsymbol{\pi}^\circ) \leqslant \mathsf{KL}(\hat{\boldsymbol{\mu}}_T^\circ \,\|\, \boldsymbol{\pi}_T^\circ) + \mathsf{KL}(\boldsymbol{\pi}_T^\circ \,\|\, \boldsymbol{\pi}^\circ) + \sqrt{\chi^2(\hat{\boldsymbol{\mu}}_T^\circ \,\|\, \boldsymbol{\pi}_T^\circ)\, \mathrm{var}_{\boldsymbol{\pi}_T^\circ} \log \frac{\boldsymbol{\pi}_T^\circ}{\boldsymbol{\pi}^\circ}}\,.$$

Note that this is intended as a substitute for the triangle inequality, and an alternative to the weak triangle inequality (Lemma 6.2.5).

2 Recall that in the strongly log-concave case, if $C_{\mathsf{LSI}}(\hat{\boldsymbol{\mu}}_0) < \infty$, then $(\boldsymbol{\pi}_t)_{t\geqslant 0}$ satisfies a uniform-in-time LSI (Exercise 5.10). Use this, together with the Fisher information convergence implied by Theorem 5.3.14, to control the variance term.

3 Combine this with Theorem 6.3.1 to give a complete KL divergence guarantee for ULMC in the strongly log-concave case.

⊵ Exercise 6.9 (Discrete-time version of hypocoercive regularization)

In this exercise, we outline a simpler, discrete-time alternative to the hypocoercive regularization result (Theorem 6.3.3) which suffices to prove Theorem 6.3.6. Let Σ denote the covariance matrix for a single step of ULMC, as given in Lemma 5.3.8. Recall, however, that for the contraction result we work with the twisted coordinates $(x, x + \frac{2}{\gamma}\, p)$. Let $\mathcal{M}$ denote the matrix corresponding to the change of coordinates $(x, p) \mapsto (x, x + \frac{2}{\gamma}\, p)$, so that the covariance matrix in this system of coordinates is $\bar{\Sigma} := \mathcal{M}\Sigma\mathcal{M}^\mathsf{T}$. Also, let $\hat{\boldsymbol{P}}$ denote the Markov kernel corresponding to one iteration of ULMC.

1 Prove that for $h \lesssim 1/\gamma$, $\lambda_{\min}(\bar{\Sigma}) \gtrsim \gamma h^3$.

2 Using the explicit expression for the Rényi divergence between Gaussians, prove the following regularization bound:

$$\mathcal{R}_q(\delta_{x,p}\hat{\boldsymbol{P}}, \delta_{\bar{x},\bar{p}}\hat{\boldsymbol{P}}) \lesssim \frac{q}{\gamma h^3}\, |||(x,p) - (\bar{x},\bar{p})|||^2\,.$$

3 Although we proved that the diffusion contracts in this coordinate system (Theorem 5.3.5), our aim now is to work with the discrete-time algorithm. Adapt the proof of Theorem 5.3.5, using Lemma 5.3.8, to show that if $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $h \lesssim 1/\beta^{1/2}\kappa$, then $\hat{\boldsymbol{P}}$ is an almost-sure contraction:

$$W_{\infty,|||\cdot|||}(\delta_{x,p}\hat{\boldsymbol{P}}, \delta_{\bar{x},\bar{p}}\hat{\boldsymbol{P}}) \leqslant \Bigl(1 - \Omega\bigl(\frac{\alpha h}{\sqrt{\beta}}\bigr)\Bigr)\, |||(x,p) - (\bar{x},\bar{p})|||\,,$$

where $W_{\infty,|||\cdot|||}$ denotes the W_∞ metric with respect to $|||\cdot|||$.

4 Similarly to the proof of Theorem 6.3.4, prove that $\mathcal{R}_q(\hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{P}}^N \,\|\, \boldsymbol{\pi}\hat{\boldsymbol{P}}^N) \leqslant \varepsilon^2$, where $\hat{\boldsymbol{\mu}}_0 = \mathsf{normal}(x_\star, \beta^{-1}I_d) \otimes \mathsf{normal}(0, I_d)$, provided that

$$N \gtrsim \frac{\sqrt{\beta}}{\alpha h} \log \frac{\beta^{3/2}\kappa d q}{\varepsilon^2 h^3}\,.$$

5 Combine this with Theorem 6.3.1 to give another proof of Theorem 6.3.6.

⊵ Exercise 6.10 (Rényi guarantees in the log-concave case)

In Altschuler et al. (2025), the following variant of Theorem 6.3.4 was established. Let $(\boldsymbol{P}_t)_{t\geqslant 0}$ denote the Markov semigroup for the underdamped Langevin diffusion, and assume that $0 \preceq \nabla^2 V \preceq \beta I_d$ and $\gamma \geqslant \sqrt{32\beta}$. Then, for all $q \geqslant 1$, $T > 0$, and $x, \bar{x}, p, \bar{p} \in \mathbb{R}^d$,

$$\mathcal{R}_q(\delta_{x,p}\boldsymbol{P}_T \,\|\, \delta_{\bar{x},\bar{p}}\boldsymbol{P}_T) \lesssim q\left(\frac{1}{\gamma T^3} + \frac{\gamma}{T}\right)\left(\|x - \bar{x}\|^2 + \frac{1}{\gamma^2 + 1/T^2}\,\|p - \bar{p}\|^2\right).$$

Can you combine this with Theorem 6.3.1 to establish Rényi guarantees for ULMC in this case?

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; we say that 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 $\operatorname{poly}(1/\varepsilon)$. In this section, we address the problem of designing *high-accuracy* samplers, i.e., samplers whose complexity scales as $\operatorname{polylog}(1/\varepsilon)$. 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 $\widetilde{\mu} \geqslant \widetilde{\pi}$. 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 (Rejection sampling)**.** *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 $Z_\mu := \int \widetilde{\mu}$ and $Z_\pi := \int \widetilde{\pi}$.*

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,

$$\begin{aligned}
\mathbb{P}(X \in A) &= \sum_{n=0}^\infty \mathbb{P}\Bigl(X_{n+1} \in A,\ U_i > \frac{\widetilde{\pi}(X_i)}{\widetilde{\mu}(X_i)}\ \forall i \in [n],\ U_{n+1} \leqslant \frac{\widetilde{\pi}(X_{n+1})}{\widetilde{\mu}(X_{n+1})}\Bigr) \\
&= \sum_{n=0}^\infty \mathbb{P}\Bigl(X_{n+1} \in A,\ U_{n+1} \leqslant \frac{\widetilde{\pi}(X_{n+1})}{\widetilde{\mu}(X_{n+1})}\Bigr)\,\mathbb{P}\Bigl(U_1 > \frac{\widetilde{\pi}(X_1)}{\widetilde{\mu}(X_1)}\Bigr)^n \\
&= \sum_{n=0}^\infty \Bigl(\int_A \frac{\widetilde{\pi}}{\widetilde{\mu}}\,\mathrm{d}\mu\Bigr)\Bigl(\int \bigl(1 - \frac{\widetilde{\pi}}{\widetilde{\mu}}\bigr)\,\mathrm{d}\mu\Bigr)^n = \frac{Z_\pi}{Z_\mu}\,\pi(A)\sum_{n=0}^\infty \Bigl(1 - \frac{Z_\pi}{Z_\mu}\Bigr)^n = \pi(A)\,. \qquad \square
\end{aligned}$$

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 $\widetilde{\pi} \leqslant \exp(-\frac{\alpha}{2}\,\|\cdot\|^2)$, and we can take $\widetilde{\mu} := \exp(-\frac{\alpha}{2}\,\|\cdot\|^2)$, 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,

$$\frac{Z_\mu}{Z_\pi} = \frac{(2\pi/\alpha)^{d/2}}{\int \exp(-V)} \leqslant \frac{(2\pi/\alpha)^{d/2}}{\int \exp(-\frac{\beta}{2}\,\|\cdot\|^2)} = \frac{(2\pi/\alpha)^{d/2}}{(2\pi/\beta)^{d/2}} = \kappa^{d/2}\,,$$

with $\kappa := \beta/\alpha$. We summarize this result in the following proposition.

Proposition 7.1.2 (Rejection sampling under strong log-concavity). *Let the target* $\pi \propto \exp(-V)$ *on* $\mathbb{R}^d$ *satisfy* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$, $V(0) = 0$, *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 $\|\mu_N - \pi\|_{\mathrm{TV}} \leqslant \varepsilon$. If we flip this around and fix the target accuracy ε, we see that the number of iterations required to achieve this guarantee is $N \geqslant \kappa^{d/2} \log(1/\varepsilon)$.

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 $n = 1, 2, 3, \dots$:

1 Propose a new point $Y_n \sim Q(X_{n-1}, \cdot)$.
2 With probability $A(X_{n-1}, Y_n)$, set $X_n := Y_n$; otherwise, set $X_n := X_{n-1}$. Here, A is the acceptance probability defined via (7.2.1).

This algorithm defines a discrete-time Markov chain whose transition kernel P can be written explicitly as

$$P(x, \mathrm{d}y) = Q(x, \mathrm{d}y)\, A(x, y) + \underbrace{\Bigl(1 - \int Q(x, \mathrm{d}y')\, A(x, y')\Bigr)}_{\text{rejection probability}}\, \delta_x(\mathrm{d}y)\,. \tag{7.2.2}$$

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 (Reversibility of Metropolis–Hastings)**.** *The Metropolis–Hastings algorithm with proposal Q is reversible with respect to π.*

Proof We want to check that $\pi(x)\, P(x, y) = \pi(y)\, P(y, x)$ for all $x, y \in \mathbb{R}^d$ with $x \neq y$. We can write

$$\begin{aligned}\pi(x)\, P(x, y) = \pi(x)\, Q(x, y)\, A(x, y) &= \pi(x)\, Q(x, y) \min\Bigl\{1, \frac{\pi(y)\, Q(y, x)}{\pi(x)\, Q(x, y)}\Bigr\} \\ &= \min\{\pi(x)\, Q(x, y),\ \pi(y)\, Q(y, x)\}\end{aligned}$$

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 Billera and Diaconis (2001). Given a proposal kernel Q, let $P(Q)$ denote the Metropolis–Hastings kernel obtained from Q (see (7.2.2)). Then, the mapping $Q \mapsto P(Q)$ 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, defined as follows:

$$\mathsf{d}(P,P') := \int_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus\mathsf{diag}} |P(x,y) - P'(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y \tag{7.2.4}$$

where $\mathsf{diag} := \{(x,x) \mid x \in \mathbb{R}^d\}$ is the diagonal in $\mathbb{R}^d \times \mathbb{R}^d$.

Theorem 7.2.5 (Billera and Diaconis (2001))**.** *Let $\mathscr{R}(\pi)$ denote the space of kernels P which are reversible with respect to π, and such that for each $x \in \mathbb{R}^d$, $P(x,\cdot)$ admits a density with respect to Lebesgue measure (except possibly having an atom at x). Then,*

$$P(Q) \in \operatorname*{arg\,min}_{P\in\mathscr{R}(\pi)} \mathsf{d}(P,Q)\,.$$

Proof Let $P \in \mathscr{R}(\pi)$, and let $S := \{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d \mid \pi(x)\,Q(x,y) > \pi(y)\,Q(y,x)\}$. Then,

$$\begin{aligned}\mathsf{d}(P,Q) &= \int_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus\mathsf{diag}} |P(x,y) - Q(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y\\ &= \int_S |P(x,y) - Q(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y + \int_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus(S\cup\mathsf{diag})} |P(x,y) - Q(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y\\ &= \int_S |P(x,y) - Q(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y + \int_S |P(y,x) - Q(y,x)|\,\pi(\mathrm{d}y)\,\mathrm{d}x\,.\end{aligned}$$

Using reversibility of P, the second term is

$$\begin{aligned}\int_S |P(y,x) - Q(y,x)|\,\pi(\mathrm{d}y)\,\mathrm{d}x &= \int_S |\pi(x)\,P(x,y) - \pi(y)\,Q(y,x)|\,\mathrm{d}x\,\mathrm{d}y\\ &\geqslant \int_S |\pi(x)\,Q(x,y) - \pi(y)\,Q(y,x)|\,\mathrm{d}x\,\mathrm{d}y - \int_S |P(x,y) - Q(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y\,.\end{aligned} \tag{7.2.6}$$

Putting this together, $\mathsf{d}(P,Q) \geqslant \int_S |\pi(x)\,Q(x,y) - \pi(y)\,Q(y,x)|\,\mathrm{d}x\,\mathrm{d}y$, so we have obtained a lower bound which does not depend on P. On the other hand, we can check that the only inequality (7.2.6) that we used is an equality for $P = P(Q)$. □

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 take $Q(x, \cdot) = \mathsf{normal}(x, hI_d)$, 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 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**.

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 $k = 0, 1, 2, \dots, K-1$:

1 Set $p_{(k+\frac{1}{2})h} := p_{kh} - \frac{h}{2}\,\nabla V(x_{kh})$.
2 Set $x_{(k+1)h} := x_{kh} + h\, p_{(k+\frac{1}{2})h}$.
3 Set $p_{(k+1)h} := p_{(k+\frac{1}{2})h} - \frac{h}{2}\,\nabla V(x_{(k+1)h})$.

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.

---

[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.

Metropolized Hamiltonian Monte Carlo (MHMC): Initialize at $X_0 \sim \mu_0$. For iterations $n = 0, 1, 2, \ldots$:

1 *Refresh* the momentum: draw $P_n \sim \mathsf{normal}(0, I_d)$.
2 *Propose* a trajectory: let $(X_n', P_n') \coloneqq F_{\mathrm{leap}}(X_n, P_n)$.
3 *Accept* the trajectory with probability $1 \wedge \exp\{H(X_n, P_n) - H(X_n', P_n')\}$. If the trajectory is accepted, set $X_{n+1} \coloneqq X_n'$; otherwise, we set $X_{n+1} \coloneqq X_n$.

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 (Correctness of MHMC)**.** *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, $Q((x, p), \cdot) = \delta_{F_T(x,p)}$.[2]

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 5.5 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 $P_n \sim \mathsf{normal}(0, I_d)$, then $-P_n \sim \mathsf{normal}(0, I_d)$ 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 5.5) 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 $F_{\mathrm{leap}}^{-1} = R \circ F_{\mathrm{leap}} \circ R$, and this is where we use the specific form of the leapfrog integrator. We leave the verification as Exercise 7.2. □

---

[2] 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.

***Remark* 7.3.3.** The proof shows that the proposal of MHMC should really be thought of as $F_{\mathrm{leap}} \circ R$, instead of F_{leap}. In fact, if we did not refresh the momentum, then repeatedly applying the idempotent operator $F_{\mathrm{leap}} \circ R$ 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 $\mathsf{Refresh}$ denotes the momentum refreshment step, we should think of MHMC as the composition

$$\mathsf{MHMC} = R \circ \mathsf{MH}(F_{\mathrm{leap}} \circ R) \circ \mathsf{Refresh}\,.$$

This is simplified to

$$\mathsf{MHMC} = \mathsf{MH}(F_{\mathrm{leap}} \circ R) \circ \mathsf{Refresh}$$

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 P with the modified kernel P_ℓ given by

$$P_\ell(x, \mathrm{d}y) = (1-\ell)\, P(x, \mathrm{d}y) + \ell\, \delta_x(\mathrm{d}y)\,.$$

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 chapter, 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.

Cold start vs. warm start.

When discussing Metropolis–Hastings algorithms, we must distinguish between convergence rates when initialized at a **cold start**, vs. a **warm start**. These terms are not precisely defined, but loosely speaking a cold start refers to an easily computable distribution which works well uniformly over the class of target distributions under consideration. In this section, a cold 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 $\chi^2(\mu_0 \parallel \pi) = O(1)$. Unsurprisingly, the rates are faster with a warm start.

One obvious benefit of a warm start is simply that we start closer to stationarity, so it takes less time to approach π. This aspect of the situation is similar to the discussion in Section 1.5. Basically, the simplest way to study Markov chains is via spectral theory, which is related to Poincaré inequalities and hence to the chi-squared divergence at initialization. Since the initial chi-squared divergence tends to be quite large—of order $\exp(d)$—the resulting analysis incurs slow mixing. This source of slow mixing, however, is sometimes illusory, as it can be addressed via stronger methods such as a log-Sobolev inequality. For example, in the case of LMC, the guarantee of Theorem 4.2.6 only improves by a logarithmic factor if we assume the warm start condition $\mathsf{KL}(\mu_0 \parallel \pi) = O(1)$.

On the other hand, for Metropolis–Hastings algorithms, we will see another crucial benefit of a warm start: it *permits us choose a larger step size while maintaining a high acceptance probability*. Indeed, let us perform a thought experiment: for each step size h, let $A_h \subseteq \mathbb{R}^d$ denote the region of space where a proposal starting from A_h has high acceptance probability. If the choice of step size h_{cold} allows the algorithm to succeed from a cold start, it means that the algorithm should begin accepting the proposals from a potentially poor initialization, so corresponding acceptance region is nearly full: $A_{h_{\text{cold}}} \approx \mathbb{R}^d$. However, this condition on the step size may be too stringent, and optimistically we could hope that it suffices to take a step size h_{warm} for which the corresponding acceptance region has high probability at stationarity: $\pi(A_{h_{\text{warm}}}) \approx 1$. In this case, in order for the algorithm to be initialized in $A_{h_{\text{warm}}}$, we require $A_{h_{\text{warm}}}$ to also have high probability under μ_0: $\mu_0(A_{h_{\text{warm}}}) \approx 1$. In turn, this is implied by the warm start condition $\chi^2(\mu_0 \,\|\, \pi) = O(1)$. It turns out that the step size thresholds h_{cold} and h_{warm} can differ substantially, and as we shall see, $h_{\text{cold}} \asymp 1/d$ and $h_{\text{warm}} \asymp 1/d^{1/2}$ for MALA. This latter effect is more fundamental and necessitates carefully delineating the two regimes (cold vs. warm).

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 (Cold start case, Dwivedi et al. (2019); Chen et al. (2020); Lee et al. (2020); Andrieu et al. (2024))**.** *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* $\sqrt{\chi^2(\mu_N \,\|\, \pi)} \leqslant \varepsilon$ *after*

$$N = \widetilde{O}\big(\kappa d \operatorname{polylog} \frac{1}{\varepsilon}\big) \qquad \textit{iterations}\,.$$

2. *MALA outputs a measure* μ_N *satisfying* $\sqrt{\chi^2(\mu_N \,\|\, \pi)} \leqslant \varepsilon$ *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, i.e.,* $\|\nabla^2 V(x) - \nabla^2 V(y)\|_{\text{op}} \lesssim \|x - y\|$ *for* $x, y \in \mathbb{R}^d$. *Then, MHMC outputs a measure* μ_N *satisfying* $\sqrt{\chi^2(\mu_N \,\|\, \pi)} \leqslant \varepsilon$ *after*

$$N = \widetilde{O}\big((\kappa^{3/4} d + \kappa^{13/12} d^{2/3} + \kappa^{7/4} d^{1/2}) \operatorname{polylog} \frac{1}{\varepsilon}\big) \qquad \textit{gradient queries}\,.$$

Note that the result for MHMC is not directly comparable because it makes a stronger higher-order smoothness assumption. Also, if we focus on the dimension dependence, all of the results incur at least a linear dependence on d.

Next, we present the results under a warm start.

Theorem 7.3.5 (Warm start case, Chewi et al. (2021); Wu et al. (2021); Chen and Gatmiry (2023a,b)). *Suppose that* $\pi \propto \exp(-V)$ *satisfies* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. *Consider the following Metropolis–Hastings algorithms initialized at a distribution satisfying* $\chi^2(\mu_0 \,\|\, \pi) = O(1)$ *and with an appropriately tuned choice of parameters.*

1. *MALA outputs a measure* μ_N *satisfying* $\|\mu_N - \pi\|_{\mathrm{TV}} \leqslant \varepsilon$ *after*
$$N = \widetilde{O}\Big(\kappa d^{1/2} \operatorname{polylog} \frac{1}{\varepsilon}\Big) \qquad \textit{iterations}\,.$$
2. *Assume in addition that* $\nabla^3 V$ *is bounded. Then, MHMC outputs a measure* μ_N *satisfying* $\|\mu_N - \pi\|_{\mathrm{TV}} \leqslant \varepsilon$ *after*
$$N = \widetilde{O}\Big(\kappa d^{1/4} \operatorname{polylog} \frac{1}{\varepsilon}\Big) \qquad \textit{iterations}\,.$$

In keeping with the theme of the rest of the book, we will focus on the results which do not assume higher-order smoothness. As we shall see, the results for MALA in both the feasible and warm start cases are sharp in a suitable sense. In other words, the improved guarantees under a warm start are not a failure of the analysis technique, but are intrinsic to the problem. Hence, it is also important to study how we can *obtain* a warm start, i.e., a measure satisfying $\chi^2(\mu_0 \,\|\, \pi) = O(1)$. In Chapter 6, or more specifically Corollary 6.3.2, we accomplished this goal via our analysis of ULMC in Rényi divergence. Critically, the cost of computing the warm start is not larger (at least with respect to dimension dependence) than the cost of running MALA afterwards using the warm start, leading to the following high-accuracy guarantee.

Corollary 7.3.6 (Altschuler and Chewi (2024a)). *Suppose that* $\pi \propto \exp(-V)$ *satisfies* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ *and* $\nabla V(0) = 0$. *There is sampler which outputs* μ *satisfying* $\sqrt{\mathsf{KL}(\mu \,\|\, \pi)} \leqslant \varepsilon$ *using*
$$N = \widetilde{O}\Big(\kappa^{3/2} d^{1/2} + \kappa d^{1/2} \operatorname{polylog} \frac{1}{\varepsilon}\Big) \qquad \textit{first-order queries}\,.$$

This result will be extended and refined in various ways through the use of the proximal sampler in Chapter 8: we will relax the assumption of strong log-concavity to a log-Sobolev inequality; we will eliminate the $\kappa^{3/2} d^{1/2}$ term from the rate; and we will show how to obtain this result using the proximal sampler alone.

The goal for the rest of the chapter is to prove the convergence results for MALA.

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 $(P^n)_{n\in\mathbb{N}}$, 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 $\mathbb{E}_\pi f = 0$,

$$\langle f, (-\mathscr{L})\, f\rangle_{L^2(\pi)} \geqslant \lambda\, \|f\|^2_{L^2(\pi)}\,.$$

Equivalently, if (X_0, X_1) are two successive iterates of the chain started at stationarity, it is equivalent to require

$$2\lambda \operatorname{var} f(X_0) \leqslant \mathbb{E}[|f(X_1) - f(X_0)|^2]\,. \tag{7.4.1}$$

In analogy with Section 1.2, we also say that P satisfies a **Poincaré inequality** with constant $1/\lambda$. We already saw in Section 2.6 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 (Mixing under a spectral gap)**.** *Suppose that the spectral gap of P is $\lambda > 0$. Then, for the law $(\mu_n)_{n\in\mathbb{N}}$ of the iterates of the $\frac{1}{2}$-lazy version of P, we have*

$$\chi^2(\mu_N \,\|\, \pi) \leqslant \exp(-\lambda N)\, \chi^2(\mu_0 \,\|\, \pi)\,.$$

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 $f \geqslant 0$,

$$\operatorname{ent}_\pi f \leqslant \frac{C_{\mathsf{MLSI}}}{2}\, \mathscr{E}(f, \log f)\,.$$

We have already encountered this inequality as Definition 1.2.25, 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 $(P_t)_{t\geqslant 0}$ via $P_t := \exp(t\mathscr{L})$. Note that the generator of $(P_t)_{t\geqslant 0}$ is also $\mathscr{L}$ and hence the Dirichlet energy for $(P_t)_{t\geqslant 0}$ coincides with the Dirichlet energy for $(P^n)_{n\in\mathbb{N}}$. Now, if we apply the calculation (1.2.24) to the semigroup $(P_t)_{t\geqslant 0}$, we obtain exponential decay of the KL divergence under an MLSI:

$$\mathsf{KL}(\mu P_t \parallel \pi) \leqslant \exp\Bigl(-\frac{2t}{C_{\mathsf{MLSI}}}\Bigr)\, \mathsf{KL}(\mu \parallel \pi)\,,$$

see Theorem 1.2.26.

Moreover, the continuous-time semigroup $(P_t)_{t\geqslant 0}$ can be simulated. Namely, let $(\tau_n)_{n\in\mathbb{N}^+} \overset{\text{i.i.d.}}{\sim} \mathsf{exponential}(1)$, $T_n := \sum_{k=1}^n \tau_k$, and consider the following algorithm. Initialize at $X_0 \sim \mu_0$, and for $n = 0, 1, 2, \dots$, let $X_{T_{n+1}} \sim P(X_{T_n}, \cdot)$, so that $(X_{T_n})_{n\in\mathbb{N}}$ are the iterates of the discrete-time Markov chain with kernel P. Also, for $T_n \leqslant t < T_{n+1}$, set $X_t := X_{T_n}$. This yields a continuous-time Markov process $(X_t)_{t\geqslant 0}$, and one can check that the associated Markov semigroup is exactly $(P_t)_{t\geqslant 0}$. Moreover, by concentration of i.i.d. sums, it holds that $T_n \approx n$, so that if the semigroup $(P_t)_{t\geqslant 0}$ requires time T_{mix} in order to mix to a desired level of accuracy, then the algorithm which simulates $(X_t)_{t\geqslant 0}$ requires $\approx T_{\mathrm{mix}}$ 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 $(P^n)_{n\in\mathbb{N}}$. 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 $(X_t)_{t\geqslant 0}$) 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 $A \subseteq \mathbb{R}^d$,

$$\int_A P(x, A^{\mathsf{c}})\, \pi(\mathrm{d}x) \geqslant \mathfrak{c}\, \pi(A)\, \pi(A^{\mathsf{c}})\,.$$

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, Lawler and Sokal (1988))**.** *The conductance $\mathfrak{c}$ and the spectral gap λ satisfy the inequalities*

$$\frac{1}{8}\,\mathfrak{c}^2 \leqslant \lambda \leqslant \mathfrak{c}\,.$$

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 (Functional form of conductance)**.** *Let the conductance of the chain be* $\mathfrak{c} > 0$*. Then, for all* $f \in L^1(\pi)$*,*

$$\mathbb{E}_\pi |f - \mathbb{E}_\pi f| \leqslant \frac{1}{\mathfrak{c}} \, \mathbb{E}|f(X_1) - f(X_0)| \,, \tag{7.4.5}$$

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,

$$\mathbb{E}_\pi |f - \mathbb{E}_\pi f| = \mathbb{E}|f(X_0') - \mathbb{E} f(X_0)| \leqslant \mathbb{E}|f(X_0') - f(X_0)| \,.$$

On the other hand, by reversibility,

$$\begin{aligned}
\mathbb{E}|f(X_1) - f(X_0)| &= \iint |f(x_1) - f(x_0)| \, P(x_0, \mathrm{d}x_1) \, \pi(\mathrm{d}x_0) \\
&= 2 \iint \mathbb{1}\{f(x_1) > f(x_0)\} \, [f(x_1) - f(x_0)] \, P(x_0, \mathrm{d}x_1) \, \pi(\mathrm{d}x_0) \\
&= 2 \iiint \mathbb{1}\{f(x_1) > t \geqslant f(x_0)\} \, P(x_0, \mathrm{d}x_1) \, \pi(\mathrm{d}x_0) \, \mathrm{d}t \\
&= 2 \int \Bigl(\int_{\{f \leqslant t\}} P(x_0, \{f \leqslant t\}^{\mathrm{c}}) \, \pi(\mathrm{d}x_0) \Bigr) \, \mathrm{d}t \\
&\geqslant 2\mathfrak{c} \int \pi(\{f \leqslant t\}) \, \pi(\{f > t\}) \, \mathrm{d}t \\
&= 2\mathfrak{c} \iiint \mathbb{1}\{f(x_0') > t \geqslant f(x_0)\} \, \pi(\mathrm{d}x_0) \, \pi(\mathrm{d}x_0') \, \mathrm{d}t \\
&= 2\mathfrak{c} \iint \mathbb{1}\{f(x_0') > f(x_0)\} \, [f(x_0') - f(x_0)] \, \pi(\mathrm{d}x_0) \, \pi(\mathrm{d}x_0') \\
&= \mathfrak{c} \iint |f(x_0') - f(x_0)| \, \pi(\mathrm{d}x_0) \, \pi(\mathrm{d}x_0') = \mathfrak{c} \, \mathbb{E}|f(X_0') - f(X_0)| \,. \qquad \square
\end{aligned}$$

Compare this with the relationship between the Cheeger isoperimetric inequality and the L^1–L^1 Poincaré inequality in Theorem 2.4.18. Indeed, the trick above of passing to the level sets of f is the discrete version of the coarea inequality (Theorem 2.4.16).

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 2.4.21. 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.

We need one more lemma (Exercise 7.5).

Lemma 7.4.6 (Lawler and Sokal (1988))**.** *Let X, X' be i.i.d. random variables with finite variance σ^2. Then, it holds that*

$$\sup_{c\in\mathbb{R}} \frac{\{\mathbb{E}|(X+c)^2-(X'+c)^2|\}^2}{\mathbb{E}[(X+c)^2]} \geqslant \sigma^2\,. \tag{7.4.7}$$

Proof of Cheeger's inequality (Theorem 7.4.3) For $f : \mathbb{R}^d \to \mathbb{R}$ with mean zero, applying the lemma and Cauchy–Schwarz,

$$\begin{aligned}
\operatorname{var}_\pi f &\leqslant \sup_{c\in\mathbb{R}} \frac{\{\mathbb{E}|(f(X_0)+c)^2-(f(X_0')+c)^2|\}^2}{\mathbb{E}[(f(X_0)+c)^2]} \\
&\leqslant \sup_{c\in\mathbb{R}} \frac{\{\mathbb{E}|(f(X_1)+c)^2-(f(X_0)+c)^2|\}^2}{\mathfrak{c}^2\,\{\operatorname{var}_\pi f + c^2\}} \\
&\leqslant \sup_{c\in\mathbb{R}} \frac{\mathbb{E}[|f(X_1)-f(X_0)|^2]\,\mathbb{E}[|f(X_0)+f(X_1)+2c|^2]}{\mathfrak{c}^2\,\{\operatorname{var}_\pi f + c^2\}} \\
&= \frac{2\,\mathscr{E}(f,f)}{\mathfrak{c}^2} \sup_{c\in\mathbb{R}} \frac{\mathbb{E}[|f(X_0)+f(X_1)|^2]+4c^2}{\operatorname{var}_\pi f + c^2} \leqslant \frac{8\,\mathscr{E}(f,f)}{\mathfrak{c}^2}\,.
\end{aligned}$$

□

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 2.4.3), 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 2.4.23).

Lemma 7.4.8 (Overlap lemma)**.** *Assume the following:*

1. *The target π satisfies a Cheeger isoperimetric inequality with constant* $\mathsf{Ch} > 0$.
2. *There exists $r > 0$ such that for any points $x, y \in \mathbb{R}^d$ with $\|x-y\| \leqslant r$, it holds that* $\|P(x,\cdot)-P(y,\cdot)\|_{\mathrm{TV}} \leqslant \frac{1}{2}$.

Then, $\mathfrak{c} \geqslant \min\{\frac{1}{8}, \frac{r}{32\,\mathsf{Ch}}\}$.

Proof Let $A_0 \subseteq \mathbb{R}^d$; for symmetry of notation, write $A_1 := A_0^{\mathrm{c}}$. By reversibility,

$$\int_{A_0} P(x,A_1)\,\pi(\mathrm{d}x) = \int_{A_1} P(y,A_0)\,\pi(\mathrm{d}y) = \frac{1}{2}\Bigl(\int_{A_0} P(x,A_1)\,\pi(\mathrm{d}x) + \int_{A_1} P(y,A_0)\,\pi(\mathrm{d}y)\Bigr)\,.$$

We want to lower bound this by a constant times $\pi(A_0)\,\pi(A_1)$.

Define bad sets and a good set:

$$\begin{aligned} B_0 &:= \Bigl\{x \in A_0 \Bigm| P(x, A_1) < \frac{1}{4}\Bigr\}, \\ B_1 &:= \Bigl\{y \in A_1 \Bigm| P(y, A_0) < \frac{1}{4}\Bigr\}, \\ G &:= \mathbb{R}^d \setminus (B_0 \cup B_1)\,. \end{aligned}$$

We can assume that $\pi(B_0) \geqslant \pi(A_0)/2$ and $\pi(B_1) \geqslant \pi(A_1)/2$. Indeed, if we have, e.g., $\pi(B_0) \leqslant \pi(A_0)/2$, then $\pi(A_0 \setminus B_0) \geqslant \pi(A_0)/2$, and

$$\int_{A_0} P(x, A_1)\, \pi(\mathrm{d}x) \geqslant \int_{A_0 \setminus B_0} P(x, A_1)\, \pi(\mathrm{d}x) \geqslant \frac{1}{4}\, \pi(A_0 \setminus B_0) \geqslant \frac{1}{8}\, \pi(A_0)\,.$$

Next, suppose that $x \in B_0$ and $y \in B_1$. Then, $P(x, A_0) \geqslant \frac{3}{4}$, 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 $B_1 \subseteq (B_0^r)^{\mathsf{c}}$, or $B_1^{\mathsf{c}} \supseteq B_0^r$. On the other hand, $G = B_0^{\mathsf{c}} \cap B_1^{\mathsf{c}} \supseteq B_0^r \setminus B_0$. Integrating the isoperimetric inequality in (2.4.14) shows that

$$\pi(G) \geqslant \pi(B_0^r) - \pi(B_0) \geqslant \frac{r}{\mathsf{Ch}}\, \pi(B_0)\, \pi((B_0^r)^{\mathsf{c}}) \geqslant \frac{r}{\mathsf{Ch}}\, \pi(B_0)\, \pi(B_1) \geqslant \frac{r}{4\,\mathsf{Ch}}\, \pi(A_0)\, \pi(A_1)\,.$$

Hence,

$$\begin{aligned} &\frac{1}{2}\Bigl(\int_{A_0} P(x, A_1)\, \pi(\mathrm{d}x) + \int_{A_1} P(y, A_0)\, \pi(\mathrm{d}y)\Bigr) \\ &\qquad \geqslant \frac{1}{2}\Bigl(\int_{A_0 \cap G} P(x, A_1)\, \pi(\mathrm{d}x) + \int_{A_1 \cap G} P(y, A_0)\, \pi(\mathrm{d}y)\Bigr) \\ &\qquad \geqslant \frac{1}{8}\bigl(\pi(A_0 \cap G) + \pi(A_1 \cap G)\bigr) = \frac{1}{8}\, \pi(G) \geqslant \frac{r}{32\,\mathsf{Ch}}\, \pi(A_0)\, \pi(A_1)\,. \qquad \square \end{aligned}$$

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 7.4.8 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 $s \in [0, \frac{1}{2}]$, the **s-conductance** of P is the largest $\mathfrak{c}_s > 0$ such that for all events $A \subseteq \mathbb{R}^d$,

$$\int_A P(x, A^{\mathsf{c}})\, \pi(\mathrm{d}x) \geqslant \mathfrak{c}_s\, \bigl(\pi(A) - s\bigr)\, \bigl(\pi(A^{\mathsf{c}}) - s\bigr)\,.$$

Observe that if $\pi(A) \leqslant s$, 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 7.4.9 (Lovász and Simonovits (1993, Corollary 1.6)). *For any* $0 < s \leqslant \frac{1}{2}$, *let*

$$\Delta_s := \sup\{|\mu_0(A) - \pi(A)| : A \subseteq \mathbb{R}^d,\ \pi(A) \leqslant s\}\,.$$

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*

$$\|\mu_N - \pi\|_{\mathrm{TV}} \leqslant \Delta_s + \frac{\Delta_s}{s} \exp\Bigl(-\frac{\mathfrak{c}_s^2 N}{2}\Bigr)\,.$$

In particular,

$$\|\mu_N - \pi\|_{\mathrm{TV}} \leqslant \sqrt{s\,\chi^2(\mu_0 \,\|\, \pi)} + \sqrt{\frac{\chi^2(\mu_0 \,\|\, \pi)}{s}} \exp\Bigl(-\frac{\mathfrak{c}_s^2 N}{2}\Bigr)\,.$$

Proof The first statement is from Lovász and Simonovits (1993, Corollary 1.6) and the proof is not particularly straightforward. We sketch an alternative approach in Exercise 7.6.

The second statement follows from the first: indeed, for $A \subseteq \mathbb{R}^d$ with $\pi(A) \leqslant s$,

$$|\mu_0(A) - \pi(A)| = \Bigl|\int \mathbb{1}_A \,\mathrm{d}(\mu_0 - \pi)\Bigr| = \Bigl|\int \mathbb{1}_A \,\bigl(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi} - 1\bigr)\,\mathrm{d}\pi\Bigr| \leqslant \sqrt{\pi(A)\,\chi^2(\mu_0 \,\|\, \pi)}$$

so that $\Delta_s \leqslant \sqrt{s\,\chi^2(\mu_0 \,\|\, \pi)}$. ☐

This result says that if $s = \varepsilon^2/4\chi^2(\mu_0\|\pi)$, then we obtain $\|\mu_N - \pi\|_{\mathrm{TV}} \leqslant \varepsilon$ after

$$N = O\Bigl(\frac{1}{\mathfrak{c}_s^2} \log \frac{\chi^2(\mu_0 \,\|\, \pi)}{\varepsilon^2}\Bigr) \qquad \text{iterations}\,.$$

Although this mixing time guarantee is stated in the TV metric, the guarantee can often be "boosted" to stronger metrics, see Exercise 7.8.

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 7.7.

Lemma 7.4.10 (Overlap lemma for s-conductance). *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* $\pi(E) \geqslant 1 - \frac{rs}{16\,\mathsf{Ch}}$ *such that*

$$\forall x, y \in E,\ \|x - y\| \leqslant r \implies \|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}} \leqslant \frac{1}{2}\,.$$

Then, $\mathfrak{c}_s \gtrsim r/\mathsf{Ch}$.

7.5 Analysis of MALA for a Cold 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 7.4.10), which then upper bounds the mixing time via Theorem 7.4.9. By strong log-concavity of π, the first hypothesis of Lemma 7.4.10 is verified, so it remains to bound the overlaps. For a kernel P, we use the shorthand $P_x := P(x, \cdot)$.

By the triangle inequality, we have the decomposition

$$\|P_x - P_y\|_{\mathrm{TV}} \leqslant \|P_x - Q_x\|_{\mathrm{TV}} + \|Q_x - Q_y\|_{\mathrm{TV}} + \|P_y - Q_y\|_{\mathrm{TV}} \,. \tag{7.5.1}$$

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 P and Q 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 $\|\nabla^2 V\|_{\mathrm{op}} \leqslant \beta$. Then, provided $h \leqslant 1/\beta$, we have*

$$\|Q_x - Q_y\|_{\mathrm{TV}} \leqslant \frac{\|x-y\|}{\sqrt{2h}} \,.$$

Proof By Pinsker's inequality,

$$\|Q_x - Q_y\|_{\mathrm{TV}}^2 \leqslant \frac{1}{2}\,\mathsf{KL}(Q_x \parallel Q_y) = \frac{\|x - h\,\nabla V(x) - y + h\,\nabla V(y)\|^2}{8h} \leqslant \frac{\|x-y\|^2}{2h}$$

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 $\|P_x - Q_x\|_{\mathrm{TV}}$. Computing this term is slightly tricky because P_x has an atom at x, but in the end we obtain

$$\begin{aligned}
\|P_x - Q_x\|_{\mathrm{TV}} &= \frac{1}{2}\Bigl[\underbrace{1 - \int Q(x,\mathrm{d}y)\,A(x,y)}_{\text{from the atom of } P_x} + \int_{\mathbb{R}^d\setminus\{x\}} |P(x,y) - Q(x,y)|\,\mathrm{d}y\Bigr] \\
&= \frac{1}{2}\Bigl[1 - \int Q(x,\mathrm{d}y)\,A(x,y) + \int Q(x,\mathrm{d}y)\,\{1 - A(x,y)\}\,\mathrm{d}y\Bigr] \\
&= 1 - \int Q(x,\mathrm{d}y)\,A(x,y) \,.
\end{aligned} \tag{7.5.3}$$

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:

$$\mathbb{E}\,A(x,Y) = \mathbb{E}\min\Bigl\{1, \frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)}\Bigr\} \geqslant \lambda\,\mathbb{P}\Bigl\{\frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} \geqslant \lambda\Bigr\} \qquad \text{for all } 0 < \lambda < 1 \,.$$

The approach now is to write out the ratio more explicitly, and then carefully group together and bound the terms.

Explicitly, we have

$$\frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} = \exp\Bigl(-V(Y) - \frac{\|x - Y + h\,\nabla V(Y)\|^2}{4h} + V(x) + \frac{\|Y - x + h\,\nabla V(x)\|^2}{4h}\Bigr)\,.$$

After some careful algebra,

$$\begin{aligned} 4\log\frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} &= h\,\{\|\nabla V(x)\|^2 - \|\nabla V(Y)\|^2\} && (7.5.4)\\ &\quad - 2\,\{V(Y) - V(x) - \langle \nabla V(x), Y - x\rangle\} && (7.5.5)\\ &\quad + 2\,\{V(x) - V(Y) - \langle \nabla V(Y), x - Y\rangle\}\,. && (7.5.6) \end{aligned}$$

The terms are grouped to more easily apply the strong convexity and smoothness of V. It yields

$$(7.5.5) \geqslant -\beta\,\|x - Y\|^2 \qquad \text{and} \qquad (7.5.6) \geqslant \alpha\,\|x - Y\|^2 \geqslant 0\,.$$

Also, for $h \leqslant {}^1\!/\!{}_\beta$,

$$\begin{aligned} (7.5.4) &= h\,\langle \nabla V(x) - \nabla V(Y), \nabla V(x) + \nabla V(Y)\rangle\\ &\geqslant -h\,\|\nabla V(x) - \nabla V(Y)\|\,\|\nabla V(x) + \nabla V(Y)\|\\ &\geqslant -\beta h\,\|x - Y\|\,(2\,\|\nabla V(x)\| + \beta\,\|x - Y\|) \geqslant -\beta h^2\,\|\nabla V(x)\|^2 - 2\beta\,\|x - Y\|^2\,. \end{aligned}$$

Therefore,

$$\log\frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} \gtrsim -\beta h^2\,\|\nabla V(x)\|^2 - \beta\,\|x - Y\|^2 \gtrsim -\beta h^2\,\|\nabla V(x)\|^2 - \beta h\,\|\xi\|^2\,.$$

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, $\|\xi\|^2 \leqslant 2d$ with probability at least $1 - \exp(-d/2)$. Also, let $E_\delta := \{\|\nabla V\| \leqslant \sqrt{\beta\,(d + \log(1/\delta))}\}$. It follows that for all $x \in E_\delta$, if we take $h \lesssim {}^1\!/\!{}_{(\beta d + \beta\sqrt{\log(1/\delta)})}$ with a sufficiently small constant, then

$$\mathbb{E}\,A(x,Y) \geqslant \frac{11}{12}\,\mathbb{P}\Bigl\{\frac{\pi(Y)\,Q(Y,x)}{\pi(x)\,Q(x,Y)} \geqslant \frac{11}{12}\Bigr\} \geqslant \frac{11}{12}\,\Bigl(1 - \exp(-\frac{d}{2})\Bigr) \geqslant \frac{5}{6}\,,$$

for sufficiently large d. Hence, for $x \in E_\delta$, we have $\|P_x - Q_x\|_{\mathrm{TV}} \leqslant \frac{1}{6}$.

Completing the analysis.

We have shown: if the step size satisfies $h \lesssim {}^1\!/\!{}_{(\beta d + \beta\sqrt{\log(1/\delta)})}$, then for all $x, y \in E_\delta$ with $\|x - y\| \leqslant r$,

$$\|P_x - P_y\|_{\mathrm{TV}} \leqslant \|P_x - Q_x\|_{\mathrm{TV}} + \|Q_x - Q_y\|_{\mathrm{TV}} + \|P_y - Q_y\|_{\mathrm{TV}} \leqslant \frac{1}{6} + \frac{r}{\sqrt{2h}} + \frac{1}{6} \leqslant \frac{1}{2}$$

provided we take $r = \sqrt{2h}/6$. Applying Lemma 7.4.10 (assuming $h \lesssim {}^1\!/\!{}_\alpha$), we deduce that $\mathfrak{c}_s \gtrsim \sqrt{\alpha h}$ provided $\pi(E_\delta) \geqslant 1 - c_0 s\sqrt{\alpha h}$, where $c_0 > 0$ is a universal constant. Since we want the step size h to be as large as possible, we take $h \asymp {}^1\!/\!{}_{(\beta d + \beta\sqrt{\log(1/\delta)})}$, where δ is chosen to satisfy $\pi(E_\delta^{\mathsf{c}}) \lesssim s/(\kappa d + \kappa\log^{1/2}(1/\delta))^{1/2}$ and $s \asymp \varepsilon^2/\chi^2(\mu_0 \,\|\, \pi)$. By the gradient concentration inequality in Lemma 6.2.7, $\pi(E_\delta^{\mathsf{c}}) \lesssim \delta$, so we can choose $B = \widetilde{O}(\log({}^{\chi^2(\mu_0/\pi)}\!/\!{}_{\varepsilon^2}))$. The final mixing time bound implied by Theorem 7.4.9 is then $\widetilde{O}(\kappa d\log^2({}^{\chi^2(\mu_0\|\pi)}\!/\!{}_{\varepsilon^2}))$.

From warm start to cold start.

The factor of $\log \chi^2(\mu_0 \,\|\, \pi)$ in the bound incurs additional dimension dependence under a cold start, since with a Gaussian initialization we can only show $\chi^2(\mu_0 \,\|\, \pi) \leqslant \kappa^{O(d)}$. 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 2.4.5). 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 $\pi^+(A) \gtrsim p\sqrt{\log(1/p)}$. Using this stronger assumption, Chen et al. (2020) 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* (Kannan et al., 2006), which was used in Lee et al. (2020). 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 7.5.7 (Lee et al. (2021a))**.** *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 $\chi^2(\mu_0 \,\|\, \pi) \leqslant \exp d$, 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 Chewi et al. (2021), with subsequent refinements in Wu et al. (2021); Chen and Gatmiry (2023a) via completely different techniques. In this section, we follow Chewi et al. (2021) because the proof is more conceptual and closer to the stochastic calculus arguments of Chapter 6.

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 $\|P_x - Q_x\|_{\mathrm{TV}}$, 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 $\|P_x - Q_x\|_{\mathrm{TV}}$ 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 P 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,

$$\iint_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus\mathsf{diag}} |P(x,y) - Q(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y \leqslant \iint_{(\mathbb{R}^d\times\mathbb{R}^d)\setminus\mathsf{diag}} |Q(x,y) - \bar Q(x,y)|\,\pi(\mathrm{d}x)\,\mathrm{d}y\,.$$

Now supposing that $\bar{Q}$ has no atoms, this inequality is the same as

$$\int \|P_x - Q_x\|_{\mathrm{TV}}\, \pi(\mathrm{d}x) \leqslant 2 \int \|Q_x - \bar{Q}_x\|_{\mathrm{TV}}\, \pi(\mathrm{d}x)\,. \tag{7.6.1}$$

Thus, we can indirectly bound $\|Q_x - P_x\|_{\mathrm{TV}}$, 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 $\|P_x - Q_x\|_{\mathrm{TV}}$ in average, whereas we really need high-probability bounds. Thankfully, we can extend the projection property to a pointwise bound.

Theorem 7.6.2 (Pointwise projection property, Chewi et al. (2021, Theorem 6))**.** *Let Q be an atomless proposal kernel and let P 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 $x \in \mathbb{R}^d$,*

$$\|P_x - Q_x\|_{\mathrm{TV}} \leqslant 2\,\|Q_x - \bar{Q}_x\|_{\mathrm{TV}} + \int \frac{\pi(y)\,\bar{Q}(y,x)}{\pi(x)} \Bigl|\frac{Q(y,x)}{\bar{Q}(y,x)} - 1\Bigr|\,\mathrm{d}y\,.$$

Consequently, for any convex increasing function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$,

$$\begin{aligned}\int \Phi(\|P_x - Q_x\|_{\mathrm{TV}})\,\pi(\mathrm{d}x) &\leqslant \frac{1}{2}\int \Phi(4\,\|Q_x - \bar{Q}_x\|_{\mathrm{TV}})\,\pi(\mathrm{d}x)\\ &\qquad + \frac{1}{2}\iint \Phi\Bigl(2\,\Bigl|\frac{Q(x,y)}{\bar{Q}(x,y)} - 1\Bigr|\Bigr)\,\bar{Q}(x,\mathrm{d}y)\,\pi(\mathrm{d}x)\,.\end{aligned} \tag{7.6.3}$$

We will not need the inequality (7.6.3), so the proof is left as Exercise 7.10. 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 $\|P_x - Q_x\|_{\mathrm{TV}}$. 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

$$\begin{aligned}\|P_x - Q_x\|_{\mathrm{TV}} = 1 - \int Q(x,\mathrm{d}y)\,A(x,y) &= \int \Bigl[1 - \Bigl(1 \wedge \frac{\pi(y)\,Q(y,x)}{\pi(x)\,Q(x,y)}\Bigr)\Bigr]\,Q(x,\mathrm{d}y)\\ &\leqslant \int \Bigl|1 - \frac{\pi(y)\,Q(y,x)}{\pi(x)\,Q(x,y)}\Bigr|\,Q(x,\mathrm{d}y)\\ &\leqslant \int \Bigl|1 - \frac{\pi(y)\,\bar{Q}(y,x)}{\pi(x)\,Q(x,y)}\Bigr|\,Q(x,\mathrm{d}y) + \int \frac{\pi(y)\,\bar{Q}(y,x)}{\pi(x)}\Bigl|\frac{Q(y,x)}{\bar{Q}(y,x)} - 1\Bigr|\,\mathrm{d}y\,.\end{aligned}$$

Using reversibility of $\bar{Q}$, the first term is

$$\int \Bigl|1 - \frac{\pi(y)\,\bar{Q}(y,x)}{\pi(x)\,Q(x,y)}\Bigr|\,Q(x,\mathrm{d}y) = \int |Q(x,y) - \bar{Q}(x,y)|\,\mathrm{d}y = 2\,\|Q_x - \bar{Q}_x\|_{\mathrm{TV}}\,,$$

which completes the proof. □

Applying the pointwise projection property.

Our goal is to bound $\|P_x - Q_x\|_{\mathrm{TV}}$ 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.

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 follows from Theorem 4.4.1 that $\|Q_x - \bar Q_x\|_{\mathrm{TV}} \lesssim \beta h^{3/2}\,\|\nabla V(x)\| + \beta d^{1/2} h$. By the gradient concentration bound in Lemma 6.2.7, for $h \lesssim 1/\beta$, we have $\|Q_x - \bar Q_x\|_{\mathrm{TV}} \lesssim \beta d^{1/2} h + \beta^{3/2} h^{3/2} \sqrt{\log(1/\delta)}$ with probability $1-\delta$ under π. In particular, we expect a step size of $h \asymp 1/\beta d^{1/2}$ to suffice to control this term, which is certainly encouraging.

To control the second term with high probability, it suffices to control the moments of this quantity under π: for $p \geqslant 1$,

$$\int \Big| \int \frac{\pi(y)\,\bar Q(y,x)}{\pi(x)} \Big|\frac{Q(y,x)}{\bar Q(y,x)} - 1\Big|\,\mathrm{d}y\Big|^p \,\pi(\mathrm{d}x) \leqslant \iint \Big|\frac{Q(y,x)}{\bar Q(y,x)} - 1\Big|^p\,\bar Q(y,\mathrm{d}x)\,\pi(\mathrm{d}y)\,.$$

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

$$\iint \Big|\frac{Q(y,x)}{\bar Q(y,x)} - 1\Big|^p\,\bar Q(y,\mathrm{d}x)\,\pi(\mathrm{d}y) \leqslant \int \mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Big[\Big|\frac{\mathrm{d}\boldsymbol{Q}_x}{\mathrm{d}\bar{\boldsymbol{Q}}_x} - 1\Big|^p\Big]\,\pi(\mathrm{d}x)\,.$$

We can again approach this term via stochastic calculus, although the arguments become more involved. First, via Girsanov's theorem (Theorem 3.2.8),

$$\frac{\mathrm{d}\boldsymbol{Q}_x}{\mathrm{d}\bar{\boldsymbol{Q}}_x} = \exp\underbrace{\Big(\frac{1}{\sqrt{2}}\int_0^h \langle \nabla V(x) - \nabla V(X_t), \mathrm{d}B_t\rangle - \frac{1}{4}\int_0^h \|\nabla V(x) - \nabla V(X_t)\|^2\,\mathrm{d}t\Big)}_{:=L_h}\,,$$

where $(B_t)_{t\geqslant 0}$ is the $\bar{\boldsymbol{Q}}_x$-Brownian motion and $(X_t)_{t\geqslant 0}$ is the Langevin diffusion driven by $(B_t)_{t\geqslant 0}$ and started at x. By applying Itô's formula (Theorem 1.1.19),

$$\exp(L_h) - 1 = \frac{1}{\sqrt{2}}\int_0^h (\exp L_t)\,\langle \nabla V(X_t) - \nabla V(x), \mathrm{d}B_t\rangle\,.$$

We need to control the p-th moment of this stochastic integral; we can take p to be an even integer. For $p = 2$, we can apply the Itô isometry (1.1.9). For larger values of p, the appropriate substitute is the Burkholder–Davis–Gundy inequality with optimal constants.

Lemma 7.6.4 (Burkholder–Davis–Gundy, Burkholder (1973); Davis (1976))**.** *There is a universal constant $C > 0$ such that for any even integer $p \geqslant 2$,*

$$\mathbb{E}\Big[\sup_{t\in[0,T]}\Big|\int_0^t \langle \eta_s, \mathrm{d}B_s\rangle\Big|^p\Big] \leqslant (Cp)^{p/2}\,\mathbb{E}\Big[\Big|\int_0^T \|\eta_t\|^2\,\mathrm{d}t\Big|^{p/2}\Big]\,.$$

We have stated this inequality in a form that is convenient for our purposes, although it applies more generally to local martingales. The growth of the optimal constants with p implies that the stochastic integral has sub-Gaussian tails.

Applying this to the problem at hand,

$$\begin{aligned}
\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Bigl[\Bigl|\frac{\mathrm{d}\boldsymbol{Q}_x}{\mathrm{d}\bar{\boldsymbol{Q}}_x}-1\Bigr|^p\Bigr] &\leqslant (Cp)^{p/2}\,\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Bigl[\Bigl|\int_0^h \exp(2L_t)\,\|\nabla V(X_t)-\nabla V(x)\|^2\,\mathrm{d}t\Bigr|^{p/2}\Bigr] \\
&\leqslant (C\beta^2 p)^{p/2}\,\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Bigl[\Bigl|\int_0^h \exp(4L_t)\,\mathrm{d}t\Bigr|^{p/4}\Bigl|\int_0^h \|X_t-x\|^4\,\mathrm{d}t\Bigr|^{p/4}\Bigr] \\
&\leqslant (C\beta^2 p)^{p/2}\sqrt{\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Bigl[\Bigl|\int_0^h \exp(4L_t)\,\mathrm{d}t\Bigr|^{p/2}\Bigr]\,\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Bigl[\Bigl|\int_0^h \|X_t-x\|^4\,\mathrm{d}t\Bigr|^{p/2}\Bigr]} \\
&\leqslant (C\beta^2 p)^{p/2}\,h^{p/2-1}\sqrt{\Bigl(\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\int_0^h \exp(2pL_t)\,\mathrm{d}t\Bigr)\Bigl(\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\int_0^h \|X_t-x\|^{2p}\,\mathrm{d}t\Bigr)}\,.
\end{aligned}$$

We bound each of the remaining expectation terms in turn. We start by observing that $\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\exp(2pL_t) = \exp((2p-1)\,\mathcal{R}_{2p}(\boldsymbol{Q}_x|_{[0,t]} \,\|\, \bar{\boldsymbol{Q}}_x|_{[0,t]}))$, so by Theorem 6.2.1 we obtain

$$\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\int_0^h \exp(2pL_t)\,\mathrm{d}t \leqslant h\exp O\bigl(\beta^2 h^3 p^2\,\|\nabla V(x)\|^2 + \beta^2 d h^2 p^2\bigr)$$

provided $h \lesssim 1/\beta p$. Next, by (6.2.2) and Lemma 6.2.4,

$$\begin{aligned}
\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\int_0^h \|X_t-x\|^{2p}\,\mathrm{d}t &\leqslant C^p h\,\Bigl(h^{2p}\,\|\nabla V(x)\|^{2p} + \mathbb{E}^{\bar{\boldsymbol{Q}}_x}\sup_{t\in[0,h]}\|B_t\|^{2p}\Bigr) \\
&\leqslant C^p h\,\bigl(h^{2p}\,\|\nabla V(x)\|^{2p} + d^p h^p p^p\bigr)\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Bigl[\Bigl|\frac{\mathrm{d}\boldsymbol{Q}_x}{\mathrm{d}\bar{\boldsymbol{Q}}_x}-1\Bigr|^p\Bigr] \\
&\quad\leqslant (C\beta^2 h^2 p)^{p/2}\,\bigl(h^{p/2}\,\|\nabla V(x)\|^p + d^{p/2}p^{p/2}\bigr)\exp O\bigl(\beta^2 h^3 p^2\,\|\nabla V(x)\|^2 + \beta^2 d h^2 p^2\bigr)\,.
\end{aligned}$$

We integrate w.r.t. $\pi(\mathrm{d}x)$ and apply Cauchy–Schwarz:

$$\begin{aligned}
\int \mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Bigl[\Bigl|\frac{\mathrm{d}\boldsymbol{Q}_x}{\mathrm{d}\bar{\boldsymbol{Q}}_x}-1\Bigr|^p\Bigr]\,\pi(\mathrm{d}x) &\leqslant (C\beta^2 h^2 p)^{p/2}\Bigl(h^{p/2}\sqrt{\int \|\nabla V(x)\|^{2p}\,\pi(\mathrm{d}x)} + d^{p/2}p^{p/2}\Bigr) \\
&\qquad\times\int \exp O\bigl(\beta^2 h^3 p^2\,\|\nabla V(x)\|^2 + \beta^2 d h^2 p^2\bigr)\,\pi(\mathrm{d}x)\,.
\end{aligned}$$

If we again apply the gradient concentration inequality (Lemma 6.2.7), we obtain

$$\Bigl\{\int \mathbb{E}^{\bar{\boldsymbol{Q}}_x}\Bigl[\Bigl|\frac{\mathrm{d}\boldsymbol{Q}_x}{\mathrm{d}\bar{\boldsymbol{Q}}_x}-1\Bigr|^p\Bigr]\,\pi(\mathrm{d}x)\Bigr\}^{1/p} \lesssim \beta h\sqrt{p}\,\bigl(\sqrt{\beta d h p} + \sqrt{dp}\bigr) \lesssim \beta h d^{1/2} p$$

provided that $h \lesssim 1/\beta d^{1/3} p^{2/3}$. By Markov's inequality, it implies that the second term arising from the pointwise projection property is at most $O(\beta d^{1/2} h p/\delta^{1/p})$ with probability at least $1-\delta$ under π. In particular, if we take $p \asymp \log(1/\delta)$, the bound reads $O(\beta d^{1/2} h \log(1/\delta))$.

Conductance argument.

We now finish with a conductance argument, similarly to Section 7.5. Namely, we must take δ so that $\delta \lesssim s\sqrt{\alpha h} \asymp \varepsilon^2\sqrt{\alpha h}/\chi^2(\mu_0 \,\|\, \pi)$ and h so that $h \lesssim 1/(\beta d^{1/2}\log(1/\delta))$. This is satisfied by taking

$h \asymp 1/\{\beta\sqrt{d}\log({}^{\chi^2(\mu_0\|\pi)}/_{\varepsilon^2})\}$. Then, Lemma 7.4.10 yields $\mathfrak{c}_s \gtrsim \sqrt{\alpha h}$. Finally, Theorem 7.4.9 yields mixing in total variation in $\widetilde{O}(\kappa\sqrt{d}\log^2({}^{\chi^2(\mu_0\|\pi)}/_{\varepsilon^2}))$ iterations.

Lower bound.

Finally, we complement the analysis with a matching lower bound. Under a warm start, Chewi et al. (2021) showed a lower bound of roughly $\widetilde{\Omega}(\sqrt{d})$, which was improved in Wu et al. (2021) to $\widetilde{\Omega}(\kappa\sqrt{d})$. We state the result here.

Theorem 7.6.5 (Wu et al. (2021))**.** *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 variation ε from π is at least $\widetilde{\Omega}(\kappa\sqrt{d}\log(1/\varepsilon))$.*

Bibliographical Notes

The Metropolis–Hastings algorithm dates back to the birth of MCMC itself (Metropolis et al., 1953; Hastings, 1970), whereas rejection sampling can be traced back to Ulam and von Neumann (von Neumann, 1951; Eckhardt, 1987). Moreover, MALA and MHMC were introduced in Besag et al. (1995) and Duane et al. (1987) respectively; in the latter reference, the algorithm was introduced as “hybrid Monte Carlo”. The work of Neal (2011) played an important role in popularizing the use of MHMC for Bayesian applications.

Traditionally, these algorithms were studied in an asymptotic framework in which a scaling $h \propto 1/d^\theta$ is sought so that the algorithm admits a non-degenerate diffusion limit as $d \to \infty$; see Roberts and Rosenthal (1998) for a particularly influential work in this direction. Part of the appeal lies in the precise calculation of the limiting acceptance probabilities, leading to an effective tuning strategy: set the step size of the algorithm so that the proportion of accepted moves equals the limit value.

The quantity d^θ is interpreted as a proxy for the mixing time, and in this way the following predictions were made for the complexities: $\Theta(d^{1/3})$ for MALA, and $\Theta(d^{1/4})$ for MHMC (Beskos et al., 2013). There are notable limitations of the scaling limit approach, however: it imposes stringent assumptions, such as higher-order smoothness or that the target is a product measure, and the analysis is often conducted at stationarity. Indeed, the scalings are different in the non-stationary or “transient” regime (Kuntz et al., 2018; Lee, 2025). There are further works aimed at addressing these issues, but a precise determination of the mixing time requires a non-asymptotic framework. The paper Chewi et al. (2021) further emphasized this point by proving a $\widetilde{\Omega}(d^{1/2})$ lower bound for MALA (stated as Theorem 7.6.5), which seemingly contradicts the $\Theta(d^{1/3})$ dependence predicted by the scaling limit. Of course, the contradiction is illusory; the scaling limit requires more assumptions (e.g., higher-order smoothness). Still, this presents a cautionary tale for interpreting asymptotic results with the non-asymptotic, minimax lens that we adopt in this book.

The overlap lemma (Lemma 7.4.8) was refined by a suggestion from Kexin Zhang. The s-conductance mixing time argument given in Exercise 7.6 arose out of discussions with Adil Salim.

The $\widetilde{O}(\kappa d)$ rate for MALA was first obtained in Dwivedi et al. (2019) with a warm start, and then in Chen et al. (2020) for a cold start. Initially, these rates also had an additional $\widetilde{O}(\kappa^{3/2}d^{1/2})$ term, which was removed via a gradient concentration bound in Lee et al. (2020). Our analysis in Section 7.5 is based on Dwivedi et al. (2019), except that we use the gradient bound from Lemma 6.2.7

(which improves, in some respects, the bound used in Chen et al. (2020)). The rate for MRW stated in Theorem 7.3.4 was proven in Andrieu et al. (2024).

The paper Chewi et al. (2021) showed that under standard assumptions, the dimension dependence of MALA improves to $d^{1/2}$ under a warm start. However, due to a typo in the paper, the condition number dependence of the result was unclear. Subsequently, the optimal bound of $\widetilde{O}(\kappa d^{1/2})$ was obtained in Wu et al. (2021) via a different approach, which avoids stochastic calculus but relies on rather delicate calculations for the acceptance probability. These papers were later revisited: the rate of Chewi et al. (2021) was clarified to be $\widetilde{O}(\kappa d^{1/2} + \kappa^2)$ in the PhD thesis Chewi (2023), and the approach of Wu et al. (2021) was further simplified in Chen and Gatmiry (2023a). In Section 7.6, we follow the approach of Chewi et al. (2021); Chewi (2023), except that we further refine it via Lemma 6.2.7 which to reach the optimal $\widetilde{O}(\kappa d^{1/2})$ bound. The warm start for MALA used to prove Corollary 7.3.6 was attained in Altschuler and Chewi (2024a).

Currently, the state-of-the-art for MHMC is Chen and Gatmiry (2023b), which obtains a $\widetilde{O}(\kappa d^{1/4})$ rate under third-order smoothness and a warm start. It is not known if the third-order smoothness can be relaxed, or if the warm start can be obtained using $\widetilde{O}(\kappa d^{1/4})$ queries.

The following lower bounds for MALA have been obtained: $\widetilde{\Omega}(d^{1/2})$ from a warm start (Chewi et al., 2021); $\widetilde{\Omega}(\kappa d)$ from a cold start (Lee et al., 2021a); $\widetilde{\Omega}(\kappa d^{1/2})$ from a warm start (Wu et al., 2021), improving upon Chewi et al. (2021). Together with the upper bounds, we now have a thorough understanding of the complexity of MALA for cold vs. warm initializations. We note that Lee et al. (2021a) also contains some lower bounds for MHMC.

There is an extensive literature on variants of these algorithms and analyses which do not fit within the framework considered in this book, and we do not attempt to comprehensively survey it here.

Another approach to designing high-accuracy algorithms based on piecewise deterministic Markov processes (PDMPs). In Lu and Wang (2022a), it is shown that the zigzag sampler can achieve a high-accuracy complexity guarantee of $\widetilde{O}(\kappa^2 d^{1/2})$ evaluations of *partial derivatives* of V, from a warm start, in the regime $\kappa \ll d/(\log d)$.

Moreover, in Chen et al. (2026), it was shown how to simulate rejection sampling using only first-order queries (that is, queries to ∇V). These two approaches show that zeroth-order queries are in fact not necessary for achieving high-accuracy sampling guarantees.

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 $(\mu_n)_{n\in\mathbb{N}}$ are the laws of the iterates of the ℓ-lazy version of P, then the relative densities $(\frac{\mu_n}{\pi})_{n\in\mathbb{N}}$ 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 $0 \preceq \nabla^2_{L^2(\pi)}\mathscr{E} \preceq 4$. 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 $\ell \geqslant \frac{1}{2}$ 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 $\mathscr{E}(\mathbb{1}_A) \geqslant \mathfrak{c}\, \|\mathbb{1}_A - \pi(A)\|^2_{L^2(\pi)}$. 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 $\lambda \leqslant \mathfrak{c}$ 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,

$$\mathsf{KL}(\mu P \,\|\, \pi) \leqslant (1-c)\,\mathsf{KL}(\mu \,\|\, \pi)\,.$$

Prove that P satisfies a MLSI with constant $C_{\mathsf{MLSI}} \leqslant 2/c$.

⊵ **Exercise 7.5** (Completing the proof of Cheeger's inequality)
In this exercise, we prove Lemma 7.4.6. We can normalize the random variables so that $\mathbb{E}\,X = 0$ and $\sigma^2 = 1$.

1 Show that as $c \to \infty$, the ratio on the left-hand side of (7.4.7) tends to $4\,\{\mathbb{E}|X - X'|\}^2$. Thus, we are done in the case $\mathbb{E}|X| \geqslant 1/2$.
2 Otherwise, set $c = 0$ and prove that $\mathbb{E}|X^2 - X'^2| \geqslant 2\,(1 - \mathbb{E}|X|)$.
Hint: Work over the event $\{X^2 \geqslant 1,\ X'^2 < 1\} \cup \{X'^2 \geqslant 1,\ X^2 < 1\}$.

⊵ **Exercise 7.6** (Mixing time bound from s-conductance)
In this exercise, we sketch an alternative approach to Theorem 7.4.9.

1 Define the following semi-norm: for $s \in (0, 1/2)$,

$$|||f|||_s \coloneqq \sup_{\pi(A)\leqslant s} \Bigl|\int f\,\mathbb{1}_A\,\mathrm{d}\pi\Bigr|\,.$$

Show that for any $p > 1$, $|||f|||_s \leqslant s^{1-1/p}\,\|f\|_{L^p(\pi)}$.
2 We define the (s,δ)-spectral gap of P to be the largest $\lambda_{s,\delta} > 0$ such that for all $f \in L^2(\pi)$,

$$\operatorname{var}_\pi f \leqslant \frac{1}{\lambda_{s,\delta}}\,\mathscr{E}(f,f) + \delta\,|||f - \mathbb{E}_\pi f|||_s^2\,.$$

We typically think of $\delta > 0$ as a universal constant. Show that if $\delta = 1/2$, say, then $\lambda_{s,1/2} \lesssim \mathfrak{c}_s$. This is the "easy" part of Cheeger's inequality.
3 Using the approach of Exercise 7.3, show that for the iterates $(\mu_n)_{n\in\mathbb{N}}$ of the $\frac{1}{2}$-lazy chain corresponding to P,

$$\chi^2(\mu_N \,\|\, \pi) \leqslant \bigl(1 - \frac{\lambda_{s,\delta}}{2}\bigr)^n\,\chi^2(\mu_0 \,\|\, \pi) + \delta\,|||\rho_0|||_s^2\,,$$

where $\rho_0 := \frac{\mathrm{d}\mu_0}{\mathrm{d}\pi} - 1$. Therefore, to prove a mixing time bound under s-conductance, it suffices to establish an analogue of Cheeger's inequality which lower bounds the (s, δ)-spectral gap in terms of the s-conductance.

4 First, we develop a functional version of s-conductance. Show that if s is at most a sufficiently small universal constant, $0 < s \ll 1$, and $|||f - \mathbb{E}_\pi f|||_{2s} \leqslant 1$, then

$$\mathbb{E}_\pi |f - \mathbb{E}_\pi f| \lesssim \frac{1}{\mathfrak{c}_s} \mathbb{E}|f(X_1) - f(X_0)| + 1\,, \tag{7.E.1}$$

where (X_0, X_1) are two successive iterates of the chain started at stationarity.

5 Next, show that (7.E.1) implies that for all $f \in L^2(\pi)$,

$$\mathrm{var}_\pi f \lesssim \frac{1}{\mathfrak{c}_s^2} \mathscr{E}(f, f) + |||(f - \mathbb{E}_\pi f)^2|||_{2s}\,.$$

Use this to prove a mixing time bound in terms of the s-conductance.

⊵ **Exercise 7.7** (s-conductance lemma)

Prove the s-conductance lemma (Lemma 7.4.10).

⊵ **Exercise 7.8** (Boosting from TV to χ^2)

The mixing time guarantee in Theorem 7.4.9 is stated for the TV distance, but guarantees for high-accuracy samplers can often be "upgraded" to hold in stronger metrics. To illustrate, suppose that in the setting of Theorem 7.4.9, the Markov chain is initialized at a measure μ_0 with $\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi} \leqslant M$. Show that $\chi^2(\mu_N \,\|\, \pi) \leqslant \varepsilon^2$ after $N = O((1/\mathfrak{c}_s^2) \log(M/\varepsilon))$ iterations, where $s = \varepsilon^2/4M^2$.

Hint: Show that $\Delta_s \leqslant Ms$ and that $\frac{\mathrm{d}\mu_n}{\mathrm{d}\pi} \leqslant M$ for all $n \in \mathbb{N}$. Bound $\chi^2(\mu_N \,\|\, \pi)$ in terms of M and $\|\mu_N - \pi\|_{\mathrm{TV}}$, and then apply Theorem 7.4.9. More generally, this approach shows that if the initialization to the Markov chain satisfies $\mathcal{R}_q(\mu_0 \,\|\, \pi) < \infty$ for some $q > 1$, then one obtains guarantees for $\mathcal{R}_{q'}(\mu_N \,\|\, \pi)$ for all $q' < q$.

Analysis of MALA for a Cold Start

⊵ **Exercise 7.9** (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?

Analysis of MALA for a Warm Start

⊵ **Exercise 7.10** (Pointwise projection property)

Prove (7.6.3) from the pointwise projection property.

⊵ **Exercise 7.11** (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 $\widetilde{O}(d^{1/3}\,\mathrm{polylog}(1/\varepsilon))$ from a warm start. See Chewi et al. (2021, Appendix C) for hints.

CHAPTER 8

The Proximal Sampler

In this chapter, we discuss the proximal sampler, which was introduced in Lee et al. (2021b) although similar ideas date back earlier. The applications of the proximal sampler include improving the condition number dependence of high-accuracy samplers and extending the sampling guarantees beyond strong log-concavity. 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.

Optimization Box 8.0.1. Given $V : \mathbb{R}^d \to \mathbb{R}$, the **proximal oracle** associated with V and step size $h > 0$ is defined as the mapping $\operatorname{prox}_{hV} : \mathbb{R}^d \to \mathbb{R}^d$ given by $\operatorname{prox}_{hV}(x) := \arg\min_{y \in \mathbb{R}^d}\{V(y) + \frac{1}{2h}\,\|y - x\|^2\}$.

Implementing the proximal oracle. Note that the proximal oracle can be viewed as a regularized version of the original optimization problem. In particular, if $\|\nabla^2 V\|_{\rm op} \leqslant \beta$ and $h \leqslant \frac{1}{2\beta}$, then the proximal oracle is the solution to a $(\frac{1}{h} - \beta)$-strongly convex and $(\frac{1}{h} + \beta)$-smooth optimization problem, with condition number $\frac{1/h+\beta}{1/h-\beta} \leqslant 3$. Therefore, it can be approximately implemented via convex optimization.

Using the proximal oracle. Given access to the proximal oracle for V, the simplest way to use it to minimize V is to iterate it: $x_{k+1} := \operatorname{prox}_{hV}(x_k)$. This is known as the **proximal point method (PPM)**. If V is α-strongly convex, then we can obtain a convergence rate for the PPM as follows. First, note that the first-order optimality condition for the PPM can be written $\operatorname{prox}_{hV}(x) = (\mathrm{id} + h\,\nabla V)^{-1}(x)$. From this, one can show that prox_{hV} is $1/(1+\alpha h)$-Lipschitz, and $\operatorname{prox}_{hV}(x_\star) = x_\star$. Therefore, $\|x_N - x_\star\| \leqslant (1+\alpha h)^{-N}\,\|x_0 - x_\star\|$. Note that unlike gradient descent, the convergence of the PPM does *not* require smoothness of V.

Application to composite optimization. Arguably, the most well-known application of the proximal oracle is the minimization of a convex composite objective $f + g$, where f is smooth and g is non-smooth but "simple" so that the proximal oracle for g can be evaluated in closed form. The canonical example is when $g = \|\cdot\|_1$ is the ℓ_1-norm, in which case this composite objective encompasses computation of the LASSO estimator for sparse regression. In this case, the **proximal gradient** algorithm consists of GD steps for f and PPM steps for g: $x_{k+1} = \operatorname{prox}_{hg}(x_k - h\,\nabla f(x_k))$. Crucially, this algorithm can be shown to converge at the same rate as GD for convex *smooth* optimization, despite the presence of the non-smooth term.

The PPM can be written $x_{k+1} = x_k - h\,\nabla V(x_{k+1})$ which shows that it is an *implicit* discretization of the gradient flow. Implicit discretizations are more stable (indeed, note that the convergence rate for the PPM given above holds for *all* choices of $h > 0$, whereas GD typically requires $h < 2/\beta$), but at the cost of being harder to implement. We will see that similar intuitions hold in the context of sampling.

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

$$\boldsymbol{\pi}(x, y) \propto \exp\Bigl(-V(x) - \frac{\|y - x\|^2}{2h}\Bigr)\,.$$

To avoid confusion, we will explicitly write $\boldsymbol{\pi}^X = \pi$ for the X-marginal, and $\boldsymbol{\pi}^Y$ for the Y-marginal. Similarly, $\boldsymbol{\pi}^{X|Y}$ and $\boldsymbol{\pi}^{Y|X}$ 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 $Y_k \sim \boldsymbol{\pi}^{Y|X}(\cdot \mid X_k) = \mathsf{normal}(X_k, hI_d)$.
2 Draw $X_{k+1} \sim \boldsymbol{\pi}^{X|Y}(\cdot \mid Y_k)$.

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

$$\pi^{X|Y}(x \mid y) \propto_x \exp\Bigl(-V(x) - \frac{\|y-x\|^2}{2h}\Bigr)\,.$$

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 $\pi^{X|Y}$, which is known as the **restricted Gaussian oracle (RGO)**, is precisely the computation of the proximal oracle. See Exercise 8.1 for another connection between the proximal sampler and the PPM.

Interpretation of the proximal sampler.

At this stage, it may be unclear why the proximal sampler should be interpreted as the PPM for sampling. Part of the justification lies in the convergence results we establish in subsequent sections that are exactly analogous to optimization guarantees for the PPM.

Here, we present another justification in the spirit of the discussion from Section 4.3. There, we observed that sampling, which can be viewed as the minimization of the KL divergence $\mathsf{KL}(\cdot \parallel \pi)$, is a composite optimization problem over the space of probability measures consisting of a smooth term (the "energy") and a non-smooth term (the "entropy"). The optimization wisdom from Optimization Box 8.0.1 suggests considering a proximal gradient method in the Wasserstein space, which would require computation of $\operatorname{prox}_{h\mathcal{H}}(\mu) := \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\{\mathcal{H}(\nu) + \frac{1}{2}\,W_2^2(\mu,\nu)\}$. Such a scheme can be shown to enjoy strong convergence guarantees—see Salim et al. (2020)—but unfortunately the proximal oracle for the entropy functional is not easily implementable.

Recall from Section 4.3 that LMC is a splitting scheme which applies GD to the energy, and *gradient flow* to the entropy, but that this scheme is biased. The paper Salim et al. (2020) shows that if we keep the GD step for the energy, then the "natural" (but intractable) method to apply to the entropy to make the whole algorithm unbiased is the proximal oracle $\operatorname{prox}_{h\mathcal{H}}$. On the other hand, what if we keep the gradient flow for the entropy—i.e., the heat flow—and ask what method we can apply to the energy to keep the algorithm unbiased? In a sense, the proximal sampler is the answer to this question, since the first step of sampling from a Gaussian is indeed the heat flow. See Exercise 8.2 for an interpretation of the RGO as the *adjoint* to the heat flow in a precise sense.

This discussion highlights that even when V is smooth, there are potential advantages to be gained by consideration of proximal methodology, since the sampling problem carries with it an *inherent* non-smoothness from the entropy functional.

Implementability of the RGO.

In order to obtain an actual algorithm from the proximal sampler, an implementation of the RGO must be provided. Obviously, we have

$$\boxed{\begin{array}{c}\text{complexity of}\\ \text{the proximal sampler}\end{array}} = \boxed{\begin{array}{c}\text{number of iterations of}\\ \text{the proximal sampler}\end{array}} \times \boxed{\begin{array}{c}\text{complexity of}\\ \text{implementing the RGO}\end{array}}\,.$$

The next few sections will be devoted to studying the iteration complexity for the proximal sampler under various assumptions, including strong log-concavity and functional inequalities such a log-Sobolev inequality.

Implementations of the RGO will be discussed in Section 8.6. Similarly as in Optimization Box 8.0.1, as soon as $\|\nabla^2 V\|_{\mathrm{op}} \leqslant \beta$ and $h \leqslant \frac{1}{2\beta}$, the RGO is a strongly log-concave and log-smooth distribution with condition number at most 3. Therefore, the cost of implementing the RGO will boil down to the cost of sampling from *well-conditioned strongly log-concave* distributions.

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 $(Q_t)_{t\geqslant 0}$ denotes the standard heat semigroup, i.e. $\mu Q_t = \mu * \mathsf{normal}(0, tI_d)$, 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 1.4.11). 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,*

$$W_2(\mu_k^X, \bar\mu_k^X) \leqslant \frac{W_2(\mu_0^X, \bar\mu_0^X)}{(1+\alpha h)^k}\,.$$

The contraction factor matches the contraction for the proximal point method in optimization, see Exercise 8.3. 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 $W_2(\mu_1^X, \bar\mu_1^X) \leqslant \frac{1}{1+\alpha h} W_2(\mu_0^X, \bar\mu_0^X)$. Next, since the heat flow is a Wasserstein contraction (which follows from (**??**) but can also be proven by a straightforward coupling), it holds that $W_2(\mu_0^Y, \bar\mu_0^Y) \leqslant W_2(\mu_0^X, \bar\mu_0^X)$, so it suffices to show $W_2(\mu_1^X, \bar\mu_1^X) \leqslant \frac{1}{1+\alpha h} W_2(\mu_0^Y, \bar\mu_0^Y)$.

We will use the following coupling lemma.

Lemma 8.2.2. *Suppose that for all $y, \bar y \in \mathbb{R}^d$, we have*

$$W_2(\pi^{X|Y=y}, \pi^{X|Y=\bar y}) \leqslant C\,\|y - \bar y\|\,. \tag{8.2.3}$$

Then, $W_2(\mu_1^X, \bar\mu_1^X) \leqslant C\, W_2(\mu_0^Y, \bar\mu_0^Y)$.

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 8.4.

The first proof we present is from Lee et al. (2021c).

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 1.2.30) and the Otto–Villani theorem (Exercise 1.16) this implies the log-Sobolev inequality (1.4.14) and Talagrand's T_2 inequality (1.4.15). Applying these inequalities,

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar y}) \leqslant \frac{2}{\alpha + \frac{1}{h}} \, \mathsf{KL}(\pi^{X|Y=y} \,\|\, \pi^{X|Y=\bar y}) \leqslant \frac{1}{(\alpha + \frac{1}{h})^2} \, \mathsf{FI}(\pi^{X|Y=y} \,\|\, \pi^{X|Y=\bar y})\,.$$

We can compute the Fisher information explicitly. Indeed,

$$\nabla \log \frac{\pi^{X|Y=y}}{\pi^{X|Y=\bar y}} = \nabla\Bigl(\frac{\|\bar y - \cdot\|^2}{2h} - \frac{\|y - \cdot\|^2}{2h}\Bigr) = \frac{y - \bar y}{h}$$

so that

$$\mathsf{FI}(\pi^{X|Y=y} \,\|\, \pi^{X|Y=\bar y}) = \mathbb{E}_{\pi^{X|Y=y}}\Bigl[\Bigl\|\nabla \log \frac{\pi^{X|Y=y}}{\pi^{X|Y=\bar y}}\Bigr\|^2\Bigr] = \frac{\|y - \bar y\|^2}{h^2}\,.$$

Hence,

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar y}) \leqslant \frac{1}{(\alpha + \frac{1}{h})^2} \, \frac{\|y - \bar y\|^2}{h^2} = \frac{1}{(1 + \alpha h)^2} \, \|y - \bar y\|^2\,.$$

□

The next proof, from Chen et al. (2022), 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 $\mathcal{F}(\mu) := \mathsf{KL}(\mu \,\|\, \pi^X)$. Then, by Exercise 8.1,

$$\pi^{X|Y=y} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \Bigl\{\mathcal{F}(\mu) + \frac{1}{2h} \, W_2^2(\mu, \delta_y)\Bigr\} =: \operatorname{prox}_{h\mathcal{F}}(\delta_y)\,.$$

The first-order optimality conditions on Wasserstein space (Ambrosio et al., 2008, 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:

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar y}) \leqslant \|y - \bar y\|^2 - 2\alpha h\, W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar y}) - \alpha^2 h^2\, W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar y})$$

and hence

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar y}) \leqslant \frac{1}{(1+\alpha h)^2}\,\|y-\bar y\|^2\,. \qquad \square$$

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 8.3).

8.3 Simultaneous Flow and Time Reversal

We now introduce two new techniques in order to further analyze the proximal sampler.

Simultaneous 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 Vempala and Wibisono (2019), and was carried out for the proximal sampler in Chen et al. (2022). Here, we place the calculation in the framework of abstract Markov semigroup theory in order to show that the final result is a consequence of general principles rather than surprising coincidences.

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 prove the following abstract result.

Theorem 8.3.1 (simultaneous flow)**.** *Suppose that we have two processes evolving via*

$$\partial_t \mu_t = \mathscr{L}_t^* \mu_t\,, \qquad \partial_t \nu_t = \mathscr{L}_t^* \nu_t\,,$$

where * *denotes the adjoint with respect to some reference measure* $\mathfrak{m}$*. Let* Γ_t *denote the carré du champ operator associated with* $\mathscr{L}_t$*,*

$$\Gamma_t(f,g) := \frac{1}{2}\,\bigl(\mathscr{L}_t(fg) - f\,\mathscr{L}_t g - g\,\mathscr{L}_t f\bigr)\,, \qquad \Gamma_t(f) := \Gamma_t(f,f)\,.$$

Assume that $\mathscr{L}_t$ *satisfies the diffusion chain rule (Definition 2.2.13) for all* $t \geqslant 0$*. Then,*

$$\partial_t \mathcal{D}_f(\mu_t \,\|\, \nu_t) = -\int f''\bigl(\frac{\mathrm{d}\mu_t}{\mathrm{d}\nu_t}\bigr)\,\Gamma_t\bigl(\frac{\mathrm{d}\mu_t}{\mathrm{d}\nu_t}\bigr)\,\mathrm{d}\nu_t\,.$$

Proof We identify μ_t and ν_t with their densities w.r.t. $\mathfrak{m}$, and all of the following integrals are w.r.t. $\mathfrak{m}$ (although we do not write this explicitly to keep the notation light). By the definition of the adjoint,

$$\begin{aligned}
\partial_t \int f\bigl(\frac{\mu_t}{\nu_t}\bigr)\,\nu_t &= \int f'\bigl(\frac{\mu_t}{\nu_t}\bigr)\,\bigl(\partial_t\mu_t - \frac{\mu_t}{\nu_t}\,\partial_t\nu_t\bigr) + \int f\bigl(\frac{\mu_t}{\nu_t}\bigr)\,\partial_t\nu_t \\
&= \int f'\bigl(\frac{\mu_t}{\nu_t}\bigr)\,\bigl(\mathscr{L}_t^*\mu_t - \frac{\mu_t}{\nu_t}\,\mathscr{L}_t^*\nu_t\bigr) + \int f\bigl(\frac{\mu_t}{\nu_t}\bigr)\,\mathscr{L}_t^*\nu_t \\
&= \int \Bigl(\mathscr{L}_t f'\bigl(\frac{\mu_t}{\nu_t}\bigr)\,\mu_t - \mathscr{L}_t\bigl(f'\bigl(\frac{\mu_t}{\nu_t}\bigr)\,\frac{\mu_t}{\nu_t}\bigr)\,\nu_t + \mathscr{L}_t f\bigl(\frac{\mu_t}{\nu_t}\bigr)\,\nu_t\Bigr)\,.
\end{aligned}$$

For the second term, we apply the definition of the carré du champ to obtain

$$\mathscr{L}_t\big(f'\big(\frac{\mu_t}{\nu_t}\big)\,\frac{\mu_t}{\nu_t}\big)\,\nu_t = 2\Gamma_t\big(f'\big(\frac{\mu_t}{\nu_t}\big),\,\frac{\mu_t}{\nu_t}\big)\,\nu_t + \mu_t\,\mathscr{L}_t f'\big(\frac{\mu_t}{\nu_t}\big) + f'\big(\frac{\mu_t}{\nu_t}\big)\,\nu_t\,\mathscr{L}_t\frac{\mu_t}{\nu_t}\,.$$

For the third term, the diffusion chain rule says that for any function f,

$$\mathscr{L}_t f(\rho) = f'(\rho)\,\mathscr{L}_t\rho + f''(\rho)\,\Gamma_t(\rho)$$

hence

$$\mathscr{L}_t f\big(\frac{\mu_t}{\nu_t}\big) = f'\big(\frac{\mu_t}{\nu_t}\big)\,\mathscr{L}_t\frac{\mu_t}{\nu_t} + f''\big(\frac{\mu_t}{\nu_t}\big)\,\Gamma_t\big(\frac{\mu_t}{\nu_t}\big)\,.$$

Substituting this in and taking advantage of a cancellation,

$$\partial_t\int f\big(\frac{\mu_t}{\nu_t}\big)\,\nu_t = -2\int\Gamma_t\big(f'\big(\frac{\mu_t}{\nu_t}\big),\,\frac{\mu_t}{\nu_t}\big)\,\nu_t + \int f''\big(\frac{\mu_t}{\nu_t}\big)\,\Gamma_t\big(\frac{\mu_t}{\nu_t}\big)\,\nu_t\,.$$

Finally, to simplify this further, the diffusion chain rule implies

$$\Gamma_t(f(\rho),\sigma) = f'(\rho)\,\Gamma_t(\rho,\sigma)\,,$$

and it finishes the proof. □

Let us now parse the content of the theorem in several important cases. First, suppose that $(\mu_t)_{t\geqslant 0}$ is the Langevin diffusion and $\nu_t = \pi$ is the stationary distribution for all $t \geqslant 0$. In this case, $\mathscr{L}_t = \mathscr{L}$ is the Langevin generator for all $t \geqslant 0$, $\Gamma_t(\rho) = \|\nabla\rho\|^2$, and we obtain the following formula: $\partial_t\mathcal{D}_f(\mu_t \,\|\, \pi) = -\int f''(\mu_t/\pi)\,\|\nabla(\mu_t/\pi)\|^2\,\mathrm{d}\pi$. In particular, for $f(x) = x\log x$, it recovers $\partial_t\,\mathsf{KL}(\mu_t \,\|\, \pi) = -\,\mathsf{FI}(\mu_t \,\|\, \pi)$.

Now suppose that $(\mu_t)_{t\geqslant 0}$, $(\nu_t)_{t\geqslant 0}$ are two copies of the Langevin diffusion, where $(\nu_t)_{t\geqslant 0}$ is not necessarily stationary. Then, we can obtain the formula $\partial_t\mathcal{D}_f(\mu_t \,\|\, \nu_t) = -\int f''(\mu_t/\nu_t)\,\|\nabla(\mu_t/\nu_t)\|^2\,\mathrm{d}\nu_t$, which contains $\partial_t\,\mathsf{KL}(\mu_t \,\|\, \nu_t) = -\,\mathsf{FI}(\mu_t \,\|\, \nu_t)$ as a special case. The evolution of the KL divergence as *both* arguments evolve is hardly any more complicated than the case where only *one* of the arguments evolves!

A further special case is obtained when $(\mu_t)_{t\geqslant 0}$, $(\nu_t)_{t\geqslant 0}$ evolve by the heat flow, which corresponds to the Langevin diffusion with potential $V = 0$ (up to time rescaling). The theorem still applies, we obtain $\partial_t\,\mathsf{KL}(\mu_t \,\|\, \nu_t) = -\frac{1}{2}\,\mathsf{FI}(\mu_t \,\|\, \nu_t)$ (the factor of $\frac{1}{2}$ arises because the generator for the heat flow is $\mathscr{L} = \frac{1}{2}\,\Delta$, which then yields $\Gamma = \frac{1}{2}\,\|\nabla\cdot\|^2$). Note that Theorem 8.3.1 explains why: the formula for the derivative in time of the f-divergence only depends on the diffusion through its *carré du champ*, which is the same (up to rescaling) for both the heat flow and for the Langevin diffusion. This shows that the heat flow is not really that special and that much of the following discussion on the proximal sampler would equally well apply if we replaced the heat flow with, e.g., the OU process, but the heat flow is convenient for both calculation and implementation.

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 "forward step" and the "backward 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 backward 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 and simultaneous backward flow.

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

$$\mathrm{d}Z_t^{\leftarrow} = \nabla \log(\pi^X Q_{h-t})(Z_t^{\leftarrow})\,\mathrm{d}t + \mathrm{d}B_t \tag{8.3.2}$$

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 $\int \pi^{X|Y=y}\,\mu_k^Y(\mathrm{d}y) = \mu_{k+1}^X$. 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 $\mu_t^{\leftarrow} := \mu Q_t^{\leftarrow}$ for the law of $Z_t^{\leftarrow}$ initialized at $Z_0^{\leftarrow} \sim \mu$. If we have two such processes $(\mu_t^{\leftarrow})_{t\in[0,h]}$, $(\nu_t^{\leftarrow})_{t\in[0,h]}$, then instantaneously at time t they evolve according a generator $\mathscr{L}_t$, which corresponds to the carré du champ $\Gamma_t(\rho) = \frac{1}{2}\,\|\nabla\rho\|^2$. So once again, by Theorem 8.3.1 we obtain $\partial_t \operatorname{KL}(\mu_t^{\leftarrow} \,\|\, \nu_t^{\leftarrow}) = -\frac{1}{2}\operatorname{FI}(\mu_t^{\leftarrow} \,\|\, \nu_t^{\leftarrow})$, which leads to a pleasing symmetry between the forward and backward steps of the proximal sampler.

8.4 Convergence under Log-Concavity

Next, we present a convergence proof for the proximal sampler under log-concavity, following Chen et al. (2022). The proof can be compared to the $1/t$ convergence rate for the Langevin diffusion under log-concavity (1.4.13), 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,*

$$\operatorname{KL}(\mu_k^X \,\|\, \pi^X) \leqslant \frac{W_2^2(\mu_0^X, \pi^X)}{kh}\,.$$

Proof **Forward step**. Along the simultaneous heat flow, Theorem 8.3.1 shows that

$$\partial_t \operatorname{KL}(\mu_0^X Q_t \,\|\, \pi^X Q_t) = -\frac{1}{2}\operatorname{FI}(\mu_0^X Q_t \,\|\, \pi^X Q_t)$$

so we need to lower bound the Fisher information. Also, log-concavity is preserved by convolution, so $\pi^X Q_t$ is log-concave (Exercise 2.11). Hence, by convexity of the KL divergence to a log-concave target along Wasserstein geodesics (**??**),

$$0 = \operatorname{KL}(\pi^X Q_t \,\|\, \pi^X Q_t) \geqslant \operatorname{KL}(\mu_0^X Q_t \,\|\, \pi^X Q_t) + \mathbb{E}_{\mu_0^X Q_t}\Big\langle \nabla \log \frac{\mu_0^X Q_t}{\pi^X Q_t}, T_{\mu_0^X Q_t \to \pi^X Q_t} - \mathrm{id}\Big\rangle\,.$$

Rearranging this and using the Cauchy–Schwarz inequality,

$$\underbrace{\mathbb{E}_{\mu_0^X Q_t}\Big[\Big\|\nabla \log \frac{\mu_0^X Q_t}{\pi^X Q_t}\Big\|^2\Big]}_{\mathsf{FI}(\mu_0^X Q_t \,\|\, \pi^X Q_t)} W_2^2(\mu_0^X Q_t, \pi^X Q_t) \geqslant \mathsf{KL}(\mu_0^X Q_t \,\|\, \pi^X Q_t)^2 .$$

Combining this with the fact that the Wasserstein distance is decreasing along the simultaneous heat flow,

$$\partial_t \mathsf{KL}(\mu_0^X Q_t \,\|\, \pi^X Q_t) \leqslant -\frac{1}{2}\,\frac{\mathsf{KL}(\mu_0^X Q_t \,\|\, \pi^X Q_t)^2}{W_2^2(\mu_0^X, \pi^X)} .$$

Solving this differential inequality,

$$\frac{1}{\mathsf{KL}(\mu_0^Y \,\|\, \pi^Y)} = \frac{1}{\mathsf{KL}(\mu_0^X Q_h \,\|\, \pi^X Q_h)} \geqslant \frac{1}{\mathsf{KL}(\mu_0^X \,\|\, \pi^X)} + \frac{h}{2\, W_2^2(\mu_0^X, \pi^X)} .$$

Backward step. Along the simultaneous backward heat flow, Theorem 8.3.1 gives

$$\partial_t \mathsf{KL}(\mu_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow}) = -\frac{1}{2}\,\mathsf{FI}(\mu_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow}) .$$

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 (8.3.2)), a similar calculation as the forward step leads to the inequality

$$\frac{1}{\mathsf{KL}(\mu_1^X \,\|\, \pi^X)} = \frac{1}{\mathsf{KL}(\mu_0^Y Q_h^{\leftarrow} \,\|\, \pi^Y Q_h^{\leftarrow})} \geqslant \frac{1}{\mathsf{KL}(\mu_0^Y \,\|\, \pi^Y)} + \frac{h}{2\, W_2^2(\mu_0^X, \pi^X)} .$$

We iterate these inequalities, using the fact that $W_2(\mu_k^X, \pi^X) \leqslant W_2(\mu_0^X, \pi^X)$ for all $k \in \mathbb{N}$ (which follows from Theorem 8.2.1) to obtain

$$\frac{1}{\mathsf{KL}(\mu_k^X \,\|\, \pi^X)} \geqslant \frac{1}{\mathsf{KL}(\mu_0^X \,\|\, \pi^X)} + \frac{kh}{W_2^2(\mu_0^X, \pi^X)}$$

or

$$\mathsf{KL}(\mu_k^X \,\|\, \pi^X) \leqslant \frac{\mathsf{KL}(\mu_0^X \,\|\, \pi^X)}{1 + kh\,\mathsf{KL}(\mu_0^X \,\|\, \pi^X)/W_2^2(\mu_0^X, \pi^X)} \leqslant \frac{W_2^2(\mu_0^X, \pi^X)}{kh} . \qquad \square$$

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 Chen et al. (2022).

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,*

$$\chi^2(\mu_k^X \,\|\, \pi^X) \leqslant \frac{\chi^2(\mu_0^X \,\|\, \pi^X)}{(1 + h/C_{\mathsf{PI}})^{2k}} .$$

Proof **Forward step**. Along the simultaneous heat flow, by Theorem 8.3.1,

$$\partial_t \chi^2(\mu_0^X Q_t \,\|\, \pi^X Q_t) = -\mathbb{E}_{\pi^X Q_t}\Big[\Big\|\nabla \frac{\mu_0^X Q_t}{\pi^X Q_t}\Big\|^2\Big]\,.$$

Since π^X satisfies a Poincaré inequality with constant C_{PI}, by subadditivity of the Poincaré constant under convolution (Proposition 2.3.8), $\pi^X Q_t$ satisfies a Poincaré inequality with constant at most $C_{\mathsf{PI}} + t$. It therefore yields

$$\partial_t \chi^2(\mu_0^X Q_t \,\|\, \pi^X Q_t) \leqslant -\frac{1}{C_{\mathsf{PI}} + t} \operatorname{var}_{\pi^X Q_t} \frac{\mu_0^X Q_t}{\pi^X Q_t} = -\frac{1}{C_{\mathsf{PI}} + t}\, \chi^2(\mu_0^X Q_t \,\|\, \pi^X Q_t)$$

and hence

$$\chi^2(\mu_0^Y \,\|\, \pi^Y) = \chi^2(\mu_0^X Q_h \,\|\, \pi^X Q_h) \leqslant \exp\Big(-\int_0^h \frac{1}{C_{\mathsf{PI}} + t}\,\mathrm{d}t\Big)\, \chi^2(\mu_0^X \,\|\, \pi^X) = \frac{\chi^2(\mu_0^X \,\|\, \pi^X)}{1 + h/C_{\mathsf{PI}}}\,.$$

Backward step. Along the simultaneous backward heat flow, Theorem 8.3.1 yields

$$\partial_t \chi^2(\mu_0^Y Q_t^{\leftarrow} \,\|\, \pi^Y Q_t^{\leftarrow}) = -\mathbb{E}_{\pi^Y Q_t^{\leftarrow}}\Big[\Big\|\nabla \frac{\mu_0^Y Q_t^{\leftarrow}}{\pi^Y Q_t^{\leftarrow}}\Big\|^2\Big]\,.$$

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

$$\chi^2(\mu_1^X \,\|\, \pi^X) = \chi^2(\mu_0^Y Q_h^{\leftarrow} \,\|\, \pi^Y Q_h^{\leftarrow}) \leqslant \frac{\chi^2(\mu_0^Y \,\|\, \pi^Y)}{1 + h/C_{\mathsf{PI}}}\,.$$

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,*

$$\mathsf{KL}(\mu_k^X \,\|\, \pi^X) \leqslant \frac{\mathsf{KL}(\mu_0^X \,\|\, \pi^X)}{(1 + h/C_{\mathsf{LSI}})^{2k}}\,.$$

Recall that if π^X is α-strongly log-concave, then it satisfies a log-Sobolev inequality with constant $C_{\mathsf{LSI}} \leqslant 1/\alpha$ (Theorem 1.2.30). 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 backward step.

Similarly to Theorem 2.2.25, 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 backward step, whereas in Theorem 8.5.2 the forward and backward steps each contribute a contraction factor of $\frac{1}{1+\alpha h}$.

8.6 Implementations of the RGO and Applications

In this section, we discuss various implementation strategies for the RGO, thereby obtaining concrete sampling guarantees.

Implementation via naïve rejection sampling.

First, we 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. Recall that for $V_y(x) := V(x) + \frac{1}{2h}\,\|y-x\|^2$, we have $(\frac{1}{h}-\beta)\,I_d \preceq \nabla^2 V_y \preceq (\frac{1}{h}+\beta)\,I_d$. In particular, for $h < \frac{1}{\beta}$, 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 $\kappa \leqslant \exp(4/d)$ for $d \geqslant 2$. 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 oracle for hV), we can construct an upper envelope for which the average number of iterations of rejection sampling is bounded by $\kappa^{d/2} \leqslant \exp(2) \leqslant 8$. 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 a proximal oracle for 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 8.6.1. *Assume that V is β-smooth. Then, if $h \leqslant \frac{1}{\beta d}$, rejection sampling implements the RGO for π^X exactly using one computation to a proximal oracle for V and $O(1)$ expected calls to an evaluation oracle for V.*

Although the rejection sampling strategy is naïve, it yields surprisingly strong results with a simple analysis. If we assume that π^X satisfies an LSI with constant $1/\alpha$, then by Theorem 8.5.2 yields an iteration complexity of roughly $\widetilde{O}(\frac{1}{\alpha h}\log(1/\varepsilon)) = \widetilde{O}(\kappa d \log(1/\varepsilon))$, where $\kappa := \beta/\alpha$ and each iteration requires $O(1)$ queries in expectation. This result already matches the guarantees for MALA under a feasible start (Theorem 7.3.4), and under a weaker assumption (LSI rather than strong log-concavity)! It is also straightforward to obtain guarantees in Rényi divergence (Exercise 8.6), which are stronger than Theorem 6.1.2 (by a factor of κ, and because the result only has a logarithmic dependence on $1/\varepsilon$). Finally, under a Poincaré inequality, Theorem 8.5.1 yields an iteration complexity of roughly $\widetilde{O}(\kappa d^2 \log(1/\varepsilon))$ (assuming that $\chi^2(\mu_0 \,\|\, \pi) \leqslant \exp(\widetilde{O}(d))$), which moreover can be shown to hold in any Rényi divergence, and which can be compared to the $\widetilde{O}(\kappa^2 d^3/\varepsilon^2)$ rate obtained via LMC (Theorem 6.2.9) or the $\widetilde{O}(\kappa^{3/2} d^2/\varepsilon)$ rate obtained via ULMC (Corollary 6.3.2) in this setting.

Implementation via high-accuracy samplers.

Next, we consider the RGO with step size $h = \frac{1}{2\beta}$, so that the RGO is a strongly log-concave and log-smooth distribution with condition number κ. In this regime, the iteration complexity of the proximal sampler is roughly $\widetilde{O}(\kappa \log(1/\varepsilon))$ under an LSI, or $\widetilde{O}(\kappa d \log(1/\varepsilon))$ under a PI, but we also have to take into account the per-iteration complexity of implementing the RGO.

Here, the implementation of the RGO will be handled by one of the high-accuracy samplers from Chapter 7. However, in order to complete the analysis, we also need to analyze how the error propagates due to the imperfect implementation. Here, we state a very simple version of the error analysis, which is proven via coupling (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,*

$$\|\hat{\mu}_N^X - \mu_N^X\|_{\mathrm{TV}} \leqslant N\delta\,.$$

In any case, we only need to run the proximal sampler for a number of iterations N which is polynomial in the various problem parameters, which means δ can be taken to be inverse polynomially small. But the high-accuracy guarantees from Chapter 7 can sample from the RGO to accuracy δ with a number of iterations that only scales logarithmically with $1/\delta$, which ultimately contributes a negligible overhead to the analysis if we ignore logarithmic factors. Therefore, the total complexity of the proximal sampler is roughly given by its iteration complexity, multiplied by the dimension dependence of the high-accuracy sampler (note that the condition number of the RGO is $O(1)$ so it does not play a role here). In summary, the use of a high-accuracy sampler allows us to safely neglect the errors arising from inexact implementation, but one could also consider other implementations with a more careful error analysis.

In particular, let us consider the state-of-the-art high-accuracy sampler—MALA initialized with a ULMC warm start—from Corollary 7.3.6. That result required knowing the mode of the distribution, which for the RGO amounts to a call to a proximal oracle for V. The complexity of implementing the RGO is then $\widetilde{O}(\sqrt{d}\,\mathrm{polylog}(1/\varepsilon))$, which readily yields the following corollaries. We focus solely on total variation guarantees for simplicity; see Altschuler and Chewi (2024a) for more detailed results.

Corollary 8.6.3. *Let $\pi^X \propto \exp(-V)$, where V is β-smooth. Take $h = \frac{1}{2\beta}$ and assume we have an evaluation and proximal oracle 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 Corollary 7.3.6.*

1. *(Theorem 8.4.1) If V is convex, we obtain the guarantee $\|\mu_N^X - \pi^X\|_{\mathrm{TV}} \leqslant \varepsilon$ using $\widetilde{O}(\beta\sqrt{d}\, W_2^2(\mu_0^X, \pi^X)/\varepsilon^2)$ queries to the oracle.*
2. *(Theorem 8.5.1) If π^X satisfies a Poincaré inequality with constant $1/\alpha$, and we write $\kappa := \beta/\alpha$ for the condition number, we obtain $\|\mu_N^X - \pi^X\|_{\mathrm{TV}} \leqslant \varepsilon$ using $O(\kappa\sqrt{d}\,\{\log \chi^2(\mu_0^X \,\|\, \pi) + \mathrm{polylog}(1/\varepsilon)\})$ queries to the oracle.*
3. *(Theorem 8.5.2) If π^X satisfies a log-Sobolev inequality with constant $1/\alpha$, and we write $\kappa := \beta/\alpha$ for the condition number, we obtain $\|\mu_N^X - \pi^X\|_{\mathrm{TV}} \leqslant \varepsilon$ using $\widetilde{O}(\kappa\sqrt{d}\,\mathrm{polylog}(\mathsf{KL}(\mu_0^X \,\|\, \pi)/\varepsilon^2))$ queries to the oracle.*

These guarantees are state-of-the-art for each of their corresponding settings. Besides extending the guarantees of Corollary 7.3.6 to more general assumptions, note that the proximal sampler automatically improves the dependence on the condition number to $\widetilde{O}(\kappa)$ (whereas the dependence in Corollary 7.3.6 was $\widetilde{O}(\kappa^{3/2})$).

Implementation via approximate rejection sampling.

In a surprising work, Fan et al. (2023) developed an approximate rejection sampling scheme, inspired by the one in Gopi et al. (2022), which succeeds at implementing the RGO with nearly constant query complexity, for step size $h \asymp 1/(\beta\sqrt{d})$. By combining this with the convergence results for the proximal sampler established in this chapter, this yields another approach to obtaining the guarantees of Corollary 8.6.3 while completely bypassing the analysis of MALA!

TODO: Present this result in detail.

Bibliographical Notes

The reader is encouraged to read the original paper Lee et al. (2021c) 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 $F := n^{-1} \sum_{i=1}^n f_i$ 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 Liang and Chen (2022a,b), and it has been applied to the problem of differentially private convex optimization Gopi et al. (2022). A variant of the proximal sampler tailored to handle non-Euclidean norms was proposed in Gopi et al. (2023), although it is currently less well-understood than its Euclidean counterpart.

The interpretation of the proximal sampler in Section 8.1 arose out of conversations with Adil Salim, and was inspired by the interpretation of LMC in Wibisono (2018). The simultaneous flow calculation (Theorem 8.3.1) has appeared in many works, but the version that appears here, emphasizing the Markov semigroup perspective and the role of the carré du champ, was anticipated by Ramon van Handel. The main results in this chapter, on the convergence of the proximal sampler, are from Lee et al. (2021b); Chen et al. (2022).

As we have seen, the proximal sampler can be interpreted as the time reversal of the heat flow, which itself was studied in Klartag and Putterman (2021); this paper also introduced the adjoint of the heat semigroup (Exercise 8.2) and connected it with the Föllmer process Föllmer (1985) and Eldan's stochastic localization Eldan (2013). This was further explored in Chen and Eldan (2022), which interpreted stochastic localization as the "localization scheme" (a concept introduced in the paper) associated with the proximal sampler and used this to give another proof of convergence in KL divergence under strong log-concavity. The adjoint of the heat semigroup was also important for recent progress on the KLS conjecture Klartag and Lehec (2022).

The optimization results in Exercise 8.3 obtained in analogy with the proximal sampler are given in Chen et al. (2022).

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

$$\operatorname{prox}_{\mathcal{F}}(\mu) := \operatorname*{arg\,min}_{\mu' \in \mathcal{P}_2(\mathbb{R}^d)} \Big\{ \mathcal{F}(\mu') + \frac{1}{2}\, W_2^2(\mu, \mu') \Big\}\,.$$

The proximal operator was used in the seminal work Jordan et al. (1998) in order to rigorously make sense of gradient flows on the Wasserstein space. Prove that the RGO satisfies

$$\pi^{X|Y=y} = \operatorname{prox}_{h\, \mathsf{KL}(\cdot \,\|\, \pi)}(\delta_y)\,.$$

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.

⊵ **Exercise 8.2** (RGO as an adjoint)
Let Q_h denote the Markov kernel corresponding to the heat flow for time h, i.e., $Q_h f(x) := \mathbb{E}\, f(x + \sqrt{h}\, \xi)$, where $\xi \sim \mathsf{normal}(0, I_d)$. Show that Q_h defines a bounded linear map $L^2(\pi^Y) \to L^2(\pi^X)$ (in fact, with operator norm equal to 1). Show that the *adjoint* Q_h^* of this linear map is precisely the RGO.

Convergence under Strong Log-Concavity

⊵ **Exercise 8.3** (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 $\operatorname{prox}_{hV} = (\mathrm{id} + h\nabla V)^{-1}$. 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

$$\|\nabla V(x)\|^2 \geqslant 2\alpha\, \{V(x) - \inf V\}\,, \qquad \text{for all } x \in \mathbb{R}^d\,.$$

Also, let $x' := \operatorname{prox}_{hV}(x)$. Inspired by Theorem 8.5.2, we can ask whether or not it holds that

$$V(x') - \inf V \leqslant \frac{1}{(1+\alpha h)^2}\, \{V(x) - \inf V\}\,.$$

Prove that this is indeed the case.
Hint: Define $V_{t,x}(z) := V(z) + \frac{1}{2t}\, \|z - x\|^2$ and let $x_t := \operatorname{arg\,min} V_{t,x}$; then, differentiate $t \mapsto V_{t,x}(x_t)$.

⊵ **Exercise 8.4** (first coupling lemma)
Prove Lemma 8.2.2.

Simultaneous Flow and Time Reversal

⊵ **Exercise 8.5** (convergence under LSI)
Verify that the result under LSI (Theorem 8.5.2) holds.

⊵ **Exercise 8.6** (convergence in Rényi divergence)
Following Theorem 2.2.25, 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 $\pi^X = \mathsf{normal}(0, \Sigma)$ and $\mu_0^X = \mathsf{normal}(m_0, \Sigma_0)$. Show that the iterates of the proximal sampler all have Gaussian distributions, and explicitly compute the means and variances.
2 In particular, show that $\mu_1^X = \mathsf{normal}(m_1, \Sigma_1)$, where the mean satisfies $m_1 = \operatorname{prox}_{hV}(m_0)$ 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. Use this to show that the contraction factors in Theorem 8.2.1 and Theorem 8.5.2 are sharp.

Implementations of the RGO and 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 matching lower bounds. However, the problem of establishing query complexity lower bounds for sampling has been challenging and the work on this topic is nascent. Here, we will give an overview of the current progress in this direction.

Optimization Box 9.0.1. In convex optimization, the situation is better understood. For example, in the high-dimensional regime, the complexity of finding an ε-minimizer of a function $V : \mathbb{R}^d \to \mathbb{R}$ with $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and minimizer $x_\star$ satisfying $\|x_\star\| \leqslant R$ is $\Omega(\sqrt{\kappa}\log(R^2/(\alpha\varepsilon)))$, where $\kappa := \beta/\alpha$ Nemirovsky and Yudin (1983). This matches the complexity of accelerated gradient descent (Optimization Box 5.3.2), which is therefore optimal for this function class. In this case, the lower bound was actually a motivating force which led to the discovery of the optimal algorithm.

Subsequently, lower bounds for convex optimization have been established for a host of function classes and oracle models.

9.1 A Brief Discussion of Query Complexity

Before turning to the results, recall the discussion and interpretation of *query complexity* (or *oracle complexity*) from the Preface, although here we present the model in more detail. Let Π be a family of probability measures over $\mathbb{R}^d$ with continuously differentiable densities. A *first-order oracle* for $\pi \in \Pi$ with density $\pi \propto \exp(-V)$ is the mapping $\mathcal{O}_\pi : \mathbb{R}^d \to \mathbb{R} \times \mathbb{R}^d$ given by $x \mapsto (V(x) - V(0), \nabla V(x))$. A *first-order sampling algorithm* interacts with the oracle $\mathcal{O}_\pi$ by querying N points $x_1, \dots, x_N$, where each query point x_j depends on the previous query points and query values $\{x_i, \mathcal{O}_\pi(x_i)\}_{i\in[j-1]}$ as well as an independent source of randomness, and outputs a "sample" x_{out}. If x_{out} is the random output of the sampling algorithm when run on oracle $\mathcal{O}_\pi$, we say that the sampling algorithm computes the measure $\operatorname{law}(x_{\text{out}})$.

The *first-order complexity* $\mathscr{C}(\Pi, \mathsf{d}, \varepsilon)$ of the class Π is the minimum number N for which there exists a sampling algorithm with the following property: for any $\pi \in \Pi$, when given N queries to oracle $\mathscr{O}_\pi$, the sampling algorithm computes $\hat\pi$ with $\mathsf{d}(\hat\pi, \pi) \leqslant \varepsilon$. Here, d can be taken to be total variation distance or the Wasserstein distance, or even an asymmetric measure such as $\sqrt{\mathsf{KL}}$. Our goal is to characterize $\mathscr{C}(\Pi, \mathsf{d}, \varepsilon)$ up to universal constants for important classes Π. In particular, we will be interested in the class of strongly log-concave and log-smooth distributions over $\mathbb{R}^d$ with known mode at 0, since we view this class as a fundamental benchmark with which to test our understanding of acceleration, discretization, etc.

One can also consider various extensions, e.g., by changing the class of measures Π or by changing the oracle to allow for different information structures (zeroth-order oracles, second-order oracles, etc.), finite sum structures, stochasticity, and so on. It is important to eventually understand the complexities for all of these settings—and indeed they have been carefully mapped out in the sister field of optimization—but we emphasize that for sampling, even the canonical case of well-conditioned log-concave measures with an exact oracle is still not fully understood.

We also recall that query complexity is not computational complexity: our model allows the algorithm to perform an unbounded amount of computation in between queries to the oracle. This is precisely the strength of this model: due to its information-theoretic nature, it is possible to establish unconditional lower bounds. In practice, the two are closely connected. If the only access one has to the target density is through the oracle (as opposed to other forms of access, such as integrals of certain test functions, etc.), then query complexity is a lower bound for the computational complexity. On the other hand, reasonable algorithms with small query complexity tend to also have small computational complexity (indeed this is true for all of the algorithms we have encountered in this book thus far), and so we treat query complexity as a useful proxy for computation time.

Finally, we briefly discuss differences with the setting of optimization. Optimization lower bounds are often first established for simple models of deterministic algorithms, e.g., algorithms whose iterates always lie in the span of the queried gradients Nesterov (2018). These lower bounds can sometimes be extended to randomized algorithms with additional tricks. In sampling, we do not have this luxury, since randomness is an integral part of any sampler. Moreover, whereas lower bound constructions for optimization must hide the location of the minimizer from the algorithm, lower bound constructions for sampling must hide *the bulk of the distribution's mass* from the algorithm, which may explain the difficulty of the constructions.

9.2 Query Complexity in One Dimension

The first query complexity result for well-conditioned log-concave sampling is from the work of Chewi et al. (2022), which established a sharp result in dimension one. In this section, we present this result in detail because it is simple enough to do so.

Define the class

$$\Pi_{\kappa,1} \coloneqq \{\pi \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}) \mid \pi \propto \exp(-V),\ 1 \leqslant V'' \leqslant \kappa,\ V'(0) = 0\}\,.$$

In our definition of $\Pi_{\kappa,1}$, 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. Note that some restriction of this nature is needed, since without *any* prior knowledge of the mode it is impossible to find it.

Theorem 9.2.1 (Chewi et al. (2022))**.** *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* $\pi \in \Pi_{\kappa,1}$*, is* $\Theta(\log\log\kappa)$*. In other words,* $\mathscr{C}(\Pi_{\kappa,1}, \|\cdot\|_{\mathrm{TV}}, \frac{1}{64}) \asymp \log\log\kappa$*.*

Lower bound.

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 $\pi_1, \dots, \pi_m \in \Pi_{\kappa,1}$, and suppose that $\boldsymbol{i} \sim \mathsf{uniform}([m])$ 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 $\exp(-\frac{2^{2m-2}}{2\kappa}) \geqslant \frac{1}{2}$ (and note that $m = \Theta(\log\kappa)$). 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. The second derivative is

$$V_i''(x) := \begin{cases} \kappa\,, & \sqrt{\kappa}\,|x| \in [2^{i-1}, 2^i) \cup \bigcup_{j=i+1}^{m+1} [2^j, \frac{5}{4}\,2^j)\,, \\ 1\,, & \text{otherwise}\,. \end{cases}$$

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.

There are two key properties of this construction. First, we will show in Lemma 9.2.2 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 9.2.4 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 $(V_i(x), V_i'(x))$ 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 9.2.2. *For each* $i \in [m]$*,*

$$\pi_i\big((\kappa^{-\frac{1}{2}}\,2^{i-2}, \kappa^{-\frac{1}{2}}\,2^{i-1}]\big) \geqslant \frac{1}{32}\,.$$

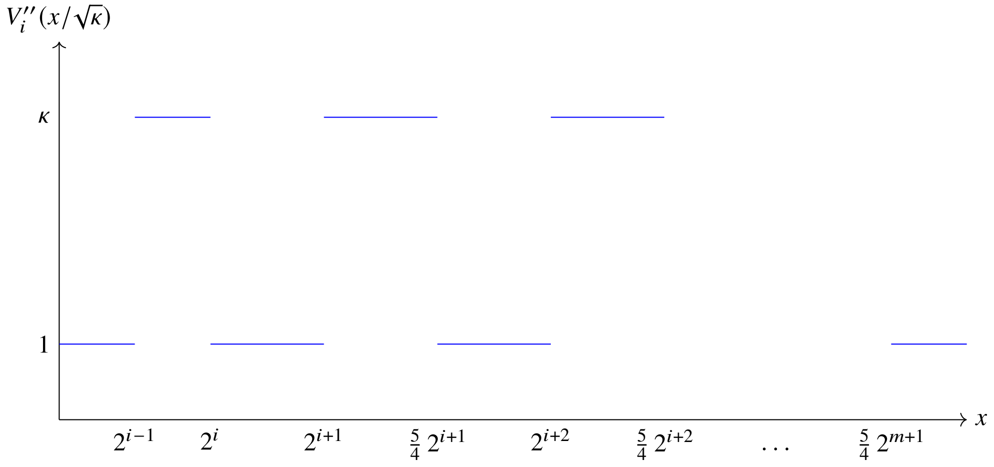

Figure 9.1 The horizontal axis is distorted for clarity.

Proof According to the definition of π_i, we have

$$\pi_i\big((\kappa^{-\frac{1}{2}}\,2^{i-2}, \kappa^{-\frac{1}{2}}\,2^{i-1}]\big) = \frac{\int_{\kappa^{-\frac{1}{2}}\,2^{i-2}}^{\kappa^{-\frac{1}{2}}\,2^{i-1}} \exp(-x^2/2)\,\mathrm{d}x}{Z_{\pi_i}}\,, \qquad Z_{\pi_i} := \int \exp(-V_i)\,.$$

Recalling that m is chosen so that $\exp(-x^2/2) \geqslant 1/2$ whenever $|x| \leqslant \kappa^{-\frac{1}{2}}\,2^{m-1}$,

$$\int_{\kappa^{-\frac{1}{2}}\,2^{i-2}}^{\kappa^{-\frac{1}{2}}\,2^{i-1}} \exp\big(-\frac{x^2}{2}\big)\,\mathrm{d}x \geqslant \frac{1}{2}\,\kappa^{-\frac{1}{2}}\,2^{i-2}\,.$$

For the normalizing constant, observe that

$$\int_0^\infty \exp(-V_i) = \int_0^{\kappa^{-\frac{1}{2}}\,2^{i}} \exp(-V_i) + \int_{\kappa^{-\frac{1}{2}}\,2^{i}}^\infty \exp(-V_i) \leqslant \kappa^{-\frac{1}{2}}\,2^{i} + \int_{\kappa^{-\frac{1}{2}}\,2^{i}}^\infty \exp(-V_i)\,.$$

Since $V_i'' = \kappa$ on $[\kappa^{-\frac{1}{2}}\,2^{i-1}, \kappa^{-\frac{1}{2}}\,2^{i}]$, it follows that $V_i'(\kappa^{-\frac{1}{2}}\,2^{i}) \geqslant \kappa^{-\frac{1}{2}}\,2^{i-1}$, and so

$$V_i(x) \geqslant \kappa^{-\frac{1}{2}}\,2^{i-1}\,(x - \kappa^{-\frac{1}{2}}\,2^{i}) + \frac{(x - \kappa^{-\frac{1}{2}}\,2^{i})^2}{2}\,, \qquad x \geqslant \kappa^{-\frac{1}{2}}\,2^{i}\,.$$

Therefore,

$$\begin{aligned}\int_{\kappa^{-\frac{1}{2}}\,2^{i}}^\infty \exp(-V_i) &\leqslant \int_{\kappa^{-\frac{1}{2}}\,2^{i}}^\infty \exp\Big(-\kappa^{-\frac{1}{2}}\,2^{i-1}\,(x - \kappa^{-\frac{1}{2}}\,2^{i}) - \frac{(x - \kappa^{-\frac{1}{2}}\,2^{i})^2}{2}\Big)\,\mathrm{d}x \\ &\leqslant \frac{1}{\kappa^{-\frac{1}{2}}\,2^{i-1}} \leqslant \frac{1}{\sqrt{\kappa}}\,,\end{aligned}$$

where we applied a standard tail estimate for Gaussian densities (Lemma 9.2.3). Then,

$$\pi_i\big((\kappa^{-\frac{1}{2}}\,2^{i-2}, \kappa^{-\frac{1}{2}}\,2^{i-1}]\big) \geqslant \frac{2^{i-3}}{2\,(2^i + 1)} \geqslant \frac{1}{32}\,,$$

which proves the result. □

In the above proof, we used the following elementary lemma.

Lemma 9.2.3. *Let* $a, x_0 > 0$. *Then,*

$$\int_{x_0}^{\infty} \exp\bigl(-a\,(x - x_0) - \frac{1}{2}\,(x - x_0)^2\bigr)\,\mathrm{d}x \leqslant \int_0^{\infty} \exp(-ax)\,\mathrm{d}x = \frac{1}{a}\,.$$

The next lemma is the main reason why we used an oscillating construction for V_i''.

Lemma 9.2.4. *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 $\{x \in \mathbb{R} : \kappa^{-\frac{1}{2}}\,2^{i-1} \leqslant |x| \leqslant \frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\}$.

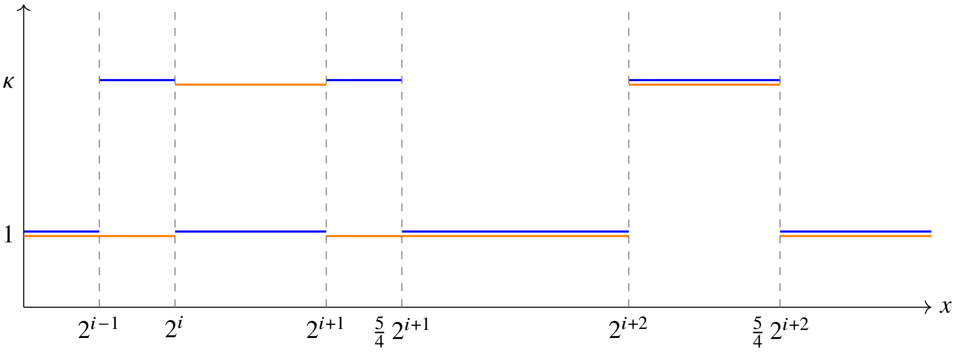

Figure 9.2 We plot $V_i''(x/\sqrt{\kappa})$ (in blue) and $V_{i+1}''(x/\sqrt{\kappa})$ (in orange). In this figure, we do not distort the horizontal axis lengths to make it easier to visually compare the relative lengths of intervals on which the second derivatives are constant.

Proof Refer to Figure 9.2 for a visual aid for the proof.

Clearly the potentials and derivatives match when $|x| \leqslant \kappa^{-\frac{1}{2}}\,2^{i-1}$. Since the second derivatives match when $|x| \geqslant \frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}$, 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 $x \geqslant 0$,

$$V_{i+1}''(x) - V_i''(x) = \begin{cases} -(\kappa - 1)\,, & \kappa^{-\frac{1}{2}}\,2^{i-1} \leqslant x \leqslant \kappa^{-\frac{1}{2}}\,2^{i}\,, \\ +(\kappa - 1)\,, & \kappa^{-\frac{1}{2}}\,2^{i} \leqslant x \leqslant \kappa^{-\frac{1}{2}}\,2^{i+1}\,, \\ -(\kappa - 1)\,, & \kappa^{-\frac{1}{2}}\,2^{i+1} \leqslant x \leqslant \frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

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,

$$\begin{aligned}
V_{i+1}\Big(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\Big) - V_i\Big(\frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\Big) = {} & \underbrace{-\frac{\kappa-1}{2}\,(\kappa^{-\frac{1}{2}}\,2^{i-1})^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}\,2^{i-1},\ \kappa^{-\frac{1}{2}}\,2^{i}]} \\
& \underbrace{-\,(\kappa-1)\,\kappa^{-\frac{1}{2}}\,2^{i-1}\,\kappa^{-\frac{1}{2}}\,2^{i} + \frac{\kappa-1}{2}\,(\kappa^{-\frac{1}{2}}\,2^{i})^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}\,2^{i},\ \kappa^{-\frac{1}{2}}\,2^{i+1}]} \\
& \underbrace{+\,(\kappa-1)\,\kappa^{-\frac{1}{2}}\,2^{i-1}\,\frac{1}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1} - \frac{\kappa-1}{2}\,\Big(\frac{1}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}\Big)^2}_{\text{integral on } [\kappa^{-\frac{1}{2}}\,2^{i+1},\ \frac{5}{4}\,\kappa^{-\frac{1}{2}}\,2^{i+1}]} \\
= {} & \frac{\kappa-1}{\kappa}\,\{-2^{2i-3} - 2^{2i-1} + 2^{2i-1} + 2^{2i-2} - 2^{2i-3}\} \\
= {} & 0\,.
\end{aligned}$$

□

We need one final ingredient: **Fano's inequality**, which is the standard tool for establishing information-theoretic lower bounds.

Theorem 9.2.5 (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,*

$$\mathbb{P}\{\widehat{\boldsymbol{i}} \ne \boldsymbol{i}\} \geqslant 1 - \frac{\mathsf{I}(\boldsymbol{i};Y) + \log 2}{\log m}\,,$$

*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 $\mathsf{H}(X) = \sum_{x\in\mathcal{X}} p(x)\log(1/p(x))$. We refer to (Cover and Thomas, 2006, §2) and Exercise 9.1 for the basic properties of entropy (and related quantities).

Let $E := \mathbb{1}\{\widehat{\boldsymbol{i}} \ne \boldsymbol{i}\}$ 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, $\mathsf{H}(E \mid \widehat{\boldsymbol{i}}) \leqslant \mathsf{H}(E) \leqslant \log 2$. Also,

$$\mathsf{H}(\boldsymbol{i} \mid E, \widehat{\boldsymbol{i}}) = \mathbb{P}\{\widehat{\boldsymbol{i}} = \boldsymbol{i}\}\,\underbrace{\mathsf{H}(\boldsymbol{i} \mid \widehat{\boldsymbol{i}}, E = 0)}_{=0} + \mathbb{P}\{\widehat{\boldsymbol{i}} \ne \boldsymbol{i}\}\,\mathsf{H}(\boldsymbol{i} \mid \widehat{\boldsymbol{i}}, E = 1) \leqslant \mathbb{P}\{\widehat{\boldsymbol{i}} \ne \boldsymbol{i}\}\log m\,.$$

Hence,

$$\mathbb{P}\{\widehat{\boldsymbol{i}} \ne \boldsymbol{i}\}\log m + \log 2 \geqslant \mathsf{H}(\boldsymbol{i} \mid \widehat{\boldsymbol{i}}) = \mathsf{H}(\boldsymbol{i}) - \mathsf{I}(\boldsymbol{i}\,;\widehat{\boldsymbol{i}}) \geqslant \log m - \mathsf{I}(\boldsymbol{i}\,;Y)$$

where the last inequality is the data-processing inequality. Rearranging the inequality completes the proof of Fano's inequality. □

Proof of Theorem 9.2.1, 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 $\|\hat{\pi}_i - \pi_i\|_{\mathrm{TV}} \leqslant \frac{1}{64}$. 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 9.2.2, a good candidate estimator $\widehat{\boldsymbol{i}}$ for $\boldsymbol{i}$ is

$$\widehat{\boldsymbol{i}} := i \in \mathbb{N} \quad \text{such that} \quad X \in (\kappa^{-\frac{1}{2}}\, 2^{i-2}, \kappa^{-\frac{1}{2}}\, 2^{i-1}] \quad \text{if such an } i \text{ exists}\,.$$

The probability that the estimator is correct is at least

$$\begin{aligned}\mathbb{P}\{\widehat{\boldsymbol{i}} = \boldsymbol{i}\} &= \frac{1}{m}\sum_{i=1}^m \mathbb{P}\{\widehat{\boldsymbol{i}} = i \mid \boldsymbol{i} = i\} = \frac{1}{m}\sum_{i=1}^m \mathbb{P}\{X \in (\kappa^{-\frac{1}{2}}\, 2^{i-2}, \kappa^{-\frac{1}{2}}\, 2^{i-1}] \mid \boldsymbol{i} = i\} \\ &= \frac{1}{m}\sum_{i=1}^m \hat{\pi}_i\big((\kappa^{-\frac{1}{2}}\, 2^{i-2}, \kappa^{-\frac{1}{2}}\, 2^{i-1}]\big) \geqslant \frac{1}{m}\sum_{i=1}^m \pi_i\big((\kappa^{-\frac{1}{2}}\, 2^{i-2}, \kappa^{-\frac{1}{2}}\, 2^{i-1}]\big) - \frac{1}{64} \geqslant \frac{1}{64}\,. \end{aligned} \tag{9.2.6}$$

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 $\mathbb{P}\{\widehat{\boldsymbol{i}} \ne \boldsymbol{i}\}$.

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 9.2.5) yields

$$\mathbb{P}\{\widehat{\boldsymbol{i}} \ne \boldsymbol{i}\} \geqslant 1 - \frac{\mathsf{I}(\boldsymbol{i}; \{x_j, \mathcal{O}_{\boldsymbol{i}}(x_j)\}_{j\in[n]}) + \log 2}{\log m}\,.$$

By the chain rule for mutual information,

$$\mathsf{I}\big(\boldsymbol{i}; \{x_j, \mathcal{O}_{\boldsymbol{i}}(x_j)\}_{j\in[n]}\big) = \sum_{j=1}^n \mathsf{I}\big(\boldsymbol{i}; x_j, \mathcal{O}_{\boldsymbol{i}}(x_j) \bigm| \{x_{j'}, \mathcal{O}_{\boldsymbol{i}}(x_{j'})\}_{j'\in[j-1]}\big)\,. \tag{9.2.7}$$

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 9.2.4 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 (9.2.7),

$$\mathsf{I}\big(\boldsymbol{i}; \{x_j, \mathcal{O}_{\boldsymbol{i}}(x_j)\}_{j\in[n]}\big) \leqslant n \log 5\,.$$

Fano's inequality then yields

$$\mathbb{P}\{\widehat{\boldsymbol{i}} \ne \boldsymbol{i}\} \geqslant 1 - \frac{n \log 5 + \log 2}{\log m}\,. \tag{9.2.8}$$

In general, for a possibly randomized algorithm, we can still deduce (9.2.8) 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 (9.2.6) and (9.2.8), 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 9.2.4 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 $\widetilde{\mu} \geqslant \widetilde{\pi}$, 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 $i_- \in \{0, 1, \dots, \lceil \frac{1}{2}\log_2\kappa \rceil\}$ such that $V(-2^{i_-}/\sqrt{\kappa}) \geqslant \frac{1}{2}$.
2 Find the first index $i_+ \in \{0, 1, \dots, \lceil \frac{1}{2}\log_2\kappa \rceil\}$ such that $V(+2^{i_+}/\sqrt{\kappa}) \geqslant \frac{1}{2}$.
3 Set $x_- := -2^{i_-}/\sqrt{\kappa}$ and $x_+ := +2^{i_+}/\sqrt{\kappa}$; then, set

$$\widetilde{\mu}(x) := \begin{cases} \exp\Bigl(-\dfrac{x - x_-}{2x_-} - \dfrac{(x-x_-)^2}{2}\Bigr), & x \leqslant x_-\,, \\ 1\,, & x_- \leqslant x \leqslant x_+\,, \\ \exp\Bigl(-\dfrac{x - x_+}{2x_+} - \dfrac{(x-x_+)^2}{2}\Bigr), & x \geqslant x_+\,. \end{cases}$$

To see why i_- and i_+ exist, from $V'' \geqslant 1$ and $V(0) = V'(0) = 0$ we have $V(x) \geqslant x^2/2$. Hence, if $|x| = 2^i/\sqrt{\kappa}$ where $i \geqslant \frac{1}{2}\log\kappa$, we have $V(x) \geqslant 1/2$.

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 9.2.1, 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 $\widetilde{\pi} \leqslant 1 \leqslant \widetilde{\mu}$ on $[0, x_+]$. Next, since $V(x_+) \geqslant 1/2$ (by the definition of x_+), convexity of V yields

$$V'(x_+) \geqslant \frac{V(x_+) - V(0)}{x_+} \geqslant \frac{1}{2x_+}\,.$$

Thus, for $x \geqslant x_+$,

$$V(x) \geqslant V(x_+) + V'(x_+)\,(x - x_+) + \frac{1}{2}\,(x - x_+)^2 \geqslant \frac{1}{2x_+}\,(x - x_+) + \frac{1}{2}\,(x-x_+)^2\,,$$

which shows that $\widetilde{\pi}(x) \leqslant \widetilde{\mu}(x)$. By a symmetric argument on $\mathbb{R}_-$, we conclude that $\widetilde{\pi} \leqslant \widetilde{\mu}$.

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

$$\int_0^{x_+} \widetilde{\mu} = \int_0^{1/\sqrt{\kappa}} \exp(-V) \geqslant \int_0^{1/\sqrt{\kappa}} \exp\Big(-\frac{\kappa x^2}{2}\Big)\, \mathrm{d}x \geqslant \frac{1}{3\sqrt{\kappa}} = \frac{x_+}{3}\,.$$

When $i_+ > 0$, then by the definition of i_+ we have $V(x_+/2) \leqslant 1/2$, so

$$\int_0^{x_+} \widetilde{\mu} \geqslant \int_0^{x_+/2} \exp(-V) \geqslant \frac{x_+}{4}\,.$$

On the other hand, by Lemma 9.2.3,

$$\int_{\mathbb{R}_+} \widetilde{\mu} = \int_0^{x_+} \widetilde{\mu} + \int_{x_+}^{\infty} \widetilde{\mu} \leqslant x_+ + \int_{x_+}^{\infty} \exp\Big(-\frac{1}{2x_+}\,(x - x_+) - \frac{1}{2}\,(x - x_+)^2\Big)\, \mathrm{d}x \leqslant 3x_+\,.$$

Hence, $\int_{\mathbb{R}_+} \widetilde{\mu} \leqslant 3x_+ \leqslant 12 \int_{\mathbb{R}_+} \widetilde{\pi}$, and similarly $\int_{\mathbb{R}_-} \widetilde{\mu} \leqslant 12 \int_{\mathbb{R}_-} \widetilde{\pi}$. Therefore, $Z_\mu/Z_\pi \leqslant 12$. □

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.

9.3 Query Complexity in Constant Dimension

What about higher dimension? In particular, consider

$$\Pi_{\kappa,d} \coloneqq \{\pi \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \mid \pi \propto \exp(-V),\ I_d \preceq \nabla^2 V \preceq \kappa I_d,\ \nabla V(0) = 0\}\,.$$

Theorem 9.3.1 (Chewi et al. (2023)). *There exists a universal constant* $\varepsilon_0 > 0$ *such that for any* $d \geqslant 2$, $\mathscr{C}(\Pi_{\kappa,d}, \|\cdot\|_{\mathrm{TV}}, \varepsilon_0) \asymp_d \log\kappa$.

This result characterizes the *low-dimensional* complexity of log-concave sampling. The lower bound is based on a two-dimensional construction witnessing $\mathscr{C}(\Pi_{\kappa,2}, \|\cdot\|_{\mathrm{TV}}, \varepsilon_0) \gtrsim \log\kappa$, which implies the lower bound in any dimension $d \geqslant 2$ because we can freely embed lower-dimensional distributions into higher-dimensional space (take the product with the Gaussian measure, for instance). On the other hand, the lower bound is tight in constant dimension because there is an algorithm witnessing $\mathscr{C}(\Pi_{\kappa,d}, \|\cdot\|_{\mathrm{TV}}, \varepsilon_0) \lesssim_d \log\kappa$ for any d (here, the notation $\lesssim_d$ hides a large dependence on d).

Since the construction is rather technical, we will not present this result in detail and instead proceed on a more intuitive level.

Lower bound.

Let us revisit the intuition behind the one-dimensional construction. Due to Lemma 9.2.2, π_i concentrates its mass on the interval $\kappa^{-1/2}\,[-2^{i-1}, +2^{i-1}]$. Therefore, consider instead the idealization $\tilde{\pi}_i = \mathsf{uniform}(\kappa^{-1/2}\,[-2^{i-1}, +2^{i-1}])$ for $i \in [m]$. This family has the following desirable properties: (1) each $\tilde{\pi}_i$ is log-concave; (2) the distributions are well-separated, in the sense that $\|\tilde{\pi}_i - \tilde{\pi}_j\|_{\mathrm{TV}} \gtrsim 1$ for $i, j \in [m]$, $i \neq j$ (this follows from Lemma 9.2.2 and it is the reason why the intervals are

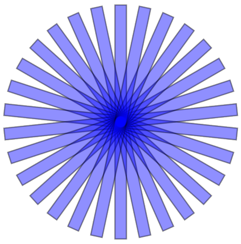

Figure 9.3 Uniform distributions over thin rectangles emanating from the origin.

chosen to double in size as i increases); (3) querying a distribution $\tilde{\pi}_i$ only yields a single bit of information (whether or not the query lands in the support of $\tilde{\pi}_i$). Moreover, since a distribution in $\Pi_{\kappa,1}$ is "effectively supported" on an interval of length at least $\Omega(\kappa^{-1/2})$ and at most $O(1)$ (think about the upper bound algorithm from Section 9.2), we can take $m = \Theta(\log \kappa)$ and still have the family $\{\tilde{\pi}_i\}_{i\in[m]}$ "resemble" members of $\Pi_{\kappa,1}$. Since the number of queries needed to identify a member of the family is the logarithm of the size of the family, i.e., $\Theta(\log\log\kappa)$, this predicts the correct lower bound in dimension one. Of course, the fatal flaw is that the family $\{\tilde{\pi}_i\}_{i\in[m]}$ is *not* a subset of $\Pi_{\kappa,1}$, being neither strongly log-concave nor log-smooth, and most of the work in the construction is to find a true subset of $\Pi_{\kappa,1}$ that suitably mimics the idealized family $\{\tilde{\pi}_i\}_{i\in[m]}$.

Now we move to two dimensions. Here, our model for an element of $\Pi_{\kappa,2}$ is the uniform distribution over a rectangle of dimensions $\kappa^{-1/2} \times 1$. If we place these rectangles pointing away from the origin and spreading out like a fan, we can now fit $\Theta(\kappa^{1/2})$ of them in $\mathbb{R}^2$ while remaining well-separated (Figure 9.3). This family of uniform distributions still satisfies the three points listed above, and now we expect a query lower bound of $\Omega(\log\kappa)$ due to the larger size of the family.

The challenge is to now modify this family so that it lies in $\Pi_{\kappa,2}$, which is unfortunately quite involved. The first step is to note that strong log-concavity does not pose an issue: if we have a family of potentials which are convex and κ-smooth, then we can add a quadratic term $\kappa^{-O(1)}\,\|\cdot\|^2/2$ to each of the potentials, where $\kappa^{-O(1)}$ is chosen so that this term has a negligible effect in the region of interest. The condition number now becomes $\kappa^{O(1)}$, but this polynomial blow-up is acceptable since we are shooting for a $\Omega(\log\kappa)$ lower bound (it only affects the bound by a constant factor).

Let us restrict to rectangles in the positive quadrant and index them via bit strings $b \in \{0,1\}^N$ corresponding to slope $[b] := 0.b_1 \dots b_N$ (represented in binary). Thus, we have 2^N distributions (eventually, $N \asymp \log\kappa$), where the b-th distribution will be concentrated on the sector $\mathscr{Z}_b := \{(x,y) \in \mathbb{R}^2 \mid x \geqslant 0,\ |y - [b]\,x| \leqslant 2^{-N},\ \|(x,y)\| \leqslant 1\}$. In other words, $\mathscr{Z}_b$ will be the *zero set* of the potential V_b corresponding to the distribution $\pi_b \propto \exp(-V_b)$. In Figure 9.3, we essentially took π_b to be uniform over $\mathscr{Z}_b$ (with the immaterial difference that we now work with sectors rather than rectangles), but now we need to smooth out these distributions. We can do this by by letting V_b grow slowly away from $\mathscr{Z}_b$, or by mollifying V_b; eventually we will need to do both. As we will see, however, the former is convenient for controlling information leakage.

Indeed, as a first attempt, suppose we let $V_b = \operatorname{dist}(\cdot, \mathscr{Z}_b)$. Then, V_b is convex because $\mathscr{Z}_b$ is convex, but it is non-smooth. We can fix the latter problem later via smoothing, but the bigger problem is that this distribution reveals too much information via queries. Indeed, the gradient of V_b immediately reveals the direction toward $\mathscr{Z}_b$, and once we locate $\mathscr{Z}_b$ we can identify b and thus π_b. (Note that since our lower bound rests on finding a subfamily $\{\pi_b\}_{b\in\{0,1\}^N} \subseteq \Pi_{\kappa,2}$ that is already hard to sample, when proving our lower bound we have to contend with algorithms which "know" the family $\{\pi_b\}_{b\in\{0,1\}^N}$ in advance, hence identifying b suffices to output an exact sample from π_b.)

The main idea is to design the potential V_b so that querying at a distance of roughly 2^{-k} away from $\mathscr{Z}_b$ reveals only the first k bits of b. If this property holds, then heuristically the "probability" of learning k bits is only 2^{-k}, so in expectation we only learn $O(1)$ bits per query. Moreover, we can ensure this property as follows: let $[b]_k \coloneqq 0.b_1 \ldots b_k$ denote the slope corresponding to the first k bits of b, and define

$$\mathscr{Z}_{b,k} \coloneqq \{(x,y) \in \mathbb{R}^2 \mid x \geqslant 0,\ |y - [b]_k x| \leqslant 2^{-k},\ \|(x,y)\| \leqslant 1\}\,.$$

Then, we define

$$V_b(x,y) \coloneqq \max_{k=1,\ldots,N} 2^{-k} \operatorname{dist}((x,y), \mathscr{Z}_{b,k})\,.$$

To see why this works, note that if we query at a large distance away from all of the $\mathscr{Z}_{b,k}$'s, then "the larger slope wins" in the maximum defining V_b, so $V_b = 2^{-1} \operatorname{dist}(\cdot, \mathscr{Z}_{b,1})$: at this distance, the query to V_b only depends on the first bit of b. Similarly, convince yourself that a query at distance $\approx 2^{-k}$ away from $\mathscr{Z}_b$ only depends on $k + O(1)$ bits of b. Importantly, since V_b is a maximum of convex functions, it is convex.

Since V_b is still not smooth (its gradient is discontinuous), we still need to smooth further. For this, we can use the standard analysis trick of *mollification*, i.e., convolution with a smooth bump function which is supported on a ball of radius $\kappa^{-O(1)}$. After mollifying and adding a quadratic term as discussed above, all potentials are $\kappa^{-O(1)}$-strongly convex and $\kappa^{O(1)}$-smooth as desired. Unfortunately, mollification introduces a further complication: it "blurs" out the distribution, which causes information to "leak" between adjacent parts of the distribution. This is particularly harmful at the origin, where all of the zero sets intersect (Figure 9.3). As a result, it is conceivable that a single query near the origin, within the scale of the mollification, could reveal all of the bits of b at once.

Handling this last issue requires care, and this is addressed in Chewi et al. (2023) with another multiscale construction. We refer to the paper for the full proof of the lower bound.

Upper bound.

Finally, to check that the lower bound is tight, we need to exhibit an algorithm with query complexity $O(\log \kappa)$ in constant dimension. Such a result is folklore in the literature on sampling from convex bodies, which shows how to "round" a convex body with a number of steps that is logarithmic in the condition number of the body (see, e.g., Kannan et al. (1997)). We refer to Chewi et al. (2023) for a self-contained analysis based on rejection sampling for our setting of interest, albeit with exponential dependence on d.

9.4 Query Complexity for Gaussians

Bibliographical Notes

The lower bounds in this chapter are from Chewi et al. (2021, 2023). Below, we survey lower bounds in related settings and other approaches.

The paper Gopi et al. (2022) obtained a zeroth-order query complexity result for distributions over the unit ball whose potentials V are of the form $V = \sum_{i=1}^{\infty} f_i + \frac{\alpha}{2} \|\cdot\|^2$, where each f_i is L-Lipschitz and the complexity is measured in terms of the number of queries to the individual f_i's. In the regime $d \ll L^2/\alpha$, the query complexity is $\widetilde{\Theta}(L^2/\alpha)$, where the upper bound is based on the proximal sampler (Section 8.1), and the lower bound is based on a reduction to optimization.

In Chatterji et al. (2022), the authors obtained 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 $\frac{\alpha}{2}$-strongly log-concave and α-log-smooth distribution whose mean lies in the ball $\mathsf{B}(0, 1/\alpha)$, with an oracle that given $x \in \mathbb{R}^d$ outputs $\nabla V(x) + \xi$ with $\xi \sim \mathsf{normal}(0, \Sigma)$ and $\operatorname{tr} \Sigma \leqslant \sigma^2 d$, 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.

Other works have focused on obtaining lower bounds against specific families of algorithms. As already discussed in Section 7.3, the works Chewi et al. (2021); Lee et al. (2021a); Wu et al. (2021) proved lower bounds for MALA, culminating in a precise understanding of the runtime of MALA both from feasible and warm start initializations. On the other hand, the paper Cao et al. (2021) studied the complexity of discretization. In their setup, there is a stochastic process $(Z_t)_{t \geqslant 0}$, driven by some underlying Brownian motion $(B_t)_{t \geqslant 0}$; for example, the process $(Z_t)_{t \geqslant 0}$ could be 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 $(B_t)_{t \geqslant 0}$, and the goal of the algorithm is to output a point $\widehat{Z}_T$ which is close to Z_T in mean squared error: $\mathbb{E}[\|Z_T - \widehat{Z}_T\|^2] \leqslant \varepsilon^2$. Within this framework, they prove that the randomized midpoint discretization (Section 5.1) is optimal for simulating the underdamped Langevin dynamics (see Theorem 5.3.11 for the upper bound). It is unclear whether their framework encompasses all possible discretization schemes.

Finally, another line of work considers the complexity of estimating the normalization constant (or partition function). In Rademacher and Vempala (2008), the authors studied the number of membership queries needed to estimate the volume of a convex body $K \subseteq \mathbb{R}^d$ such that $\mathsf{B}(0, 1) \subseteq K \subseteq \mathsf{B}(0, O(d^8))$ 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 Cousins and Vempala (2018); Jia et al. (2021)).

In Ge et al. (2020), the authors considered 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 showed 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 upper bound from Theorem 5.3.11, they showed 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 proved that $\Omega(d^{1-o(1)}/\varepsilon^{2-o(1)})$ queries are necessary. 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.

Exercises

Query Complexity in One Dimension

⊵ **Exercise 9.1** (basics of information theory)

Let X, Y be random variables taking values in finite alphabets $\mathcal{X}$, $\mathcal{Y}$ respectively. Let $p_{X,Y}$ denote the joint distribution, p_X denote the marginal distribution of X, etc. We make the following definitions.

- The entropy of X is $\mathsf{H}(X) := \sum_{x\in\mathcal{X}} p_X(x)\log(1/p_X(x))$.
- The joint entropy of (X,Y) is $\mathsf{H}(X,Y) := \sum_{x\in\mathcal{X},\,y\in\mathcal{Y}} p_{X,Y}(x,y)\log(1/p_{X,Y}(x,y))$.
- The conditional entropy of X given Y is $\mathsf{H}(X \mid Y) := \sum_{y\in\mathcal{Y}} p_Y(y)\,\mathsf{H}(X \mid Y = y)$, where we define $\mathsf{H}(X \mid Y = y) := \sum_{x\in\mathcal{X}} p_{X|Y}(x \mid y)\log(1/p_{X|Y}(x \mid y))$.
- The mutual information of X and Y is $\mathsf{I}(X;Y) := \mathsf{KL}(p_{X,Y} \parallel p_X \otimes p_Y)$.

Prove the following facts.

1. $0 \leqslant \mathsf{H}(X) \leqslant \log|\mathcal{X}|$.
2. Chain rule for entropy: $\mathsf{H}(X,Y) = \mathsf{H}(X) + \mathsf{H}(Y \mid X) = \mathsf{H}(Y) + \mathsf{H}(X \mid Y)$.
3. The mutual information satisfies $\mathsf{I}(X;Y) = \mathsf{H}(X) - \mathsf{H}(X \mid Y) = \mathsf{H}(Y) - \mathsf{H}(Y \mid X)$.
4. Chain rule for mutual information: if $X_1,\dots,X_n$ are random variables taking values in $\mathcal{X}$, then $\mathsf{I}(X_1,\dots,X_n;Y) = \sum_{i=1}^n \mathsf{I}(X_i;Y \mid X_1,\dots,X_{i-1})$, where the conditional mutual information is given by $\mathsf{I}(X_2;Y \mid X_1) = \mathsf{H}(Y \mid X_1) - \mathsf{H}(Y \mid X_1,X_2)$.

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

$$x_{k+1} := x_k - h\,\nabla V(x_k)\,, \qquad k = 0, 1, 2, \dots \tag{10.2.1}$$

we can instead consider the **mirror descent** iteration

$$\nabla\phi(x_{k+1}) := \nabla\phi(x_k) - h\,\nabla V(x_k)\,, \qquad k = 0, 1, 2, \dots \tag{10.2.2}$$

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 Nemirovsky and Yudin (1983) 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 $(Z_t)_{t \geqslant 0}$ is the solution to the stochastic differential equation

$$Z_t^* = \nabla\phi(Z_t)\,, \qquad \mathrm{d}Z_t^* = -\nabla V(Z_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2\phi(Z_t)]^{1/2}\,\mathrm{d}B_t\,. \tag{10.2.3}$$

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 1.1.19), one can write down an SDE for $(Z_t)_{t \geqslant 0}$ 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

$$\int f\,\partial_t\pi_t = \partial_t \int f\,\mathrm{d}\pi_t = \int \mathscr{L}f\,\frac{\pi_t}{\pi}\,\mathrm{d}\pi = -\int \Big\langle \nabla f, [\nabla^2\phi]^{-1}\,\nabla\frac{\pi_t}{\pi}\Big\rangle\,\mathrm{d}\pi \tag{10.2.5}$$

$$= -\int \Big\langle \nabla f, [\nabla^2\phi]^{-1}\,\nabla\log\frac{\pi_t}{\pi}\Big\rangle\,\mathrm{d}\pi_t = \int f\,\operatorname{div}\Big(\pi_t\,[\nabla^2\phi]^{-1}\nabla\log\frac{\pi_t}{\pi}\Big) \tag{10.2.6}$$

from which we deduce the Fokker–Planck equation

$$\partial_t\pi_t = \operatorname{div}\Big(\pi_t\,[\nabla^2\phi]^{-1}\,\nabla\log\frac{\pi_t}{\pi}\Big)\,.$$

From the interpretation as a continuity equation (see Theorem 1.3.17), we deduce that $(\pi_t)_{t \geqslant 0}$ describes the evolution of a particle which travels according to the family of vector fields $t \mapsto -[\nabla^2\phi]^{-1}\,\nabla\log(\pi_t/\pi)$. Recalling that $\nabla\log(\pi_t/\pi)$ is the Wasserstein gradient of $\mathsf{KL}(\cdot \,\|\, \pi)$ 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 2.5.1).

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

$$\mathrm{d}Z_t^* = -Z_t^*\,\mathrm{d}t + \sqrt{2}\,[\nabla^2 V^*(Z_t^*)]^{-1/2}\,\mathrm{d}B_t\,, \tag{10.2.7}$$

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 2.2.9): if $\pi \propto \exp(-V)$ is strictly log-concave, then for all $f : \mathbb{R}^d \to \mathbb{R}$,

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

Below, we will also give an alternative proof of the Bregman transport inequality (Theorem 2.2.12) 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 1.2.23) 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 $x \in \mathcal{X}$,

$$\nabla^2 V(x) \succeq \alpha\,\nabla^2\phi(x)\,.$$

2 V is **β-relatively smooth** (w.r.t. ϕ) if for all $x \in \mathcal{X}$,

$$\nabla^2 V(x) \preceq \beta\,\nabla^2\phi(x)\,.$$

Observe that when $\phi = \frac{\|\cdot\|^2}{2}$, these definitions reduce to the usual definitions of strong convexity and smoothness. Recall from Definition 2.2.11 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

$$D_\phi(x, y) := \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle\,.$$

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*

$$D_V \geqslant \alpha\, D_\phi\,.$$

Similarly, V is β-relatively smooth w.r.t. ϕ if and only if

$$D_V \leqslant \beta\, D_\phi\,.$$

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, Chewi et al. (2020))**.** *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 $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\operatorname{var}_\pi f \leqslant \frac{1}{\alpha}\, \mathbb{E}_\pi\langle \nabla f, [\nabla^2\phi]^{-1}\, \nabla f\rangle = \frac{1}{\alpha}\, \mathscr{E}(f, f)\,.$$

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 1.2.30), 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 $\mathsf{CD}(\alpha, \infty)$ condition $\nabla^2\widetilde{V} + \mathrm{Ric} \succeq \alpha$. 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 2.5. Such a calculation was performed in, e.g., Kolesnikov (2014), which implies that if $(\nabla\phi)_\#\pi$ is log-concave, then the Newton Langevin diffusion satisfies $\mathsf{CD}(\frac{1}{2}, \infty)$. 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 $\mathsf{CD}(\alpha, \infty)$, 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 Ahn and Chewi (2021), we consider the following discretization of (10.2.3).

$$\begin{aligned} \nabla\phi(X^+_{kh}) &\coloneqq \nabla\phi(X_{kh}) - h\,\nabla V(X_{kh})\,, \\ X_{(k+1)h} &\coloneqq \nabla\phi^*(X^*_{(k+1)h})\,, \end{aligned} \tag{MLMC}$$

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 Wibisono (2018) (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):

$$\langle \nabla\phi(\tilde x) - \nabla\phi(x), \tilde x - z\rangle = D_\phi(\tilde x, x) + D_\phi(z, \tilde x) - D_\phi(z, x)\,, \qquad \text{for all } x, \tilde x, z \in \mathcal{X}\,. \tag{10.2.12}$$

Lemma 10.2.13 (Bregman proximal lemma, Chen and Teboulle (1993))**.** *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 Nesterov and Nemirovskii (1994): 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 $v \in \mathbb{R}^d$,

$$\nabla^3\phi(x)[v,v,v] \leqslant 2M_\phi\,\|v\|^3_{\nabla^2\phi(x)} := 2M_\phi\,\langle v, \nabla^2\phi(x)\,v\rangle^{3/2}\,.$$

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 ((Nesterov, 2018, Corollary 5.1.1))**.** *Suppose that ϕ is M_ϕ-self-concordant. Then, for all $x \in \mathcal{X}$ and $u \in \mathbb{R}^d$,*

$$\nabla^3\phi(x)\,u \preceq 2M_\phi\,\|u\|_{\nabla^2\phi(x)}\,\nabla^2\phi(x)\,.$$

Self-concordant functions are well-studied due to their central role in the theory of interior-point methods for optimization, see the monograph Nesterov and Nemirovskii (1994). A key example of a self-concordant mirror map is when the constraint set is a polytope,

---

[1] The proof is surprisingly difficult. The reader can try to prove the result with a worse constant factor.

$$\mathcal{X} = \{x \in \mathbb{R}^d : \langle a_i, x\rangle < b_i \text{ for all } i \in [N]\}\,,$$

in which case $\phi(x) = \log(1/\sum_{i=1}^N (b_i - \langle a_i, x\rangle))$ 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 2.2.11). To study these costs, we first state an analogue of Brenier's theorem (Theorem 1.3.8).

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 1.10). 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 (Villani, 2009b, 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 $\mu_t := \operatorname{law}(X_t)$, we obtain the W_2 geodesic.

Now consider the above theorem. If we let $W := \nabla\phi(Y) = \nabla\phi(X) - \nabla h(X)$, then we have the coupling $(X_0, X_1) := (\nabla(\phi - h)^*(W), \nabla\phi^*(W))$ of $\mu_0 := \mu$ and $\mu_1 := \nu$. Then, we can interpolate by setting $X_t := (1 - t)\, X_0 + t\, X_1$ and $\mu_t := \operatorname{law}(X_t)$, 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

$$\mu_t = \operatorname{law}(X_t)\,, \qquad X_t \coloneqq (1-t)\,X_0 + t\,X_1\,,$$

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 $\mathcal{H}(\mu) \coloneqq \int \mu \log \mu$. 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

$$\mathcal{H}(\nu) \geqslant \mathcal{H}(\mu) + \mathbb{E}\langle \nabla \log \mu(X), Y - X\rangle\,. \tag{10.2.19}$$

As an aside, we remark that this observation leads to another proof of the Bregman transport inequality.

Proof of the Bregman transport inequality (Theorem 2.2.12) 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

$$\mathcal{H}(\mu) \geqslant \mathcal{H}(\pi) + \mathbb{E}\langle \nabla \log \pi(Z), X - Z\rangle\,.$$

Hence,

$$\begin{aligned}\mathsf{KL}(\mu \,\|\, \pi) &\geqslant \underbrace{\mathbb{E}\, V(Z) + \mathcal{H}(\pi)}_{=\mathsf{KL}(\pi\|\pi)=0} + \mathbb{E}\langle \underbrace{\nabla V(Z) + \nabla \log \pi(Z)}_{=0}, X - Z\rangle + \mathbb{E}\, D_V(X, Z)\\ &= \mathcal{D}_V(\mu \,\|\, \pi)\,,\end{aligned}$$

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 (Ahn and Chewi (2021))**.** *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.,* $\|\nabla V(x)\|_{[\nabla^2\phi(x)]^{-1}} \leqslant L$ *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* $\sqrt{\mathsf{KL}(\bar\mu_{Nh} \,\|\, \pi)} \leqslant \varepsilon$ *after*

$$N = O\Big(\frac{\beta' d\, \mathcal{D}_\phi(\pi,\mu_0)}{\varepsilon^4}\Big) \qquad \textit{iterations}\,.$$

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* $h \asymp \frac{\alpha\varepsilon^2}{\beta' d}$ *we obtain* $\sqrt{\alpha}\,\mathcal{D}_\phi(\pi,\mu_{Nh}) \leqslant \varepsilon$ *and* $\sqrt{\mathsf{KL}(\bar\mu_{Nh,2Nh} \,\|\, \pi)} \leqslant \varepsilon$ *after*

$$N = O\Big(\frac{\kappa d}{\varepsilon^2}\log\frac{\alpha\,\mathcal{D}_\phi(\pi,\mu_0)}{\varepsilon^2}\Big) \qquad \textit{iterations}\,,$$

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*

$$h\,\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) \leqslant (1-\alpha h)\,\mathcal{D}_\phi(\pi,\mu_{kh}) - \mathcal{D}_\phi(\pi,\mu_{(k+1)h}) + \beta' d h^2\,.$$

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,

$$\begin{aligned}
\mathcal{E}(\mu_{kh}^+) - \mathcal{E}(\pi) &= \mathbb{E}[V(X_{kh}^+) - V(X_{kh}) + V(X_{kh}) - V(Z)] \\
&\leqslant \mathbb{E}\big[\langle \nabla V(X_{kh}), X_{kh}^+ - X_{kh}\rangle + \beta\, D_\phi(X_{kh}^+, X_{kh}) \\
&\qquad\qquad + \langle \nabla V(X_{kh}), X_{kh} - Z\rangle - \alpha\, D_\phi(Z, X_{kh})\big] \\
&= \mathbb{E}\big[\langle \nabla V(X_{kh}), X_{kh}^+ - Z\rangle + \beta\, D_\phi(X_{kh}^+, X_{kh}) - \alpha\, D_\phi(Z, X_{kh})\big]\,. \qquad (10.2.22)
\end{aligned}$$

Next, by the Bregman proximal lemma (Lemma 10.2.13),

$$\langle \nabla V(X_{kh}), X_{kh}^+ - Z\rangle = \frac{1}{h}\,\{D_\phi(Z, X_{kh}) - D_\phi(Z, X_{kh}^+) - D_\phi(X_{kh}^+, X_{kh})\}\,.$$

Substituting this into (10.2.22) and using $h \leqslant \frac{1}{\beta}$, it yields

$$\mathcal{E}(\mu_{kh}^+) - \mathcal{E}(\pi) \leqslant \frac{1}{h}\,\{(1-\alpha h)\,\mathcal{D}_\phi(\mu_{kh},\pi) - \mathcal{D}_\phi(\mu_{kh}^+,\pi)\}\,. \qquad (10.2.23)$$

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.

$$\begin{aligned}
&\mathbb{E}[f(X_{(k+1)h}^*) - f(X_{kh}^*)]\\
&\quad = \mathbb{E}\int_{kh}^{(k+1)h} \big\langle \nabla^2 V(X_t)\,[\nabla^2\phi(X_t)]^{-2}, \nabla^2\phi(X_t)\big\rangle\,\mathrm{d}t\\
&\qquad + \mathbb{E}\int_{kh}^{(k+1)h} \big\langle \nabla V(X_t)^\mathsf{T}\,[\nabla^2\phi(X_t)]^{-1}\,[\nabla^3\phi(X_t)]\,[\nabla^2\phi(X_t)]^{-2}, \nabla^2\phi(X_t)\big\rangle\,\mathrm{d}t\\
&\quad = \mathbb{E}\int_{kh}^{(k+1)h} \big\langle \nabla^2 V(X_t), [\nabla^2\phi(X_t)]^{-1}\big\rangle\,\mathrm{d}t \qquad (10.2.24)\\
&\qquad + \mathbb{E}\int_{kh}^{(k+1)h} \operatorname{tr}\big(\nabla V(X_t)^\mathsf{T}\,[\nabla^2\phi(X_t)]^{-1}\,[\nabla^3\phi(X_t)]\,[\nabla^2\phi(X_t)]^{-1}\big)\,\mathrm{d}t\,. \qquad (10.2.25)
\end{aligned}$$

By relative smoothness, since $\nabla^2 V \preceq \beta\,\nabla^2\phi$,

$$(10.2.24) \leqslant \beta dh\,.$$

For (10.2.25), we use Lemma 10.2.15, which implies

$$\begin{aligned}
(10.2.25) &\leqslant 2M_\phi \int_{kh}^{(k+1)h} \mathbb{E}\big[\big\|[\nabla^2\phi(X_t)]^{-1}\,\nabla V(X_t)\big\|_{\nabla^2\phi(X_t)}\,\operatorname{tr}\big([\nabla^2\phi(X_t)]\,[\nabla^2\phi(X_t)]^{-1}\big)\big]\,\mathrm{d}t\\
&\leqslant 2M_\phi d \int_{kh}^{(k+1)h} \mathbb{E}\big[\|\nabla V(X_t)\|_{[\nabla^2\phi(X_t)]^{-1}}\big]\,\mathrm{d}t \leqslant 2LM_\phi dh\,.
\end{aligned}$$

Hence, we have proven

$$\mathcal{E}(\mu_{(k+1)h}) - \mathcal{E}(\mu_{kh}^+) \leqslant \beta' dh\,. \qquad (10.2.26)$$

3. The flow step dissipates the entropy. Let $\mu_t := \operatorname{law}(X_t) = \operatorname{law}(\nabla\phi^*(X_t^*))$. The mirror diffusion (10.2.11) evolves according to the vector field $-[\nabla^2\phi]^{-1}\,\nabla\log\mu_t$. Also, note that $\nabla_y D_\phi(x,y) = -\nabla^2\phi(x)\,(y-x)$. Using these, one can show that

$$\partial_t \mathcal{D}_\phi(\pi,\mu_t) \leqslant \mathbb{E}\langle[\nabla^2\phi(X_t)]^{-1}\,\nabla\log\mu(X_t), [\nabla^2\phi(X_t)]\,(Z-X_t)\rangle = \mathbb{E}\langle\nabla\log\mu(X_t), (Z-X_t)\rangle\,,$$

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),

$$\mathcal{H}(\pi) - \mathcal{H}(\mu_t) \geqslant \mathbb{E}\langle\nabla\log\mu(X_t), (Z-X_t)\rangle\,.$$

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

$$\mathcal{D}_\phi(\pi, \mu_{(k+1)h}) - \mathcal{D}_\phi(\pi, \mu_{kh}^+) \leqslant h\,\{\mathcal{H}(\pi) - \mathcal{H}(\mu_{(k+1)h})\}\,. \tag{10.2.27}$$

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 Chewi (2021); Lee and Yue (2021)). 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 $\phi(x) := \sum_{i=1}^d x_i \log x_i$ 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 (Ahn and Chewi (2021))**.** *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*

$$|||\nabla V(x)|||_* \leqslant L \qquad \text{for all } x \in \mathcal{X}\,.$$

- *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 $\bar\mu_{Nh} := N^{-1}\sum_{k=1}^N \mu_{kh}$ it holds that $\sqrt{\mathsf{KL}(\bar\mu_{Nh} \,\|\, \pi)} \leqslant \varepsilon$ after

$$N = O\Big(\frac{L^2\,\mathcal{D}_\phi(\pi, \mu_0^+)}{\varepsilon^4}\Big) \qquad \textit{iterations}\,.$$

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 Nesterov and Nemirovskii (1994); Bubeck (2015); Nesterov (2018). Relative convexity and relative smoothness were introduced in Bauschke et al. (2017); Lu et al. (2018).

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 Hsieh et al. (2018). The mirror Langevin diffusion was introduced in an earlier draft of Hsieh et al. (2018), as well as in Zhang et al. (2020). In Zhang et al. (2020), 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, Chewi et al. (2020) studied the mirror Langevin diffusion in continuous time, and Ahn and Chewi (2021) introduced and studied the MLMC discretization, which *does* lead to vanishing bias (as $h \searrow 0$).

Since then, there have been further works studying the non-vanishing bias issue: Jiang (2021) 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 et al. (2022a) showed that the Euler–Maruyama discretization has vanishing bias under stronger assumptions; and Gatmiry and Vempala (2022) 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 Ahn and Chewi (2021) because it is the cleanest, and because it relies on assumptions which are well-motivated from convex optimization.

More generally, the mirror Langevin diffusion is an example of a Riemannian diffusion. See Hsu (2002) for general theory, and Girolami and Calderhead (2011) for applications to Bayesian problems.

Exercises

Mirror Langevin

⊵ **Exercise 10.1** (the mirror Langevin diffusion in the primal space)
Use Itô's formula (Theorem 1.1.19) 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 $(Z_t^*)_{t\geqslant 0}$ (we write $\mathscr{L}_{\mathrm{dual}}$ instead of $\mathscr{L}^*$ to avoid confusion with the adjoint of $\mathscr{L}$). By computing $\partial_t\,\mathbb{E}\,g(Z_t^*)$, show that

$$\mathscr{L}_{\mathrm{dual}}\,g = \mathscr{L}(g\circ\nabla\phi)\circ(\nabla\phi)^{-1}\,.$$

Then, via a similar calculation to (10.2.5) and (10.2.6), show that the Dirichlet energy for $(Z_t^*)_{t\geqslant 0}$ 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 $(Z_t^*)_{t\geqslant 0}$.

3 Show that the mirror Poincaré inequality in Corollary 10.2.10 implies the following Poincaré inequality in the dual space: for all $f : \mathbb{R}^d \to \mathbb{R}$,

$$\operatorname{var}_{\pi^*} f \leqslant \frac{1}{\alpha}\, \mathbb{E}_{\pi^*}\langle \nabla f, [\nabla^2 \phi^*]^{-1}\, \nabla f\rangle = \frac{1}{\alpha}\, \mathscr{E}_{\text{dual}}(f, f)\,.$$

⊵ **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 Balasubramanian et al. (2022).

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 $\|\nabla V(x)\| \leqslant \varepsilon$.

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 $\mathsf{KL}(\cdot \,\|\, \pi)$ w.r.t. the Wasserstein geometry (Section 1.4). Moreover, the gradient of the KL divergence at μ is $\nabla \log(\mu/\pi)$, and the squared norm of the gradient is the Fisher information $\mathsf{FI}(\mu \,\|\, \pi) = \mathbb{E}_\mu[\|\nabla \log(\mu/\pi)\|^2]$. Hence, a reasonable definition of finding an approximate first-order stationary point in sampling is to output a sample from μ satisfying $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \leqslant \varepsilon$. 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: $\mathsf{FI}(\mu \,\|\, \pi) = 0$ implies $\mu = \pi$. This is a specific feature of the sampling problem.[1] The intuition behind the proof is straightforward: if $\mathsf{FI}(\mu \,\|\, \pi) = 0$, then $\nabla \log(\mu/\pi) = 0$ (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 $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \leqslant \varepsilon$ 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 $\mathsf{FI}(\mu \,\|\, \pi)$ implies μ is close to π, but how small must $\mathsf{FI}(\mu \,\|\, \pi)$ 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).

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

$$\mathsf{FI}(\mu \,\|\, \pi) \lesssim m^2 \exp\Bigl(-\frac{m^2}{2}\Bigr) \to 0 \qquad \text{as } m \to \infty\,.$$

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 $\mathsf{FI}(\mu \,\|\, \pi)$ 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.

---

[1] In fact, the KL divergence $\mathsf{KL}(\cdot \,\|\, \pi)$ is always strictly convex with respect to taking convex combinations of measures, and hence always has a unique global minimum.

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 $\nabla \log(\mu/\pi) = 0$ (μ-a.e.); from the expression $\mathsf{FI}(\mu \,\|\, \pi) = \mathbb{E}_\mu[\|\nabla \log(\mu/\pi)\|^2]$, it suggests that $\mathsf{FI}(\mu \,\|\, \pi) = 0$, 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 $\mathsf{FI}(\mu \,\|\, \pi) = \mathbb{E}_\mu[\|\nabla \log(\mu/\pi)\|^2]$. 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: $\mathsf{FI}(\mu \,\|\, \pi) = \infty$.

In our theorem below, the Fisher information bound should be interpreted as follows: $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \leqslant \varepsilon$ means that μ has enough regularity w.r.t. π *and* $\mathbb{E}_\mu[\|\nabla \log(\mu/\pi)\|^2] \leqslant \varepsilon^2$.

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 (Balasubramanian et al. (2022)). *Let $(\mu_t)_{t\geqslant 0}$ 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 $0 < h \leqslant \frac{1}{4\beta}$, for all $N \in \mathbb{N}$,*

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leqslant \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{Nh} + 6\beta^2 dh\,.$$

In particular, if $\mathsf{KL}(\mu_0 \,\|\, \pi) \leqslant K_0$ and we choose $h = \sqrt{K_0}/(2\beta\sqrt{dN})$, then provided that $N \geqslant 9K_0/d$,

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leqslant \frac{8\beta\sqrt{dK_0}}{\sqrt{N}}\,.$$

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* $\mathsf{FI}(\cdot \,\|\, \pi)$ *is convex.*

Proof Let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)$ be such that $\mathsf{FI}(\mu_0 \,\|\, \pi) \vee \mathsf{FI}(\mu_1 \,\|\, \pi) < \infty$. For $t \in (0,1)$, let $\mu_t := (1-t)\,\mu_0 + t\,\mu_1$, and write $f_t := \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} = (1-t)\,f_0 + t\,f_1$. Then,

$$\begin{aligned}\mathsf{FI}(\mu_t \,\|\, \pi) = \int \frac{\|\nabla f_t\|^2}{f_t}\,\mathrm{d}\pi &\leqslant (1-t)\int \frac{\|\nabla f_0\|^2}{f_0}\,\mathrm{d}\pi + t\int \frac{\|\nabla f_1\|^2}{f_1}\,\mathrm{d}\pi \\ &= (1-t)\,\mathsf{FI}(\mu_0 \,\|\, \pi) + t\,\mathsf{FI}(\mu_1 \,\|\, \pi)\end{aligned}$$

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 $\bar\mu_{Nh} := \frac{1}{Nh}\int_0^{Nh} \mu_t\,\mathrm{d}t$, we have

$$\mathsf{FI}(\bar\mu_{Nh} \,\|\, \pi) \leqslant \frac{1}{Nh}\int \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \tag{11.2.3}$$

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 $kh \leqslant t$, 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 $\mathsf{KL}(\mu_0 \,\|\, \pi) = O(d)$, we conclude that it is possible to algorithmically obtain a sample from a measure μ with $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \leqslant \varepsilon$ 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

$$\partial_t \,\mathsf{KL}(\mu_t \,\|\, \pi) \leqslant -\frac{1}{2}\,\mathsf{FI}(\mu_t \,\|\, \pi) + 6\beta^2 d\,(t-kh)\,.$$

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 $t \in [kh, (k+1)h]$,

$$\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) - \mathsf{KL}(\mu_{kh} \,\|\, \pi) \leqslant -\frac{1}{2}\int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 3\beta^2 d h^2\,.$$

Summing over $k = 0, 1, \dots, N-1$ and dividing by Nh,

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leqslant \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{Nh} + 6\beta^2 dh\,.$$

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 $\mathsf{FI}(\mu \,\|\, \pi) = 0$ 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 $(\mu_t)_{t\geqslant 0}$ 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 $\mathsf{KL}(\mu_0 \,\|\, \pi) < \infty$, 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*

$$\sum_{k=1}^\infty h_k = \infty\,, \qquad \textit{and} \qquad \sum_{k=1}^\infty h_k^2 < \infty\,.$$

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 $t \in [\tau_n, \tau_{n+1}]$

$$\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) \leqslant -\frac{1}{2}\,\mathsf{FI}(\mu_t \,\|\, \pi) + 6\beta^2 d\,(t - \tau_n)\,. \tag{11.3.3}$$

By integrating this inequality and summing,

$$\mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) \leqslant \mathsf{KL}(\mu_0 \,\|\, \pi) - \frac{1}{2}\int_0^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 3\beta^2 d \sum_{k=1}^n h_k^2\,. \tag{11.3.4}$$

Rearranging and using the convexity of the Fisher information, it yields

$$\mathsf{FI}(\bar\mu_{\tau_n} \,\|\, \pi) \leqslant \frac{1}{\tau_n}\int_0^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leqslant \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{\tau_n} + \frac{6\beta^2 d}{\tau_n}\sum_{k=1}^\infty h_k^2\,. \tag{11.3.5}$$

On the other hand, if $t \in [\tau_n, \tau_{n+1}]$, then integrating (11.3.3) and combining with (11.3.4) yields

$$\mathsf{KL}(\mu_t \,\|\, \pi) \leqslant \mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) + 3\beta^2 d\,(t - \tau_n)^2 \leqslant \mathsf{KL}(\mu_0 \,\|\, \pi) + 6\beta^2 d \sum_{k=1}^\infty h_k^2 < \infty\,.$$

Therefore, $\{\mathsf{KL}(\mu_t \,\|\, \pi) \mid t \geqslant 0\}$ 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 $\mu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\|\mu - \pi\|_{\mathrm{TV}}^2 \leqslant \frac{C_{\mathsf{PI}}}{4}\,\mathsf{FI}(\mu \,\|\, \pi)\,.$$

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,

$$\begin{aligned}\int \sqrt{f}\,\mathrm{d}\pi &= \int \sqrt{(f\wedge 1)\,(f\vee 1)}\,\mathrm{d}\pi \leqslant \sqrt{\int (f\wedge 1)\,\mathrm{d}\pi \int (f \vee 1)\,\mathrm{d}\pi}\\ &= \sqrt{(1-\|\mu-\pi\|_{\mathrm{TV}})\,(1+\|\mu-\pi\|_{\mathrm{TV}})} = \sqrt{1 - \|\mu-\pi\|_{\mathrm{TV}}^2}\,.\end{aligned}$$

Therefore,

$$\|\mu-\pi\|_{\mathrm{TV}}^2 \leqslant 1 - \Bigl(\int \sqrt{f}\,\mathrm{d}\pi\Bigr)^2 = \operatorname{var}_\pi \sqrt{f}\,.$$

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,

$$\|\mu-\pi\|_{\mathrm{TV}}^2 \leqslant C_{\mathsf{PI}}\,\mathbb{E}_\pi[\|\nabla\sqrt{f}\|^2] = \frac{C_{\mathsf{PI}}}{4}\,\mathsf{FI}(\mu \,\|\, \pi)\,. \qquad \square$$

Corollary 11.3.7. *Let* $(\mu_t)_{t\geqslant 0}$ *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* $\mathsf{KL}(\mu_0 \parallel \pi) \leqslant K_0$ *and we choose step size* $h = \sqrt{K_0}/(2\beta\sqrt{dN})$, *then*

$$\|\bar{\mu}_{Nh} - \pi\|_{\mathrm{TV}}^2 := \left\| \frac{1}{Nh} \int_0^t \mu_t \, \mathrm{d}t - \pi \right\|_{\mathrm{TV}}^2 \leqslant \frac{2C_{\mathsf{PI}}\beta\sqrt{dK_0}}{\sqrt{N}} .$$

Bibliographical Notes

Exercises

⊵ **Exercise 11.1** (Fisher information for mixtures of Gaussians)
Prove Proposition 11.1.1.

References

Ahn, Kwangjun, and Chewi, Sinho. 2021. Efficient constrained sampling via the mirror-Langevin algorithm. Pages 28405–28418 of: Ranzato, M., Beygelzimer, A., Nguyen, K., Liang, P. S., Vaughan, J. W., and Dauphin, Y. (eds), *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc.

Albritton, Dallas, Armstrong, Scott, Mourrat, Jean-Christophe, and Novack, Matthew. 2024. Variational methods for the kinetic Fokker–Planck equation. *Anal. PDE*, **17**(6), 1953–2010.

Alonso-Gutiérrez, David, and Bastero, Jesús. 2015. *Approaching the Kannan-Lovász-Simonovits and variance conjectures*. Lecture Notes in Mathematics, vol. 2131. Springer, Cham.

Alquier, Pierre. 2024. User-friendly introduction to PAC–Bayes bounds. *Foundations and Trends® in Machine Learning*, **17**(2), 174–303.

Altschuler, Jason M., and Chewi, Sinho. 2023. Shifted composition II: shift Harnack inequalities and curvature upper bounds. *arXiv preprint 2401.00071*.

Altschuler, Jason M., and Chewi, Sinho. 2024a. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 3.

Altschuler, Jason M., and Chewi, Sinho. 2024b. Shifted composition I: Harnack and reverse transport inequalities. *IEEE Transactions on Information Theory*, 1–1.

Altschuler, Jason M., and Chewi, Sinho. 2024c. Shifted composition III: local error framework for KL divergence. *arXiv preprint 2412.17997*.

Altschuler, Jason M., and Talwar, Kunal. 2023. Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling. Pages 2509–2510 of: Neu, Gergely, and Rosasco, Lorenzo (eds), *Proceedings of Thirty Sixth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 195. PMLR.

Altschuler, Jason M., Chewi, Sinho, and Zhang, Matthew S. 2025. Shifted composition IV: underdamped Langevin and numerical discretizations with partial acceleration. *arXiv preprint 2506.23062*.

Ambrosio, Luigi, Gigli, Nicola, and Savaré, Giuseppe. 2008. *Gradient flows in metric spaces and in the space of probability measures*. Second edn. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel.

Andrieu, Christophe, Lee, Anthony, Power, Sam, and Wang, Andi Q. 2024. Explicit convergence bounds for Metropolis Markov chains: isoperimetry, spectral gaps and profiles. *Ann. Appl. Probab.*

Arnaudon, Marc, Thalmaier, Anton, and Wang, Feng-Yu. 2006. Harnack inequality and heat kernel estimates on manifolds with curvature unbounded below. *Bull. Sci. Math.*, **130**(3), 223–233.

Bakry, Dominique, Gentil, Ivan, and Ledoux, Michel. 2014. *Analysis and geometry of Markov diffusion operators*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 348. Springer, Cham.

Balasubramanian, Krishna, Chewi, Sinho, Erdogdu, Murat A., Salim, Adil, and Zhang, Matthew. 2022. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. Pages 2896–2923 of: Loh, Po-Ling, and Raginsky, Maxim (eds), *Proceedings of Thirty Fifth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 178. PMLR.

Bardet, Jean-Baptiste, Gozlan, Nathaël, Malrieu, Florent, and Zitt, Pierre-André. 2018. Functional inequalities for Gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence. *Bernoulli*, **24**(1), 333–353.

Barthe, Franck, and Cordero-Erausquin, Dario. 2013. Invariances in variance estimates. *Proc. Lond. Math. Soc. (3)*, **106**(1), 33–64.

Baudoin, Fabrice. 2017. Bakry–Émery meet Villani. *Journal of Functional Analysis*, **273**(7), 2275–2291.

Bauerschmidt, Roland, Bodineau, Thierry, and Dagallier, Benoit. 2024. Stochastic dynamics and the Polchinski equation: an introduction. *Probab. Surv.*, **21**, 200–290.

Bauschke, Heinz H., Bolte, Jérôme, and Teboulle, Marc. 2017. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, **42**(2), 330–348.

Benamou, Jean-David, and Brenier, Yann. 1999. A numerical method for the optimal time-continuous mass transport problem and related problems. Pages 1–11 of: *Monge Ampère equation: applications to geometry and optimization (Deerfield Beach, FL, 1997)*. Contemp. Math., vol. 226. Amer. Math. Soc., Providence, RI.

Besag, Julian, Green, Peter, Higdon, David, and Mengersen, Kerrie. 1995. Bayesian computation and stochastic systems. *Statistical Science*, **10**(1), 3–66.

Beskos, Alexandros, Pillai, Natesh, Roberts, Gareth, Sanz-Serna, Jesus-Maria, and Stuart, Andrew. 2013. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, **19**(5A), 1501–1534.

Billera, Louis J., and Diaconis, Persi. 2001. A geometric interpretation of the Metropolis–Hastings algorithm. *Statist. Sci.*, **16**(4), 335–339.

Blanchet, Adrien, and Bolte, Jérôme. 2018. A family of functional inequalities: Lojasiewicz inequalities and displacement convex functions. *J. Funct. Anal.*, **275**(7), 1650–1673.

Bobkov, Sergey G. 1997. An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in Gauss space. *Ann. Probab.*, **25**(1), 206–214.

Bobkov, Sergey G., and Ledoux, Michel. 2000. From Brunn–Minkowski to Brascamp–Lieb and to logarithmic Sobolev inequalities. *Geom. Funct. Anal.*, **10**(5), 1028–1052.

Bobkov, Sergey G., Gentil, Ivan, and Ledoux, Michel. 2001. Hypercontractivity of Hamilton–Jacobi equations. *J. Math. Pures Appl. (9)*, **80**(7), 669–696.

Bobkov, Serguei G., and Houdré, Christian. 1997. Some connections between isoperimetric and Sobolev-type inequalities. *Mem. Amer. Math. Soc.*, **129**(616), viii+111.

Bolley, François, Gentil, Ivan, and Guillin, Arnaud. 2018. Dimensional improvements of the logarithmic Sobolev, Talagrand and Brascamp–Lieb inequalities. *Ann. Probab.*, **46**(1), 261–301.

Borell, Christer. 1975. The Brunn–Minkowski inequality in Gauss space. *Invent. Math.*, **30**(2), 207–216.

Borell, Christer. 1985. Geometric bounds on the Ornstein–Uhlenbeck velocity process. *Z. Wahrsch. Verw. Gebiete*, **70**(1), 1–13.

Borell, Christer. 2000. Diffusion equations and geometric inequalities. *Potential Anal.*, **12**(1), 49–71.

Bou-Rabee, Nawaf, Eberle, Andreas, and Zimmer, Raphael. 2020. Coupling and convergence for Hamiltonian Monte Carlo. *Ann. Appl. Probab.*, **30**(3), 1209–1250.

Boucheron, Stéphane, Lugosi, Gábor, and Massart, Pascal. 2013. *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Brenier, Yann. 1991. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, **44**(4), 375–417.

Brigati, Giovanni. 2023. Time averages for kinetic Fokker–Planck equations. *Kinet. Relat. Models*, **16**(4), 524–539.

Brigati, Giovanni, and Pedrotti, Francesco. 2024. Heat flow, log-concavity, and Lipschitz transport maps. *arXiv preprint 2404.15205*.

Brigati, Giovanni, Stoltz, Gabriel, Wang, Andi Q., and Wang, Lihan. 2025. Explicit convergence rates of underdamped Langevin dynamics under weighted and weak Poincaré–Lions inequalities. *arXiv preprint 2407.16033*.

Bubeck, Sébastien. 2015. Convex optimization: algorithms and complexity. *Foundations and Trends® in Machine Learning*, **8**(3-4), 231–357.

Bunne, Charlotte, Hsieh, Ya-Ping, Cuturi, Marco, and Krause, Andreas. 2023. The Schrödinger bridge between Gaussian measures has a closed form. Pages 5802–5833 of: Ruiz, Francisco, Dy, Jennifer, and van de Meent, Jan-Willem (eds), *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research, vol. 206. PMLR.

Burkholder, Donald L. 1973. Distribution function inequalities for martingales. *Ann. Probability*, **1**, 19–42.

Caffarelli, Luis A. 2000. Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.*, **214**(3), 547–563.

Cao, Yu, Lu, Jianfeng, and Lu, Yulong. 2019. Exponential decay of Rényi divergence under Fokker–Planck equations. *J. Stat. Phys.*, **176**(5), 1172–1184.

Cao, Yu, Lu, Jianfeng, and Wang, Lihan. 2021. Complexity of randomized algorithms for underdamped Langevin dynamics. *Commun. Math. Sci.*, **19**(7), 1827–1853.

Cao, Yu, Lu, Jianfeng, and Wang, Lihan. 2023. On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. *Arch. Ration. Mech. Anal.*, **247**(5), 90.

Carrillo, José A., and Vaes, Urbain. 2021. Wasserstein stability estimates for covariance-preconditioned Fokker–Planck equations. *Nonlinearity*, **34**(4), 2275–2295.

Cattiaux, Patrick, Conforti, Giovanni, Gentil, Ivan, and Léonard, Christian. 2023. Time reversal of diffusion processes under a finite entropy condition. *Ann. Inst. Henri Poincaré Probab. Stat.*, **59**(4), 1844–1881.

Chafaï, Djalil. 2004. Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto Univ.*, **44**(2), 325–363.

Chafaï, Djalil, and Malrieu, Florent. 2010. On fine properties of mixtures with respect to concentration of measure and Sobolev type inequalities. *Ann. Inst. Henri Poincaré Probab. Stat.*, **46**(1), 72–96.

Chatterji, Niladri S., Bartlett, Peter L., and Long, Philip M. 2022. Oracle lower bounds for stochastic gradient sampling algorithms. *Bernoulli*, **28**(2), 1074–1092.

Chen, August Y., and Sridharan, Karthik. 2025. Optimization, isoperimetric inequalities, and sampling via Lyapunov potentials. *arXiv preprint 2410.02979*.

Chen, Fan, Chewi, Sinho, Daskalakis, Constantinos, and Rakhlin, Alexander. 2026. High-accuracy sampling for diffusion models and log-concave distributions. *arXiv preprint 2602.01338*.

Chen, Gong, and Teboulle, Marc. 1993. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, **3**(3), 538–543.

Chen, Hong-Bin, Chewi, Sinho, and Niles-Weed, Jonathan. 2021a. Dimension-free log-Sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, **281**(11), 109236.

Chen, Yongxin, Georgiou, Tryphon T., and Pavon, Michele. 2021b. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *SIAM Rev.*, **63**(2), 249–313.

Chen, Yongxin, Chewi, Sinho, Salim, Adil, and Wibisono, Andre. 2022. Improved analysis for a proximal algorithm for sampling. Pages 2984–3014 of: Loh, Po-Ling, and Raginsky, Maxim (eds), *Proceedings of Thirty Fifth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 178. PMLR.

Chen, Yuansi. 2021. An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture. *Geom. Funct. Anal.*, **31**(1), 34–61.

Chen, Yuansi, and Eldan, Ronen. 2022. Localization schemes: a framework for proving mixing bounds for Markov chains (extended abstract). Pages 110–122 of: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*.

Chen, Yuansi, and Gatmiry, Khashayar. 2023a. A simple proof of the mixing of Metropolis-adjusted Langevin algorithm under smoothness and isoperimetry. *arXiv preprint 2304.04095*.

Chen, Yuansi, and Gatmiry, Khashayar. 2023b. When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm? *arXiv preprint 2304.04724*.

Chen, Yuansi, Dwivedi, Raaz, Wainwright, Martin J., and Yu, Bin. 2020. Fast mixing of Metropolized Hamiltonian Monte Carlo: benefits of multi-step gradients. *J. Mach. Learn. Res.*, **21**, Paper No. 92, 71.

Chen, Zongchen, and Vempala, Santosh S. 2019. Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. Pages Art. No. 64, 12 of: *Approximation, randomization, and combinatorial optimization. Algorithms and techniques*. LIPIcs. Leibniz Int. Proc. Inform., vol. 145. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern.

Cheng, Xiang, Chatterji, Niladri S., Bartlett, Peter L., and Jordan, Michael I. 2018. Underdamped Langevin MCMC: a non-asymptotic analysis. Pages 300–323 of: Bubeck, Sébastien, Perchet, Vianney, and Rigollet, Philippe (eds), *Proceedings of the 31st Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 75. PMLR.

Chewi, Sinho. 2021. The entropic barrier is n-self-concordant. *arXiv e-prints*.

Chewi, Sinho. 2023. *An optimization perspective on log-concave sampling and beyond*. Ph.D. thesis, MIT.

Chewi, Sinho. 2024. *Log-concave sampling*. Forthcoming.

Chewi, Sinho. 2025. *Lectures on optimization*. Available at `https://chewisinho.github.io/opt_notes_final.pdf`.

Chewi, Sinho, and Pooladian, Aram-Alexandre. 2023. An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *Reports. Mathematical*, **361**, 1471–1482.

Chewi, Sinho, and Stromme, Austin J. 2024. The ballistic limit of the log-Sobolev constant equals the Polyak–Lojasiewicz constant. *arXiv preprint 2411.11415*.

Chewi, Sinho, Le Gouic, Thibaut, Lu, Chen, Maunu, Tyler, Rigollet, Philippe, and Stromme, Austin J. 2020. Exponential ergodicity of mirror-Langevin diffusions. Pages 19573–19585 of: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds), *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc.

Chewi, Sinho, Lu, Chen, Ahn, Kwangjun, Cheng, Xiang, Gouic, Thibaut Le, and Rigollet, Philippe. 2021. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. Pages 1260–1300 of: Belkin, Mikhail, and Kpotufe, Samory (eds), *Proceedings of Thirty Fourth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 134. PMLR.

Chewi, Sinho, Gerber, Patrik R., Lu, Chen, Gouic, Thibaut Le, and Rigollet, Philippe. 2022. The query complexity of sampling from strongly log-concave distributions in one dimension. Pages 2041–2059 of: Loh, Po-Ling, and Raginsky, Maxim (eds), *Proceedings of Thirty Fifth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 178. PMLR.

Chewi, Sinho, De Dios Pont, Jaume, Li, Jerry, Lu, Chen, and Narayanan, Shyam. 2023. Query lower bounds for log-concave sampling. Pages 2139–2148 of: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*.

Chewi, Sinho, Erdogdu, Murat A., Li, Mufan (Bill), Shen, Ruoqi, and Zhang, Matthew S. 2024a. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *Found. Comput. Math.*, **24**(4).

Chewi, Sinho, Nitanda, Atsushi, and Zhang, Matthew S. 2024b. Uniform-in-N log-Sobolev inequality for the mean-field Langevin dynamics with convex energy. *arXiv preprint 2409.10440*.

Chewi, Sinho, Niles-Weed, Jonathan, and Rigollet, Philippe. 2025. *Statistical optimal transport*. Lecture Notes in Mathematics, vol. 2364. Springer, Cham. École d'Été de Probabilités de Saint-Flour XLIX – 2019.

Conforti, Giovanni, Durmus, Alain, and Gentiloni Silveri, Marta. 2025. KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM J. Math. Data Sci.*, **7**(1), 86–109.

Cordero-Erausquin, Dario. 2002. Some applications of mass transport to Gaussian-type inequalities. *Arch. Ration. Mech. Anal.*, **161**(3), 257–269.

Cordero-Erausquin, Dario. 2017. Transport inequalities for log-concave measures, quantitative forms, and applications. *Canad. J. Math.*, **69**(3), 481–501.

Cordero-Erausquin, Dario, McCann, Robert J., and Schmuckenschläger, Michael. 2001. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Invent. Math.*, **146**(2), 219–257.

Cordero-Erausquin, Dario, Fradelizi, Matthieu, and Maurey, Bernard. 2004. The (B) conjecture for the Gaussian measure of dilates of symmetric convex sets and related problems. *J. Funct. Anal.*, **214**(2), 410–427.

Cousins, Ben, and Vempala, Santosh. 2018. Gaussian cooling and $O^*(n^3)$ algorithms for volume and Gaussian volume. *SIAM J. Comput.*, **47**(3), 1237–1273.

Cover, Thomas M., and Thomas, Joy A. 2006. *Elements of information theory*. Second edn. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Cuesta, Juan Antonio, and Matrán, Carlos. 1989. Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab.*, **17**(3), 1264–1276.

Cuturi, Marco. 2013. Sinkhorn distances: lightspeed computation of optimal transport. In: Burges, C.J., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K.Q. (eds), *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc.

Dalalyan, Arnak S. 2017a. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. Pages 678–689 of: Kale, Satyen, and Shamir, Ohad (eds), *Proceedings of the 2017 Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 65. PMLR.

Dalalyan, Arnak S. 2017b. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **79**(3), 651–676.

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

Dalalyan, Arnak S., and Tsybakov, Alexandre B. 2012. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, **78**(5), 1423–1443.

Davis, Burgess. 1976. On the L^p norms of stochastic integrals and other martingales. *Duke Math. J.*, **43**(4), 697–704.

Dembo, Amir, and Zeitouni, Ofer. 2010. *Large deviations techniques and applications*. Stochastic Modelling and Applied Probability, vol. 38. Springer-Verlag, Berlin. Corrected reprint of the second (1998) edition.

do Carmo, Manfredo P. 1992. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA. Translated from the second Portuguese edition by Francis Flaherty.

Dolbeault, Jean, Mouhot, Clément, and Schmeiser, Christian. 2009. Hypocoercivity for kinetic equations with linear relaxation terms. *Comptes Rendus Mathématique*, **347**(9-10), 511–516.

Duane, Simon, Kennedy, Anthony D., Pendleton, Brian J., and Roweth, Duncan. 1987. Hybrid Monte Carlo. *Phys. Lett. B*, **195**(2), 216–222.

Durmus, Alain, and Moulines, Éric. 2017. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, **27**(3), 1551–1587.

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

Durmus, Alain, Majewski, Szymon, and Miasojedow, B łażej. 2019. Analysis of Langevin Monte Carlo via convex optimization. *J. Mach. Learn. Res.*, **20**, Paper No. 73, 46.

Dwivedi, Raaz, Chen, Yuansi, Wainwright, Martin J., and Yu, Bin. 2019. Log-concave sampling: Metropolis–Hastings algorithms are fast. *Journal of Machine Learning Research*, **20**(183), 1–42.

Eberle, Andreas. 2011. Reflection coupling and Wasserstein contractivity without convexity. *C. R. Math. Acad. Sci. Paris*, **349**(19-20), 1101–1104.

Eberle, Andreas. 2016. Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, **166**(3-4), 851–886.

Eberle, Andreas, and Lörler, Francis. 2024a. Non-reversible lifts of reversible diffusion processes and relaxation times. *Probability Theory and Related Fields*, 1–31.

Eberle, Andreas, and Lörler, Francis. 2024b. Space-time divergence lemmas and optimal non-reversible lifts of diffusions on Riemannian manifolds with boundary. *arXiv preprint 2412.16710*.

Eberle, Andreas, Guillin, Arnaud, and Zimmer, Raphael. 2019. Couplings and quantitative contraction rates for Langevin dynamics. *Ann. Probab.*, **47**(4), 1982–2010.

Eberle, Andreas, Guillin, Arnaud, Hahn, Leo, Lörler, Francis, and Michel, Manon. 2025. Convergence of non-reversible Markov processes via lifting and flow Poincaré inequality. *arXiv preprint 2503.04238*.

Eckhardt, Roger. 1987. Stan Ulam, John von Neumann, and the Monte Carlo method. With contributions by Tony Warnock, Gary D. Doolen and John Hendricks, Stanislaw Ulam 1909–1984.

El Alaoui, Ahmed, Montanari, Andrea, and Sellke, Mark. 2022. Sampling from the Sherrington–Kirkpatrick Gibbs measure via algorithmic stochastic localization. Pages 323–334 of: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science—FOCS 2022*. IEEE Computer Soc., Los Alamitos, CA.

Eldan, Ronen. 2013. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geom. Funct. Anal.*, **23**(2), 532–569.

Eldan, Ronen. 2015. A two-sided estimate for the Gaussian noise stability deficit. *Invent. Math.*, **201**(2), 561–624.

Eldan, Ronen. 2016. Skorokhod embeddings via stochastic flows on the space of Gaussian measures. *Ann. Inst. Henri Poincaré Probab. Stat.*, **52**(3), 1259–1280.

Eldan, Ronen. 2018. Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geom. Funct. Anal.*, **28**(6), 1548–1596.

Eldan, Ronen. 2020. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probab. Theory Related Fields*, **176**(3-4), 737–755.

Eldan, Ronen. 2023. Analysis of high-dimensional distributions using pathwise methods. Pages 4246–4270 of: *ICM—International Congress of Mathematicians. Vol. 6. Sections 12–14*. EMS Press, Berlin.

Eldan, Ronen, and Lee, James R. 2018. Regularization under diffusion and anticoncentration of the information content. *Duke Math. J.*, **167**(5), 969–993.

Eldan, Ronen, and Shamir, Omer. 2022. Log concavity and concentration of Lipschitz functions on the Boolean hypercube. *J. Funct. Anal.*, **282**(8), Paper No. 109392, 22.

Eldan, Ronen, Lee, James R., and Lehec, Joseph. 2017. Transport-entropy inequalities and curvature in discrete-space Markov chains. Pages 391–406 of: *A journey through discrete mathematics*. Springer, Cham.

Eldan, Ronen, Mikulincer, Dan, and Zhai, Alex. 2020. The CLT in high dimensions: quantitative bounds via martingale embedding. *Ann. Probab.*, **48**(5), 2494–2524.

Eldan, Ronen, Koehler, Frederic, and Zeitouni, Ofer. 2022. A spectral condition for spectral gap: fast mixing in high-temperature Ising models. *Probab. Theory Related Fields*, **182**(3-4), 1035–1051.

Erbar, Matthias, and Maas, Jan. 2012. Ricci curvature of finite Markov chains via convexity of the entropy. *Arch. Ration. Mech. Anal.*, **206**(3), 997–1038.

Erdogdu, Murat A., Mackey, Lester, and Shamir, Ohad. 2018. Global non-convex optimization with discretized diffusions. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds), *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc.

Erdogdu, Murat A., Hosseinzadeh, Rasa, and Zhang, Shunshi. 2022. Convergence of Langevin Monte Carlo in chi-squared and Rényi divergence. Pages 8151–8175 of: Camps-Valls, Gustau, Ruiz, Francisco J. R., and Valera, Isabel (eds), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research, vol. 151. PMLR.

Evans, Lawrence C. 2010. *Partial differential equations*. Second edn. Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence, RI.

Eyring, Henry. 1935. The activated complex in chemical reactions. *J. Chem. Phys.*, 107–115.

Fan, Jiaojiao, Yuan, Bo, and Chen, Yongxin. 2023. Improved dimension dependence of a proximal algorithm for sampling. Pages 1473–1521 of: Neu, Gergely, and Rosasco, Lorenzo (eds), *Proceedings of Thirty Sixth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 195. PMLR.

Fathi, Max, and Shu, Yan. 2018. Curvature and transport inequalities for Markov chains in discrete spaces. *Bernoulli*, **24**(1), 672–698.

Fathi, Max, Gozlan, Nathael, and Prod'homme, Maxime. 2020. A proof of the Caffarelli contraction theorem via entropic regularization. *Calc. Var. Partial Differential Equations*, **59**(3), Paper No. 96, 18.

Fathi, Max, Mikulincer, Dan, and Shenfeld, Yair. 2024. Transportation onto log-Lipschitz perturbations. *Calc. Var. Partial Differential Equations*, **63**(3), Paper No. 61, 25.

Folland, Gerald B. 1999. *Real analysis*. Second edn. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York. Modern techniques and their applications, A Wiley-Interscience Publication.

Föllmer, Hans. 1985. An entropy approach to the time reversal of diffusion processes. Pages 156–163 of: *Stochastic differential systems (Marseille-Luminy, 1984)*. Lect. Notes Control Inf. Sci., vol. 69. Springer, Berlin.

Foster, James, Lyons, Terry, and Oberhauser, Harald. 2021. The shifted ODE method for underdamped Langevin MCMC. *arXiv e-prints*.

Ganesh, Arun, and Talwar, Kunal. 2020. Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC. Pages 7222–7233 of: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., and Lin, H. (eds), *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc.

Gangbo, Wilfrid, and McCann, Robert J. 1996. The geometry of optimal transportation. *Acta Math.*, **177**(2), 113–161.

Gatmiry, Khashayar, and Vempala, Santosh S. 2022. Convergence of the Riemannian Langevin algorithm. *arXiv e-prints*.

Ge, Rong, Lee, Holden, and Lu, Jianfeng. 2020. Estimating normalizing constants for log-concave distributions: algorithms and lower bounds. Pages 579–586 of: *STOC'20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York.

Gentil, Ivan. 2008. From the Prékopa–Leindler inequality to modified logarithmic Sobolev inequality. *Ann. Fac. Sci. Toulouse Math. (6)*, **17**(2), 291–308.

Gentil, Ivan, Léonard, Christian, Ripani, Luigia, and Tamanini, Luca. 2020. An entropic interpolation proof of the HWI inequality. *Stochastic Process. Appl.*, **130**(2), 907–923.

Girolami, Mark, and Calderhead, Ben. 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **73**(2), 123–214. With discussion and a reply by the authors.

Gong, Yun, He, Niao, and Shen, Zebang. 2025. Poincare inequality for local log-Polyak–Lojasiewicz measures: non-asymptotic analysis in low-temperature regime. *arXiv preprint 2501.00429*.

Gopi, Sivakanth, Lee, Yin Tat, and Liu, Daogao. 2022. Private convex optimization via exponential mechanism. Pages 1948–1989 of: Loh, Po-Ling, and Raginsky, Maxim (eds), *Proceedings of Thirty Fifth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 178. PMLR.

Gopi, Sivakanth, Lee, Yin Tat, Liu, Daogao, Shen, Ruoqi, and Tian, Kevin. 2023. Algorithmic aspects of the log-Laplace transform and a non-Euclidean proximal sampler. Pages 2399–2439 of: Neu, Gergely, and Rosasco, Lorenzo (eds), *Proceedings of Thirty Sixth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 195. PMLR.

Gozlan, Nathael, and Sylvestre, Maxime. 2025. Global regularity estimates for optimal transport via entropic regularisation. *arXiv preprint 2501.11382*.

Gozlan, Nathael, Roberto, Cyril, Samson, Paul-Marie, and Tetali, Prasad. 2014. Displacement convexity of entropy and related inequalities on graphs. *Probab. Theory Related Fields*, **160**(1-2), 47–94.

Guédon, Olivier, and Milman, Emanuel. 2011. Interpolating thin-shell and sharp large-deviation estimates for isotropic log-concave measures. *Geom. Funct. Anal.*, **21**(5), 1043–1068.

Guillin, Arnaud, and Wang, Feng-Yu. 2012. Degenerate Fokker–Planck equations: Bismut formula, gradient estimate and Harnack inequality. *J. Differential Equations*, **253**(1), 20–40.

Hastings, W. Keith. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97–109.

He, Ye, Balasubramanian, Krishnakumar, and Erdogdu, Murat A. 2020. On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method. Pages 7366–7376 of: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds), *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc.

Helffer, Bernard, and Sjöstrand, Johannes. 1994. On the correlation for Kac-like models in the convex case. *J. Statist. Phys.*, **74**(1-2), 349–409.

Hoffman, Matthew D., and Gelman, Andrew. 2014. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, **15**, 1593–1623.

Holley, Richard, and Stroock, Daniel. 1987. Logarithmic Sobolev inequalities and stochastic Ising models. *J. Statist. Phys.*, **46**(5-6), 1159–1194.

Hörmander, Lars. 2007. *Notions of convexity*. Modern Birkhäuser Classics. Birkhäuser Boston, Inc., Boston, MA. Reprint of the 1994 edition.

Hsieh, Ya-Ping, Kavis, Ali, Rolland, Paul, and Cevher, Volkan. 2018. Mirrored Langevin dynamics. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds), *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc.

Hsu, Elton P. 2002. *Stochastic analysis on manifolds*. Graduate Studies in Mathematics, vol. 38. American Mathematical Society, Providence, RI.

Isaksson, Marcus, and Mossel, Elchanan. 2012. Maximally stable Gaussian partitions with discrete applications. *Israel J. Math.*, **189**, 347–396.

Janati, Hicham, Muzellec, Boris, Peyré, Gabriel, and Cuturi, Marco. 2020. Entropic optimal transport between unbalanced Gaussian measures has a closed form. Pages 10468–10479 of: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., and Lin, H. (eds), *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc.

Jia, He, Laddha, Aditi, Lee, Yin Tat, and Vempala, Santosh. 2021. *Reducing isotropy and volume to KLS: an $O^*(n^3\psi^2)$ volume algorithm*. New York, NY, USA: Association for Computing Machinery. Page 961–974.

Jiang, Qijia. 2021. Mirror Langevin Monte Carlo: the case under isoperimetry. Pages 715–725 of: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., and Vaughan, J. Wortman (eds), *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc.

Jordan, Richard, Kinderlehrer, David, and Otto, Felix. 1998. The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, **29**(1), 1–17.

Kandasamy, Saravanan, and Nagaraj, Dheeraj. 2024. The Poisson midpoint method for Langevin dynamics: provably efficient discretization for diffusion models. Pages 65972–66024 of: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds), *Advances in Neural Information Processing Systems*, vol. 37. Curran Associates, Inc.

Kannan, Ravi, Lovász, László, and Simonovits, Miklós. 1995. Isoperimetric problems for convex bodies and a localization lemma. *Discrete Comput. Geom.*, **13**(3-4), 541–559.

Kannan, Ravi, Lovász, László, and Simonovits, Miklós. 1997. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures Algorithms*, **11**(1), 1–50.

Kannan, Ravi, Lovász, László, and Montenegro, Ravi. 2006. Blocking conductance and mixing in random walks. *Combin. Probab. Comput.*, **15**(4), 541–570.

Karimi, Hamed, Nutini, Julie, and Schmidt, Mark. 2016. Linear convergence of gradient and proximal-gradient methods under the Polyak–Lojasiewicz condition. Pages 795–811 of: *European Conference on Machine Learning and Knowledge Discovery in Databases—Volume 9851*. ECML PKDD 2016. Berlin, Heidelberg: Springer-Verlag.

Kim, Kyurae, Gruffaz, Samuel, Park, Ji Won, and Durmus, Alain O. 2025. Analysis of kinetic Langevin Monte Carlo under the stochastic exponential Euler discretization from underdamped all the way to overdamped. *arXiv preprint 2510.03949*.

Kim, Young-Heon, and Milman, Emanuel. 2012. A generalization of Caffarelli's contraction theorem via (reverse) heat flow. *Math. Ann.*, **354**(3), 827–862.

Kindler, Guy, Kirshner, Naomi, and O'Donnell, Ryan. 2018. Gaussian noise sensitivity and Fourier tails. *Israel J. Math.*, **225**(1), 71–109.

Kinoshita, Yuri, and Suzuki, Taiji. 2022. Improved convergence rate of stochastic gradient Langevin dynamics with variance reduction and its application to optimization. Pages 19022–19034 of: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds), *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc.

Klartag, Bo'az. 2018. Eldan's stochastic localization and tubular neighborhoods of complex-analytic sets. *J. Geom. Anal.*, **28**(3), 2008–2027.

Klartag, Bo'az. 2023. Logarithmic bounds for isoperimetry and slices of convex sets. *Ars Inven. Anal.*, Paper No. 4, 17.

Klartag, Bo'az, and Lehec, Joseph. 2022. Bourgain's slicing problem and KLS isoperimetry up to polylog. *Geom. Funct. Anal.*, **32**(5), 1134–1159.

Klartag, Bo'az, and Putterman, Eli. 2021. Spectral monotonicity under Gaussian convolution. *arXiv preprint 2107.09496*.

Klartag, Bo'az, Kozma, Gady, Ralli, Peter, and Tetali, Prasad. 2016. Discrete curvature and abelian groups. *Canad. J. Math.*, **68**(3), 655–674.

Kolesnikov, Alexander V. 2014. Hessian metrics, $CD(K, N)$-spaces, and optimal transportation of log-concave measures. *Discrete Contin. Dyn. Syst.*, **34**(4), 1511–1532.

Kook, Yunbum, and Vempala, Santosh S. 2025. The localization method for high-dimensional inequalities. *arXiv preprint 2512.10848*.

Kramers, Hendrik A. 1940. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, **7**, 284–304.

Kuntz, Juan, Ottobre, Michela, and Stuart, Andrew M. 2018. Non-stationary phase of the MALA algorithm. *Stoch. Partial Differ. Equ. Anal. Comput.*, **6**(3), 446–499.

Latala, Rafa l, and Oleszkiewicz, Krzysztof. 2000. Between Sobolev and Poincaré. Pages 147–168 of: *Geometric aspects of functional analysis*. Lecture Notes in Math., vol. 1745. Springer, Berlin.

Lawler, Gregory F., and Sokal, Alan D. 1988. Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.*, **309**(2), 557–580.

Le Gall, Jean-François. 2016. *Brownian motion, martingales, and stochastic calculus*. French edn. Graduate Texts in Mathematics, vol. 274. Springer, [Cham].

Ledoux, Michel. 2000. The geometry of Markov diffusion generators. vol. 9. Probability theory.

Ledoux, Michel. 2001. *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, vol. 89. American Mathematical Society, Providence, RI.

Ledoux, Michel. 2018. *Remarks on some transportation cost inequalities*.

Lee, Hyun S. 2025. *Scaling for MCMC on heterogeneous targets and nonstationary phases*. Ph.D. thesis, University of Warwick.

Lee, Yin Tat, and Vempala, Santosh S. 2017. Eldan's stochastic localization and the KLS hyperplane conjecture: an improved lower bound for expansion. Pages 998–1007 of: *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*. IEEE Computer Soc., Los Alamitos, CA.

Lee, Yin Tat, and Vempala, Santosh S. 2018. Stochastic localization + Stieltjes barrier = tight bound for log-Sobolev. Pages 1122–1129 of: *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York.

Lee, Yin Tat, and Yue, Man-Chung. 2021. Universal barrier is n-self-concordant. *Math. Oper. Res.*, **46**(3), 1129–1148.

Lee, Yin Tat, Shen, Ruoqi, and Tian, Kevin. 2020. Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo. Pages 2565–2597 of: Abernethy, Jacob, and Agarwal, Shivani (eds), *Proceedings of Thirty Third Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 125. PMLR.

Lee, Yin Tat, Shen, Ruoqi, and Tian, Kevin. 2021a. Lower bounds on Metropolized sampling methods for well-conditioned distributions. Pages 18812–18824 of: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., and Vaughan, J. Wortman (eds), *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc.

Lee, Yin Tat, Shen, Ruoqi, and Tian, Kevin. 2021b. Structured logconcave sampling with a restricted Gaussian oracle. Pages 2993–3050 of: Belkin, Mikhail, and Kpotufe, Samory (eds), *Proceedings of Thirty Fourth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 134. PMLR.

Lee, Yin Tat, Shen, Ruoqi, and Tian, Kevin. 2021c. Structured logconcave sampling with a restricted Gaussian oracle. *arXiv e-prints*.

Lehec, Joseph. 2013. Representation formula for the entropy and functional inequalities. *Ann. Inst. Henri Poincaré Probab. Stat.*, **49**(3), 885–899.

Lehec, Joseph. 2025. Convergence in total variation for the kinetic Langevin algorithm. *Math. Stat. Learn.*, **8**(1-2), 71–104.

Léonard, Christian. 2017. On the convexity of the entropy along entropic interpolations. Pages 194–242 of: *Measure theory in non-smooth spaces*. Partial Differ. Equ. Meas. Theory. De Gruyter Open, Warsaw.

Li, Lei, Wang, Mengchao, and Wang, Yuliang. 2025. Estimates of the numerical density for stochastic differential equations with multiplicative noise. *arXiv preprint 2409.04991*.

Li, Mufan (Bill), and Erdogdu, Murat A. 2023. Riemannian Langevin algorithm for solving semidefinite programs. *Bernoulli*, **29**(4), 3093–3113.

Li, Ruilin, Tao, Molei, Vempala, Santosh S., and Wibisono, Andre. 2022a. The mirror Langevin algorithm converges with vanishing bias. Pages 718–742 of: Dasgupta, Sanjoy, and Haghtalab, Nika (eds), *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*. Proceedings of Machine Learning Research, vol. 167. PMLR.

Li, Ruilin, Zha, Hongyuan, and Tao, Molei. 2022b. Sqrt(d) dimension dependence of Langevin Monte Carlo. In: *International Conference on Learning Representations*.

Li, Xuechen, Wu, Yi, Mackey, Lester, and Erdogdu, Murat A. 2019. Stochastic Runge–Kutta accelerates Langevin Monte Carlo and beyond. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds), *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.

Liang, Jiaming, and Chen, Yongxin. 2022a. A proximal algorithm for sampling. *arXiv e-prints*.

Liang, Jiaming, and Chen, Yongxin. 2022b. A proximal algorithm for sampling from non-smooth potentials. *arXiv e-prints*.

Liang, Jiaming, Mitra, Siddharth, and Wibisono, Andre. 2024. On independent samples along the Langevin diffusion and the unadjusted Langevin algorithm. *arXiv preprint 2402.17067*.

Lott, John, and Villani, Cédric. 2009. Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math. (2)*, **169**(3), 903–991.

Lovász, László, and Simonovits, Miklós. 1993. Random walks in a convex body and an improved volume algorithm. *Random Structures Algorithms*, **4**(4), 359–412.

Lu, Haihao, Freund, Robert M., and Nesterov, Yurii. 2018. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, **28**(1), 333–354.

Lu, Jianfeng. 2025. Quantitative hypocoercivity and lifting of classical and quantum dynamics. *arXiv preprint 2510.22305*.

Lu, Jianfeng, and Wang, Lihan. 2022a. Complexity of zigzag sampling algorithm for strongly log-concave distributions. *Stat. Comput.*, **32**(3), Paper No. 48, 12.

Lu, Jianfeng, and Wang, Lihan. 2022b. On explicit L^2-convergence rate estimate for piecewise deterministic Markov processes in MCMC algorithms. *Ann. Appl. Probab.*, **32**(2), 1333–1361.

Ma, Yi-An, Chatterji, Niladri S., Cheng, Xiang, Flammarion, Nicolas, Bartlett, Peter L., and Jordan, Michael I. 2021. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, **27**(3), 1942 – 1992.

Maas, Jan. 2011. Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.*, **261**(8), 2250–2292.

Mallasto, Anton, Gerolin, Augusto, and Minh, Hà Quang. 2022. Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Inf. Geom.*, **5**(1), 289–323.

Mangoubi, Oren, and Smith, Aaron. 2019. Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: numerical integrators. Pages 586–595 of: Chaudhuri, Kamalika, and Sugiyama, Masashi (eds), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research, vol. 89. PMLR.

Mangoubi, Oren, and Vishnoi, Nisheeth. 2018. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds), *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc.

Marton, Katalin. 1996. Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, **24**(2), 857–866.

Menz, Georg, and Schlichting, André. 2014. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, **42**(5), 1809–1884.

Metropolis, Nicholas, Rosenbluth, Arianna W., Rosenbluth, Marshall N., Teller, Augusta H., and Teller, Edward. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.

Mielke, Alexander. 2013. Geodesic convexity of the relative entropy in reversible Markov chains. *Calc. Var. Partial Differential Equations*, **48**(1-2), 1–31.

Mikulincer, Dan, and Shenfeld, Yair. 2023. On the Lipschitz properties of transportation along heat flows. Pages 269–290 of: *Geometric aspects of functional analysis*. Lecture Notes in Math., vol. 2327. Springer, Cham.

Mikulincer, Dan, and Shenfeld, Yair. 2024. The Brownian transport map. *Probab. Theory Related Fields*, **190**(1-2), 379–444.

Milman, Emanuel. 2009. On the role of convexity in isoperimetry, spectral gap and concentration. *Invent. Math.*, **177**(1), 1–43.

Milstein, Grigori N., and Tretyakov, Michael V. 2021. *Stochastic numerics for mathematical physics*. Second edn. Scientific Computation. Springer, Cham.

Mironov, Ilya. 2017. Rényi differential privacy. Pages 263–275 of: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*.

Montanari, Andrea. 2023. Sampling, diffusions, and stochastic localization. *arXiv preprint 2305.10690*.

Mossel, Elchanan, and Neeman, Joe. 2015. Robust optimality of Gaussian noise stability. *J. Eur. Math. Soc. (JEMS)*, **17**(2), 433–482.

Mou, Wenlong, Flammarion, Nicolas, Wainwright, Martin J., and Bartlett, Peter L. 2022. Improved bounds for discretization of Langevin diffusions: near-optimal rates without convexity. *Bernoulli*, **28**(3), 1577–1601.

Neal, Radford M. 2011. MCMC using Hamiltonian dynamics. Pages 113–162 of: *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handb. Mod. Stat. Methods. CRC Press, Boca Raton, FL.

Negrea, Jeffrey. 2022. *Approximations and scaling limits of Markov chains with applications to MCMC and approximate inference*. Ann Arbor, MI: ProQuest LLC. Thesis (Ph.D.)–University of Toronto (Canada).

Nemirovsky, Arkadii S., and Yudin, David B. 1983. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Inc., New York. Translated from the Russian and with a preface by E. R. Dawson.

Nesterov, Yu. E. 1983. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, **269**(3), 543–547.

Nesterov, Yurii. 2018. *Lectures on convex optimization*. Springer Optimization and Its Applications, vol. 137. Springer, Cham.

Nesterov, Yurii, and Nemirovskii, Arkadii S. 1994. *Interior-point polynomial algorithms in convex programming*. SIAM Studies in Applied Mathematics, vol. 13. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Nickl, Richard, and Wang, Sven. 2024. On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *J. Eur. Math. Soc. (JEMS)*, **26**(3), 1031–1112.

Nutz, Marcel, and Wiesel, Johannes. 2022. Entropic optimal transport: convergence of potentials. *Probab. Theory Related Fields*, **184**(1-2), 401–424.

Ollivier, Yann. 2007. Ricci curvature of metric spaces. *C. R. Math. Acad. Sci. Paris*, **345**(11), 643–646.

Ollivier, Yann. 2009. Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.*, **256**(3), 810–864.

Ollivier, Yann, and Villani, Cédric. 2012. A curved Brunn–Minkowski inequality on the discrete hypercube, or: what is the Ricci curvature of the discrete hypercube? *SIAM J. Discrete Math.*, **26**(3), 983–996.

Otto, Felix. 2001. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, **26**(1-2), 101–174.

Otto, Felix, and Villani, Cédric. 2000. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, **173**(2), 361–400.

Otto, Felix, and Villani, Cédric. 2001. Comment on: “Hypercontractivity of Hamilton–Jacobi equations” [J. Math. Pures Appl. (9) **80** (2001), no. 7, 669–696] by S. G. Bobkov, I. Gentil and M. Ledoux. *J. Math. Pures Appl. (9)*, **80**(7), 697–700.

Pavliotis, Grigorios A. 2014. *Stochastic processes and applications*. Texts in Applied Mathematics, vol. 60. Springer, New York. Diffusion processes, the Fokker–Planck and Langevin equations.

Rademacher, Luis, and Vempala, Santosh. 2008. Dispersion of mass and the complexity of randomized geometric algorithms. *Adv. Math.*, **219**(3), 1037–1069.

Rassoul-Agha, Firas, and Seppäläinen, Timo. 2015. *A course on large deviations with an introduction to Gibbs measures*. Graduate Studies in Mathematics, vol. 162. American Mathematical Society, Providence, RI.

Roberts, Gareth O., and Rosenthal, Jeffrey S. 1998. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **60**(1), 255–268.

Rockafellar, R. Tyrrell. 1966. Characterization of the subdifferentials of convex functions. *Pacific J. Math.*, **17**, 497–510.

Rothaus, Oscar S. 1985. Analytic inequalities, isoperimetric inequalities and logarithmic Sobolev inequalities. *J. Funct. Anal.*, **64**(2), 296–313.

Roussel, Julien, and Stoltz, Gabriel. 2018. Spectral methods for Langevin dynamics and associated error estimates. *ESAIM. Mathematical Modelling and Numerical Analysis*, **52**(3), 1051–1083.

Salim, Adil, Korba, Anna, and Luise, Giulia. 2020. The Wasserstein proximal gradient algorithm. Pages 12356–12366 of: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., and Lin, H. (eds), *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc.

Santambrogio, Filippo. 2015. *Optimal transport for applied mathematicians*. Progress in Nonlinear Differential Equations and their Applications, vol. 87. Birkhäuser/Springer, Cham. Calculus of variations, PDEs, and modeling.

Sanz-Serna, Jesús María, and Zygalakis, Konstantinos C. 2021. Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *Journal of Machine Learning Research*, **22**(242), 1–37.

Schlichting, André. 2019. Poincaré and log–Sobolev inequalities for mixtures. *Entropy*, **21**(1).

Schuh, Katharina. 2024. Global contractivity for Langevin dynamics with distribution-dependent forces and uniform in time propagation of chaos. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, **60**(2), 753–789.

Shen, Ruoqi, and Lee, Yin Tat. 2019. The randomized midpoint method for log-concave sampling. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds), *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.

Shi, Bobby, Tian, Kevin, and Zhang, Matthew S. 2025. Perspectives on stochastic localization. *arXiv preprint 2510.04460*.

Srinivasan, Rishikesh, and Nagaraj, Dheeraj. 2025. Poisson midpoint method for log concave sampling: beyond the strong error lower bounds. *arXiv preprint 2506.07614*.

Steele, J. Michael. 2001. *Stochastic calculus and financial applications*. Applications of Mathematics (New York), vol. 45. Springer-Verlag, New York.

Sturm, Karl-Theodor. 2006a. On the geometry of metric measure spaces. I. *Acta Math.*, **196**(1), 65–131.

Sturm, Karl-Theodor. 2006b. On the geometry of metric measure spaces. II. *Acta Math.*, **196**(1), 133–177.

Su, Weijie, Boyd, Stephen, and Candès, Emmanuel J. 2016. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *J. Mach. Learn. Res.*, **17**, Paper No. 153, 43.

Sudakov, Vladimir N., and Cirel'son, Boris S. 1974. Extremal properties of half-spaces for spherically invariant measures. *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, **41**, 14–24, 165. Problems in the theory of probability distributions, II.

Talagrand, Michel. 1991. A new isoperimetric inequality and the concentration of measure phenomenon. Pages 94–124 of: *Geometric aspects of functional analysis (1989–90)*. Lecture Notes in Math., vol. 1469. Springer, Berlin.

van Handel, Ramon. 2016. *Probability in high dimension*.

Vempala, Santosh, and Wibisono, Andre. 2019. Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices. Pages 8094–8106 of: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds), *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.

Vershynin, Roman. 2018. *High-dimensional probability*. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 47. Cambridge University Press, Cambridge. An introduction with applications in data science, With a foreword by Sara van de Geer.

Villani, Cédric. 2003. *Topics in optimal transportation*. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence, RI.

Villani, Cédric. 2009a. Hypocoercivity. *Mem. Amer. Math. Soc.*, **202**(950), iv+141.

Villani, Cédric. 2009b. *Optimal transport*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338. Springer-Verlag, Berlin. Old and new.

von Neumann, John. 1951. Various techniques used in connection with random digits. Chap. 13, pages 36–38 of: Householder, A. S., Forsythe, G. E., and Germond, H. H. (eds), *Monte Carlo method*. National Bureau of Standards Applied Mathematics Series, vol. 12. Washington, DC: US Government Printing Office.

Wang, Feng-Yu. 1997. Logarithmic Sobolev inequalities on noncompact Riemannian manifolds. *Probab. Theory Related Fields*, **109**(3), 417–424.

Wang, Feng-Yu. 2011. Equivalent semigroup properties for the curvature-dimension condition. *Bull. Sci. Math.*, **135**(6-7), 803–815.

Wang, Feng-Yu. 2013. *Harnack inequalities for stochastic partial differential equations*. SpringerBriefs in Mathematics. Springer, New York.

Wang, Feng-Yu. 2014. *Analysis for diffusion processes on Riemannian manifolds*. Advanced Series on Statistical Science & Applied Probability, vol. 18. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ.

Wang, Songbo. 2024. Uniform log-Sobolev inequalities for mean field particles with flat-convex energy. *arXiv preprint 2408.03283*.

Wibisono, Andre. 2018. Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem. Pages 2093–3027 of: Bubeck, Sébastien, Perchet, Vianney, and Rigollet, Philippe (eds), *Proceedings of the 31st Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 75. PMLR.

Wu, Keru, Schmidler, Scott, and Chen, Yuansi. 2021. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *arXiv e-prints*.

Yu, Lu, Karagulyan, Avetik, and Dalalyan, Arnak S. 2024. Langevin Monte Carlo for strongly log-concave distributions: randomized midpoint revisited. In: *The Twelfth International Conference on Learning Representations*.

Zhang, Kelvin S., Peyré, Gabriel, Fadili, Jalal, and Pereyra, Marcelo. 2020. Wasserstein control of mirror Langevin Monte Carlo. Pages 3814–3841 of: Abernethy, Jacob, and Agarwal, Shivani (eds), *Proceedings of Thirty Third Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 125. PMLR.

Zhang, Matthew S. 2025. Analysis of Langevin midpoint methods using an anticipative Girsanov theorem. *arXiv preprint 2507.12791*.

Zhang, Matthew S., Chewi, Sinho, Li, Mufan (Bill), Balasubramanian, Krishnakumar, and Erdogdu, Murat A. 2023. Improved discretization analysis for underdamped Langevin Monte Carlo. Pages 36–71 of: Neu, Gergely, and Rosasco, Lorenzo (eds), *Proceedings of Thirty Sixth Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 195. PMLR.

Zimmermann, David. 2013. Logarithmic Sobolev inequalities for mollified compactly supported measures. *J. Funct. Anal.*, **265**(6), 1064–1083.

Zimmermann, David. 2016. Elementary proof of logarithmic Sobolev inequalities for Gaussian convolutions on $\mathbb{R}$. *Ann. Math. Blaise Pascal*, **23**(1), 129–140.

Index

c-concavity, 44
c-conjugate potentials, 44
f-divergence, 39, 213

absolute continuity, 26
abstract Wiener space, 101
acceleration, 148

Bakry–Émery theorem, 19
Benamou–Brenier formula, 44
blow-up, 68
Bochner identity, 50
Bonnet–Myers diameter bound, 83
Brascamp–Lieb inequality, 52, 238
Bregman divergence, 53, 238
Bregman proximal lemma, 241
Bregman transport cost, 53, 242
Bregman transport inequality, 53, 87, 243
Brenier potential, 24
Brenier's theorem, 22
  Bregman transport costs, 242
Brownian bridge, 105
Brownian motion, 4
Brunn–Minkowski inequality, 90

Caffarelli's contraction theorem, 61, 112
Cameron–Martin
  space, 100
  theorem, 100
carré du champ, 14
  iterated operator, 18, 50
Cheeger's inequality, 193
chi-squared divergence, 17
coarea inequality, 72
cold start, 189
concentration function, 68
conductance, 193
  s-conductance, 196
continuity equation, 26
covariant derivative, 78
Cramér–Rao inequality, 112

curvature
  Gaussian, 79
  Ricci, 80
    coarse, 84
  Riemann, 79
  scalar, 80
  sectional, 80
curvature-dimension condition, 19, 50, 82

data-processing inequality, 39
descent lemma, 126
diffusion coefficient, 8
diffusion semigroup, 19, 54
Dirichlet energy, 14, 192
displacement interpolation, 30
divergence, 78
divergence lemma, 158
Donsker–Varadhan variational principle, 40
Doob martingale, 41
Doob's transform, 103
drift coefficient, 8

Efron–Stein inequality, 42
elementary process, 5
entropy, 228
Euler–Maruyama discretization, 121
evolution variational inequality (EVI), 130
exponential map, 30
Eyring–Kramers formula, 60

Föllmer drift, 107
Fano's inequality, 228
finite variation, 95
first variation, 31
Fisher information, 18, 55
flow map, 144
Fokker–Planck equation, 12
friction, 147

generalized geodesic, 243
geodesic, 29, 78
geodesic convexity, 31

Girsanov's theorem, 103, 133
  shifted Girsanov, 115, 133
Grönwall's lemma, 16
gradient, 77
gradient descent, 122
  accelerated, 148

Hörmander's L^2 method, 52
Hamilton's equations of motion, 144
Hamiltonian, 144
Hamiltonian Monte Carlo (HMC)
  ideal, 143
  Metropolized (MHMC), 187
Hamilton–Jacobi equation, 44
Harnack inequality, 36, 47, 92, 115, 137
  log-Harnack, 92
Helffer–Sjöstrand identity, 88
Hellinger distance, 254
Holley–Stroock perturbation, 60
Hopf–Lax semigroup, 45
hypercontractivity, 89, 169
hypocoercivity, 153
  entropic, 156

infinitesimal generator, 11, 192
isoperimetric profile, 71
isoperimetry
  Cheeger, 72, 195
  Gaussian, 70, 76, 200
  sphere, 70, 93
Itô integral, 3
Itô isometry, 6
Itô process, 8
Itô's formula, 9, 98

Kannan–Lovász–Simonovits (KLS) conjecture, 59
Kantorovich problem, 20
Kazamaki's condition, 103
Kim–Milman map, 116
Kolmogorov's backward equation, 12
Kolmogorov's forward equation, 12
Kullback–Leibler (KL) divergence, 17
  chain rule, 40

Langevin diffusion, 3
Langevin Monte Carlo (LMC), 121
Laplace–Beltrami operator, 79
Latala–Oleszkiewicz inequality (LOI), 85
lazy chain, 189
Le Cam's inequality, 254
leapfrog integrator, 187
Levi–Civita connection, 78
Lichnerowicz inequality, 94
Lie bracket, 79
local error analysis, 140
local martingale, 7
localization, 6
localizing sequence, 6
log-Sobolev inequality (LSI), 18, 48, 126
  defective, 64
  local, 56
  modified, 84, 192
logarithmic map, 30

Malliavin calculus, 114
manifold, 29
  Hessian, 77
Markov semigroup, 10
martingale, 5
  exponential, 102
Marton's tensorization, 67
McCann's interpolation, 30
mean-squared error analysis, *see* local error analysis
mesh, 95
metastability, 250
metric derivative, 26
Metropolis-adjusted Langevin algorithm (MALA), 187
Metropolis–Hastings (MH) filter, 185
Metropolized random walk (MRW), 187
Minkowski content, 71
mirror descent, 236
mirror Langevin, 236
mirror Langevin Monte Carlo (MLMC), 240
mirror map, 236
Monge problem, 20
Monge–Ampère equation, 87
mutual information, 228

Newton Langevin diffusion, 238
no-U-turn sampler (NUTS), 161
Novikov's condition, 103

optimal transport, 20
  dual, 21
  entropic regularization, 110
  fundamental theorem, 22
  optimal transport plan, 20
Ornstein–Uhlenbeck (OU) process, 42, 153
Otto calculus, 33
Otto–Villani theorem, 36

parallel transport, 78
Pinsker's inequality, 37
Poincaré inequality (PI), 16, 48, 192
  L^p–L^q, 73
  local, 55
  mirror, 239
  space-time, 157
Poisson bracket, 163
Poisson equation, 52, 87
Polyak–Lojasiewicz (PL) inequality, 35, 59, 126, 221
Prékopa–Leindler inequality, 89
proximal oracle, 209
proximal point method (PPM), 209
proximal sampler, 209

quadratic variation, 96

Rényi divergence, 57, 166
randomized midpoint discretization, 139
reflection coupling, 136
rejection sampling, 183
relative convexity, 238
relative smoothness, 238
restricted Gaussian oracle (RGO), 210

reverse transport inequality, *see* Harnack inequality
reversibility, 13, 185
Ricci curvature, 51
Riemannian metric, 29, 77

Schrödinger bridge, 109
second-order lift, 157
self-concordance, 241
semimartingale, 97
shifted ODE algorithm, 161
Sobolev inequality, 82
Sobolev norm, 87
  inverse, 87
stochastic calculus, 3
stochastic differential equation (SDE), 9
stochastic localization, 108
stopping time, 6
symplectic integrator, 187
synchronous coupling, 136

Talagrand's T_1 inequality, 48
Talagrand's T_2 inequality, 36, 48
tangent bundle, 77
tangent space, 29, 77
time reversal, 105, 116, 215
total variation, 95
  total variation (TV) distance, 38, 96
  total variation norm, 95

underdamped Langevin diffusion, 147

vector field, 77
volume measure, 79

warm start, 189
Wasserstein gradient flow, 32
Wasserstein metric, 20, 44
weak triangle inequality, 173
Wiener measure, 99