

# Statistical Optimal Transport

Sinho Chewi  
Yale

Jonathan Niles-Weed  
NYU

Philippe Rigollet  
MIT

*Ecole d'Eté de Probabilités de Saint-Flour XLIX*



---

## Contents

<b>Preface</b> .....	1
<b>1 Optimal transport</b> .....	9
1.1 The optimal transport problem .....	9
1.2 Wasserstein distances .....	15
1.3 Optimal transport in one dimension .....	21
1.4 Brenier's theorem .....	23
1.5 Kantorovich duality .....	29
1.6 Duality for $p = 1$ .....	39
1.7 Discussion .....	42
1.8 Exercises .....	43
<b>2 Estimation of Wasserstein distances</b> .....	47
2.1 The Wasserstein law of large numbers .....	47
2.2 The dyadic partitioning argument .....	48
2.3 Dual chaining bounds .....	54
2.4 A finer analysis for $d = 2$ .....	58
2.5 Applications .....	63
2.6 Optimality .....	66
2.7 Faster rates for smooth measures .....	70
2.8 Regularization of Wasserstein distances .....	74
2.9 Discussion .....	82
2.10 Exercises .....	85
<b>3 Estimation of transport maps</b> .....	89
3.1 Problem formulation .....	89
3.2 The semidual problem and its stability .....	91

VIII    Contents

3.3	A special case: affine transport maps	94
3.4	Obtaining the slow rate	95
3.5	The fixed point argument	99
3.6	Obtaining the fast rate	100
3.7	Discussion	104
3.8	Exercises	105
<b>4</b>	<b>Entropic optimal transport</b>	109
4.1	Derivation of entropic optimal transport	110
4.2	Duality	114
4.3	Statistical rates for dual solutions	117
4.4	Statistical rates for primal solutions	123
4.5	Discussion	129
4.6	Exercises	131
<b>5</b>	<b>Wasserstein gradient flows: theory</b>	135
5.1	Metric derivative and the continuity equation	136
5.2	Elements of Riemannian geometry	141
5.3	The Riemannian structure of Wasserstein space	143
5.4	Otto calculus	145
5.5	Bures–Wasserstein	149
5.6	Gaussian mixtures	152
5.7	Wasserstein–Fisher–Rao	154
5.8	Mean-field particle systems	159
5.9	Discussion	162
5.10	Exercises	163
<b>6</b>	<b>Wasserstein gradient flows: applications</b>	167
6.1	Variational inference	167
6.2	Sampling	177
6.3	Interacting particle systems	187
6.4	Non-parametric maximum likelihood	191
6.5	Mean-field neural networks	195
6.6	Transformers	197
6.7	Discussion	201
6.8	Exercises	203
<b>7</b>	<b>Metric geometry of the Wasserstein space</b>	205
7.1	Geodesics	205
7.2	Curvature	211

7.3	Tangent cones	218
7.4	Discussion	227
7.5	Exercises	228
<b>8</b>	<b>Wasserstein barycenters</b>	229
8.1	The Hilbert case	231
8.2	Barycenters on positively curved spaces	233
8.3	Parametric rates for Wasserstein barycenters	240
8.4	Discussion	243
8.5	Exercises	244
<b>A</b>	<b>Convex analysis</b>	247
A.1	Convex functions, subdifferentials, and duality	247
A.2	Strong convexity and smoothness	250
A.3	Convex conjugate of a quadratic function	253
<b>B</b>	<b>Probability</b>	255
<b>References</b>		259
<b>Index</b>		283



---

## Preface

The history of optimal transport begins in 1781 with a memoir by Gaspard Monge that he submitted to the Académie des Sciences [Mon81]. Since then, it has grown into a mature mathematical field with many important discoveries, such as Kantorovich’s duality theory, Brenier’s theorem, Otto’s calculus, the JKO scheme, and the Lott–Sturm–Villani definition of the Ricci curvature of geodesic spaces, to name a few. The first comprehensive treatment of optimal transport dates back to the seminal volumes of Rachev and Rüschendorf [RR98a, RR98b]. We also refer the reader to the excellent texts of Villani [Vil03, Vil09b], Ambrosio, Gigli, and Savaré [AGS08], and more recently, Santambrogio [San15] for a comprehensive treatment of this subject from the mathematical perspective, and to the notes of Ambrosio and Gigli [AG13], Ambrosio, Brué, and Semola [ABS21], and the short monograph of Figalli and Glaudo [FG23] for quicker introductions. Even a quick inspection of their tables of contents reveals that Monge’s question was the gateway to many more, and that the field of optimal transport has many unexpected connections, ranging from geometry to partial differential equations.

More recently, optimal transport has made a resounding entrance into the field of machine learning under the impetus of Marco Cuturi, who showed that Wasserstein distances could be computed efficiently using the Sinkhorn algorithm [Cut13]. This initial spark was followed by an extensive toolbox, built on optimal transport, that covered multiple tasks across various areas of machine learning and graphics. The development of this toolbox, now called *computational optimal transport*, was led by Marco Cuturi and Gabriel Peyré and collected in their inspiring manuscript [PC19b]. The far-reaching scope of optimal transport across machine learning and data science rests on the fact that many

objects such as point clouds, polygonal meshes, or even documents can be encoded as probability measures. In turn, the Wasserstein metric between these probability measures offers a semantically meaningful notion of distance.

So what is *statistical* optimal transport? This modifier comes largely as an echo to Peyré and Cuturi’s *computational* optimal transport. It is an umbrella term that captures the remarkably diverse points of contact between statistics and optimal transport. The aim of the present monograph is to provide an introduction to a selection of topics within this burgeoning field according to our tastes (see [PZ20] for a complementary treatment).

Historically, Wasserstein distances have been employed in statistics as a tool to quantify the rate of convergence of empirical probability measures  $\mu_n$  to their limit  $\mu$ . This line of work was inaugurated in a celebrated work of Dudley [Dud69], who provided bounds for  $W_1(\mu_n, \mu)$ . Wasserstein distances are particularly well-suited for quantifying this convergence for several reasons. First, unlike the total variation distance or Kullback–Leibler divergence, the Wasserstein distance between a discrete distribution  $\mu_n$  and a potentially continuous one  $\mu$  remains finite and informative. Second, by definition, bounding the Wasserstein distance amounts to exhibiting a coupling between the two measures. Third and finally, thanks to Kantorovich duality, a bound on the Wasserstein distance translates into a strong uniform bound on test functions. For example, when  $p = 1$ ,  $W_1(\mu_n, \mu) \leq \varepsilon$  implies that  $|\int f \, d\mu_n - \int f \, d\mu| \leq \varepsilon$  for all functions  $f$  that are 1-Lipschitz; in fact the two statements are equivalent (see Section 1.6 for details).

In the last decade, following the impetus of machine learning, optimal transport has percolated to many more aspects of statistics. One of the most exciting directions is a new avenue of research that had largely been out of reach of classical methods in the past. In this class of problems, the *coupling* of data is the main obstacle to statistical analysis. More concretely, consider a classical statistical setup where one observes independent copies of a pair of random variables  $(X, Y)$  where  $X$  is thought of as input and  $Y$  output. Regression falls in this framework, as does, more generally, all of supervised learning. In particular, the observed  $X$  and  $Y$  are coupled. A more challenging model arises when  $X$  and  $Y$  are observed in an *uncoupled* fashion: independent copies of  $X$  and independent copies of  $Y$ . Such a setup arises naturally in single-cell genomics where the destructive nature of the prevailing sequencing

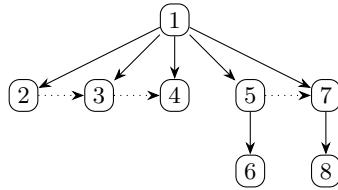
process does not allow for taking multiple measurements of the same cell. This conundrum is a key obstacle to cellular trajectory reconstruction where one aims at reconstructing the time evolution of a cell in a genetic landscape; see [SST<sup>+</sup>19, BMPKC22] for more details. However, under a natural physical hypothesis that the distribution of  $Y$  is obtained from the distribution of  $X$  by evolving it in a conservative vector field for a short time, it is natural to model the unobserved coupling between  $X$  and  $Y$  as arising from optimal transport. These problems and more raise fundamental questions about the estimation of Wasserstein distances and the corresponding couplings, which are taken up in the first part of this monograph.

The aforementioned applications make use of the role of optimal transport in endowing the space of probability measures with an interpretable and useful notion of distance. A deeper study of this space, however, uncovers a rich underlying geometrical structure admitting descriptions of curvature, geodesics, gradient flows, etc. This geometric perspective, first advocated by Felix Otto in his seminal article [Ott01], provides statisticians with powerful new tools for the design and analysis of algorithms for manipulating probability distributions. A key development in this regard was the edifying interpretation, by Jordan, Kinderlehrer, and Otto [JKO98], of the Langevin diffusion as a Wasserstein gradient flow of the KL divergence. Since the Langevin diffusion is popularly employed as a sampling algorithm in Bayesian statistics, this discovery has ushered in a decade of research linking sampling to optimization over the Wasserstein space. More broadly, Wasserstein gradient flows and their variants yield new algorithmic paradigms and fresh perspectives for diverse problems including the nonparametric MLE, the dynamics and training of neural networks, and variational inference (see Chapter 6). Geometric considerations also lead to novel applications, such as the geometric averaging of data which can be cast as probability measures, e.g., images or speech. The subject of Wasserstein geometry is studied in the second half of this monograph.

*How to read this book.*

This monograph aims to offer a concise introduction to optimal transport, quickly transitioning to its applications in statistics and machine learning. It is primarily tailored for students and researchers in these fields, yet it remains accessible to a broader audience of applied mathematicians and computer scientists.

Chapter 1 serves as the gateway to the subsequent chapters by presenting the foundational concepts of optimal transport that will be used throughout. The remaining chapters are largely independent, with the exceptions of chapters 5 and 6, and chapters 7 and 8, which should be studied together. Figure 0.1 illustrates the dependencies between the various chapters.



**Fig. 0.1.** Dependencies between chapters. Solid arrows show prerequisites; dotted arrows indicate references.

Each chapter concludes with a series of exercises, allowing readers to apply the concepts learned to questions not addressed in the main text.

*Acknowledgments.*

**Sinho Chewi.** Optimal transport was not what I originally planned to study as a graduate student, but my research career has been infinitely richer as a result. For this, I am grateful to my advisor and co-author Philippe for fearlessly inducting me into this beautiful area, and my academic sibling and co-author Jon for paving the way and constantly acting as my source of inspiration.

I owe my understanding of optimal transport to many sources, beginning with Villani's expertly written monograph [Vil03] and including countless hours of discussion with collaborators and members of Philippe's group. I am happy to have the opportunity to contribute to the growth of the community through this pedagogical text.

I would like to acknowledge the Institute for Advanced Study for the hospitable working environment and support through the Eric and Wendy Schmidt Fund.

*New Haven, CT, July 2024.*

**Jonathan Niles-Weed.** I have been lucky to learn about optimal transport from many brilliant collaborators; among them, Jason Altschuler,

Francis Bach, Sivaraman Balakrishnan, Quentin Berthet, Marco Cuturi, Vincent Divol, Alberto González-Sanz, Tudor Manole, Aram-Alexandre Pooladian, Austin Stromme, and Larry Wasserman have had an especially large impact on me. I am deeply grateful to them for their knowledge and insight. It is an honor to be able to share some of their work in this book.

I am indebted to students at NYU who participated in a graduate seminar on portions of this material in 2021 and whose feedback significantly improved the exposition. I also thank Sivaraman Balakrishnan (again) for his perceptive comments on an early version of this manuscript.

Of course, I would be nowhere without Sinho Chewi and Philippe Rigollet, my colleagues and friends. You both have inspired me beyond measure. Collaborating with you has been one of the highlights of my career, and I hope to have the privilege of many more collaborations ahead.

My work on these notes was primarily supported by an National Science Foundation CAREER award DMS-2339829 and several other NSF grants (DMS-2015291, DMS-2210583), along with a Sloan Foundation fellowship and gifts from Apple and Google.

*New York, NY, July 2024.*

**Philippe Rigollet.** These notes have grown significantly since I was honored to give the 2019 St. Flour lecture series, which initially covered roughly Chapters 1, 7, and 8. Much of the material presented here was unfamiliar to me when I first delivered the lectures.

First and foremost, I extend my heartfelt thanks to the organizers (especially Christophe Bahadoran) and attendees of the 2019 St. Flour Summer School. The valuable feedback I've received from the audience has been instrumental in motivating the extensive expansion of topics covered in these notes. My only regret is the delay in producing these notes, which prevents them from being published concurrently with those of my wonderful co-lecturers, Nicolas Curien and Elchanan Mossel. Though we won't be sharing a volume, we will forever be bonded as Chevaliers du Taste-fourme.

I have learned most from others and would like to thank my collaborators who have taught me so much on this topic: Jason Altschuler, Julien Clancy, Max Daniels, Aiden Forrow, Borjan Geshkovski, Florian Gunsilius, Jan-Christian Hütter, Cyril Letrouit, Jean-Michel Loubes,

Chen Lu, Tyler Maunu, Mor Nitzan, Quentin Paris, Geoff Schiebinger, Austin Stromme, George Stepaniants, Felipe Suarez, William Torous, Kaizheng Wang, Yuling Yan, Aleksandr Zimin. In particular, for enlightening discussions on optimal transport and other topics, I would like to specifically acknowledge Sébastien Bubeck, Ramon van Handel, Thibaut Le Gouic, Vianney Perchet, Yury Polyanskiy, Maxim Raginsky, and Justin Solomon. Many of the ideas in Chapters 5 and 6 were core discussion topics during the program on *Geometric Methods in Optimization and Sampling* at the Simons Institute in Berkeley during Fall 2021. I extend my gratitude to all the participants, my co-organizers, and Peter Bartlett, who made this program so wonderful and stimulating.

Since then, the community around statistical optimal transport has grown significantly, and these notes have benefited from interactions with many people. Starting with Marco Cuturi, who introduced me to optimal transport when we were in Princeton together, I also learned a lot of optimal transport from Guillaume Carlier, Victor Chernozhukov, Lenaïc Chizat, Simone Di Marino, Augusto Gerolin, Promit Ghosal, Marc Hallin, Zaid Harchaoui, Kengo Kato, Anna Korba, Alexei Kroshnin, Qin Li, Jan Maas, Axel Munk, Robert McCann, Dan Mikulincer, Youssef Marzouk, Soumik Pal, Victor Panaretos, Gabriel Peyré, Filippo Santambrogio, Bodhi Sen, Vladimir Spokoiny, Yair Shenfeld, and Jia-Jie Zhu among others.

Part of this material has been taught at MIT in 2023 and 2024, at Université Paris Sorbonne in 2022, and two other summer schools in 2023: The Princeton Machine Learning Theory Summer School and the CIME summer school in Cetraro, Italy. The audience has given me excellent feedback, contributing to the improvement of these notes. Special thanks to Théo Dumont, Max Daniels, and Giulia Bertagnolli for typing up the respective material.

Some typos and errors were fixed by Michael Diao, Haruki Kono, Hugo Lavenant, Aimee Maurais, Yaroslav Mukhin, Madhav Sankaranarayanan, Sabarish Sainathan, Yucheng Shang, Vishwak Srinivasan, Panos Tsimpos, Oliver Wang, Liane Xu, and Julie Zhu. We thank them for helping make this manuscript more readable.

Last but not least, I would like to thank my wonderful co-authors, Sinho Chewi and Jonathan Niles-Weed. Thank you both for joining me on this epic adventure. I've learned more from you than anyone else, and you have been as much students as you have been teachers to me.

My work on these notes was primarily supported by NSF grant CCF-1838071 and several other grants from the National Science Foundation during the period 2019-2024 (DMS-1712596, CCF-1740751, DMS-2022448, CCF-2106377). I am also thankful for a gift from Apple. A first draft was conceived in Spring 2019, when I was supported by the Eric and Wendy Schmidt Fund at the Institute for Advanced Study.

*Cambridge, MA, July 2024.*



---

## Optimal transport

### 1.1 The optimal transport problem

In his 1781 memoir [Mon81], Monge formulated the following problem: how can one transport a given pile of sand to fill a given ditch so as to minimize the cost of transporting the sand? This problem can be modeled using probability distributions. Indeed, note first that for this task to be solvable, the pile and the ditch must occupy the same volume. Without loss of generality, let us normalize this volume to be 1. We are therefore led to consider two probability measures,  $\mu$  and  $\nu$  over  $\mathbb{R}^d$ . It is often convenient to reason about two random variables,  $X \sim \mu$ ,  $Y \sim \nu$ . This is our **input** to a **constrained optimization problem**.

#### 1.1.1 The Monge and Kantorovich problems

Back to our sand analogy, transporting the pile means finding a (measurable) function, called a *transport map*  $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , which indicates that the sand located at  $x \in \mathbb{R}^d$  should be moved to  $T(x) \in \mathbb{R}^d$ . For the transport map to actually complete the job (filling the ditch), one needs to ensure that  $T(X) \sim \nu$  whenever  $X \sim \mu$ . We say that  $T$  *pushes*  $\mu$  to  $\nu$  or that  $\nu$  is the *pushforward measure* of  $\mu$  (through  $T$ ) and write  $T_\# \mu = \nu$ . This is our **constraint**.

Turning now to our **objective** function, recall that Monge's question involved minimizing the cost of transporting the sand. There are many ways to measure this cost (effort, fuel consumption, etc.) so to simplify our exposition, we simply measure it in terms of the Euclidean distance travelled by the sand. The sand at location  $x$  travels a distance of  $\|T(x) - x\|$ . Therefore, the average distance travelled is

$$\int \|T(x) - x\| \mu(dx).$$

The Monge formulation of the optimal transport problem is therefore to minimize the above objective subject to the constraint that  $T$  pushes  $\mu$  to  $\nu$ :

$$\inf_{T:T_\# \mu = \nu} \int \|T(x) - x\| \mu(dx).$$

Note that many choices for the transport cost may be considered. In full generality, it is customary to consider a general cost  $c(X, T(X))$ , where  $c(x, y)$  measures the cost of transporting  $x \in \mathbb{R}^d$  to  $y \in \mathbb{R}^d$ . In this general framework, we may even allow  $X$  and  $Y$  to be defined on two different spaces, not necessarily  $\mathbb{R}^d$ . In these notes, we focus primarily on the cases where  $c(x, y) = \|x - T(x)\|$  or  $c(x, y) = \|x - T(x)\|^2$ , which give rise to the *Wasserstein distances*. The space  $\mathbb{R}^d$  may also be replaced with more complex spaces such as Riemannian manifolds, but this is generally beyond the scope of these lectures (with the exception of Section 5.6).

While the Monge problem is easy to formulate, we need to ask several questions:

- Does there always exist such a valid transport map or, conversely, is the constraint set empty?
- If there is a minimizer, is it unique? How to characterize it? Note that our constraint is not convex, which makes finding an answer to this question rather difficult.

A simple example gives an answer to the first question. Indeed, take  $d = 1$ , assume that  $\mu = \delta_0$  is a point mass at 0, and that  $\nu = \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_1$  is a mixture of two point masses. Whatever our choice of the transport map  $T$ , the pushforward  $T_\# \mu$  is the point mass  $\delta_{T(0)}$  at  $T(0)$ , so we cannot achieve the transport at all, at least with a deterministic map.

Intuitively, we would like:

$$T(0) = \begin{cases} -1 & \text{w.p. } \frac{1}{2} \\ 1 & \text{w.p. } \frac{1}{2} \end{cases} \quad \text{and} \quad T(x) = x, \quad \forall x \neq 0.$$

Such a  $T$  is not a function but a Markov kernel: it assigns a probability distribution to each point  $x \in \mathbb{R}$ .

The second question remained without a satisfactory answer for almost two centuries until the Soviet mathematician Leonid Kantorovich [Kan42] introduced a relaxation of the problem that exactly

allows for Markov kernels, as discussed in the example above, in a groundbreaking two-pager. Equivalently, this formulation involves *couplings* as opposed to maps.

Let  $\mu, \nu$  be two probability measures over  $\mathbb{R}^d$  and let  $\gamma$  be a *coupling* between these two distributions, that is, a joint distribution over  $\mathbb{R}^d \times \mathbb{R}^d$  such that its first marginal is  $\mu$  and its second marginal is  $\nu$ : for any Borel set  $A \in \mathbb{R}^d$ , we have

$$\gamma(A \times \mathbb{R}^d) = \mu(A) \quad \text{and} \quad \gamma(\mathbb{R}^d \times A) = \nu(A).$$

The terminology *coupling* comes from the fact that while  $X \sim \mu$  and  $Y \sim \nu$  were random variables that had nothing to do with each other, the coupling forces them to live on the same probability space by describing their probabilistic dependence. Here and throughout these notes, we use the notation  $\Gamma_{\mu, \nu}$  for the set of couplings of  $\mu$  and  $\nu$ .

Let  $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$  be a measurable cost function. The general Kantorovich formulation of the optimal transport problem consists of the following optimization problem:

$$\inf_{\gamma \in \Gamma_{\mu, \nu}} \int c(x, y) \gamma(dx, dy). \quad (\text{KOT})$$

### 1.1.2 Couplings

To get a better understanding of the Kantorovich problem, it is informative to explore the set  $\Gamma_{\mu, \nu}$ .

Perhaps the simplest coupling is the independent coupling  $\gamma = \mu \otimes \nu$  where  $X \sim \mu$  and  $Y \sim \nu$  are simply assumed to be independent: for any Borel sets  $A, B \subseteq \mathbb{R}^d$ ,

$$\gamma(A \times B) = \mu(A) \cdot \nu(B).$$

The next proposition collects preliminary facts about  $\Gamma_{\mu, \nu}$ .

**Proposition 1.1.** *Let  $\mu, \nu$  be two probability measures on  $\mathbb{R}^d$ . The set  $\Gamma_{\mu, \nu}$  of couplings between  $\mu$  and  $\nu$  is non-empty, convex, and compact with respect to the topology of weak convergence.*

*Proof.* Because the independent coupling always exists, we know that  $\Gamma_{\mu, \nu} \neq \emptyset$ .

To show that  $\Gamma_{\mu, \nu}$  is convex, consider two couplings  $\gamma_0, \gamma_1 \in \Gamma_{\mu, \nu}$  and for any  $\lambda \in (0, 1)$  define the mixture  $\gamma_\lambda = (1 - \lambda) \gamma_0 + \lambda \gamma_1$ . Observe that for any Borel set  $A \in \mathbb{R}^d$ ,

$$\begin{aligned}\gamma_\lambda(A \times \mathbb{R}^d) &= (1 - \lambda)\gamma_0(A \times \mathbb{R}^d) + \lambda\gamma_1(A \times \mathbb{R}^d) \\ &= (1 - \lambda)\mu(A) + \lambda\mu(A) = \mu(A).\end{aligned}$$

Hence the first marginal of  $\gamma_\lambda$  is given by  $\mu$  and by the same argument its second marginal is given by  $\nu$ . Thus  $\gamma_\lambda \in \Gamma_{\mu,\nu}$  for any  $\lambda \in (0, 1)$ , whence  $\Gamma_{\mu,\nu}$  is convex.

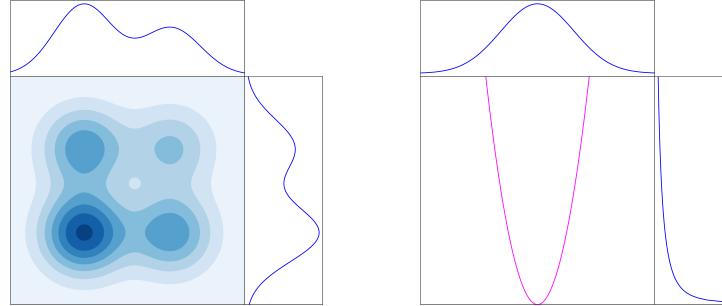
To complete the proof of our proposition, we show that  $\Gamma_{\mu,\nu}$  is compact. By Prokhorov's theorem (Theorem B.3), it is sufficient to show that  $\Gamma_{\mu,\nu}$  is closed and (uniformly) tight. To that end, recall that from Prokhorov's theorem, the constant sequences  $(\mu)_n, (\nu)_n$  are both tight, so that for any  $\varepsilon > 0$ , there exists a compact set  $K \subset \mathbb{R}^d$  such that  $\mu(K^c) + \nu(K^c) < \varepsilon$ . Then the set  $K \times K$  is also compact and for any  $\gamma \in \Gamma_{\mu,\nu}$ ,

$$\gamma((K \times K)^c) \leq \gamma(\mathbb{R}^d \times K^c) + \gamma(K^c \times \mathbb{R}^d) = \mu(K^c) + \nu(K^c) < \varepsilon.$$

Hence,  $\Gamma_{\mu,\nu}$  is tight. Moreover, since  $\gamma \in \Gamma_{\mu,\nu}$  is equivalent to

$$\int f(x) \gamma(dx, dy) = \int f d\mu \quad \text{and} \quad \int f(y) \gamma(dx, dy) = \int f d\nu$$

for all bounded continuous  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , by the definition of weak convergence (Theorem B.4) it follows that  $\Gamma_{\mu,\nu}$  is closed. Therefore, Prokhorov's theorem yields that  $\Gamma_{\mu,\nu}$  is compact.  $\square$



**Fig. 1.1.** (Left) Independent coupling of a mixture of two Gaussians. (Right) Deterministic coupling of  $X \sim \mathcal{N}(0, 1)$  with  $Y \sim \chi^2_1$ .

A coupling  $\gamma \in \Gamma_{\mu,\nu}$  captures the dependence between two random variables  $X \sim \mu$  and  $Y \sim \nu$ . As mentioned above, one option is to assume that  $X$  and  $Y$  are independent, which gives rise to the independent coupling. In Figure 1.1 (Left), we plot the independent coupling between two mixtures of Gaussians. At the opposite extreme of the independent coupling, suppose by way of example that  $X \sim \mathcal{N}(0, 1)$  and  $Y \sim \chi_1^2$  and observe that  $Y$  has the same distribution as  $X^2$ . Then we can take the deterministic coupling such that  $Y = X^2$ :

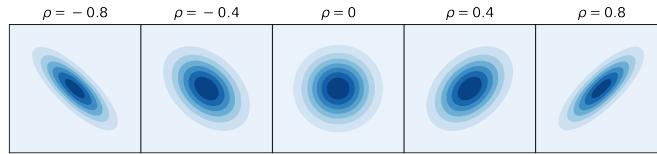
$$\gamma(dx, dy) = \mu(dx) \delta_{x^2}(dy).$$

We plot this coupling in Figure 1.1 (Right); observe that it is degenerate.

To continue our exploration of couplings, assume that  $X \sim \mathcal{N}(0, 1)$  and  $Y \sim \mathcal{N}(0, 1)$ , then we can take any coupling where

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \quad (1.1)$$

and  $\rho \in [-1, 1]$  is the correlation between  $X$  and  $Y$ . See Figure 1.2.



**Fig. 1.2.** The bivariate Gaussian coupling (1.1) for five different values of  $\rho$ .

The content of Brenier's theorem later in this chapter is that under mild regularity conditions, the solution of the Kantorovich problem with quadratic cost is achieved by a deterministic coupling. As we discuss next, when  $\mu$  and  $\nu$  are discrete measures, this fact can be understood by examining the geometry of the set  $\Gamma_{\mu,\nu}$ .

### 1.1.3 Discrete optimal transport

The case where  $\mu$  and  $\nu$  are two discrete distributions is of special practical relevance. For example,  $\mu, \nu$  can be empirical measures on a point cloud. Consider the case where

$$\mu = \sum_{i=1}^m p_i \delta_{x_i}, \quad \text{and} \quad \nu = \sum_{j=1}^n q_j \delta_{y_j}.$$

In this case, a coupling  $\gamma$  of  $X \sim \mu$  and  $Y \sim \nu$  is characterized by a non-negative matrix  $P \in \mathbb{R}^{m \times n}$  where  $P_{i,j} = \gamma(X = x_i, Y = y_j)$ . The marginal constraints on  $\gamma \in \Gamma_{\mu,\nu}$  readily translate into

$$\forall i \in [m], \sum_{j \in [n]} P_{i,j} = p_i, \quad \text{and} \quad \forall j \in [n], \sum_{i \in [m]} P_{i,j} = q_j.$$

Introducing  $\mathbf{1}_m, \mathbf{1}_n$  for the all-ones vectors of sizes  $m$  and  $n$ , respectively, these constraints can be represented concisely as  $P\mathbf{1}_n = p$ ,  $P^\top \mathbf{1}_m = q$ , where  $p = (p_1, \dots, p_n)^\top$  and  $q = (q_1, \dots, q_m)^\top$ .

Like the coupling, the cost  $c$  can also be captured by an  $m \times n$  matrix  $C$  where  $C_{i,j} = c(x_i, y_j)$  for  $i \in [m]$ ,  $j \in [n]$ . The Kantorovich optimal transport problem (KOT) is therefore equivalent to

$$\min_{P \in \mathbb{R}_+^{m \times n}} \sum_{i,j \in [n]} C_{i,j} P_{i,j} \quad \text{s.t.} \quad P\mathbf{1}_n = p, \quad P^\top \mathbf{1}_m = q,$$

which can also be written more concisely as

$$\min_{P \in \mathbb{R}_+^{m \times n}} \langle C, P \rangle \quad \text{s.t.} \quad P\mathbf{1}_n = p, \quad P^\top \mathbf{1}_m = q,$$

where  $\langle C, P \rangle = \text{tr}(C^\top P)$  is the Frobenius inner product on the set of  $m \times n$  real matrices.

In particular, when  $m = n$  and all of the weights  $p_i, q_j$  are equal to  $1/n$ , the set of valid coupling matrices  $P$  is (a multiple of) the set of doubly stochastic matrices, also known as the *Birkhoff polytope*:

$$\text{Birk} := \{\gamma \in \mathbb{R}_+^{n \times n} : \gamma\mathbf{1}_n = \mathbf{1}_n, \mathbf{1}_n^\top \gamma = \mathbf{1}_n^\top\}. \quad (1.2)$$

Then, (KOT) reduces to

$$\min_{P \in n^{-1}\text{Birk}} \langle C, P \rangle. \quad (1.3)$$

The extreme points of the Birkhoff polytope are permutation matrices: they are binary matrices  $\pi \in \{0, 1\}^{n \times n}$  with exactly one non-zero entry in each row and column. In particular, general principles of convex geometry imply that the solution to any linear program of the form (1.3) can be taken to be a matrix of the form  $n^{-1}\pi$ . A transport plan of this form is induced by a deterministic map (the permutation), and hence in this case there is a solution to the Monge problem. As we shall see in the subsequent sections, extreme points of  $\Gamma_{\mu,\nu}$  play a special role more generally in the geometry of the optimal transport problem.

## 1.2 Wasserstein distances

The Kantorovich problem (KOT) makes sense for a wide variety of cost functions, with different interpretations in each case. For instance, one natural example comes from taking  $c(x, y) = \mathbb{1}_{x \neq y}$  to be the trivial metric. In this case, (KOT) gives:

$$\inf_{\gamma \in \Gamma_{\mu, \nu}} \gamma(X \neq Y)$$

which is a well-known formulation of the total variation distance. (See Exercise 9.) Note, however, that the trivial distance  $\mathbb{1}_{x \neq y}$  is unrelated to the geometry of  $\mathbb{R}^d$ . In particular, it does not say whether  $x$  and  $y$  are far from each other but only if they are different. This limitation manifests itself in the total variation. Indeed, if  $\mu = \delta_x$  and  $\nu = \delta_y$ , then the objective of (KOT) is equal to 1 as soon as  $x \neq y$ .

To obtain a geometrically meaningful quantity from the Kantorovich problem, we need to choose a cost that reflects the actual distance between  $x$  and  $y$ . This idea gives rise to the *Wasserstein distances*. For any  $p \geq 1$ , let  $\mathcal{P}_p(\mathbb{R}^d)$  be the set of probability measures over  $\mathbb{R}^d$  equipped with the Euclidean norm  $\|\cdot\|$  that have finite  $p$ -th moment:

$$\mu \in \mathcal{P}_p(\mathbb{R}^d) \Leftrightarrow \int \|x\|^p \mu(dx) < \infty.$$

The  $p$ -Wasserstein distance between two probability measures  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  is defined by

$$W_p(\mu, \nu) = \inf_{\gamma \in \Gamma_{\mu, \nu}} \left( \int \|x - y\|^p \gamma(dx, dy) \right)^{1/p},$$

where we recall that  $\Gamma_{\mu, \nu}$  is the set of couplings between  $\mu$  and  $\nu$ .

We first show that in fact the above infimum is attained. To that end, define

$$I(\gamma) := \int \|x - y\|^p \gamma(dx, dy)$$

and observe that by definition, there exists a sequence  $(\gamma_n)_n$  in  $\Gamma_{\mu, \nu}$  such that  $I(\gamma_n) \rightarrow W_p^p(\mu, \nu)$ . Since  $\Gamma_{\mu, \nu}$  is compact (Proposition 1.1), there is a subsequence of  $(\gamma_n)_n$  which converges to some  $\bar{\gamma} \in \Gamma_{\mu, \nu}$ . By definition  $W_p(\mu, \nu) \leq I(\bar{\gamma})$ . Since  $(x, y) \mapsto \|x - y\|^p$  is unbounded,  $I$  is not continuous, but it is lower semicontinuous, so  $I(\bar{\gamma}) \leq \liminf_{n \rightarrow \infty} I(\gamma_n) = W_p^p(\mu, \nu)$  (part three of the portmanteau theorem, Theorem B.4). Hence

$I(\bar{\gamma}) = W_p^p(\mu, \nu)$ . Note that the only property of the cost function we used in this proof is lower semicontinuity so this argument readily extends to more general costs.

We can therefore adopt the following definition of Wasserstein distances. Note that these distances should really be called Kantorovich–Rubinstein distances but we stick to the modern trend of “Wassersteinification”.

**Definition 1.2.** *The  $p$ -Wasserstein distance between two probability measures  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  is defined by*

$$W_p(\mu, \nu) = \min_{\gamma \in \Gamma_{\mu, \nu}} \left( \int \|x - y\|^p \gamma(dx, dy) \right)^{1/p}.$$

**Proposition 1.3.** *The  $p$ -Wasserstein distance defines a metric over  $\mathcal{P}_p(\mathbb{R}^d)$ , that is for every  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ , it holds*

1.  $W_p(\mu, \nu) \geq 0$
2.  $W_p(\mu, \nu) = W_p(\nu, \mu)$
3.  $W_p(\mu, \nu) = 0$  iff  $\mu = \nu$
4.  $W_p(\mu, \nu) \leq W_p(\mu, \rho) + W_p(\rho, \nu)$  for any  $\rho \in \mathcal{P}_p(\mathbb{R}^d)$ .

*Proof.* Note first that 1. and 2. hold trivially.

We now turn to the proof of 3. If  $\mu = \nu$ , then the measure  $\gamma(dx, dy) = \mu(dx) \delta_x(dy)$  is a valid coupling:  $\gamma \in \Gamma_{\mu, \nu}$ . Concretely,  $\gamma$  is the law of  $(X, X)$  for  $X \sim \mu$ . Therefore

$$0 \leq W_p^p(\mu, \mu) \leq \int \|x - y\|^p \gamma(dx, dy) = \int \|x - x\|^p \mu(dx) = 0.$$

To show the other direction of 3., observe that if  $W_p(\mu, \nu) = 0$ , there exists  $\bar{\gamma} \in \Gamma_{\mu, \nu}$  such that  $(X, Y) \sim \bar{\gamma}$ , and  $X = Y$  almost surely; in particular, they must have the same distribution:  $\mu = \nu$ .

To complete the proof, we check the triangle inequality 4. To that end, we employ the *gluing lemma* (Lemma B.5) which ensures that there exists  $X, Y, Z$  such that  $X \sim \mu$ ,  $Y \sim \nu$ ,  $Z \sim \rho$  and such that  $(X, Z)$  and  $(Z, Y)$  are optimally coupled.

Then

$$\begin{aligned} W_p(\mu, \nu) &\leq (\mathbb{E} \|X - Y\|^p)^{1/p} \\ &= (\mathbb{E} \|X - Z + Z - Y\|^p)^{1/p} \\ &\leq (\mathbb{E} \|X - Z\|^p)^{1/p} + (\mathbb{E} \|Z - Y\|^p)^{1/p} \end{aligned}$$

$$= W_p(\mu, \rho) + W_p(\rho, \nu),$$

where in the first line we used the suboptimality of the coupling  $(X, Y)$ , in the third line we used the triangle inequality for  $L^p$  norms, and in the last line, we used the optimality of the couplings  $(X, Z)$  and  $(Z, Y)$ .  $\square$

*Example 1.4 (Wasserstein distances in simple cases).*

1. Fix  $x, y \in \mathbb{R}^d$ . Then

$$W_p(\delta_x, \delta_y) = \|x - y\|.$$

Therefore,  $(\mathbb{R}^d, \|\cdot\|)$  is isometrically embedded in  $(\mathcal{P}_p(\mathbb{R}^d), W_p)$  via  $x \mapsto \delta_x$ .

2. Fix  $x, y \in \mathbb{R}^d$  and  $0 \leq \lambda, \tau \leq 1$ . Then

$$W_p(\lambda \delta_x + (1 - \lambda) \delta_y, \tau \delta_x + (1 - \tau) \delta_y) = |\lambda - \tau|^{1/p} \|x - y\|.$$

Note that it follows from ordering of the  $L^p$  norms that  $W_p(\mu, \nu) \leq W_q(\mu, \nu)$  whenever  $p \leq q$ . In particular, the smallest of the Wasserstein distances is  $W_1$ .

Wasserstein distances induce a useful topology on random variables: they *metrize weak convergence* on compact spaces; see Appendix B for background. More specifically, a sequence  $(\mu_n)_n$  satisfies  $W_p(\mu_n, \mu) \rightarrow 0$  if and only if it converges weakly to  $\mu$ , denoted  $\mu_n \xrightarrow{\text{w}} \mu$ , and the  $p$ -th moment converges:  $\int \|\cdot\|^p d\mu_n \rightarrow \int \|\cdot\|^p d\mu$ . This “metrization” property can be found in all of the main texts on optimal transport and has often been employed as a justification for the use of Wasserstein distance as opposed to other distances. This is hardly a discriminating feature, however, and many other distances (Lévy–Prokhorov, Fortet–Mourier, etc.) also have this property; see [Vil09b, Chapter 6]. In fact, this folklore result does not do justice to the quantitative meaning of  $W_p(\mu, \nu) \leq \varepsilon$  for some  $\varepsilon$ .

For example, the following statement implies that if two random variables with sufficiently light tails are close in  $p$ -Wasserstein distance for any  $p > 1$ , then all of their moments must also be close. We formalize the assumption that the tails are light by considering *sub-exponential* random variables, that is, random variables  $Z$  satisfying

$$\mathbb{E}e^{|Z|} \leq 2. \tag{1.4}$$

For such random variables, we have the following bound.

**Proposition 1.5.** *Let  $X \sim \mu$  and  $Y \sim \nu$  be two sub-exponential random variables. Then, for any  $p > 1$ , there exists a constant  $C_p > 0$  such that for any integer  $\ell \geq 1$ , it holds*

$$|\mathbb{E}|X|^\ell - \mathbb{E}|Y|^\ell| \leq (C_p \ell)^\ell W_p(\mu, \nu).$$

*Proof.* By convexity of the function  $x \mapsto |x|^\ell$ ,  $\ell \geq 1$ , it holds

$$|X|^\ell - |Y|^\ell \leq \ell |X - Y| (|X|^{\ell-1} \vee |Y|^{\ell-1}).$$

Taking expectation on both sides and applying Hölder's inequality yields for any coupling  $(X, Y)$ ,

$$|\mathbb{E}|X|^\ell - \mathbb{E}|Y|^\ell| \leq \ell (\mathbb{E}|X - Y|^p)^{1/p} (\mathbb{E}(|X|^{\ell-1} \vee |Y|^{\ell-1})^q)^{1/q}, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Taking the optimal coupling between  $X$  and  $Y$  yields

$$|\mathbb{E}|X|^\ell - \mathbb{E}|Y|^\ell| \leq \ell W_p(\mu, \nu) (\mathbb{E}(|X|^{\ell-1} \vee |Y|^{\ell-1})^q)^{1/q}.$$

To conclude, recall that that it is a standard property of sub-exponential random variables [Ver18] that if  $Z$  is sub-exponential, then  $(\mathbb{E}[|Z|^k])^{1/k} \leq k$  for all  $k \geq 1$ . Thus

$$(\mathbb{E}(|X|^{\ell-1} \vee |Y|^{\ell-1})^q)^{\frac{1}{(\ell-1)q}} \leq 2(\ell-1)q \leq 2 \frac{\ell p}{p-1}.$$

□

The above result is encouraging: obtaining bounds on the Wasserstein distance between two measures implies quantitative bounds on the distance between their moments. It could be the case, though, that the Wasserstein distance tends to be quite large compared to other commonly used distances or divergences such as total variation or the Kullback–Leibler divergence; see [Tsy09, Chapter 2] for a list of such distances, their comparison, and relevance to statistical problems.

It turns out, however, that the Wasserstein distance can often be controlled by other commonly used distances. For example, the next result shows that on a bounded domain, the Wasserstein distance is dominated by the total variation distance (see Exercise 9 for background). Moreover, its proof is our first illustration of how to bound Wasserstein distances—it suffices to exhibit a (suboptimal) coupling  $\gamma$  such that  $\mathbb{E}_\gamma \|X - Y\|^p$  is controlled appropriately.

**Theorem 1.6.** Let  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  be two distributions with densities  $f$  and  $g$  respectively. Then, for any  $p \geq 1$ , it holds

$$W_p^p(\mu, \nu) \leq 2^{p-1} \inf_{x_0 \in \mathbb{R}^d} \int \|x - x_0\|^p |f(x) - g(x)| \, dx.$$

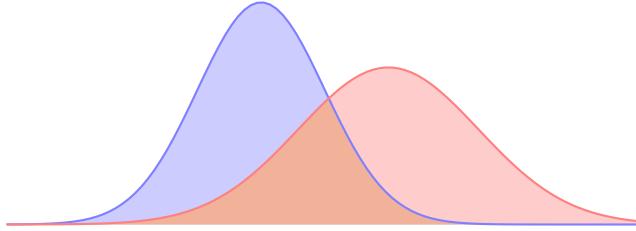
In particular, if the supports of both  $\mu$  and  $\nu$  are included in the same ball of diameter  $D$ , then

$$W_p^p(\mu, \nu) \leq D^p d_{\text{TV}}(\mu, \nu),$$

where  $d_{\text{TV}}(\mu, \nu)$  is the total variation distance between  $\mu$  and  $\nu$  and is defined by

$$d_{\text{TV}}(\mu, \nu) = \frac{1}{2} \int |f(x) - g(x)| \, dx.$$

*Proof.* Assume that  $\mu \neq \nu$  as otherwise the statement is trivial. As mentioned before the statement of the theorem, we construct an explicit coupling between  $\mu$  and  $\nu$ . To that end, consider the three positive functions  $(f - g)_+$ ,  $(f - g)_-$ , and  $f \wedge g$  (see Figure 1.3 for reference) and observe that:



**Fig. 1.3.** The integrals  $\int (f - g)_+$ ,  $\int (f - g)_-$ , and  $\int f \wedge g$  correspond to the blue, red, and orange regions, respectively.

$$\int ((f - g)_+ - (f - g)_-) = \int f - \int g = 1 - 1 = 0$$

so that

$$\int (f - g)_+ = \int (f - g)_- =: t > 0$$

and

$$\int f \wedge g = \frac{1}{2} \left( \int f + \int g - \int (f - g)_+ - \int (f - g)_- \right) = 1 - t.$$

Next, we normalize these functions to obtain three densities

$$h_+ = \frac{1}{t} (f - g)_+, \quad h_- = \frac{1}{t} (f - g)_-, \quad h_\wedge = \frac{1}{1-t} f \wedge g.$$

We can rewrite  $f$  and  $g$  as mixtures of the above densities:

$$f = th_+ + (1-t)h_\wedge, \quad g = th_- + (1-t)h_\wedge.$$

Next, let  $Z_+$ ,  $Z_-$ , and  $Z_\wedge$  be three independent random variables with densities  $h_+$ ,  $h_-$ , and  $h_\wedge$  respectively and let  $B$  be a Bernoulli random variable with parameter  $t \in (0, 1]$ , independent of  $Z_+$ ,  $Z_-$ , and  $Z_\wedge$ .

We are now in a position to define our coupling between  $\mu$  and  $\nu$ . To that end, let  $(X, Y)$  be a random pair such that

$$\begin{aligned} X &= BZ_+ + (1-B)Z_\wedge, \\ Y &= BZ_- + (1-B)Z_\wedge, \end{aligned}$$

and observe that the distribution  $\gamma$  of  $(X, Y)$  is indeed a valid coupling between  $\mu$  and  $\nu$ . Using this fact together with the inequality  $\|x - y\|^p \leq 2^{p-1} (\|x - x_0\|^p + \|y - x_0\|^p)$ , we get

$$\begin{aligned} W_p^p(\mu, \nu) &\leq \mathbb{E}_\gamma \|X - Y\|^p \\ &= \mathbb{P}(B = 0) \cdot 0 + \mathbb{P}(B = 1) \int \|x - y\|^p h_+(x) h_-(y) dx dy \\ &\leq t 2^{p-1} \left( \int \|x - x_0\|^p h_+(x) dx + \int \|y - x_0\|^p h_-(y) dy \right) \\ &= t 2^{p-1} \left( \int \|x - x_0\|^p (h_+(x) + h_-(x)) dx \right) \\ &= 2^{p-1} \left( \int \|x - x_0\|^p ((f - g)_+(x) + (f - g)_-(x)) dx \right) \\ &= 2^{p-1} \left( \int \|x - x_0\|^p |f(x) - g(x)| dx \right) \end{aligned}$$

and the result follows by minimizing the right-hand side with respect to  $x_0$ . The second statement follows easily by taking  $x_0$  to be the center of said ball.  $\square$

The assumption that  $\mu$  and  $\nu$  are absolutely continuous is superfluous and the exact same proof follows by manipulating measures rather than densities, albeit with slightly more opaque notation; see [Vil09b, Theorem 6.15].

### 1.3 Optimal transport in one dimension

To gain a bit of insight into optimal transport, we look at the simpler case where  $\mu$  and  $\nu$  are probability measures on the real line. In this case, we may define their associated cumulative distribution functions.

Recall that the *cumulative distribution function* (CDF) of a random variables  $Z$  is the function  $F : \mathbb{R} \rightarrow [0, 1]$  defined by

$$F(t) := \mathbb{P}(Z \leq t), \quad t \in \mathbb{R}.$$

Since  $F$  is monotonically non-decreasing, we may define its *pseudo-inverse*  $F^\dagger$  by

$$F^\dagger(u) = \inf\{t \in \mathbb{R} : F(t) \geq u\}, \quad u \in [0, 1],$$

with the convention that  $\inf \emptyset = \infty$ . While  $F^\dagger$  is not an inverse per se, it does satisfy the following property:

$$F^\dagger(u) \leq t \Leftrightarrow u \leq F(t) \tag{1.5}$$

This property is often used to simulate random variables. Let  $U \sim \text{Unif}([0, 1])$  be a uniform random variable, then  $Z \sim F^\dagger(U)$  has CDF  $F$ . Indeed, for any  $t \in \mathbb{R}$ ,

$$\mathbb{P}(Z \leq t) = \mathbb{P}(F^\dagger(U) \leq t) = \mathbb{P}(U \leq F(t)) = F(t). \tag{1.6}$$

The following theorem characterizes optimal transport in one dimension in terms of CDFs.

**Theorem 1.7.** *Let  $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$  be two probability distributions with CDFs  $F_\mu$  and  $F_\nu$  respectively. Let  $U \sim \text{Unif}([0, 1])$  be a uniform random variable and denote by  $\bar{\gamma}$  the distribution of  $(F_\mu^\dagger(U), F_\nu^\dagger(U))$ . Then  $\bar{\gamma} \in \Gamma_{\mu, \nu}$  is a valid coupling between  $\mu$  and  $\nu$  and it is optimal:*

$$W_1(\mu, \nu) = \int |x - y| \bar{\gamma}(dx, dy) = \min_{\gamma \in \Gamma_{\mu, \nu}} \int |x - y| \gamma(dx, dy).$$

Moreover,

$$W_1(\mu, \nu) = \int_{-\infty}^{\infty} |F_\mu(t) - F_\nu(t)| dt.$$

*Proof.* It follows from (1.6) that  $\bar{\gamma} \in \Gamma_{\mu,\nu}$  and it remains to check that it is optimal. To that end, observe that for any  $\gamma \in \Gamma_{\mu,\nu}$ , it follows from Fubini's theorem that for  $(X, Y) \sim \gamma$ ,

$$\begin{aligned} \int |x - y| \gamma(dx, dy) &= \iint_{-\infty}^{\infty} (\mathbb{1}_{x \leq t < y} + \mathbb{1}_{y \leq t < x}) dt \gamma(dx, dy) \\ &= \int_{-\infty}^{\infty} (\gamma(X \leq t < Y) + \gamma(Y \leq t < X)) dt \\ &= \int_{-\infty}^{\infty} (\gamma(X \leq t) + \gamma(Y \leq t) - 2\gamma(X \leq t, Y \leq t)) dt \\ &\geq \int_{-\infty}^{\infty} (F_{\mu}(t) + F_{\nu}(t) - 2(F_{\mu}(t) \wedge F_{\nu}(t))) dt \\ &= \int_{-\infty}^{\infty} |F_{\mu}(t) - F_{\nu}(t)| dt. \end{aligned}$$

To show that the above inequality becomes an equality when  $\gamma = \bar{\gamma}$ , observe that

$$\begin{aligned} \bar{\gamma}(X \leq t, Y \leq t) &= \mathbb{P}(F_{\mu}^{\dagger}(U) \leq t, F_{\nu}^{\dagger}(U) \leq t) \\ &= \mathbb{P}(U \leq F_{\mu}(t), U \leq F_{\nu}(t)) \\ &= \mathbb{P}(U \leq F_{\mu}(t) \wedge F_{\nu}(t)) \\ &= F_{\mu}(t) \wedge F_{\nu}(t). \end{aligned}$$

We have proved

$$\int |x - y| \gamma(dx, dy) \geq \int_{-\infty}^{\infty} |F_{\mu}(t) - F_{\nu}(t)| dt = \int |x - y| \bar{\gamma}(dx, dy)$$

so that  $\bar{\gamma}$  is an optimal coupling.  $\square$

If  $Z$  admits a density, then its CDF  $F$  is actually a left inverse of  $F^{\dagger}$ , i.e.,  $F \circ F^{\dagger} = \text{Id}$ . If  $\mu$  has a density, this fact implies that the optimal coupling  $\bar{\gamma}$  takes the following special form. If  $X \sim \mu$ , then

$$(X, F_{\nu}^{\dagger} \circ F_{\mu}(X)) \sim \bar{\gamma}. \quad (1.7)$$

In other words, the solution to the Monge problem and the Kantorovich problem coincide since we have found a transport map  $\bar{T} = F_{\nu}^{-1} \circ F_{\mu}$  such that  $\bar{T}_{\#}\mu = \nu$  and

$$\int |x - \bar{T}(x)| \mu(dx) = \min_{\gamma \in \Gamma_{\mu,\nu}} \int |x - y| \gamma(dx, dy)$$

$$= \min_{T: T_\# \mu = \nu} \int |x - T(x)| \mu(dx).$$

Although we have focused on the  $W_1$  distance in this section, the coupling  $\bar{\gamma}$  given in (1.7) turns out to be universally optimal, in the sense that it is optimal for the Kantorovich problem for any strictly convex cost (a cost of the form  $c(x, y) = h(x - y)$  where  $h: \mathbb{R} \rightarrow \mathbb{R}$  is strictly convex); this includes all  $W_p$  distances for  $p > 1$ . See Exercise 8.

Note that  $\bar{T}$  is a *monotone* increasing function as the composition of two increasing functions. Continuous monotone increasing functions in one dimension are derivatives of convex functions, suggesting that this property may be generalized to higher dimension by considering gradients of convex functions. Existence of such monotone transport maps in higher dimensions is the content of the influential result of Brenier [Bre87], which we explore next.

## 1.4 Brenier's theorem

The  $p$ -Wasserstein distance is a natural object for any  $p \geq 1$ . However, the cases  $p = 1, 2$  possess remarkable special structure, and we focus on them in much of what follows. We first explore the case  $p = 2$ , which is notable for its close connection to convex analysis.

Recall that

$$W_2^2(\mu, \nu) = \min_{\gamma \in \Gamma_{\mu, \nu}} \int \|x - y\|^2 \gamma(dx, dy). \quad (W_2^2)$$

**Theorem 1.8 (Brenier).** *Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  be two probability measures such that  $\mu$  has a density and let  $X \sim \mu$ . If  $\bar{\gamma}$  is an optimal coupling for  $(W_2^2)$ ,*

$$\int \|x - y\|^2 \bar{\gamma}(dx, dy) = \min_{\gamma \in \Gamma_{\mu, \nu}} \int \|x - y\|^2 \gamma(dx, dy) = W_2^2(\mu, \nu),$$

*then there exists a convex function  $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $(X, \nabla \varphi(X)) \sim \bar{\gamma} \in \Gamma_{\mu, \nu}$ .*

The crucial assumption of Brenier's theorem is that  $\mu$  has a density, that is, that  $\mu$  is absolutely continuous with respect to the Lebesgue measure. In particular, since convex functions are differentiable Lebesgue-almost everywhere, this assumption guarantees that  $\nabla \varphi(X)$  makes sense.

Before turning to the proof, we first consider a first statistical implication of Brenier's theorem. Brenier's theorem asserts that, as long as  $\mu$  has a density, for any  $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ , there exists a convex function  $\phi$  so that  $\nabla\phi_{\#}\mu = \nu$ . Since gradients of convex functions are natural analogues of monotone functions in higher dimensions, this theorem is therefore a significant generalization of the classical univariate fact mentioned in Section 1.3, that if  $U \sim \text{Unif}([0, 1])$ , then  $F_{\nu}^{\dagger}(U) \sim \nu$ .

In one dimension, the function  $F_{\nu}^{\dagger}$  is known as the quantile function of  $\nu$ , and is of fundamental statistical significance. Brenier's theorem therefore can be used to define a multivariate notion of quantiles [CGHH17, HdBCAM21]. This point of view has proven to be extremely fruitful and has led to a wide range of statistical applications [Hal22]. (For more details, see the discussion section.)

Returning to the content of Brenier's theorem, at this point it is not obvious what optimal transport has to do with gradients of convex functions. We therefore begin by studying such gradients to gain intuition.

#### 1.4.1 Gradients of convex functions

Note first that a continuous function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is such that  $f = \varphi'$  for some differentiable convex function  $\varphi$  if and only if  $f$  is non-decreasing. Indeed, convexity of  $\varphi$  implies that for any  $x, y \in \mathbb{R}$ :

$$\varphi(x) - \varphi(y) \leq (x - y) \varphi'(x), \quad (1.8)$$

$$\varphi(y) - \varphi(x) \leq (y - x) \varphi'(y). \quad (1.9)$$

Summing the above two inequalities yields

$$(x - y) (\varphi'(x) - \varphi'(y)) \geq 0,$$

so that  $\varphi'$  is non-decreasing.

Is there an analogue of this statement for functions on  $\mathbb{R}^d$ ? Of course we immediately get that for any  $x, y \in \mathbb{R}^d$

$$\langle x - y, \tilde{\nabla}\varphi(x) - \tilde{\nabla}\varphi(y) \rangle \geq 0,$$

where  $\tilde{\nabla}\varphi(x) \in \partial\varphi(x)$  denotes a subgradient of  $\varphi$  at  $x$  (see Appendix A for preliminaries on convex analysis). Unfortunately, while in dimension 1, the two-point inequalities (1.8)–(1.9) imply inequalities for any arrangement of points, in higher dimension this is no longer the case and we need to capture additional information.

In fact, convexity implies many such inequalities: for any integer  $k \geq 2$ , and any collection of points  $x_1, \dots, x_k \in \mathbb{R}^d$ , we have

$$\begin{aligned}\varphi(x_i) - \varphi(x_{i+1}) &\leq \langle x_i - x_{i+1}, \tilde{\nabla} \varphi(x_i) \rangle, \quad i = 1, \dots, k-1, \\ \varphi(x_k) - \varphi(x_1) &\leq \langle x_k - x_1, \tilde{\nabla} \varphi(x_k) \rangle.\end{aligned}$$

Summing these inequalities yields:

$$\sum_{i=1}^k \langle x_i - x_{i+1}, \tilde{\nabla} \varphi(x_i) \rangle \geq 0, \quad (1.10)$$

with the convention that  $x_{k+1} = x_1$ .<sup>1</sup>

### 1.4.2 Cyclical monotonicity

Since there may exist several points in the subdifferential of  $\varphi$  at  $x$ , we first describe the graph  $\{(x, \tilde{\nabla} \varphi(x)) : x \in \mathbb{R}^d\}$  before thinking about  $\tilde{\nabla} \varphi(\cdot)$  as a map from  $\mathbb{R}^d$  to  $\mathbb{R}^d$ . We first define an important property of such graphs.

**Definition 1.9.** A set  $A \subset \mathbb{R}^d \times \mathbb{R}^d$  is said to be *cyclically monotone* if for any integer  $k \geq 2$ , and points  $(x_i, y_i) \in A$ ,  $i = 1, \dots, k$ , it holds

$$\sum_{i=1}^k \langle x_i - x_{i+1}, y_i \rangle \geq 0, \quad (1.11)$$

with the convention that  $x_{k+1} = x_1$ .

In light of (1.10), the set  $\partial \varphi \subset \mathbb{R}^d \times \mathbb{R}^d$  is cyclically monotone whenever  $\varphi$  is convex. It turns out that all cyclically monotone subsets of  $\mathbb{R}^d \times \mathbb{R}^d$  are of this form.

**Theorem 1.10 (Rockafellar).** A set  $A \subset \mathbb{R}^d \times \mathbb{R}^d$  is cyclically monotone if and only if there exists a closed convex function  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  such that

$$A \subseteq \partial \varphi.$$

The proof of this classical theorem of convex analysis can be found in Appendix A.

Note that condition (1.11) is equivalent to the requirement that

---

<sup>1</sup> With this convention, the points  $x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_k \rightarrow x_1$  form a “cycle.”

$$\sum_{i=1}^k \|x_i - y_i\|^2 \leq \sum_{i=1}^k \|x_{i+1} - y_i\|^2 \quad (1.12)$$

for any points  $(x_i, y_i) \in A$ ,  $i = 1, \dots, k$ . This formulation enables us to see the connection with optimal transport. Indeed, consider the following example for illustration purposes. Let

$$\mu = \frac{1}{n} \sum_{j=1}^n \delta_{a_j}, \quad \nu = \frac{1}{n} \sum_{j=1}^n \delta_{b_j}.$$

In this discrete case, the set of all couplings between  $\mu$  and  $\nu$  can be identified with the Birkhoff polytope (see Section 1.1.3), whose extreme points are rescaled permutation matrices. Since the discrete optimal transport problem is a linear program, solutions can be taken to be extreme points. We may therefore restrict our attention to couplings given by permutations. Concretely, a permutation  $\sigma$  of  $\{1, \dots, n\}$  corresponds to the coupling

$$\gamma = \frac{1}{n} \sum_{j=1}^n \delta_{(a_j, b_{\sigma(j)})}.$$

Such a coupling is optimal if its cost is minimal among all permutations, that is, if

$$\sum_{j=1}^n \|a_j - b_{\sigma(j)}\|^2 \leq \sum_{j=1}^n \|a_j - b_{\tau(j)}\|^2, \quad \forall \tau. \quad (1.13)$$

This condition is precisely equivalent to the support of  $\gamma$  being cyclically monotone. Indeed, by relabeling the atoms of  $\nu$ , we may assume without loss of generality that  $\sigma$  is the identity permutation. Then the support of  $\gamma$  consists of the pairs  $(a_j, b_j)$ ,  $j \in \{1, \dots, n\}$ . Given any subset of  $k$  distinct points  $(x_i, y_i) = (a_{j_i}, b_{j_i}) \in \text{supp}(\gamma)$ ,  $i = 1, \dots, k$ , let  $\tau$  be the cyclic permutation of  $\{j_1, \dots, j_k\}$  that leaves other indices unchanged

$$\tau(j) = \begin{cases} j_{i-1} & \text{if } j = j_i, i \in \{2, \dots, k\} \\ j_k & \text{if } j = j_1 \\ j & \text{otherwise.} \end{cases}$$

Then (1.13) implies (1.12). In fact, since *any* permutation  $\tau$  can be decomposed into cycles, similar reasoning then shows that (1.12) is also a sufficient condition for (1.13) to hold.

The preceding discussion indicates that, in the discrete case, the support of an optimal coupling is cyclically monotone. A similar phenomenon holds in the general case; however, the argument given above is not valid when  $\gamma$  does not assign positive mass to points in its support. Nevertheless, the following result shows that a similar strategy can be made to work by reasoning about small neighborhoods (e.g., balls) rather than individual points. Some care is required to ensure that it is possible to modify  $\gamma$  on such neighborhoods while maintaining the constraint  $\gamma \in \Gamma_{\mu,\nu}$ .

**Proposition 1.11.** *Let  $\bar{\gamma} \in \Gamma_{\mu,\nu}$  be an optimal coupling between  $\mu$  and  $\nu$  in the sense that*

$$\int \|x - y\|^2 \bar{\gamma}(dx, dy) = \min_{\gamma \in \Gamma_{\mu,\nu}} \int \|x - y\|^2 \gamma(dx, dy) = W_2^2(\mu, \nu).$$

*Then  $\text{supp}(\bar{\gamma})$  is cyclically monotone.*

*Proof.* Suppose that  $S := \text{supp}(\bar{\gamma})$  is *not* cyclically monotone. Then there exists  $k \geq 2$  and  $(x_i, y_i) \in S$ ,  $i = 1, \dots, k$  such that

$$\sum_{i=1}^k \|x_i - y_i\|^2 > \sum_{i=1}^k \|x_{i+1} - y_i\|^2,$$

and by continuity of the Euclidean norm, there exist neighborhoods  $U_i, V_i$  of  $x_i, y_i$  respectively for  $i = 1, \dots, k$  such that  $\bar{\gamma}(U_i \times V_i) > 0$  and

$$\sum_{i=1}^k \|\tilde{x}_i - \tilde{y}_i\|^2 > \sum_{i=1}^k \|\tilde{x}'_{i+1} - \tilde{y}'_i\|^2, \quad (1.14)$$

for all  $\tilde{x}_i, \tilde{x}'_i \in U_i, \tilde{y}_i, \tilde{y}'_i \in V_i$ ,  $i = 1, \dots, k$ .

Now, let  $\gamma_i, i = 1, \dots, k$  be a family of (conditional) probability distributions on  $\mathbb{R}^d \times \mathbb{R}^d$  defined such that  $\gamma_i(A) = \bar{\gamma}(A \mid U_i \times V_i)$  for any Borel set  $A \subset \mathbb{R}^d \times \mathbb{R}^d$ . Next, let  $\gamma_i^{(1)}$  and  $\gamma_i^{(2)}$  denote the first and second marginal of  $\gamma_i$  respectively and define the mixture:

$$\gamma = \bar{\gamma} + \frac{c}{k} \sum_{i=1}^k (\gamma_{i+1}^{(1)} \otimes \gamma_i^{(2)} - \gamma_i),$$

where  $c > 0$  is to be chosen later and with the convention that  $\gamma_{k+1} = \gamma_1$ .

Note that for any Borel set  $A \subset \mathbb{R}^d \times \mathbb{R}^d$ , it holds

$$\begin{aligned}
\gamma(A) &\geq \bar{\gamma}(A) - \frac{c}{k} \sum_{i=1}^k \gamma_i(A) \\
&= \bar{\gamma}(A) - \frac{c}{k} \sum_{i=1}^k \bar{\gamma}(A \mid U_i \times V_i) \\
&= \bar{\gamma}(A) - \frac{c}{k} \sum_{i=1}^k \frac{\bar{\gamma}(A \cap (U_i \times V_i))}{\bar{\gamma}(U_i \times V_i)} \\
&\geq \bar{\gamma}(A) - \frac{c\bar{\gamma}(A)}{k} \sum_{i=1}^k \frac{1}{\bar{\gamma}(U_i \times V_i)}.
\end{aligned}$$

Thus  $\gamma(A) \geq 0$  if  $c < \min_{i \in [k]} \bar{\gamma}(U_i \times V_i)$ . Moreover,  $\gamma(\mathbb{R}^d \times \mathbb{R}^d) = 1$  so that  $\gamma$  is indeed a probability distribution over  $\mathbb{R}^d \times \mathbb{R}^d$ .

To check that  $\gamma \in \Gamma_{\mu, \nu}$  observe that for any Borel set  $B \subset \mathbb{R}^d$ ,

$$\begin{aligned}
\gamma(B \times \mathbb{R}^d) &= \mu(B) + \frac{c}{k} \sum_{i=1}^k (\gamma_{i+1}^{(1)}(B) - \gamma_i(B \times \mathbb{R}^d)) \\
&= \mu(B) + \frac{c}{k} \sum_{i=1}^k (\gamma_{i+1}^{(1)}(B) - \gamma_i^{(1)}(B)) \\
&= \mu(B) + \frac{c}{k} (\gamma_{k+1}^{(1)}(B) - \gamma_1^{(1)}(B)) = \mu(B).
\end{aligned}$$

Similarly

$$\begin{aligned}
\gamma(\mathbb{R}^d \times B) &= \nu(B) + \frac{c}{k} \sum_{i=1}^k (\gamma_i^{(2)}(B) - \gamma_i(\mathbb{R}^d \times B)) \\
&= \nu(B) + \frac{c}{k} \sum_{i=1}^k (\gamma_i^{(2)}(B) - \gamma_i^{(2)}(B)) = \nu(B).
\end{aligned}$$

Next observe that

$$\begin{aligned}
&\int \|x - y\|^2 \gamma(dx, dy) - \int \|x - y\|^2 \bar{\gamma}(dx, dy) \\
&= \frac{c}{k} \sum_{i=1}^k \left( \int_{U_{i+1} \times V_i} \|x - y\|^2 \gamma_{i+1}^{(1)}(dx) \gamma_i^{(2)}(dy) \right. \\
&\quad \left. - \int_{U_i \times V_i} \|x - y\|^2 \gamma_i(dx, dy) \right)
\end{aligned}$$

$$< 0,$$

by (1.14). This contradicts optimality of  $\bar{\gamma}$ .  $\square$

### 1.4.3 Proof of Brenier's theorem

We are now in a position to prove Brenier's theorem.

Let  $\bar{\gamma}$  be an optimal coupling. In light of Proposition 1.11,  $\text{supp}(\bar{\gamma})$  is cyclically monotone. By Rockafeller's Theorem 1.10, this implies that there exists a convex function  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  such that  $\bar{\gamma}(Y \in \partial\varphi(X)) = 1$ . But since  $\varphi$  is convex, it is locally Lipschitz on the interior of its domain, and the boundary of the domain has measure zero by convexity. Hence,  $\varphi$  is almost everywhere differentiable with respect to the Lebesgue measure over its domain by Rademacher's theorem. Since  $\mu$  has a density, this implies that  $\varphi$  is differentiable  $\mu$  almost everywhere. Therefore,  $\bar{\gamma}(Y = \nabla\varphi(X)) = 1$  or in other words, if  $X \sim \mu$ , then  $(X, \nabla\varphi(X)) \sim \bar{\gamma}$ .

## 1.5 Kantorovich duality

Brenier's theorem shows that an optimal coupling for  $(W_2^2)$  if  $\mu$  has a density is a deterministic coupling given by the gradient of a convex function  $\varphi$ . This result raises the question of whether it is possible to solve an optimization problem to find  $\varphi$  directly, or whether it is possible to certify that a convex function  $\varphi$  corresponds to an optimal coupling. These questions can be answered by employing tools from convex duality.

In the fully discrete setting (see Section 1.1.3),  $(W_2^2)$  is a linear program or LP (linear objective & linear constraints), which admits a useful theory of duality. This intuition carries over to the general setting (and the link can be made precise through approximation arguments, see [Dud02, Chapter 11]). In fact, it was through optimal transport that Kantorovich actually introduced LP duality, which has furnished algorithmic advances continuously since its inception.

### 1.5.1 The dual Kantorovich problem

The dual problem to  $(W_2^2)$  is a maximization problem. To find its expression, encode the constraint  $\gamma \in \Gamma_{\mu, \nu}$  as

$$\begin{aligned} & \sup_{f,g \in C_b} \left\{ \int f(x) \mu(dx) + \int g(y) \nu(dy) - \int (f(x) + g(y)) \gamma(dx, dy) \right\} \\ &= \begin{cases} 0, & \text{if } \gamma \in \Gamma_{\mu,\nu}, \\ \infty, & \text{otherwise,} \end{cases} \end{aligned}$$

where the supremum is taken over the set  $C_b$  of continuous and bounded functions over  $\mathbb{R}^d$ . Thus,  $(W_2^2)$  is equivalent to

$$\begin{aligned} & \inf_{\gamma \in \mathcal{M}_+} \left\{ \int \|x - y\|^2 \gamma(dx, dy) \right. \\ & \quad \left. + \sup_{f,g \in C_b} \int f(x) \mu(dx) + \int g(y) \nu(dy) - \int (f(x) + g(y)) \gamma(dx, dy) \right\}, \end{aligned}$$

where the infimum is taken over the set  $\mathcal{M}_+$  of all positive measures on  $\mathbb{R}^d \times \mathbb{R}^d$  (unrestricted). Note that for  $\gamma \notin \Gamma_{\mu,\nu}$  this new objective is infinite so the problem is strictly equivalent.

Next, we switch the inf and sup to get the following lower bound on the value of  $(W_2^2)$ :

$$\begin{aligned} & \sup_{f,g \in C_b} \left\{ \int f(x) \mu(dx) + \int g(y) \nu(dy) \right. \\ & \quad \left. + \inf_{\gamma \in \mathcal{M}_+} \left\{ \int (\|x - y\|^2 - f(x) - g(y)) \gamma(dx, dy) \right\} \right\} \end{aligned} \tag{1.15}$$

Next observe that since  $\gamma$  is a positive measure, it holds,

$$\begin{aligned} & \inf_{\gamma \in \mathcal{M}_+} \left\{ \int (\|x - y\|^2 - f(x) - g(y)) \gamma(dx, dy) \right\} \\ &= \begin{cases} 0, & \text{if } f(x) + g(y) \leq \|x - y\|^2, \forall x, y \in \mathbb{R}^d, \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

Indeed, if there exists a pair  $(x, y)$  that violates the above constraint then we can take the sequence of measures  $\gamma_n = n\delta_{(x,y)}$  and the integral would converge to  $-\infty$ .

Hence we have shown that  $(W_2^2)$  is bounded below by

$$\sup_{\substack{f,g \in C_b \\ f(x) + g(y) \leq \|x - y\|^2}} \left\{ \int f d\mu + \int g d\nu \right\}.$$

Though the choice  $f, g \in C_b$  above is motivated by reference to the weak topology on the set of probability measures, this lower bound

on  $W_2^2(\mu, \nu)$  holds for any pair  $(f, g)$  of integrable functions satisfying  $f(x) + g(y) \leq \|x - y\|^2$  almost everywhere.

**Lemma 1.12.** *Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ , then*

$$\begin{aligned} W_2^2(\mu, \nu) &= \inf_{\gamma \in \Gamma_{\mu, \nu}} \int \|x - y\|^2 \gamma(dx, dy) \\ &\geq \sup_{\substack{f \in L^1(\mu), g \in L^1(\nu) \\ f(x) + g(y) \leq \|x - y\|^2}} \left\{ \int f d\mu + \int g d\nu \right\}. \end{aligned}$$

*Proof.* Let  $f \in L^1(\mu), g \in L^1(\nu)$  be such that  $f(x) + g(y) \leq \|x - y\|^2$  for  $\mu$ -a.e.  $x$ ,  $\nu$ -a.e.  $y$ , and fix  $\gamma \in \Gamma_{\mu, \nu}$ . Then

$$\begin{aligned} \int f(x) \mu(dx) + \int g(y) \nu(dy) &= \int (f(x) + g(y)) \gamma(dx, dy) \\ &\leq \int \|x - y\|^2 \gamma(dx, dy). \end{aligned}$$

The proof follows by taking the supremum on the left-hand side and the infimum on the right-hand side.  $\square$

The dual Kantorovich problem is given by

$$\sup_{\substack{f \in L^1(\mu), g \in L^1(\nu) \\ f(x) + g(y) \leq \|x - y\|^2}} \left\{ \int f d\mu + \int g d\nu \right\}. \quad (\text{D-}W_2^2)$$

It is the dual problem to the *primal* problem  $(W_2^2)$ .

In particular, Lemma 1.12 describes a phenomenon known as *weak* duality, in which the dual is only shown to be a lower bound on the primal problem. This terminology is to be contrasted with *strong* duality, where the inequality becomes an equality so that the primal and dual objectives take the same optimal value. While strong duality is, strictly speaking, only a statement about objective values, it is often the case that the solutions to the primal and dual problems are related to each other; see [BV04, Chapter 5] for a treatment of duality in the context of convex optimization.

We show in Subsection 1.5.3 that strong duality in fact holds and it leads to important consequences for our problem of interest.

### 1.5.2 The semidual

Before moving to strong duality, we make a quick detour to define the *semidual* problem, a partially solved version of the dual problem  $(\mathbf{D}\text{-}\mathbf{W}_2^2)$ . It plays an important role in the estimation of optimal transport maps (Chapter 3).

Consider  $(\mathbf{D}\text{-}\mathbf{W}_2^2)$  and suppose that we hold the first dual potential  $f$  fixed; given this choice of  $f$ , what is the optimal choice of  $g$ ? Since the dual problem is a supremum, we want to make  $g$  as large as possible, but we must respect the constraint  $f(x) + g(y) \leq \|x - y\|^2$ . The optimal function  $g$  is therefore given by

$$g(y) = \inf_{x \in \mathbb{R}^d} \{\|x - y\|^2 - f(x)\}. \quad (1.16)$$

The function defined in (1.16) is called the *c-conjugate* or *c-transform* of  $f$ , denoted  $f^c$ , associated with the cost  $c(x, y) = \|x - y\|^2$ . This reasoning shows that we can reformulate the dual as

$$(\mathbf{D}\text{-}\mathbf{W}_2^2) = \sup_{f \in L^1(\mu)} \left\{ \int f \, d\mu + \int f^c \, d\nu \right\}. \quad (1.17)$$

This is a version of the *semidual* problem, and it is applicable to optimal transport for any cost function  $c$  provided that we replace  $\|x - y\|^2$  with  $c(x, y)$  in (1.16).

However, for the quadratic cost, we can go one step further and explicitly link the semidual with convex analysis. In this case, the semidual is given by

$$\inf_{\phi \in L^1(\mu)} \left\{ \int \phi \, d\mu + \int \phi^* \, d\nu \right\} \quad (\mathbf{SD})$$

where  $\phi^*$  denotes the *convex conjugate* of  $\phi$ ; see Appendix A.

**Proposition 1.13.** *Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  be probability measures. Then, the dual problem  $(\mathbf{D}\text{-}\mathbf{W}_2^2)$  is equivalent to the semidual problem  $(\mathbf{SD})$  in the following sense:*

1. **Objective values:** Write  $\mathbf{S}$  and  $\mathbf{D}$  for the optimal objective values of  $(\mathbf{SD})$  and  $(\mathbf{D}\text{-}\mathbf{W}_2^2)$  respectively. Then

$$\mathbf{D} = \int \|\cdot\|^2 \, d\mu + \int \|\cdot\|^2 \, d\nu - 2 \cdot \mathbf{S}.$$

**2. Solutions:** A pair of functions  $(f, g)$  is optimal for  $(\mathbf{D-W}_2^2)$  if and only if  $f = \|\cdot\|^2 - 2\varphi$  and  $g = \|\cdot\|^2 - 2\varphi^*$  where  $\varphi$  is optimal for  $(\mathbf{SD})$ .

*Proof.* Let us write  $f = \|\cdot\|^2 - 2\phi$  and  $g = \|\cdot\|^2 - 2\psi$ ; this is simply a reparametrization of the dual potentials. Then,

$$\int f \, d\mu + \int g \, d\nu = \int \|\cdot\|^2 \, d\mu + \int \|\cdot\|^2 \, d\nu - 2 \left( \int \phi \, d\mu + \int \psi \, d\nu \right).$$

The constraint  $f(x) + g(y) \leq \|x - y\|^2$  translates into

$$\|x\|^2 - 2\phi(x) + \|y\|^2 - 2\psi(y) \leq \|x - y\|^2 \Leftrightarrow \phi(x) + \psi(y) \geq \langle x, y \rangle.$$

Hence,  $(\mathbf{D-W}_2^2)$  is equivalent to

$$\inf_{\substack{\phi \in L^1(\mu), \psi \in L^1(\nu) \\ \phi(x) + \psi(y) \geq \langle x, y \rangle}} \left\{ \int \phi \, d\mu + \int \psi \, d\nu \right\}.$$

Next, let us apply the same trick as described above: for fixed  $\phi$ , the optimal choice of  $\psi$  obeying the constraint is given by

$$\psi(y) = \sup_{x \in \mathbb{R}^d} \{ \langle x, y \rangle - \phi(x) \},$$

which is precisely the definition of the convex conjugate  $\phi^*$ . Substituting this in yields the equivalence.  $\square$

In the preceding proof, we showed that for fixed  $\phi_0$ , the optimal choice of  $\psi$  is  $\psi = \phi_0^*$ . Due to the symmetry of the problem, for fixed  $\psi = \phi_0^*$ , the optimal choice of  $\phi$  is then  $\phi^* = \phi_0^{**}$ . One could imagine iterating this process, obtaining better and better dual potentials, but actually the process halts here. Since  $\phi_0^*$  is a closed convex function, it is self-dual, so that  $\phi_0^{***} = \phi_0^*$ ; see Appendix A. In the end, this argument shows that the optimal potential  $\varphi$  in  $(\mathbf{SD})$  can be taken to be a closed convex function.

Thus far, we have seen two convex functions  $\varphi$  arise from the optimal transport problem. From the primal standpoint, Brenier's Theorem 1.8 shows that the optimal transport plan is supported on the subdifferential of a convex function. From the dual standpoint, a minimizer of  $(\mathbf{SD})$  can be taken to be convex. In the next section, we show that these two convex functions are one and the same.

### 1.5.3 The fundamental theorem of optimal transport

Recall from Brenier's theorem that if a measure  $\mu$  has a density then any optimal coupling between  $\mu$  and  $\nu$  is supported on the graph of the gradient of a convex function. It turns out that the converse holds: any coupling  $\gamma \in \Gamma_{\mu,\nu}$  supported on the graph of the gradient of a convex function has to be optimal. This equivalence follows from the fundamental theorem of optimal transport stated below. In fact, this theorem contains another fundamental result about strong duality between the primal problem  $(W_2^2)$  and its dual  $(D-W_2^2)$  which is the key to establishing this equivalence.

**Theorem 1.14 (Fundamental theorem of optimal transport).** *Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  be two probability measures such that  $\mu$  has a density and let  $X \sim \mu$ . Then the following are equivalent:*

(i)  $\bar{\gamma} \in \Gamma_{\mu,\nu}$  is an optimal coupling in the sense that:

$$\int \|x - y\|^2 \bar{\gamma}(dx, dy) = W_2^2(\mu, \nu).$$

(ii) There exists a proper convex function  $\varphi$  such that  $(X, \nabla \varphi(X)) \sim \bar{\gamma} \in \Gamma_{\mu,\nu}$ .

(iii) Strong duality holds between  $(W_2^2)$  and  $(D-W_2^2)$ :

$$\int \|x - y\|^2 \bar{\gamma}(dx, dy) = \sup_{\substack{f \in L^1(\mu), g \in L^1(\nu) \\ f(x) + g(y) \leq \|x - y\|^2}} \left\{ \int f d\mu + \int g d\nu \right\}.$$

Moreover, the above supremum is achieved for

$$\bar{f}(x) := \|x\|^2 - 2\varphi(x) \quad \text{and} \quad \bar{g}(y) := \|y\|^2 - 2\varphi^*(y).$$

*Proof.* We have already proved that (i)  $\Rightarrow$  (ii) in Subsection 1.4 so it remains to prove that (ii)  $\Rightarrow$  (iii) and (iii)  $\Rightarrow$  (i).

We first prove that (ii)  $\Rightarrow$  (iii). To that end, observe that for  $\mu$  almost every  $x$

$$\|x - \nabla \varphi(x)\|^2 = \|x\|^2 + \|\nabla \varphi(x)\|^2 - 2 \langle x, \nabla \varphi(x) \rangle.$$

Moreover, the convex conjugate  $\varphi^*$  of  $\varphi$  satisfies for  $\mu$  almost every  $x$ ,

$$\varphi(x) + \varphi^*(\nabla \varphi(x)) = \langle \nabla \varphi(x), x \rangle.$$

This is the optimality condition for the Fenchel–Young inequality (Theorem A.6). The above two displays yield

$$\|x - \nabla\varphi(x)\|^2 = \underbrace{\|x\|^2 - 2\varphi(x)}_{\bar{f}(x)} + \underbrace{\|\nabla\varphi(x)\|^2 - 2\varphi^*(\nabla\varphi(x))}_{\bar{g}(\nabla\varphi(x))}.$$

Integrating with respect to  $\mu$  yields

$$\begin{aligned} \int \|x - y\|^2 \bar{\gamma}(dx, dy) &= \int \|x - \nabla\varphi(x)\|^2 \mu(dx) \\ &= \int \bar{f}(x) \mu(dx) + \int \bar{g}(\nabla\varphi(x)) \mu(dx) \\ &= \int \bar{f}(x) \mu(dx) + \int \bar{g}(y) \nu(dy). \end{aligned} \quad (1.18)$$

We now check that the pair  $(\bar{f}, \bar{g})$  satisfies the constraints of **(D-W<sub>2</sub><sup>2</sup>)**. It follows from the Fenchel–Young inequality (Theorem A.6) that for any  $x, y \in \mathbb{R}^d$ ,

$$\begin{aligned} \bar{f}(x) + \bar{g}(y) &= \|x\|^2 + \|y\|^2 - 2(\varphi(x) + \varphi^*(y)) \\ &\leq \|x\|^2 + \|y\|^2 - 2\langle x, y \rangle = \|x - y\|^2. \end{aligned} \quad (1.19)$$

To check integrability, note that from the definition of convex conjugation,  $\varphi = \varphi^{**}$  and  $\varphi^*$  are both lower bounded by affine functions. Therefore, since  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  we have that  $f_+ \in L^1(\mu)$  and  $g_+ \in L^1(\nu)$ . Moreover, (1.18) yields  $\int f d\mu + \int g d\nu \geq 0$  so that  $\int f d\mu > -\infty$  and  $\int g d\nu > -\infty$ . It yields

$$\int |f| d\mu = 2 \int f_+ d\mu - \int f d\mu < \infty$$

so that  $f \in L^1(\mu)$  and similarly  $g \in L^1(\nu)$ . This completes the proof of  $(ii) \Rightarrow (iii)$ .

We now turn to the proof of  $(iii) \Rightarrow (i)$ . We have by (1.18) that for any  $\gamma \in \Gamma_{\mu, \nu}$

$$\begin{aligned} \int \|x - y\|^2 \bar{\gamma}(dx, dy) &= \int \bar{f} d\mu + \int \bar{g} d\nu \\ &= \int (\bar{f}(x) + \bar{g}(y)) \gamma(dx, dy) \\ &\leq \int \|x - y\|^2 \gamma(dx, dy), \end{aligned}$$

where in the last inequality, we used (1.19). Therefore,  $\bar{\gamma}$  is an optimal coupling and  $(i)$  follows.  $\square$

Some implications of this theorem are still true even when  $\mu$  does not have a density. In particular, it can be shown using these tools that strong duality still holds, and moreover that the converse of Proposition 1.11 holds: for any two measures  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ , if  $\gamma \in \Gamma_{\mu, \nu}$  is supported on a cyclically monotone set, then it must be an optimal coupling (see [AG13] for example).

The optimal  $f$  and  $g$  arising in Theorem 1.14 play a central role in the sequel.

**Definition 1.15 (Kantorovich potentials).** *The functions*

$$\bar{f}(x) = \|x\|^2 - 2\varphi(x) \quad \text{and} \quad \bar{g}(y) = \|y\|^2 - 2\varphi^*(y)$$

*that realize the optimum of the dual Kantorovich problem  $(D-W_2^2)$  are called Kantorovich potentials for the pair  $(\mu, \nu)$ .*

Even though *a priori* solutions to  $(D-W_2^2)$  are only defined almost everywhere,  $\bar{f}$  and  $\bar{g}$  are *bona fide* functions defined everywhere on  $\mathbb{R}^d$ . Note that by symmetry of  $(D-W_2^2)$  and the form of the Kantorovich potentials, it is easy to check that if  $\nu$  admits a density, then  $\nabla\varphi^*$  is an optimal transport map from  $\nu$  to  $\mu$ .

#### 1.5.4 An improved Brenier theorem

With the fundamental theorem, we can state an improved version of Brenier's theorem, which is often useful.

**Theorem 1.16 (Improved Brenier).** *Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  be two probability measures such that  $\mu$  has a density and let  $X \sim \mu$ . Then there exists a convex function  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $(X, \nabla\varphi(X)) \sim \bar{\gamma} \in \Gamma_{\mu, \nu}$  and  $\bar{\gamma}$  is an optimal coupling for  $(W_2^2)$ :*

$$\int \|x - y\|^2 \bar{\gamma}(dx, dy) = \min_{\gamma \in \Gamma_{\mu, \nu}} \int \|x - y\|^2 \gamma(dx, dy) = W_2^2(\mu, \nu).$$

*Moreover,  $\nabla\varphi$  is unique in the sense that if there exists a convex function  $\psi$  such that  $\nabla\psi(X) \sim \nu$ , then  $\nabla\psi(X) = \nabla\varphi(X)$ , almost surely.*

*In particular, any valid coupling  $\gamma \in \Gamma_{\mu, \nu}$  of the form  $(X, \nabla\psi(X)) \sim \gamma$  for some convex function  $\psi$ , must be the unique optimal coupling between  $\mu$  and  $\nu$ .*

*Proof.* We have already proved the existence of  $\varphi$  in the previous subsection and we need to prove uniqueness of  $\nabla\varphi$ .

It turns out that as soon as every optimal coupling is induced by a transport map, this transport map (and hence the optimal coupling) must be unique.

To see this, let  $\gamma_1$  and  $\gamma_2$  be two optimal couplings induced by the transport maps  $T_1$  and  $T_2$  respectively:

$$\gamma_1(Y = T_1(X)) = 1, \quad \text{and} \quad \gamma_2(Y = T_2(X)) = 1.$$

Then consider the coupling  $\bar{\gamma} = (\gamma_1 + \gamma_2)/2 \in \Gamma_{\mu,\nu}$ . Then  $\bar{\gamma}$  is also optimal since

$$\begin{aligned} & \int \|x - y\|^2 \bar{\gamma}(dx, dy) \\ &= \frac{1}{2} \int \|x - y\|^2 \gamma_1(dx, dy) + \frac{1}{2} \int \|x - y\|^2 \gamma_2(dx, dy) = W_2^2(\mu, \nu). \end{aligned}$$

In particular, it follows from Brenier's theorem that  $\bar{\gamma}$  is also induced by a transport map  $T$  (which happens to be the gradient of a convex function but we do not need this fact here). Therefore, if  $(X, Y) \sim \bar{\gamma}$ , it must be the case that the conditional distribution of  $Y$  given  $X$  is the Dirac  $\delta_{T(X)}$ . But by construction this conditional distribution is the mixture of two Diracs  $(\delta_{T_1(X)} + \delta_{T_2(X)})/2$  and the two may only be the same when  $T_1(X) = T_2(X) = T(X)$ , almost surely.

Therefore, if there exists a convex function  $\psi$  such that  $\nabla\psi(X) \sim \nu$ , then by Theorem 1.14, it must be that  $\gamma$  such that  $(X, \nabla\psi(X)) \sim \gamma \in \Gamma_{\mu,\nu}$  is an optimal coupling and therefore that  $\nabla\psi(X) = \nabla\varphi(X)$ , almost everywhere in light of the above discussion.

The last statement of the theorem follows by observing that the equivalence (ii)  $\Leftrightarrow$  (i) in Theorem 1.14 implies that optimal couplings are supported on the graph of the gradient of a convex function, which has to be unique from the above argument.  $\square$

The improved Brenier theorem is very useful since it characterizes optimality of a transport map: if a transport map is the gradient of a convex function, then it is optimal and is unique! We call this map the *Brenier map*. We can use Theorem 1.16 to characterize optimal transport maps in two fundamental instantiations of the optimal transport problem: the one-dimensional case and the Gaussian case.

*Example 1.17 (One-dimensional optimal transport).* Recall that the cumulative distribution function (CDF)  $F$  of a random variable  $X$  is given by the map  $t \mapsto \mathbb{P}(X \leq t)$ .

**Proposition 1.18.** *Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$  be two univariate distributions with CDFs,  $F_\mu$  and  $F_\nu$  respectively and such that  $\mu$  admits a density. Then*

$$W_2^2(\mu, \nu) = \int_0^1 |F_\mu^\dagger(u) - F_\nu^\dagger(u)|^2 du$$

*and the optimal coupling between  $\mu$  and  $\nu$  is induced by the Brenier map  $F_\nu^\dagger \circ F_\mu$ .*

*Proof.* Since  $\mu$  has a density,  $F_\mu \circ F_\mu^\dagger = \text{id}$ . Let  $U \sim \text{Unif}([0, 1])$  be a uniform random variable and define  $X = F_\mu^\dagger(U)$ ,  $Y = F_\nu^\dagger(U)$  so that  $X \sim \mu$  and  $Y \sim \nu$ . Next observe that  $Y = F_\nu^\dagger \circ F_\mu(X)$  and that  $F_\nu^\dagger \circ F_\mu$  is an increasing function. In light of Theorem 1.16, this defines the unique optimal coupling between  $\mu$  and  $\nu$ .  $\square$

The geometric consequence of this identity, as explored further in Chapter 7, is that the metric space  $\mathcal{P}_2(\mathbb{R})$  equipped with the 2-Wasserstein distance is flat. Indeed the map  $\mu \mapsto F_\mu^\dagger$  is an isometric embedding of  $(\mathcal{P}_2(\mathbb{R}), W_2)$  into the (flat) Hilbert space  $L^2([0, 1])$ .

*Example 1.19 (Gaussian optimal transport).* We can also use the improved Brenier theorem to derive the optimal transport map between two Gaussian measures. Let  $m_1, m_2 \in \mathbb{R}^d$  and let  $\Sigma_1, \Sigma_2$  be positive definite  $d \times d$  matrices. Let  $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$  and  $\mu_2 = \mathcal{N}(m_2, \Sigma_2)$ .

Recall that affine maps preserve Gaussianity. Namely, if  $X_1 \sim \mu_1$  and  $T(x) = Ax + b$ , where  $A \in \mathbb{R}^{d \times d}$  and  $b \in \mathbb{R}^d$ , then  $T(X_1)$  is also Gaussian. To calculate the distribution of  $T(X_1)$ , it suffices to compute the mean and covariance, and we find that

$$\mathbb{E}T(X_1) = Am_1 + b, \quad \text{cov } T(X_1) = A\Sigma_1 A^\top.$$

It is therefore a reasonable guess that the optimal transport map from  $\mu_1$  to  $\mu_2$  is affine, and this can be verified using Brenier's Theorem 1.16. For this, we require that  $Am_1 + b = m_2$  and  $A\Sigma_1 A^\top = \Sigma_2$ , representing the constraint that  $T_\# \mu_1 = \mu_2$ . We also require  $T$  to be the gradient of a convex function. If we set

$$\varphi(x) = \frac{1}{2} \langle x, Ax \rangle + \langle b, x \rangle,$$

then  $\nabla \varphi(x) = \frac{1}{2}(A + A^T)x + b$ , and  $\varphi$  is convex provided  $A + A^T \succeq 0$ . We conclude that  $T = \nabla \varphi$  is the gradient of a convex function (and therefore optimal) provided that  $A$  is symmetric and positive definite.

How do we choose  $A$  and  $b$  so that the pushforward constraints and the PSD constraint on  $A$  are simultaneously met? The most naïve choice for  $A$ , namely  $A = \Sigma_1^{-1/2}\Sigma_2^{1/2}$ , works when  $\Sigma_1$  and  $\Sigma_2$  commute, but in general this choice of  $A$  is not even symmetric. It takes a little ingenuity to find a PSD choice for  $A$  that works, but it can be done as follows. Starting with the idea that  $A$  is PSD and satisfies  $A\Sigma_1 A = \Sigma_2$ , then by squaring  $\Sigma_1^{1/2} A \Sigma_1^{1/2}$  we find that

$$(\Sigma_1^{1/2} A \Sigma_1^{1/2})^2 = \Sigma_1^{1/2} A \Sigma_1 A \Sigma_1^{1/2} = \Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2}.$$

Taking square roots and solving, we obtain

$$A = \Sigma_1^{-1/2} (\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \Sigma_1^{-1/2}.$$

It is seen that this is the matrix  $A$  that we are looking for. By using the other constraint  $Am_1 + b = m_2$ , we find that

$$T(x) = \Sigma_1^{-1/2} (\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \Sigma_1^{-1/2} (x - m_1) + m_2.$$

By Theorem 1.16, this is the unique optimal transport map from  $\mu_1$  to  $\mu_2$ . Moreover, by substituting this into the definition of the Wasserstein distance, we find that (exercise!)

$$W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|^2 + \text{tr}[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2}]. \quad (1.20)$$

## 1.6 Duality for $p = 1$

In the case of the 1-Wasserstein metric, the dual takes a remarkably simple form. Most textbooks derive this result as a specific instantiation of strong duality for optimal transport with a general cost  $c$ , which requires tools to generalize Theorem 1.14 beyond the case of the quadratic cost  $c(x, y) = \|x - y\|^2$ . The tools include a generalized notion of cyclical monotonicity and of Legendre transform, and the reader is invited to become familiar with them (see the discussion section for a brief overview). When specialized to the  $p$ -Wasserstein distance, they yield the following result; see, e.g., [San15, Section 3.1.1] for a proof.

**Theorem 1.20.** Fix  $p \geq 1$  and let  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  be two probability measures. Then the following holds:

$$W_p^p(\mu, \nu) = \sup_{\substack{f \in L^1(\mu), g \in L^1(\nu) \\ f(x) + g(y) \leq \|x - y\|^p}} \left\{ \int f \, d\mu + \int g \, d\nu \right\}. \quad (1.21)$$

In the case, where  $p = 1$ , this result can be simplified to eliminate one of the dual functions.

**Theorem 1.21.** Let  $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$  be two probability measures. Then the following holds:

$$W_1(\mu, \nu) = \sup_{f \in \text{Lip}_1} \left\{ \int f \, d\mu - \int f \, d\nu \right\}, \quad (1.22)$$

where  $\text{Lip}_1$  is the set of 1-Lipschitz functions.

Before proceeding to the proof of this theorem, let us see where Lipschitz functions come from. Recall that the semidual in Section 1.5.2 was also removing one of the dual functions. In particular, when  $p = 1$ , we can replace the function  $g$  in (1.21) with the  $c$ -transform

$$f^c(y) = \inf_{x \in \mathbb{R}^d} \{\|x - y\| - f(x)\}.$$

The following lemma holds.

**Lemma 1.22.** For  $c(x, y) = \|x - y\|$ , a function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is a  $c$ -transform  $g = f^c$  if and only if it is 1-Lipschitz. Moreover, any 1-Lipschitz function satisfies  $g^c = -g$ .

*Proof.* Write

$$g(y) = f^c(y) = \inf_{x \in \mathbb{R}^d} \{\|x - y\| - f(x)\}.$$

For any  $x$ , the function  $y \mapsto \|x - y\| - f(x)$  is clearly 1-Lipschitz by the reverse triangle inequality. Since the set of 1-Lipschitz functions is closed under taking infima, the function  $g$  is also 1-Lipschitz.

To prove the converse, let  $g$  be a 1-Lipschitz function so that for any  $x, y \in \mathbb{R}^d$ , it holds

$$g(y) \leq \|x - y\| + g(x).$$

Taking the infimum over  $x$  yields  $g \leq (-g)^c$ . Moreover,

$$(-g)^c(y) = \inf_{x \in \mathbb{R}^d} \{\|x - y\| + g(x)\} \leq g(y),$$

where we took  $y = x$  in the last inequality.

We have shown that  $(-g)^c = g$  and in particular that  $g$  has to be a  $c$ -transform (of  $-g$ ). The second statement of the lemma follows from the fact that if  $g$  is 1-Lipschitz, then so is  $-g$ .  $\square$

We are now in a position to prove Theorem 1.21.

*Proof of Theorem 1.21.* By the argument in Subsection 1.5.2, (1.21) is equal to the following semidual

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

We can continue optimizing the potentials further to obtain

$$\sup_{f \in L^1(\mu)} \left\{ \int f^{cc} \, d\mu + \int (f^{cc})^c \, d\nu \right\}.$$

It follows from Lemma 1.22 that

$$\{f^{cc} : f \in L^1(\mu)\}$$

only contains 1-Lipschitz functions. Moreover, since  $\mu \in \mathcal{P}_1(\mathbb{R}^d)$ , it holds that any 1-Lipschitz function is integrable against  $\mu$ :

$$\begin{aligned} \int |f| \, d\mu &\leq \int |f(x) - f(y)| \, \mu(dx) + |f(y)| \leq \int \|x - y\| \, \mu(dx) + |f(y)| \\ &< \infty, \end{aligned}$$

so that  $f \in L^1(\mu)$ . Hence, we have shown that

$$W_1(\mu, \nu) = \sup_{f \in \text{Lip}_1} \left\{ \int f \, d\mu + \int f^c \, d\nu \right\} = \sup_{f \in \text{Lip}_1} \left\{ \int f \, d\mu - \int f \, d\nu \right\}, \quad (1.23)$$

which concludes the proof of Theorem 1.21.  $\square$

## 1.7 Discussion

**§1.1.** For a historical bibliography, consult [Vil03].

**§1.2.** Villani [Vil09b] remarks that the common terminology “Wasserstein distances” is “very questionable”—though these distances are attributed to Leonid Vaserstein by Dobrushin [Dob70], Vaserstein was not their progenitor. Nevertheless, the name has stuck.

Proposition 1.5 is taken from the paper [RNW19], which also provides a converse. Further comparisons between “information divergences” (e.g., total variation, Kullback–Leibler, chi-squared divergence) and optimal transport distances are implied by so-called transport inequalities, which also connects to a larger literature on concentration of measure; see, e.g., [BV05] and [Vil09b, Chapter 22].

**§1.3.** As shown in Exercise 8, the monotone coupling in Theorem 1.7 and Proposition 1.18 is also optimal for any cost function which is a strictly convex function of  $x - y$ ; see [San15, Section 2.2].

**§1.4.** Brenier’s theorem was first used to define multivariate quantiles for  $\nu \in \mathcal{P}_2(\mathbb{R}^d)$  by [CGHH17], and this definition was extended to  $\nu$  which may not have a second moment by [Hal17]. In addition to its mathematical elegance, this definition of multivariate quantiles has important implications for nonparametric testing [GS22, DS23].

Rockafellar’s theorem is from [Roc66]. The proof of Proposition 1.11 is from [GM96].

**§1.5.** Although we deduced strong duality by explicitly exhibiting dual potentials for which the duality gap is zero (namely, the ones obtained from the characterization of optimal couplings as having cyclically monotone support), it is also possible to prove strong duality directly via appeal to an abstract min-max principle; see [Vil03, Section 1.1].

Many of the arguments in Section 1.5 generalize to general continuous costs  $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ : the  $c$ -conjugate of a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  is given by  $f^c(y) := \inf_{x \in \mathcal{X}} \{c(x, y) - f(x)\}$ , and likewise the  $c$ -conjugate of  $g : \mathcal{Y} \rightarrow \mathbb{R}$  is given by  $g^c(x) := \inf_{y \in \mathcal{Y}} \{c(x, y) - g(y)\}$ . We say that  $f$  is  $c$ -concave if  $f = g^c$  for some function  $g$ . The equality (1.17) still holds and equals the optimal transport value (strong duality). Any optimal transport plan is supported on a  $c$ -cyclically monotone set, which is defined as in (1.12) but replacing the quadratic cost with  $c$ . Then,  $c$ -cyclically monotone sets are characterized as  $c$ -subdifferentials, where the  $c$ -subdifferential of a  $c$ -concave function  $f = g^c$  is the set of  $(x, y)$  pairs such that  $f(x) + g(y) = c(x, y)$ . What does *not* generalize as easily, however, is Brenier’s theorem, which requires further conditions

to ensure the single-valuedness of the  $c$ -subdifferential. For example if  $c(x, y) = h(x - y)$  for some strictly convex function  $h$ , then a unique optimal transport *map* exists but it need not be the gradient of a convex function; see [San15, Theorem 1.17].

**§1.6.** The duality formula for  $W_1$  is classical and is closely related to the bounded Lipschitz metric [Dud02], which can be extended to define a norm over signed measures. As discussed in Subsection 2.8.1, this formula expresses  $W_1$  as an integral probability metric.

## 1.8 Exercises

1. Let  $A \subseteq \mathbb{R} \times \mathbb{R}$  be monotone:

$$\forall (x, y), (x', y') \in A, x < x' \Rightarrow y < y'.$$

For simplicity, assume that  $A$  is contained in the graph of a function. Show that  $A$  is cyclically monotone.

2. Let  $X \in \mathbb{R}$  be a random variable that admits a density supported on the whole real line. Let  $f, g$  be two monotone increasing functions such that  $f(X)$  has the same distribution as  $g(X)$ . Then,  $f = g$  almost everywhere.
3. Let  $m_1, m_2 \in \mathbb{R}^d$  and let  $\Sigma_1, \Sigma_2$  be positive definite  $d \times d$  matrices. Let  $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$  and  $\mu_2 = \mathcal{N}(m_2, \Sigma_2)$ .
  - a) Verify the equation (1.20).
  - b) Using a suboptimal coupling, prove the simple upper bound

$$W_2^2(\mu_1, \mu_2) \leq \|m_1 - m_2\|^2 + \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_{\text{HS}}^2$$

where  $\|M_1 - M_2\|_{\text{HS}}^2 := \text{tr}((M_1 - M_2)^T(M_1 - M_2))$ . Show that this is an equality when  $\Sigma_1$  and  $\Sigma_2$  commute.

- c) Prove that if  $\nu_1, \nu_2$  are probability measures with means  $m_1, m_2$  and covariance matrices  $\Sigma_1, \Sigma_2$  respectively, then

$$W_2(\nu_1, \nu_2) \geq W_2(\mu_1, \mu_2).$$

*Hint:* What are the optimal dual potentials for the optimal transport problem from  $\mu_1$  to  $\mu_2$ ?

4. a) Let  $X$  and  $Y$  be random vectors in  $\mathbb{R}^d$ . Prove that

$$W_2^2(\text{law}(X), \text{law}(Y))$$

$$= \|\mathbb{E}X - \mathbb{E}Y\|^2 + W_2^2(\text{law}(X - \mathbb{E}X), \text{law}(Y - \mathbb{E}Y)).$$

Thanks to this equality, often when we work with the  $W_2$  distance, it suffices to consider centered random variables.

- b) Let  $X, Y_1, \dots, Y_k$  be random vectors in  $\mathbb{R}^d$  and  $\lambda_1, \dots, \lambda_k \geq 0$ . Suppose that  $X$  and  $Y_i$  are optimally coupled for each  $i = 1, \dots, k$ . Show that  $X$  and  $\sum_{i=1}^k \lambda_i Y_i$  are optimally coupled.
- 5. Let  $\nu$  admit a density w.r.t. Lebesgue measure. Show that  $W_2^2(\cdot, \nu)$  is strictly convex; that is, if  $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$  are distinct and  $t \in (0, 1)$ , then

$$W_2^2((1-t)\mu_0 + t\mu_1, \nu) < (1-t)W_2^2(\mu_0, \nu) + tW_2^2(\mu_1, \nu). \quad (1.24)$$

*Hint:* Start by proving that (1.24) holds with  $\leq$  instead of  $<$ . Next, supposing that (1.24) fails, show that there is an optimal transport plan between  $(1-t)\mu_0 + t\mu_1$  and  $\nu$  which is not induced by a transport map, contradicting Brenier's theorem.

- 6. Consider the measures  $\mu = \frac{1}{2}\mathcal{N}(-m, 1) + \frac{1}{2}\mathcal{N}(+m, 1)$  and  $\nu = \frac{1}{4}\mathcal{N}(-m, 1) + \frac{3}{4}\mathcal{N}(+m, 1)$ , where  $m > 0$ . Prove that  $W_2(\mu, \nu) \asymp m$ .  
*Hint:* The point of this question is that computing the optimal transport map from  $\mu$  to  $\nu$  is painful, but it is not hard to obtain good lower and upper bounds. For the lower bound, prove and use the fact that for any 1-Lipschitz function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , it holds that  $\mathbb{E}_\mu f - \mathbb{E}_\nu f \leq W_1(\mu, \nu) \leq W_2(\mu, \nu)$ . For the upper bound, exhibit a coupling of  $\mu$  and  $\nu$ .
- 7. Recall that the chi-squared divergence between two probability measures  $\mu$  and  $\nu$  is defined as

$$\chi^2(\mu \parallel \nu) = \int \left( \frac{d\mu}{d\nu} - 1 \right)^2 d\nu.$$

Show that

$$W_1(\mu, \nu) \leq \sqrt{\chi^2(\mu \parallel \nu)},$$

when  $\nu$  has unit variance.

- 8. Let  $c : \mathbb{R} \rightarrow \mathbb{R}$  be strictly convex and consider optimal transport over  $\mathbb{R}$  with the cost function  $(x, y) \mapsto c(x - y)$ . For example, when  $c(z) = |z|^p$  for  $p > 1$ , we obtain  $W_p^p$ . In this exercise, we show that the coupling  $\bar{\gamma}$  given in Proposition 1.18 is *universally optimal* for all costs of this form. See the discussion section for background.
- a) Show that for any  $a, b \in \mathbb{R}$ ,  $\bar{\gamma}((-\infty, a] \times (-\infty, b]) = F_\mu(a) \wedge F_\nu(b)$ .

b) Show that  $\bar{\gamma}$  is the *unique* coupling of  $\mu$  and  $\nu$  such that

$$\forall (x, y), (x', y') \in \text{supp } \bar{\gamma}, x < x' \Rightarrow y \leq y'. \quad (1.25)$$

*Hint:* Consider the sets  $A = (-\infty, a] \times (b, \infty)$  and  $B = (a, \infty) \times (-\infty, b]$ . Show that any coupling satisfying (1.25) must assign one of these two sets measure zero. Use this to show that any coupling  $\gamma$  which satisfies (1.25) must agree with  $\bar{\gamma}$ .

c) Let  $(x, y), (x', y') \in \text{supp } \gamma$ , where  $\gamma$  is an optimal transport plan between  $\mu$  and  $\nu$  with cost defined by  $c$  as above. By  $c$ -cyclical monotonicity of  $\text{supp } \gamma$ ,

$$c(x - y) + c(x' - y') \leq c(x - y') + c(x' - y).$$

Show that if  $x < x'$ , then  $y \leq y'$ , hence  $\gamma = \bar{\gamma}$ .

*Hint:* Argue by contradiction. If the claim fails, then both  $u := x - y'$  and  $v := x' - y$  lie between  $w := x - y$  and  $w' := x' - y'$ . Write  $u$  and  $v$  as convex combinations of  $w$  and  $w'$ , and apply strict convexity of  $c$ .

d) Deduce that the optimal cost equals  $\mathbb{E}c(F_\mu^\dagger(U) - F_\nu^\dagger(U))$  where  $U \sim \text{Unif}([0, 1])$ .

9. Given two probability measures  $\mu, \nu$  over the same space  $\mathbb{S}$ , define the *total variation distance* between  $\mu$  and  $\nu$  to be

$$d_{\text{TV}}(\mu, \nu) = \sup_{A \subseteq \mathbb{S}} |\mu(A) - \nu(A)|,$$

where the supremum ranges over all measurable subsets. Show that the total variation distance is also equal to all of the following. (*Hint:* Consider the coupling in Theorem 1.6.)

a) If  $f$  and  $g$  denote the respective densities of  $\mu$  and  $\nu$  with respect to some common dominating measure  $\lambda$  for  $\mu$  and  $\nu$  (e.g.,  $\lambda = \mu + \nu$ ), then

$$d_{\text{TV}}(\mu, \nu) = \frac{1}{2} \int |f - g| d\lambda = 1 - \int f \wedge g d\lambda = \mu(f \geq g).$$

b)  $d_{\text{TV}}(\mu, \nu) = \inf \mathbb{P}(X \neq Y)$  where the infimum ranges over all couplings  $(X, Y)$  of  $\mu$  and  $\nu$ .  
c)  $d_{\text{TV}}(\mu, \nu) = \sup \{ \int h d(\mu - \nu) \mid h : \mathbb{S} \rightarrow [0, 1] \}$ . Compare with  $W_1$  duality from Theorem 1.21.

10. Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  admit densities with respect to Lebesgue measure. Show that if  $\nabla\varphi$  is the optimal transport map from  $\mu$  to  $\nu$ , then  $\nabla\varphi^*$  is the optimal transport map from  $\nu$  to  $\mu$ . (See Appendix A.) Apply this fact to the optimal transport map between two Gaussians and discover a non-trivial matrix identity.

---

## Estimation of Wasserstein distances

In applications of optimal transport in statistics, it is paramount to be able to obtain good upper and lower bounds on the Wasserstein distance between probability measures. This chapter describes tools to bound the Wasserstein distance. To do so, we heavily employ the primal *and* dual formulations of optimal transport. As a primary application, we consider a quantitative form of the *Wasserstein law of large numbers*, which is the statement that if  $\mu_n$  is an empirical measure consisting of  $n$  i.i.d. samples from a probability measure  $\mu$ , then  $\mathbb{E}W_p(\mu_n, \mu) \rightarrow 0$  as  $n \rightarrow \infty$ .

### 2.1 The Wasserstein law of large numbers

Suppose that  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mu$ , where  $\mu$  is a probability measure on a compact subset of  $\mathbb{R}^d$ , which we assume for convenience is equal to the unit cube  $[0, 1]^d$ . The *empirical measure* is defined to be the (random) measure

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

The law of large numbers implies that  $\mu_n \hookrightarrow \mu$  and also  $\int \|\cdot\|^p d\mu_n \rightarrow \int \|\cdot\|^p d\mu$  almost surely; therefore, the discussion in Chapter 1 implies that  $W_p(\mu_n, \mu) \rightarrow 0$ . Moreover, since  $W_p(\mu_n, \mu)$  is bounded almost surely, we also have convergence in mean:

$$\mathbb{E}W_p(\mu_n, \mu) \rightarrow 0.$$

How fast does this convergence occur? In the context of the classic law of large numbers for bounded random vectors  $X_1, \dots, X_n$  in  $\mathbb{R}^d$ , we of course have

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E} X \right\|^2 \lesssim \frac{1}{n}.$$

Note that the rate of decay  $n^{-1}$  holds irrespective of the dimension, and is true even in infinite-dimensional Hilbert spaces.

By contrast, the Wasserstein law of large numbers behaves quite differently. In this chapter, we prove the following proposition.

**Proposition 2.1.** *If the support of  $\mu$  lies in  $[0, 1]^d$ , then*

$$\mathbb{E} W_1(\mu_n, \mu) \lesssim \sqrt{d} \cdot \begin{cases} n^{-1/2} & \text{if } d = 1, \\ (\log n/n)^{1/2} & \text{if } d = 2, \\ n^{-1/d} & \text{if } d \geq 3, \end{cases} \quad (2.1)$$

and this rate is unimprovable in general.

In contrast to the standard law of large numbers, the convergence of  $\mu_n$  to  $\mu$  in Wasserstein distance degrades exponentially as the dimension grows, a phenomenon often known as the curse of dimensionality.

## 2.2 The dyadic partitioning argument

The fact that the Wasserstein distance is defined by a minimization over couplings suggests a natural strategy for proving bounds: we can show an upper bound on  $W_1$  by exhibiting a coupling with a small cost. In this section, we build such a coupling, which, perhaps surprisingly, gives rise to good bounds in many situations. The main idea is to attempt to couple  $\mu$  and  $\nu$  by recursively constructing candidate couplings at multiple scales.

Before stating the bound, let us describe the basic strategy. For simplicity, let us consider proving an upper bound on  $W_1(\mu, \nu)$  for  $\mu$  and  $\nu$  whose support lies in  $[0, 1]^d$ . We first make a trivial observation:

$$W_1(\mu, \nu) \leq \sqrt{d}. \quad (2.2)$$

Indeed, the diameter of  $[0, 1]^d$  is  $\sqrt{d}$ , so no coupling between  $\mu$  and  $\nu$  can move mass a greater distance than this.

Let us now imagine a slight sharpening of this bound. Let  $\mathcal{Q}$  be the collection of cubes of side length  $1/2$  whose corners lie at points of the form  $2^{-1}(k_1, \dots, k_d)$  for  $k_1, \dots, k_d \in \{0, 1, 2\}$ . These cubes form a partition of  $[0, 1]^d$  into  $2^d$  pieces.<sup>1</sup> Suppose for the sake of argument that  $\mu(Q) = \nu(Q)$  for all  $Q \in \mathcal{Q}$ ,  $j = 1, \dots, 2^d$ , so that  $\mu$  and  $\nu$  assign the same mass to each of the small cubes. Then, it would be possible to couple  $\mu$  and  $\nu$  by only moving mass within each small cube. Since the diameter of each small cube is  $\sqrt{d}/2$ , any such coupling improves on the bound in (2.2) by a factor of 2.

Even when  $\mu$  and  $\nu$  do not assign the same mass to each small cube, we can use the above idea to construct a coupling between  $\mu$  and  $\nu$  in two steps: first, we can match as much mass as possible between  $\mu$  and  $\nu$  within each cube. This creates a partial coupling between a portion of  $\mu$ 's mass and a portion of  $\nu$ 's. Since we only move mass within each small cube, the total cost of this partial coupling is at most  $\sqrt{d}/2$ . We then need to extend this partial coupling to a full coupling, by transporting  $\mu$ 's extra mass on any cube  $Q$  for which  $\mu(Q) > \nu(Q)$  to  $\nu$ 's extra mass on some cube  $Q'$  for which  $\nu(Q') > \mu(Q')$ . The amount of extra mass matched in this step is  $\sum_{Q \in \mathcal{Q}} (\mu(Q) - \nu(Q))_+ = \frac{1}{2} \sum_{Q \in \mathcal{Q}} |\mu(Q) - \nu(Q)|$ , at a total cost of at most  $\frac{\sqrt{d}}{2} \sum_{Q \in \mathcal{Q}} |\mu(Q) - \nu(Q)|$ .

Combining these bounds yields the refined estimate

$$W_1(\mu, \nu) \leq \frac{\sqrt{d}}{2} \sum_{Q \in \mathcal{Q}} |\mu(Q) - \nu(Q)| + \frac{\sqrt{d}}{2}. \quad (2.3)$$

This bound improves on (2.2) when  $\mu$  and  $\nu$  assign similar mass to each cube.

The proof of the following bound is based on recursing the above argument  $J$  times. At the  $j$ -th stage, we bound the discrepancy between  $\mu$  and  $\nu$  on  $2^{dj}$  cubes of side length  $2^{-j}$ . To state this bound, let us define the set  $\mathcal{Q}_j$ ,  $j \geq 0$ , to consist of a set of  $2^{dj}$  cubes of side length  $2^{-j}$  which form a partition of  $[0, 1]^d$ .<sup>2</sup>

**Theorem 2.2 (Dyadic partitioning bound).** *Let  $\mu, \nu \in \mathcal{P}([0, 1]^d)$ . For any  $J \geq 0$ ,*

---

<sup>1</sup> These cubes overlap at their boundaries, but we can easily modify these sets by removing overlaps to obtain a bona fide partition.

<sup>2</sup> As above, we assume that the elements of  $\mathcal{Q}_j$  been modified at their boundary so that  $\mathcal{Q}_j$  is a partition and so that  $\mathcal{Q}_{j+1}$  is a refinement of  $\mathcal{Q}_j$  for all  $j \geq 0$ .

$$W_1(\mu, \nu) \leq \sqrt{d} \sum_{j=0}^{J-1} \left( 2^{-j} \sum_{Q \in \mathcal{Q}_{j+1}} |\mu(Q) - \nu(Q)| \right) + \sqrt{d} 2^{-J}.$$

*Proof.* We define a sequence of positive measures  $\mu_0, \dots, \mu_J$  and  $\nu_0, \dots, \nu_J$ , which satisfy  $\sum_{j=0}^J \mu_j = \mu$  and  $\sum_{j=0}^J \nu_j = \nu$  and such that

$$\mu_j(Q) = \nu_j(Q) \quad \forall Q \in \mathcal{Q}_j, \quad j = 0, \dots, J.$$

We write for simplicity  $\Omega := [0, 1]^d$ . We first claim that

$$W_1(\mu, \nu) \leq \sqrt{d} \sum_{j=0}^J 2^{-j} \mu_j(\Omega). \quad (2.4)$$

This bound is nothing but an instantiation of the strategy described above: since  $\mu_j$  and  $\nu_j$  assign the same mass to each element of  $\mathcal{Q}_j$ , there exists a coupling  $\gamma_j$  between  $\mu_j$  and  $\nu_j$  which only moves mass within each element of  $\mathcal{Q}_j$ ; for instance, we can take the piecewise independent coupling

$$\gamma_j = \sum_{Q \in \mathcal{Q}_j: \mu_j(Q) > 0} \frac{(\mu_j)|_Q \otimes (\nu_j)|_Q}{\mu_j(Q)}.$$

The fact that  $\gamma_j \in \Gamma_{\mu_j, \nu_j}$  implies  $\gamma = \sum_{j=0}^J \gamma_j \in \Gamma_{\mu, \nu}$ , and

$$\begin{aligned} W_1(\mu, \nu) &\leq \int \|x - y\| \gamma(dx, dy) \\ &= \sum_{j=0}^J \int \|x - y\| \gamma_j(dx, dy) \\ &\leq \sqrt{d} \sum_{j=0}^J 2^{-j} \mu_j(\Omega), \end{aligned}$$

where the last inequality follows from the fact if  $(x, y) \in \text{supp}(\gamma_j)$ , then  $x$  and  $y$  lie in the same element  $Q \in \mathcal{Q}_j$ , so that  $\|x - y\| \leq \text{diam}(Q) = \sqrt{d} 2^{-j}$ .

We now exhibit the measures  $\mu_j$  and  $\nu_j$  which give rise to the final bound. Define the restriction of  $\mu_J$  on each  $Q \in \mathcal{Q}_J$  by setting

$$(\mu_J)|_Q = \frac{\mu(Q) \wedge \nu(Q)}{\mu(Q)} \mu|_Q,$$

where by convention we let  $\mu_J$  be zero on  $Q$  if  $\mu(Q) = 0$ . Similarly, set

$$(\nu_J)|_Q = \frac{\mu(Q) \wedge \nu(Q)}{\nu(Q)} \nu|_Q.$$

For  $1 \leq j < J$ , let

$$\begin{aligned} \mu'_j &= \mu - \sum_{j < k \leq J} \mu_k, \\ \nu'_j &= \nu - \sum_{j < k \leq J} \nu_k, \end{aligned}$$

and then, for each  $Q \in \mathcal{Q}_j$ , define

$$\begin{aligned} (\mu_j)|_Q &= \frac{\mu'_j(Q) \wedge \nu'_j(Q)}{\mu'_j(Q)} (\mu'_j)|_Q, \\ (\nu_j)|_Q &= \frac{\mu'_j(Q) \wedge \nu'_j(Q)}{\nu'_j(Q)} (\nu'_j)|_Q. \end{aligned}$$

Finally, we set

$$\mu_0 = \mu - \sum_{j=1}^J \mu_j \quad \text{and} \quad \nu_0 = \nu - \sum_{j=1}^J \nu_j,$$

so that

$$\sum_{j=0}^J \mu_j = \mu \quad \text{and} \quad \sum_{j=0}^J \nu_j = \nu.$$

It is easy to see that  $\mu_j(Q) = \nu_j(Q)$  for all  $Q \in \mathcal{Q}_j$  and all  $j \in \{0, \dots, J\}$ . To apply (2.4), we also need to check that  $\mu_j, \nu_j \geq 0$ .

**Lemma 2.3.** *The measures  $\mu_0, \dots, \mu_J$  and  $\nu_0, \dots, \nu_J$  are all positive.*

*Proof.* By symmetry, it suffices to verify this fact for the sequence  $\mu_0, \dots, \mu_J$ .

We first show by backwards induction on  $j$  that

$$\mu_{j+1} \geq 0 \quad \text{and} \quad 0 \leq \sum_{j < k \leq J} \mu_k \leq \mu \quad (\mathcal{A}_j)$$

for all  $j = 0, \dots, J-1$ .

For  $j = J - 1$ , these bounds follow directly from the construction of  $\mu_J$ . Next assume that  $(A_j)$  holds for some  $j$ , then

$$\mu'_j = \mu - \sum_{j < k \leq J} \mu_k \geq 0,$$

and therefore  $\mu_j \geq 0$ , since  $\mu_j$  is obtained by reweighting  $\mu'_j$  on each element of  $\mathcal{Q}_j$  by a non-negative quantity. Note also that this non-negative quantity is also bounded by one so that we also have  $\mu_j \leq \mu'_j$ . Together these two facts yields  $0 \leq \mu_j \leq \mu'_j$  so that

$$0 \leq \sum_{j-1 < k \leq J} \mu_k = \sum_{j < k \leq J} \mu_k + \mu_j \leq \sum_{j < k \leq J} \mu_k + \mu'_j = \mu.$$

We have proved that  $(A_{j-1})$  holds. By induction, we obtain that  $\mu_1, \dots, \mu_J$  are all positive. Finally, since we have also shown that

$$\sum_{0 < k \leq J} \mu_k \leq \mu,$$

we obtain  $\mu_0 \geq 0$  as well.  $\square$

In light of (2.4), it remains to bound  $\mu_j(\Omega)$  for  $j = 0, \dots, J$ . We first claim that

$$|\mu'_j(Q) - \nu'_j(Q)| = |\mu(Q) - \nu(Q)| \quad \forall Q \in \mathcal{Q}_j, j = 1, \dots, J. \quad (2.5)$$

This follows from the fact that

$$\mu'_j(Q) - \nu'_j(Q) = \mu(Q) - \nu(Q) - \sum_{j < k \leq J} (\mu_k(Q) - \nu_k(Q)),$$

since  $\mu_k$  and  $\nu_k$  assign the same mass to each element of  $\mathcal{Q}_k$  and since  $Q$  can be written as a disjoint union of elements of  $\mathcal{Q}_k$ , so the sum vanishes. We now claim that we can bound the mass that  $\mu_j$  and  $\nu_j$  assign to elements of  $\mathcal{Q}_j$  in terms of the difference between  $\mu$  and  $\nu$  on cubes in  $\mathcal{Q}_{j+1}$ .

**Lemma 2.4.** *If  $R \in \mathcal{Q}_j$  for some  $0 \leq j < J$ , then*

$$\mu_j(R) = \nu_j(R) \leq \sum_{Q \subseteq R, Q \in \mathcal{Q}_{j+1}} |\mu(Q) - \nu(Q)|.$$

*Proof.* We have already shown that  $\mu_j(R) = \nu_j(R)$ , so it suffices to show that expression holds for  $\mu_j(R)$ . For notational consistency, we set  $\mu'_0 = \mu_0$ . Then, for any  $0 \leq j < J$  and any  $R \in \mathcal{Q}_j$ ,

$$\begin{aligned} \mu_j(R) &\leq \mu'_j(R) \\ &= \sum_{Q \subseteq R, Q \in \mathcal{Q}_{j+1}} \mu'_j(Q) \\ &= \sum_{Q \subseteq R, Q \in \mathcal{Q}_{j+1}} (\mu'_{j+1}(Q) - \mu_{j+1}(Q)) \\ &= \sum_{Q \subseteq R, Q \in \mathcal{Q}_{j+1}} (\mu'_{j+1}(Q) - \nu'_{j+1}(Q))_+ \\ &\leq \sum_{Q \subseteq R, Q \in \mathcal{Q}_{j+1}} |\mu'_{j+1}(Q) - \nu'_{j+1}(Q)| \\ &= \sum_{Q \subseteq R, Q \in \mathcal{Q}_{j+1}} |\mu(Q) - \nu(Q)|, \end{aligned}$$

where the second equality comes from comparing the definitions of  $\mu'_j$  and  $\mu'_{j+1}$ , and the last equality follows from (2.5).  $\square$

Putting it all together, (2.4) implies

$$\begin{aligned} W_1(\mu, \nu) &\leq \sqrt{d} \sum_{j=0}^J 2^{-j} \mu_j(\Omega) \\ &= \sqrt{d} \sum_{j=0}^{J-1} 2^{-j} \mu_j(\Omega) + \sqrt{d} 2^{-J} \mu_J(\Omega) \\ &= \sqrt{d} \sum_{j=0}^{J-1} \left( 2^{-j} \sum_{R \in \mathcal{Q}_j} \mu_j(R) \right) + \sqrt{d} 2^{-J} \mu_J(\Omega) \\ &\leq \sqrt{d} \sum_{j=0}^{J-1} \left( 2^{-j} \sum_{Q \in \mathcal{Q}_{j+1}} |\mu(Q) - \nu(Q)| \right) + \sqrt{d} 2^{-J}. \end{aligned}$$

This concludes the proof of Theorem 2.2.  $\square$

Applying Theorem 2.2 to  $\mu$  and  $\mu_n$ , we obtain the following bound.

**Proposition 2.5.** *If the support of  $\mu$  lies in  $[0, 1]^d$ , then*

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{d} \cdot \begin{cases} n^{-1/2} & \text{if } d = 1, \\ (\log n) n^{-1/2} & \text{if } d = 2, \\ n^{-1/d} & \text{if } d \geq 3. \end{cases}$$

*Proof.* Theorem 2.2 implies that for any  $J \geq 0$ ,

$$\begin{aligned} \mathbb{E}W_1(\mu_n, \mu) &\leq \sqrt{d} \sum_{j=0}^{J-1} 2^{-j} \sum_{Q \in \mathcal{Q}_{j+1}} \mathbb{E}|\mu_n(Q) - \mu(Q)| + \sqrt{d} 2^{-J} \\ &\leq \sqrt{d} \sum_{j=0}^{J-1} 2^{-j} 2^{d(j+1)/2} \left( \sum_{Q \in \mathcal{Q}_{j+1}} \mathbb{E}(\mu_n(Q) - \mu(Q))^2 \right)^{1/2} \\ &\quad + \sqrt{d} 2^{-J} \\ &\leq \sqrt{d} \sum_{j=0}^{J-1} 2^{-j} 2^{d(j+1)/2} n^{-1/2} + \sqrt{d} 2^{-J} \\ &\lesssim \sqrt{d} \cdot \begin{cases} 2^{(J+1)(d/2-1)} n^{-1/2} + 2^{-J} & \text{if } d \geq 3, \\ J n^{-1/2} + 2^{-J} & \text{if } d = 2, \\ n^{-1/2} + 2^{-J} & \text{if } d = 1. \end{cases} \end{aligned}$$

To balance these terms, we choose  $J$  such that  $2^J \leq n^{1/2} < 2^{J+1}$  if  $d \leq 2$ , and  $J$  such that  $2^{J+1} \leq n^{1/d} < 2^{J+2}$  if  $d \geq 3$ .  $\square$

Note that bound of Proposition 2.5 is weaker than that of Proposition 2.1 when  $d = 2$ . Unfortunately, the dyadic partitioning argument does not yield a sharp bound in two dimensions. We return to this question in Section 2.4.

### 2.3 Dual chaining bounds

In this section, we present a superficially different proof of Proposition 2.5. Rather than constructing a coupling in the primal, we use the dual representation of the 1-Wasserstein distance instead. The benefit of this approach is that we can write

$$\begin{aligned} W_1(\mu_n, \mu) &= \sup_{f \in \text{Lip}_1} \left\{ \int f \, d\mu_n - \int f \, d\mu \right\} \\ &= \sup_{f \in \text{Lip}_1} \frac{1}{n} \sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\}. \end{aligned} \tag{2.6}$$

The random process  $f \mapsto \frac{1}{n} \sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\}$  is known as an *empirical process*, and bounding the expected suprema of such processes is a very common task in many areas of statistics.

To control this empirical process, we use a standard technique known as *chaining*. Given a class  $\mathcal{F}$  of real-valued functions on  $\Omega \subseteq \mathbb{R}^d$ , we call a set  $F = \{f_1, \dots, f_N\}$  an  $\varepsilon$ -cover of  $\mathcal{F}$  if, for any  $f \in \mathcal{F}$ , there exists  $f_i \in F$  such that  $\|f - f_i\|_{L^\infty(\Omega)} \leq \varepsilon$ . The  $\varepsilon$ -covering number of  $\mathcal{F}$  is

$$N(\varepsilon, \mathcal{F}) = \min\{|F| : F \text{ is an } \varepsilon\text{-cover of } \mathcal{F}\}.$$

The chaining argument shows that the covering number of a class  $\mathcal{F}$  controls the supremum of an empirical process indexed by that set. We use the following version:

**Proposition 2.6** ([vH14, Theorem 5.31]). *If  $\mathcal{F}$  is a set of real-valued functions on  $\Omega$  such that  $\|f\|_{L^\infty(\Omega)} \leq R$  for all  $f \in \mathcal{F}$ , then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\} \lesssim \inf_{\tau > 0} \left\{ \tau + \frac{1}{\sqrt{n}} \int_\tau^R \sqrt{\log N(\varepsilon, \mathcal{F})} \, d\varepsilon \right\}.$$

Proposition 2.6 and (2.6) imply that we can obtain an upper bound on  $\mathbb{E}W_1(\mu_n, \mu)$  as long as we can calculate the covering numbers of the set of Lipschitz functions on  $[0, 1]^d$ . We also notice that we can assume without loss of generality that the functions appearing in (2.6) take the value 0 at  $(0, \dots, 0)$ . Indeed, a Lipschitz function on  $[0, 1]^d$  is bounded, and since the value of  $\frac{1}{n} \sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\}$  is unaffected if we shift  $f$  by a constant, we may fix its value at  $(0, \dots, 0)$  to be 0 without loss of generality.

**Lemma 2.7.** *Denote by  $\overline{\text{Lip}}_1([0, 1]^d)$  the set of 1-Lipschitz functions on  $[0, 1]^d$  satisfying  $f(0) = 0$ . Then*

$$\log N(\varepsilon, \overline{\text{Lip}}_1([0, 1]^d)) \lesssim (4\sqrt{d}/\varepsilon)^d.$$

*Proof.* We bound the covering number by exhibiting an  $\varepsilon$ -cover of  $\overline{\text{Lip}}_1([0, 1]^d)$  of the specified size. To do so, we again use the notion of a dyadic partition of  $[0, 1]^d$  into a set  $\mathcal{Q}_j$  of cubes of side length  $2^{-j}$ . Each element of  $\mathcal{Q}_j$  is of the form  $2^{-j}([k_1, k_1 + 1] \times \dots \times [k_d, k_d + 1])$  for some integers  $k_1, \dots, k_d \in [2^j - 1] := \{0, \dots, 2^j - 1\}$ , and we denote such an element by  $Q_{\vec{k}}$  for  $\vec{k} = (k_1, \dots, k_d)$ .<sup>3</sup>

<sup>3</sup> This collection of cubes overlaps at the boundaries, but as above we may remove overlaps to obtain a disjoint partition of  $[0, 1]^d$ .

Fix an integer  $j \geq 0$  and positive  $\delta > 0$  to be specified. Consider the set  $\mathcal{H}$  of functions  $h$  satisfying the following requirements:

1.  $h$  is constant on each element of  $\mathcal{Q}_j$ , i.e., there exist constants  $(h_{\vec{k}})_{\vec{k} \in [2^j-1]^d}$  such that  $h(x) = h_{\vec{k}}$  for all  $x \in Q_{\vec{k}}$ .
2.  $h_{\vec{k}}$  is an integer multiple of  $\delta$  for all  $\vec{k} \in [2^j-1]^d$ .
3.  $h_{(0,\dots,0)} = 0$ .
4. If  $\|\vec{k} - \vec{k}'\|_\infty \leq 1$ , then  $|h_{\vec{k}} - h_{\vec{k}'}| \leq 2^{-j}\sqrt{d} + \delta$ .

We first claim that  $\mathcal{H}$  constitutes an  $\varepsilon$ -cover of  $\overline{\text{Lip}}_1([0, 1]^d)$  if  $2^{-j}\sqrt{d} + \delta \leq \varepsilon$ . Given any  $f \in \overline{\text{Lip}}_1([0, 1]^d)$ , denote by  $h_f$  the element of  $\mathcal{H}$  given by  $(h_f)_{\vec{k}} = \delta \lfloor f(2^{-j}(k_1, \dots, k_d)) / \delta \rfloor$  for all  $\vec{k} \in [2^j-1]^d$ . To see that  $h_f \in \mathcal{H}$ , note that it immediately satisfies the first three requirements by construction, and for the fourth, we have

$$\begin{aligned} |(h_f)_{\vec{k}} - (h_f)_{\vec{k}'}| &= \delta \left| \lfloor f(2^{-j}(k_1, \dots, k_d)) / \delta \rfloor - \lfloor f(2^{-j}(k'_1, \dots, k'_d)) / \delta \rfloor \right| \\ &\leq |f(2^{-j}(k_1, \dots, k_d)) - f(2^{-j}(k'_1, \dots, k'_d))| + \delta \\ &\leq 2^{-j} \|\vec{k} - \vec{k}'\|_2 + \delta, \end{aligned}$$

where the last inequality follows from the fact that  $f$  is Lipschitz. Since  $\|\vec{k} - \vec{k}'\|_2 \leq \sqrt{d}$  when  $\|\vec{k} - \vec{k}'\|_\infty = 1$ , the claim follows. Finally, for any  $x \in Q_{\vec{k}}$ , the fact that  $f$  is Lipschitz again implies

$$\begin{aligned} |f(x) - (h_f)_{\vec{k}}| &= |f(x) - \delta \lfloor f(2^{-j}(k_1, \dots, k_d)) / \delta \rfloor| \\ &\leq |f(x) - f(2^{-j}(k_1, \dots, k_d))| + \delta \\ &\leq \text{diam}(Q_{\vec{k}}) + \delta \\ &= 2^{-j}\sqrt{d} + \delta. \end{aligned}$$

Therefore  $\|f - h_f\|_\infty \leq 2^{-j}\sqrt{d} + \delta$ .

We have shown that for every  $f \in \overline{\text{Lip}}_1([0, 1]^d)$ , there exists  $h_f \in \mathcal{H}$  such that  $\|f - h_f\|_\infty \leq 2^{-j}\sqrt{d} + \delta$ . Therefore, if  $2^{-j}\sqrt{d} + \delta \leq \varepsilon$ , then  $\mathcal{H}$  is an  $\varepsilon$ -cover of  $\overline{\text{Lip}}_1([0, 1]^d)$ . We fix  $\delta = 2^{-j}\sqrt{d}$ , so that this requirement reduces to  $2^{-j}\sqrt{d} \leq \varepsilon/2$ .

To bound  $|\mathcal{H}|$ , note that if we fix the value of  $h_{\vec{k}}$  for some  $\vec{k}$ , then for any  $\vec{k}'$  such that  $\|\vec{k} - \vec{k}'\|_\infty = 1$ , there are at most 5 possible values of  $h_{\vec{k}'}$ . This follows from the fact that  $h_{\vec{k}'}$  must be an integer multiple of  $\delta = 2^{-j}\sqrt{d}$ , and there are 5 integer multiples of  $\delta$  in the interval  $[h_{\vec{k}} - 2\delta, h_{\vec{k}} + 2\delta]$ . Therefore, if we consider specifying an element  $\mathcal{H}$  by specifying the values of  $h_{\vec{k}}$  sequentially by setting  $h_{(0,\dots,0)} = 0$  and

proceeding in lexicographic order, then at each stage we have at most 5 choices for the next value of  $h_{\vec{k}}$ . This implies that  $|\mathcal{H}| \leq 5^{2^{dj}-1}$ .

For any  $j$  for which  $2^{-j}\sqrt{d} \leq \varepsilon/2$ , we have therefore obtained an  $\varepsilon$ -cover  $\mathcal{H}$  of  $\mathcal{F}$  satisfying  $\log |\mathcal{H}| \lesssim 2^{dj}$ . Choosing  $2^j$  to be the smallest power of two larger than  $2\sqrt{d}/\varepsilon$  yields the claim.  $\square$

With the bound of Lemma 2.7 in hand, we can give another proof of Proposition 2.5.

*Proof of Proposition 2.5.* Since  $\|f\|_\infty \leq \sqrt{d}$  for all  $f \in \overline{\text{Lip}}_1([0, 1]^d)$ , by Proposition 2.6 and (2.6), for any  $\tau > 0$ ,

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \tau + \frac{1}{\sqrt{n}} \int_\tau^{\sqrt{d}} \sqrt{\log N(\varepsilon, \overline{\text{Lip}}_1([0, 1]^d))} \, d\varepsilon.$$

Applying Lemma 2.7 yields

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \tau + \frac{1}{\sqrt{n}} \int_\tau^{\sqrt{d}} (4\sqrt{d}/\varepsilon)^{d/2} \, d\varepsilon.$$

We now consider the bound separately for  $d = 1$  and  $d > 1$ . If  $d = 1$ , then we may take  $\tau = 0$  to obtain

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \frac{1}{\sqrt{n}} \int_0^1 (4/\varepsilon)^{1/2} \, d\varepsilon \lesssim n^{-1/2}.$$

If  $d > 1$ , then  $\varepsilon^{-d/2}$  is no longer integrable at 0, so we take  $\tau = 4\sqrt{d}n^{-1/d}$  to obtain

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{d}n^{-1/d} + \frac{1}{\sqrt{n}} \int_{4\sqrt{d}n^{-1/d}}^{\sqrt{d}} (4\sqrt{d}/\varepsilon)^{d/2} \, d\varepsilon.$$

When  $d = 2$ , the integral is  $O(\log n)$ , and we obtain  $\mathbb{E}W_1(\mu_n, \mu) \lesssim (\log n)/\sqrt{n}$ . When  $d > 2$ , the integral is  $O(n^{1/2-1/d})$ , and we obtain  $\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{d}n^{-1/d}$ .  $\square$

Though these two proofs of Proposition 2.5 look quite different, they are in fact very similar: in both cases, we employ a multi-scale decomposition of  $[0, 1]^d$ . The dyadic partitioning argument uses this decomposition to construct a coupling in the primal; the chaining argument uses this decomposition to control the covering numbers of Lipschitz functions in the dual.

## 2.4 A finer analysis for $d = 2$

Both the dyadic partition argument presented in Section 2.2 and the chaining argument presented in Section 2.3 suffer from the defect that they fail to obtain the correct rate for the Wasserstein law of large numbers in two dimensions. This fact is related to the fact that  $d = 2$  is the “critical” case for the behavior  $\mathbb{E}W_1(\mu_n, \mu)$ —it can be shown that in  $d = 1$ , the cost of the optimal transport between  $\mu_n$  and  $\mu$  is dominated by “global” features and that when  $d \geq 3$ , the cost of optimal transport is dominated by “local” irregularities. In dimension 2, by contrast, irregularities at all scales contribute simultaneously, and bounding the optimal cost requires more care.

The correct rate for  $d = 2$  was first discovered by Ajtai, Komlós, and Tusnády [AKT84] by a somewhat delicate argument. In this section, we present an ingenious approach due to Bobkov and Ledoux [BL21] that obtains the correct rate by simpler means. This proof is based on Fourier analysis, and as a first step, we show that we can focus our attention on periodic functions, to which the tools of Fourier analysis can naturally be applied. This gives rise to the following periodic version of the Wasserstein distance: for probability measures  $\mu$  and  $\nu$  on  $\mathbb{R}^d$ , define

$$\widetilde{W}_1(\mu, \nu) = \sup_{f \in \widetilde{\text{Lip}}} \int f(\mathrm{d}\mu - \mathrm{d}\nu), \quad (2.7)$$

where  $\widetilde{\text{Lip}}$  denotes the set of 1-Lipschitz,  $2\pi$ -periodic  $C^\infty$  functions on  $\mathbb{R}^d$ . For measures on the cube, this definition actually agrees with standard Wasserstein distance.

**Lemma 2.8.** *If the supports of  $\mu$  and  $\nu$  lie in  $[0, 1]^d$ , then  $\widetilde{W}_1(\mu, \nu) = W_1(\mu, \nu)$ .*

*Proof.* The point of this lemma is that, under the restriction on the support of  $\mu$  and  $\nu$ , we can assume that the Lipschitz functions appearing in the dual representation of  $W_1$  are both periodic and smooth.

We first handle the former restriction. Define a metric  $\mathbf{d}_{\mathbb{T}^d}$  on  $\mathbb{R}^d$  by

$$\mathbf{d}_{\mathbb{T}^d}(x, y) = \min_{z \in \mathbb{Z}^d} \|x - y - 2\pi z\|.$$

The notation  $\mathbf{d}_{\mathbb{T}^d}$  is used to emphasize that this is the metric that arises from identifying the opposite faces of  $[0, 2\pi]^d$  so that it becomes a flat torus. Given any  $f \in \text{Lip}_1([0, 1]^d)$ , define the function  $\tilde{f} : \mathbb{R}^d \rightarrow \mathbb{R}$  by

$$\tilde{f}(y) = \sup_{x \in [0,1]^d} \{f(x) - d_{\mathbb{T}^d}(x, y)\}. \quad (2.8)$$

For each  $x \in [0, 1]^d$ , the function  $y \mapsto f(x) - d_{\mathbb{T}^d}(x, y)$  is  $2\pi$ -periodic and Lipschitz with respect to the Euclidean metric on  $\mathbb{R}^d$  (since both facts are true of  $d_{\mathbb{T}^d}(x, y)$ ). Both periodicity and Lipschitzness are preserved by taking pointwise suprema, so these properties are inherited by  $\tilde{f}$  as well. We also note the crucial fact that  $\tilde{f} = f$  on  $[0, 1]^d$ : indeed, for  $y \in [0, 1]^d$ , we clearly have  $\tilde{f}(y) \geq f(y)$  by choosing  $x = y$  in (2.8). On the other hand, since  $d_{\mathbb{T}^d}(x, y) = \|x - y\|$  for any  $x, y \in [0, 1]^d$ , we also have

$$f(x) - d_{\mathbb{T}^d}(x, y) = f(x) - \|x - y\| \leq f(y) \quad \forall x \in [0, 1]^d,$$

where the inequality follows from the fact that  $f$  is Lipschitz. Taking suprema on both sides yields  $\tilde{f}(y) \leq f(y)$ .

Since the supports of  $\mu$  and  $\nu$  lie in  $[0, 1]^d$ , we therefore have, for any  $f \in \text{Lip}_1([0, 1]^d)$

$$\int f(d\mu - d\nu) = \int \tilde{f}(d\mu - d\nu),$$

where the function on the right side is Lipschitz and  $2\pi$ -periodic. This implies that we can always assume that the functions appearing in the dual representation of  $W_1$  are periodic.

The restriction to smooth functions is routine: since any Lipschitz function can be uniformly approximated by a smooth function, we can always assume that the functions in question are  $C^\infty$ .  $\square$

Given a probability measure  $\mu$ , we denote by  $\phi_\mu$  its Fourier transform (or characteristic function):

$$\phi_\mu(m) = \int e^{i\langle m, z \rangle} \mu(dz), \quad m \in \mathbb{Z}^d. \quad (2.9)$$

The basis of the Bobkov–Ledoux argument is the following proposition.

**Proposition 2.9.**

$$\widetilde{W}_1(\mu, \nu)^2 \leq \sum_{m \neq 0} \|m\|^{-2} |\phi_\mu(m) - \phi_\nu(m)|^2, \quad (2.10)$$

where the sum is over all nonzero  $m \in \mathbb{Z}^d$  and  $\|m\|^2 = m_1^2 + \cdots + m_d^2$ .

Before giving the proof, we pause for a moment to compare Proposition 2.9 to Theorem 2.2. Both results give a bound on  $W_1(\mu, \nu)$  by comparing them at different scales: in the case of Theorem 2.2, this is done by calculating how much they differ on smaller and smaller cubes, in the case of Proposition 2.9, this is done by calculating how much they differ at higher and higher frequencies. In both cases, each term in the sum is weighted by the scale of the comparison ( $2^{-j}$  in the case of Theorem 2.2,  $\|m\|^{-2}$  in the case of Proposition 2.9). The key difference between these bounds is that Theorem 2.2 has an  $\ell^1$  flavor, whereas Proposition 2.9 has an  $\ell^2$  flavor. This difference turns out to be the source of the  $\sqrt{\log n}$  savings in the rate for  $d = 2$ .

*Proof of Proposition 2.9.* Given a  $2\pi$ -periodic  $C^\infty$  function  $f$ , we can expand it as a Fourier series:

$$f(x) = \sum_{m \in \mathbb{Z}^d} \hat{f}_m e^{i\langle m, x \rangle},$$

where the coefficients  $\hat{f}_m$  tend to zero faster than any polynomial as  $\|m\| \rightarrow \infty$ . We may therefore differentiate term-by-term and apply Parseval's identity to obtain

$$\frac{1}{(2\pi)^d} \int_{[0, 2\pi]^d} (\partial_i f(x))^2 dx = \sum_{m \in \mathbb{Z}^d} m_i^2 |\hat{f}(m)|^2,$$

and summing over the coordinates yields

$$\frac{1}{(2\pi)^d} \int_{[0, 2\pi]^d} \|\nabla f(x)\|^2 dx = \sum_{m \in \mathbb{Z}^d} \|m\|^2 |\hat{f}(m)|^2.$$

If we assume that  $f$  is 1-Lipschitz, then  $\|\nabla f(x)\| \leq 1$  for all  $x \in [0, 2\pi]^d$ , so

$$\sum_{m \in \mathbb{Z}^d} \|m\|^2 |\hat{f}(m)|^2 = \frac{1}{(2\pi)^d} \int_{[0, 2\pi]^d} \|\nabla f(x)\|^2 dx \leq 1. \quad (2.11)$$

Fubini's theorem therefore implies that for any 1-Lipschitz,  $2\pi$ -periodic  $C^\infty$  function  $f$ ,

$$\int f (\mathrm{d}\mu - \mathrm{d}\nu) = \int \sum_{m \in \mathbb{Z}^d} \hat{f}_m e^{i\langle m, x \rangle} (\mu(\mathrm{d}x) - \nu(\mathrm{d}x))$$

$$\begin{aligned}
&= \sum_{m \in \mathbb{Z}^d} \hat{f}_m (\phi_\mu(m) - \phi_\nu(m)) \\
&= \sum_{m \in \mathbb{Z}^d \setminus \{0\}} \hat{f}_m (\phi_\mu(m) - \phi_\nu(m)),
\end{aligned}$$

where the last equality follows from the fact that  $\phi_\mu(0) = \phi_\nu(0) = 1$ . The result then follows from the Cauchy–Schwarz inequality and (2.11).  $\square$

Unfortunately, Proposition 2.9 is often vacuous—if  $\mu$  is not absolutely continuous, then  $\phi_\mu$  is not integrable, and the sum in (2.10) can diverge. This problem is immediately apparent when attempting to apply Proposition 2.9 to the singular empirical measure  $\mu_n$ . The solution to this issue is to inject additional regularity into the problem by convolving with Gaussians. For any  $\varepsilon > 0$ , we denote by  $\nu \star \gamma_\varepsilon$  the convolution of  $\nu$  with a  $\mathcal{N}(0, \varepsilon I)$  distribution; equivalently,  $\nu \star \gamma_\varepsilon$  is the law of  $X + \sqrt{\varepsilon}Z$  where  $X \sim \nu$  and  $Z \sim \mathcal{N}(0, I)$  are independent. We first recall the effect that this smoothing has on the Fourier transform.

**Lemma 2.10.** *For all  $\varepsilon > 0$ ,*

$$\phi_{\nu \star \gamma_\varepsilon}(m) = \phi_\nu(m) e^{-\varepsilon \|m\|^2/2} \quad \forall m \in \mathbb{Z}^d.$$

*Proof.* This follows directly from the representation

$$\phi_{\nu \star \gamma_\varepsilon}(m) = \int e^{i\langle m, y \rangle} (\nu \star \gamma_\varepsilon)(dy) = \mathbb{E} e^{i\langle m, X + \sqrt{\varepsilon}Z \rangle}$$

for  $X \sim \nu$  and  $Z \sim \mathcal{N}(0, I)$  independent.  $\square$

The smoothing operation is useful because it immediately ensures that the Fourier transform of the resulting measure is well-behaved. Moreover, smoothing only changes  $\widetilde{W}_1$  by a small amount.

**Lemma 2.11.** *For all  $\varepsilon > 0$ ,*

$$\widetilde{W}_1(\mu, \nu) \leq \widetilde{W}_1(\mu, \nu \star \gamma_\varepsilon) + \sqrt{d\varepsilon}.$$

*Proof.* First, the expression  $\widetilde{W}_1$  satisfies the triangle inequality; this follows directly from the definition in (2.7):

$$\widetilde{W}_1(\mu, \nu') + \widetilde{W}_1(\nu', \nu) = \sup_{f \in \widetilde{\text{Lip}}} \int f(d\mu - d\nu') + \sup_{f \in \widetilde{\text{Lip}}} \int f(d\nu' - d\nu)$$

$$\begin{aligned} &\geq \sup_{f \in \widetilde{\text{Lip}}} \left\{ \int f(\mathrm{d}\mu - \mathrm{d}\nu') + \int f(\mathrm{d}\nu' - \mathrm{d}\nu) \right\} \\ &= \widetilde{W}_1(\mu, \nu). \end{aligned}$$

Second,  $\widetilde{W}_1$  is dominated by  $W_1$ , since the supremum in (2.7) is taken over a strict subset of  $\text{Lip}$ . Combining these facts yields

$$\widetilde{W}_1(\mu, \nu) \leq \widetilde{W}_1(\mu, \nu \star \gamma_\varepsilon) + \widetilde{W}_1(\nu, \nu \star \gamma_\varepsilon) \leq \widetilde{W}_1(\mu, \nu \star \gamma_\varepsilon) + W_1(\nu, \nu \star \gamma_\varepsilon).$$

We now use the fact that  $(X, X + \sqrt{\varepsilon}Z)$  with  $X \sim \nu, Z \sim \mathcal{N}(0, I)$  independent is a coupling between  $\nu$  and  $\nu \star \gamma_\varepsilon$ , so that

$$W_1(\nu, \nu \star \gamma_\varepsilon) \leq \mathbb{E}\|X - (X + \sqrt{\varepsilon}Z)\| = \sqrt{\varepsilon} \mathbb{E}\|Z\| \leq \sqrt{d\varepsilon}.$$

This concludes the proof.  $\square$

Combining the preceding two lemmas yields the following corollary to Proposition 2.9.

**Corollary 2.12.** *For any  $\varepsilon > 0$ ,*

$$\widetilde{W}_1(\mu, \nu) \leq \sqrt{\sum_{m \neq 0} \|m\|^{-2} e^{-\varepsilon \|m\|^2} |\phi_\mu(m) - \phi_\nu(m)|^2} + 2\sqrt{d\varepsilon}.$$

We can now prove the desired bound.

**Theorem 2.13.** *For any probability measure  $\mu$  with support in  $[0, 1]^2$ ,*

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{\log n/n}.$$

*Proof.* Since  $\mu$  and  $\mu_n$  have support lying in  $[0, 1]^2$ , we may equivalently prove an upper bound on  $\mathbb{E}\widetilde{W}_1(\mu_n, \mu)$ . Applying Corollary 2.12 and Jensen's inequality yields, for any  $\varepsilon > 0$ ,

$$\mathbb{E}\widetilde{W}_1(\mu_n, \mu) \leq \sqrt{\sum_{m \neq 0} \|m\|^{-2} e^{-\varepsilon \|m\|^2} \mathbb{E}|\phi_{\mu_n}(m) - \phi_\mu(m)|^2} + 2\sqrt{2\varepsilon}.$$

We can write  $\phi_{\mu_n}(m) - \phi_\mu(m) = \frac{1}{n} \sum_{i=1}^n \{e^{i\langle m, X_i \rangle} - \mathbb{E}e^{i\langle m, X_i \rangle}\}$ , and since  $|e^{i\langle m, X_i \rangle}| = 1$  almost surely we conclude that

$$\mathbb{E}|\phi_{\mu_n}(m) - \phi_\mu(m)|^2 \leq n^{-1} \quad \forall m \in \mathbb{Z}^d. \quad (2.12)$$

Continuing, we have for any  $\varepsilon > 0$ ,

$$\mathbb{E}\widetilde{W}_1(\mu_n, \mu) \leq n^{-1/2} \sqrt{\sum_{m \neq 0} \|m\|^{-2} e^{-\varepsilon \|m\|^2}} + 2\sqrt{2\varepsilon}. \quad (2.13)$$

By comparing the sum to the integral  $\int_{\|x\| \geq 1} \|x\|^{-2} e^{-\varepsilon \|x\|^2} dx$ , we obtain that the sum is of order  $\log(1/\varepsilon)$ . Therefore, we obtain

$$\mathbb{E}\widetilde{W}_1(\mu_n, \mu) \lesssim \sqrt{\log(1/\varepsilon)/n} + \sqrt{\varepsilon}.$$

Choosing  $\varepsilon = n^{-1}$  gives the claim.  $\square$

## 2.5 Applications

### 2.5.1 Estimation of Wasserstein distances

So far, we have focused on estimating a measure  $\mu$  in Wasserstein distance using the empirical measure  $\mu_n$ . As the title of this chapter indicates, we are often interested in the estimation of Wasserstein distances. Indeed, Wasserstein distances are central to many statistical tasks. For example, one of the first applications of the 1-Wasserstein distance (under the name “earth mover’s distance”) to machine learning was in the context of information retrieval where it was used to measure the distance between images [RTG00]. Other immediate examples include nearest neighbors [BDI<sup>+</sup>20, Pon23] and regression [GP22, CLM23] for example.

The goal of estimation is to produce an estimator  $\widehat{W}$  of  $W_1(\mu, \nu)$  using i.i.d. data  $X_1, \dots, X_m \sim \mu$  and  $Y_1, \dots, Y_n \sim \nu$ . A natural candidate is the *plug-in* estimator  $\widehat{W} := W_1(\mu_m, \nu_n)$  where  $\mu_m$  and  $\nu_n$  are the empirical measures associated to the samples above. A performance bound for this estimator can be readily obtained using the triangle inequality and Proposition 2.1: for  $d \geq 3$ ,

$$\mathbb{E}|W_1(\mu_m, \nu_n) - W_1(\mu, \nu)| \leq \mathbb{E}W_1(\mu_m, \mu) + \mathbb{E}W_1(\nu_n, \nu) \lesssim (m \wedge n)^{-1/d}.$$

This coarse bound turns out to be sharp in general. Moreover, using the more general result (2.21) presented at the end of this Chapter, we can get that for  $d > 2p$ ,

$$\mathbb{E}|W_p(\mu_m, \nu_n) - W_p(\mu, \nu)| \leq \mathbb{E}W_p(\mu_m, \mu) + \mathbb{E}W_p(\nu_n, \nu) \lesssim (m \wedge n)^{-1/d}.$$

It turns out that when  $p > 1$ , this bound is only sharp when  $\mu$  and  $\nu$  are sufficiently close. Indeed, better rates can be obtained when  $p > 1$  and  $W_p(\mu, \nu) > c > 0$ . For example, when  $p = 2$ , [MNW24] show that

$$\mathbb{E}|W_2(\mu_m, \nu_n) - W_2(\mu, \nu)| \lesssim (m \wedge n)^{-2/d}.$$

That paper also shows that this rate is essentially sharp. While significant, this improvement shows that estimation of Wasserstein distances still suffers from the curse of dimensionality.

### 2.5.2 Hypothesis testing

These upper bounds can be readily applied to two classical non-parametric hypothesis testing problems: goodness-of-fit and two-sample (a.k.a. homogeneity) testing.

Consider first the goodness-of-fit test. Given observation  $X_1, \dots, X_n$  i.i.d. from some unknown distribution  $\mu$ , and a fixed distribution  $\mu^0$ , the goal is to test

$$H_0 : \mu = \mu^0 \quad \text{vs.} \quad H_1 : \mu \neq \mu^0.$$

For example,  $\mu^0$  can be taken to be a standard Gaussian distribution on  $\mathbb{R}^d$  or the uniform distribution on  $[0, 1]^d$ . There exist many goodness-of-fit tests when  $d = 1$ , for example, the Kolmogorov–Smirnov test for continuous distributions. For discrete distributions, the  $\chi^2$ -test is another popular choice; see, e.g., [LR05, Chapter 14].

Note that the two hypotheses can be written equivalently as

$$H^0 : W_1(\mu, \mu^0) = 0 \quad \text{vs.} \quad H_1 : W_1(\mu, \mu^0) > 0.$$

To study the theoretical limits of a such a setting, we can consider a quantitative version of this testing problem, with hypotheses

$$H^0 : W_1(\mu, \mu^0) = 0 \quad \text{vs.} \quad H_1 : W_1(\mu, \mu^0) > \varepsilon$$

for some  $\varepsilon > 0$ . We then ask how large the separation  $\varepsilon$  must be in order to guarantee that the combined type I and type II errors may be kept small.

Consider a simple test, which consists in rejecting the null hypothesis at level  $\alpha \in (0, 1)$  as soon as  $W_1(\mu_n, \mu^0) > T_n^\alpha$  for some threshold  $T_n^\alpha$  such that

$$\mu^0[W_1(\mu_n, \mu^0) > T_n^\alpha] = \alpha.$$

If we observe  $X_i = x_i$ ,  $i = 1, \dots, n$ , a  $p$ -value for such a test may be computed as

$$\mu^0[W_1(\mu_n, \mu^0) > W_1(\mu_n^{\text{obs}}, \mu^0)]$$

where

$$\mu_n^{\text{obs}} = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}.$$

Both the computation of  $T_n^\alpha$  and of the  $p$ -value require understanding the actual distribution of  $W_1(\mu_n, \mu^0)$  under the null hypothesis. Our results above are actually quite far from achieving this level of precision since we only know an upper bound on  $\mathbb{E}[W_1(\mu_n, \mu^0)]$  when  $\mu_0$  is supported on the unit cube. Nevertheless, these bounds are sufficient to paint a rather disappointing picture of the potential of the Wasserstein distance in multivariate goodness-of-fit tests. Indeed, our results suggest that we must take  $T_n^\alpha \gtrsim n^{-1/d}$  in order to control the type I error of this test, and therefore that a separation of  $\varepsilon \gg n^{-1/d}$  is necessary to have reasonable power. Even in moderate dimensions, this level of separation is quite large.

It turns out that the test described above is not optimal for this problem, and that consistent testing is possible with the smaller separation  $n^{-2/d}$  via a different approach. Nevertheless, the simple fact remains that the slow convergence of  $W_1$  is an impediment to the use of goodness of fit tests based on the Wasserstein distance.

An explanation for this phenomenon is that a test based on  $W_1$  tries to be powerful against too many alternatives. Assume for the sake of discussion that  $\mu^0$  is the standard Gaussian distribution over  $\mathbb{R}^d$ . The 1-Wasserstein distance does not discriminate between distributions that are not  $\mu^0$ : Gaussian distributions, distributions with smooth densities, those with discontinuous densities, or even discrete distributions. Our test  $\{W_1(\mu_n, \mu^0) > T_n^\alpha\}$  tries to detect all of them and spreads thin, resulting in low power against all alternatives. This behavior is to be contrasted with a simple parametric test, for example the Wald test  $\{|\bar{X}_n| > \tau_n\}$  where  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ , which simply tries to detect if the mean of  $\mu$  differs from that of  $\mu^0$ . This test is clearly unable to detect even if  $\mu$  is a Rademacher distribution, which is quite far from  $\mu_0$ , but it focuses all of its efforts on shifts in means: when these happen, it can detect them very accurately.

The manifestation of the curse of dimensionality also extends to two-sample testing where one observes two samples  $X_1, \dots, X_m$  from  $\mu$  and  $Y_1, \dots, Y_n$  from  $\nu$  and the goal is to test

$$H_0 : \mu = \nu \quad \text{vs.} \quad H_1 : W_1(\mu, \nu) > \varepsilon.$$

Denote the corresponding empirical distributions by

$$\mu_m = \frac{1}{m} \sum_{i=1}^m \delta_{X_i}, \quad \nu_n = \frac{1}{n} \sum_{j=1}^n \delta_{Y_j}.$$

In this context it is natural to reject the null hypothesis if  $W_1(\mu_m, \nu_n)$  is large. Akin to the goodness-of-fit test, such tests require a sample size that is exponential in the dimension to achieve any reasonable power.

The conclusion of this section is that Wasserstein distances suffer from the curse of dimensionality and are therefore unsuitable for statistical applications of moderate dimension. In Section 2.8 we describe various regularizations of Wasserstein distances that escape the curse of dimensionality and have been successfully applied in large-scale statistical applications.

## 2.6 Optimality

We have established upper bounds on the Wasserstein distance between the empirical distribution  $\mu_n$  and the data generating distribution  $\mu$  in two different ways: using the primal and using the dual formulation of the problem. Omitting idiosyncrasies associated to low dimensions, we found that  $\mu_n$  estimates  $\mu$  in  $W_1$  distance at a rate of order  $n^{-1/d}$ . While this result readily yields consistency, the rate is slow even in moderate dimensions and is symptomatic of the curse of dimensionality that plagues most non-parametric methods. One could wonder then whether such rates can be improved.

Note that there are two ways to potentially improve these rates. The most obvious one would be to provide a tighter analysis than the one above and show that in fact,  $\mathbb{E}[W_1(\mu_n, \mu)]$  is much smaller than  $n^{-1/d}$ . Another possibility would be that while this rate is tight for the empirical measures  $\mu_n$ , there could be another estimator  $\tilde{\mu}_n$  of  $\mu$  that enjoys much faster rates. In fact, the answer to both questions, while different in nature, is negative, as illustrated by lower bounds.

While a negative answer to the second question implies a negative answer to the first one—if no estimator can estimate  $\mu$  faster than  $n^{-1/d}$  then certainly the empirical measure  $\mu_n$  cannot—we also make the negative answer to the first question explicit since it is, in some

sense stronger. Indeed, we show below that even in the case where  $\mu$  is the uniform measure on  $[0, 1]^d$  then,  $\mathbb{E}[W_1(\mu_n, \mu)] \gtrsim n^{-1/d}$ . However, in that case, there is clearly a better estimator than  $\mu_n$ : simply take  $\tilde{\mu}_n = \mu$  itself! The answer to the second question relies on the theory of minimax lower bounds as in [Tsy09, Chapter 2] and states that for any estimator, i.e., any measurable function  $\tilde{\mu}_n = \tilde{\mu}_n(X_1, \dots, X_n)$  of the data  $X_1, \dots, X_n$ , there exists  $\mu$  supported on  $[0, 1]^d$  such that  $\mathbb{E}[W_1(\tilde{\mu}_n, \mu)] \gtrsim n^{-1/d}$ . Unlike the lower bound for the empirical measure  $\mu_n$ , in the minimax lower bounds, the unfavorable distribution  $\mu$  is not explicit.

### 2.6.1 Lower bounds for the empirical measure $\mu_n$

The goal of this section is to show that any distribution supported on  $n$  points has to be far from the uniform measure on  $[0, 1]^d$  in  $W_1$  distance.

**Theorem 2.14.** *Fix  $d \geq 3$  and let  $\mu$  denote the uniform measure on  $[0, 1]^d$ . Then for any measure  $\tilde{\mu}_n$  supported on  $n$  points  $x_1, \dots, x_n \in \mathbb{R}^d$ , it holds*

$$W_1(\tilde{\mu}_n, \mu) \geq \frac{1}{108d} n^{-1/d}.$$

*Proof.* We employ the dual formulation of Theorem 1.21 since proving a lower bound on  $W_1$  can be done by simply exhibiting a 1-Lipschitz function, which we define as follows. Given  $x \in [0, 1]^d$ , let  $\xi(x) \in \{x_1, \dots, x_n\}$  denote the closest point to  $x$  in  $\{x_1, \dots, x_n\}$  (ties are broken arbitrarily). Next, consider the function

$$f_n(x) = \|x - \xi_n(x)\|,$$

which is 1-Lipschitz thanks to the reverse triangle inequality. Moreover, for any  $i = 1, \dots, n$ , we have  $f_n(x_i) = 0$  so that  $\int f_n d\tilde{\mu}_n = 0$ . Hence

$$W_1(\tilde{\mu}_n, \mu) \geq \int f_n d\mu = \int \|x - \xi_n(x)\| \mu(dx).$$

To bound this quantity from below, we show that  $\mu$  places significant mass on points that are far from *any*  $x_i$ . To that end, consider a partition  $\mathcal{Q}$  of  $[0, 1]^d$  into cubes of side length  $(2n)^{-1/d}$ . Since  $|\mathcal{Q}| = 2n$ , there exist  $n$  such cubes  $Q_1, \dots, Q_n$  that do not contain any of the  $x_i$ 's. Let  $Q \in \mathcal{Q}$  be one such cube with center  $q$  and consider its subcube  $Q' \subset Q$  also with center  $q$  but with a smaller side length than  $Q$  by a factor of  $1 - 2/d$ . Using Minkowski sum notation, we can write this as:

$$Q' = \left(1 - \frac{2}{d}\right) (Q - \{q\}) + \{q\}.$$

By construction, any  $x \in Q'$  satisfies

$$\|x - \xi_n(x)\| \geq \inf_{\substack{x \in Q' \\ y \in Q^c}} \|x - y\| = \frac{1}{d} \cdot (2n)^{-1/d}.$$

Hence

$$\int \|x - \xi_n(x)\| \mu(dx) \geq \sum_{i=1}^n \int_{Q'_i} \|x - \xi_n(x)\| \mu(dx) \geq \frac{(2n)^{-1/d}}{d} \sum_{i=1}^n \mu(Q'_i).$$

We conclude by observing that

$$\mu(Q'_i) = \left(\frac{1-2/d}{(2n)^{1/d}}\right)^d \geq \frac{1}{54n},$$

where we used the fact that  $d \mapsto (1-2/d)^d$  is increasing and that  $d \geq 3$ .  $\square$

Theorem 2.14 shows that  $W_1(\mu_n, \mu)$  is indeed of order  $n^{-1/d}$  at least for  $d \geq 3$ . In fact the lower bound holds almost surely in  $X_1, \dots, X_n$  since it only exploits the fact that  $\mu_n$  has a support of size at most  $n$ . Note that the  $d$  dependence in Theorem 2.14 is off by some polynomial factors in  $d$ . It can be shown that the  $\sqrt{d}$  factor in Proposition 2.5 cannot be improved; see Exercise 1.

### 2.6.2 Minimax lower bounds

While it is hard to think of a better estimator for  $\mu$  than  $\mu_n$  in general (in Section 2.7 we show that we can under additional assumptions on  $\mu$ ) it could be the case that there exists another estimator  $\tilde{\mu}_n$  for which  $\mathbb{E}[W_1(\tilde{\mu}_n, \mu)]$  is smaller than  $\mathbb{E}[W_1(\mu_n, \mu)]$  uniformly over all measures  $\mu$ . This possibility is ruled out by the following minimax lower bound.

**Theorem 2.15.** *Fix  $d \geq 3, n \geq 8$  and let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations from a distribution  $\mu$  on  $\mathbb{R}^d$ . For any estimator  $\tilde{\mu}_n$ , i.e., any measurable function of  $X_1, \dots, X_n$ , there exists a measure  $\mu$  supported on  $[0, 1]^d$  such that*

$$\mathbb{E}_\mu[W_1(\tilde{\mu}_n, \mu)] \geq \frac{1}{16} (2n)^{-1/d}.$$

*Proof.* Our proof relies on classical techniques for minimax lower bounds. In particular, we use Theorem 2.12 in [Tsy09]. According to this theorem, if we can find  $2^m$  probability measures indexed by  $\omega \in \{-1, 1\}^m$  each supported on  $[0, 1]^d$  such that

- (i)  $W_1(\mu^{(\omega)}, \mu^{(\omega')}) \geq \frac{r_n}{2} \sum_{j=1}^m |\omega_j - \omega'_j|$  for any  $\omega, \omega' \in \{-1, 1\}^m$ ,
- (ii) for any  $\omega, \omega' \in \{-1, 1\}^m$  differing in at most one coordinate,

$$\mathsf{KL}(\mu^{(\omega)} \parallel \mu^{(\omega')}) \leq \frac{1}{2n},$$

then for any estimator  $\tilde{\mu}_n$  based on  $n$  i.i.d. observations, there exists  $\omega \in \{-1, 1\}^m$  such that

$$\mathbb{E}_{\mu^{(\omega)}}[W_1(\tilde{\mu}_n, \mu^{(\omega)})] \geq \frac{mr_n}{4}.$$

In our construction, we take  $m = n$  and define the measures  $\mu^{(\omega)}$  to be supported on a discrete set as follows. As in the proof of Theorem 2.14, let  $\mathcal{Q}$  denote a partition of  $[0, 1]^d$  into  $2n$  cubes of side length  $(2n)^{-1/d}$  and let  $q_1, \dots, q_{2n}$  denote their centers. Let  $\mu^{(0)}$  denote the uniform measure on  $\{q_1, \dots, q_{2n}\}$ :

$$\mu^{(0)} = \frac{1}{2n} \sum_{i=1}^{2n} \delta_{q_i}.$$

For  $\omega \in \{-1, 1\}^n$ , let  $\mu^{(\omega)}$  denote a perturbation of  $\mu^{(0)}$  defined as

$$\mu^{(\omega)} = \mu^{(0)} + \frac{\alpha}{2n} \sum_{i=1}^n \omega_i (\delta_{q_i} - \delta_{q_{n+i}}),$$

where  $\omega = (\omega_1, \dots, \omega_n)$  and  $\alpha \in (0, 1)$  is to be defined later. Note that  $\mu^{(\omega)}$  is a probability measure.

Since  $\|q_j - q_k\| \geq (2n)^{-1/d}$  for  $j \neq k$  we have

$$W_1(\mu^{(\omega)}, \mu^{(\omega')}) \geq \frac{\alpha}{2n} (2n)^{-1/d} \sum_{j=1}^n |\omega_j - \omega'_j| =: \frac{r_n}{2} \sum_{j=1}^n |\omega_j - \omega'_j|$$

for any  $\omega, \omega' \in \{0\}^n \cup \{-1, 1\}^n$ .

It remains to show that (ii) holds for a suitable choice of  $\alpha$ . To that end, suppose that  $\omega$  and  $\omega'$  differ on the  $j$ th coordinate. Observe that

$$\begin{aligned}
\text{KL}(\mu^{(\omega)} \parallel \mu^{(\omega')}) &= \sum_{i=1}^{2n} \mu^{(\omega)}(q_i) \log \left( \frac{\mu^{(\omega)}(q_i)}{\mu^{(\omega')}(q_i)} \right) \\
&= \frac{1}{2n} \left\{ (1 + \alpha\omega_j) \log \frac{1 + \alpha\omega_j}{1 - \alpha\omega_j} + (1 - \alpha\omega_j) \log \frac{1 - \alpha\omega_j}{1 + \alpha\omega_j} \right\} \\
&= \frac{\alpha}{n} \log \frac{1 + \alpha}{1 - \alpha},
\end{aligned}$$

and this quantity is smaller than  $\frac{1}{2n}$  if  $\alpha = \frac{1}{4}$ . With this choice of  $\alpha$ , we obtain

$$r_n = \frac{1}{4n} (2n)^{-1/d}, \quad (2.14)$$

which implies the desired bound.  $\square$

## 2.7 Faster rates for smooth measures

The preceding section indicates that no estimator can avoid the slow  $n^{-1/d}$  rate in general.

There are multiple ways to alleviate this curse of dimensionality and the rest of this chapter illustrates two main approaches. In this section, we impose smoothness assumptions on the measure  $\mu$ . Such assumptions are classical in non-parametric statistics and known to partially mitigate the curse of dimensionality. In the next section, we describe how modifying/regularizing the Wasserstein distance into other distances that are similar in nature can be used to bypass the curse of dimensionality altogether.<sup>4</sup>

The fact that imposing smoothness conditions on  $\mu$  can lead to better rates is natural in light of the lower bound presented in Theorem 2.15. The measures used in the proof are mixtures of Dirac masses and are therefore highly “irregular” in the sense that they do not even possess densities with respect to the Lebesgue measure. By assuming that  $\mu$  is smooth, we rule out these pathological examples.

For mathematical convenience, we consider smooth densities defined on the torus  $\mathbb{T}^d := \mathbb{R}^d / (2\pi\mathbb{Z})^d$ . Concretely, this can be viewed as isomorphic to the set  $[0, 2\pi)^d$ , equipped with the metric  $d_{\mathbb{T}^d}(x, y) := \min_{z \in \mathbb{Z}^d} \|x - y - 2\pi z\|$ . On this space, the Wasserstein distance coincides with  $\tilde{W}_1$  defined in Section 2.4.

---

<sup>4</sup> Since we are interested in improvements to the exponent in the rate of decay, in the remainder of this chapter we ignore dimension-dependent constants in the bounds for clarity.

We focus on the torus so that we can again use the tools of Fourier analysis. A similar but slightly more technical argument can extend the results of this section to standard Euclidean space. Note that the  $n^{-1/d}$  minimax lower bound proved in the previous section still holds on the torus, so in moving to this setting we have not affected the fundamental statistical difficulty of the problem.

We consider a probability measure  $\mu$  on  $\mathbb{T}^d$  with a density, which we also denote by  $\mu$ . We make the assumption that the density of  $\mu$  is smooth, in the sense that it lies in a *Sobolev space*.

**Definition 2.16.** *Given a positive integer  $s$ , the Sobolev space  $\mathcal{H}^s$  consists of all functions  $f : \mathbb{T}^d \rightarrow \mathbb{R}$  such that for every multi-index  $\alpha$  with  $|\alpha| \leq s$ , the derivative  $d^\alpha f$  lies in  $L^2$ . Given  $f \in \mathcal{H}^s$ , its Sobolev norm is defined to be*

$$\|f\|_{\mathcal{H}^s}^2 = \max_{|\alpha| \leq s} \int_{\mathbb{T}^d} \|d^\alpha f\|^2 dx.$$

The importance of the Sobolev spaces lies in their close connection with Fourier series. If  $\mu \in \mathcal{H}^s$ , then its Fourier transform (2.9) satisfies

$$\sum_{m \in \mathbb{Z}^d} (1 + \|m\|^{2s}) |\phi_\mu(m)|^2 < \infty.$$

Moreover, this expression in terms of Fourier coefficients actually gives an equivalence of norms. Indeed, from the Fourier representation

$$\mu(x) \propto \sum_{m \in \mathbb{Z}^d} \phi_\mu(m) e^{-i\langle m, x \rangle}$$

we obtain, for any multi-index  $\alpha$ ,

$$d^\alpha \mu(x) \propto \sum_{m \in \mathbb{Z}^d} m^{2\alpha} \phi_\mu(m) e^{-i\langle m, x \rangle},$$

where  $m^\alpha = m_1^{\alpha_1} \cdots m_d^{\alpha_d}$ . By Parseval's identity,

$$\int_{\mathbb{T}^d} \|d^\alpha \mu\|^2 dx \asymp \sum_{m \in \mathbb{Z}^d} m^{2\alpha} |\phi_\mu(m)|^2 \tag{2.15}$$

$$\lesssim \sum_{m \in \mathbb{Z}^d} (1 + \|m\|^{2s}) |\phi_\mu(m)|^2. \tag{2.16}$$

On the other hand, by the binomial theorem,

$$\|m\|^{2s} = \sum_{|\alpha|=s} m^{2\alpha}.$$

By (2.15), it holds that

$$\begin{aligned} \sum_{m \in \mathbb{Z}^d} \|m\|^{2s} |\phi_\mu(m)|^2 &= \sum_{m \in \mathbb{Z}^d} \sum_{|\alpha|=s} m^{2\alpha} |\phi_\mu(m)|^2 \\ &\lesssim \sum_{|\alpha|=s} \int_{\mathbb{T}^d} \|d^\alpha \mu\|^2 dx. \end{aligned} \quad (2.17)$$

By (2.16) and (2.17) (applying the latter inequality also for  $s=0$ ), we have shown that

$$\|\mu\|_{\mathcal{H}^s}^2 \asymp \sum_{m \in \mathbb{Z}^d} (1 + \|m\|^{2s}) |\phi_\mu(m)|^2.$$

We construct an estimator  $\tilde{\mu}_n$  obtained by estimating the Fourier coefficients of  $\mu$  for all  $m \in \mathbb{Z}^d$  satisfying  $\|m\| \leq M$ . Concretely, we define

$$\widehat{\phi_\mu}(m) = \phi_{\mu_n}(m) = \frac{1}{n} \sum_{j=1}^n e^{i\langle m, X_j \rangle},$$

then for any  $M \geq 1$  we set

$$\tilde{\mu}_n(x) = \frac{1}{(2\pi)^d} \sum_{\|m\| \leq M} \widehat{\phi_\mu}(m) e^{-i\langle m, x \rangle}.$$

Note that while  $\tilde{\mu}_n$  is always a real-valued function on  $\mathbb{T}^d$  integrating to 1, it may not be positive everywhere; however, we ignore this issue for now. Even when the density  $\tilde{\mu}_n$  takes negative values, the definition of  $\widetilde{W}_1$  in terms of its dual representation (2.7) still gives a sensible interpretation of  $\widetilde{W}_1(\mu, \tilde{\mu}_n)$ .

We have the following result.

**Proposition 2.17.** *Assume  $\mu \in \mathcal{H}^s(\mathbb{T}^d)$  with  $\|\mu\|_{\mathcal{H}^s} \lesssim 1$ . For any  $M \geq 1$  and  $d \geq 3$ , the estimator  $\tilde{\mu}_n$  defined above satisfies*

$$\mathbb{E} \widetilde{W}_1(\mu, \tilde{\mu}_n) \lesssim n^{-1/2} M^{d/2-1} + M^{-(s+1)}.$$

*Proof.* We first note that

$$\widetilde{W}_1(\mu, \tilde{\mu}_n)^2 \lesssim \sum_{m \neq 0, \|m\| \leq M} \|m\|^{-2} |\phi_\mu(m) - \widehat{\phi_\mu}(m)|^2$$

$$+ \sum_{\|m\| > M} \|m\|^{-2} |\phi_\mu(m)|^2.$$

This follows directly from Proposition 2.9, using the fact that the (signed) measure  $\tilde{\mu}_n$  has Fourier coefficients  $\widehat{\phi_\mu}(m)$  for  $\|m\| \leq M$  and zero otherwise. As in Section 2.4, we may use the fact that  $|e^{i\langle m, X_j \rangle}| \leq 1$  to conclude that  $\mathbb{E}|\phi_\mu(m) - \widehat{\phi_\mu}(m)|^2 \leq n^{-1}$ . Therefore,

$$\mathbb{E}\widetilde{W}_1(\mu, \tilde{\mu}_n) \leq n^{-1/2} \sqrt{\sum_{m \neq 0, \|m\| \leq M} \|m\|^{-2}} + \sqrt{\sum_{\|m\| > M} \|m\|^{-2} |\phi_\mu(m)|^2}.$$

Before proceeding, we pause to compare this bound with (2.13). There are two differences: first, the smooth cut-off  $e^{-\varepsilon\|m\|^2}$  in (2.13) has been replaced by the restriction  $\|m\| \leq M$ . This change is inessential: since  $e^{-\varepsilon\|m\|^2} \ll 1$  when  $\|m\|^2 \gg \varepsilon^{-1}$ , the smooth cut off term mimics a restriction to  $\|m\| \lesssim \varepsilon^{-1/2}$ . The second difference is that the term  $2\sqrt{2\varepsilon}$  in (2.13) has been replaced by a term that depends on the higher Fourier coefficients of  $\mu$ . This change is crucial, since, as we now show, it implies that the second term automatically becomes smaller when  $\mu$  is smooth.

Since  $\sum_{m \in \mathbb{Z}^d} \|m\|^{2s} |\phi_\mu(m)|^2 \lesssim 1$ , we may write

$$\begin{aligned} \sum_{\|m\| > M} \|m\|^{-2} |\phi_\mu(m)|^2 &\leq M^{-2(s+1)} \sum_{\|m\| > M} \|m\|^{2s} |\phi_\mu(m)|^2 \\ &\lesssim M^{-2(s+1)}. \end{aligned}$$

For the first term, we can compare the sum with the integral  $(\int_{1 \leq \|x\| \leq M} \|x\|^{-2} dx)^{1/2}$ , which is of order  $M^{d/2-1}$  when  $d \geq 3$ .  $\square$

Tuning  $M$  appropriately, we arrive at the following theorem.

**Theorem 2.18.** *If  $\mu \in \mathcal{H}^s(\mathbb{T}^d)$  with  $\|\mu\|_{\mathcal{H}^s} \lesssim 1$ , then there exists an estimator  $\tilde{\mu}_n$  such that*

$$\mathbb{E}\widetilde{W}_1(\mu, \tilde{\mu}_n) \lesssim n^{-\frac{s+1}{d+2s}}.$$

*Proof.* Apply Proposition 2.17 with  $M \asymp n^{1/(d+2s)}$ .  $\square$

Theorem 2.18 shows that, when  $s > 0$ , the estimator  $\tilde{\mu}_n$  strictly improves over the empirical measure  $\mu_n$ . However, we note that the estimator  $\tilde{\mu}_n$  is a signed measure, which may be viewed as undesirable.

This is a common phenomenon in non-parametric statistics; for instance, in the design of kernel density estimators for very smooth densities, it is necessary to employ higher-order kernels which take negative values. If a positive estimator is desired, then it is possible to show that the estimator  $\bar{\mu}_n$  defined by

$$\bar{\mu}_n := \arg \min_{\nu \in \mathcal{P}(\mathbb{T}^d)} \widetilde{W}_1(\nu, \tilde{\mu}_n)$$

also achieves the bound in Theorem 2.18: indeed, since  $\mu \in \mathcal{P}(\mathbb{T}^d)$ ,

$$\widetilde{W}_1(\mu, \bar{\mu}_n) \leq \widetilde{W}_1(\mu, \tilde{\mu}_n) + \widetilde{W}_1(\bar{\mu}_n, \tilde{\mu}_n) \leq 2\widetilde{W}_1(\mu, \tilde{\mu}_n),$$

so that  $\bar{\mu}_n$  is worse than  $\tilde{\mu}_n$  by a factor of at most 2.

## 2.8 Regularization of Wasserstein distances

The curse of dimensionality that plagues statistical optimal transport has been recognized since its early days. To overcome this limitation, researchers have proposed multiple solutions which can, in retrospect, be viewed as some kind of regularization of the original optimal transport problem. In the rest of this section, we review three examples and demonstrate how they escape the curse of dimensionality.

### 2.8.1 Integral probability metrics

Recall from the dual chaining argument of Section 2.3 that the rate  $n^{-1/d}$  came directly from the entropy number of the class of 1-Lipschitz functions. Lemma 2.7 showed

$$\log N(\varepsilon, \overline{\text{Lip}}_1([0, 1]^d)) \lesssim (4\sqrt{d}/\varepsilon)^d.$$

The polynomial scaling in  $1/\varepsilon$  is characteristic of non-parametric classes, as opposed to parametric classes where this scaling is logarithmic; see e.g., [GN16]. This raises the question of potentially replacing the class of 1-Lipschitz functions with a smaller, ideally parametric, class of functions.

Take for example the class of linear functions on  $\mathbb{R}^d$ :

$$\mathcal{F}_{\text{lin}} := \left\{ f(x) = \langle \theta, x \rangle : \theta, x \in \mathbb{R}^d, \|\theta\| = 1 \right\},$$

and consider the quantity

$$\begin{aligned}\delta(\mu, \nu) &= \sup_{f \in \mathcal{F}_{\text{lin}}} \left\{ \int f \, d\mu - \int f \, d\nu \right\} \\ &= \sup_{\theta \in \mathbb{R}^d, \|\theta\|=1} \left\{ \int \langle \theta, x \rangle \, \mu(dx) - \int \langle \theta, y \rangle \, \nu(dy) \right\} \\ &= \|\mathbb{E}_\mu[X] - \mathbb{E}_\nu[Y]\|.\end{aligned}$$

In particular,  $\delta(\mu, \nu) = 0$  if and only if  $\mu$  and  $\nu$  have the same mean. This is of course not sufficient to say that the two measures are the same so the above quantity does not define a distance between probability measures like the Wasserstein distance. To do so, we need to find a class of test functions  $\mathcal{F}$  that is large enough to yield a distance but not as massive as 1-Lipschitz functions so as to escape the curse of dimensionality.

**Definition 2.19.** *A metric  $d(\cdot, \cdot)$  between two probability measures is called an integral probability metric (IPM) if it satisfies the properties of a metric and can be written in the form*

$$d(\mu, \nu) = \sup_{f \in \mathcal{F}} \left| \int f \, d\mu - \int f \, d\nu \right|. \quad (2.18)$$

Note that both the 1-Wasserstein distance  $W_1$  and the quantity  $\delta$  above are of the form (2.18) with  $\mathcal{F} = \text{Lip}_1$  and  $\mathcal{F} = \mathcal{F}_{\text{lin}}$  respectively. Indeed, the absolute value in (2.18) is implicit when  $\mathcal{F}$  is symmetric:  $\mathcal{F} = -\mathcal{F}$ . However, while  $W_1$  is an IPM, the quantity  $\delta$  is not because it fails to satisfy the properties of a metric; here: definiteness.

Another example of a choice for  $\mathcal{F}$  is the set of bounded Lipschitz functions which indeed yields an IPM, but the size of this class is the same as  $\text{Lip}_1$  for the matter at hand here. To improve the sample complexity, we need much smoother functions.

### 2.8.2 Maximum mean discrepancy

Reproducing Kernel Hilbert Spaces (RKHS) form a flexible and practical class of functions. To define this class of functions very briefly we introduce some basic definitions and key properties. We refer the reader to [MFSS17] for more details on kernel methods that are particularly relevant to this section.

Consider a reproducing kernel Hilbert space  $\mathcal{H}$  of functions  $\mathbb{R}^d \rightarrow \mathbb{R}$  associated to a bounded positive definite kernel  $k$  on  $\mathbb{R}^d$ . Denote by  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  and  $\|\cdot\|_{\mathcal{H}}$  the inner product and norm on  $\mathcal{H}$  respectively. The reproducing property of the RKHS  $\mathcal{H}$  ensures that for any  $f \in \mathcal{H}$ ,

$$\langle k(x, \cdot), f \rangle_{\mathcal{H}} = f(x).$$

In particular taking  $f = k(y, \cdot)$  yields

$$\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = k(x, y).$$

We are now in a position to define the Maximum Mean Discrepancy.

**Definition 2.20.** *Let  $\mathcal{H}$  be an RKHS. The Maximum Mean Discrepancy (MMD) between two probability measures  $\mu$  and  $\nu$  on  $\mathbb{R}^d$  is defined to be the quantity*

$$\text{MMD}(\mu, \nu) = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \left| \int f \, d\mu - \int f \, d\nu \right|.$$

Without further assumptions on the RKHS, MMD need not define a distance between probability measures. Indeed, observe that the set of linear functions on  $\mathbb{R}^d$  equipped with the Euclidean inner product is in fact an RKHS associated to the linear kernel  $k(x, y) = \langle x, y \rangle$ . Moreover, if  $f(x) = \langle \theta, x \rangle$ , then

$$\|f\|_{\mathcal{H}}^2 = \|\langle \theta, \cdot \rangle\|_{\mathcal{H}}^2 = \|k(\theta, \cdot)\|_{\mathcal{H}}^2 = k(\theta, \theta) = \|\theta\|^2.$$

Hence, the unit ball of  $\mathcal{H}$  is no other than  $\mathcal{F}_{\text{lin}}$  and we have shown that this class is not rich enough to define an IPM.

In fact, we have not addressed whether MMD is finite. To that end, we use the following useful proposition.

**Proposition 2.21.** *Let  $\mathcal{H}$  be an RKHS. The Maximum Mean Discrepancy (MMD) between two probability measures  $\mu$  and  $\nu$  on  $\mathbb{R}^d$  can be equivalently defined as*

$$\text{MMD}(\mu, \nu) = \left\| \int k(x, \cdot) \mu(dx) - \int k(x, \cdot) \nu(dx) \right\|_{\mathcal{H}}.$$

*Proof.* For any  $f \in \mathcal{H}$  it holds

$$\int f(x) (\mu - \nu)(dx) = \int \langle k(x, \cdot), f \rangle_{\mathcal{H}} (\mu - \nu)(dx)$$

$$= \left\langle \int k(x, \cdot) (\mu - \nu)(dx), f \right\rangle_{\mathcal{H}}.$$

Hence, the claim follows from Cauchy–Schwarz.  $\square$

As a corollary of Proposition 2.21, we get that

$$\begin{aligned} \text{MMD}^2(\mu, \nu) &= \iint k(x, y) \mu(dx) \mu(dy) + \iint k(x, y) \nu(dx) \nu(dy) \\ &\quad - 2 \iint k(x, y) \mu(dx) \nu(dy) \end{aligned} \quad (2.19)$$

and

$$\begin{aligned} \text{MMD}(\mu, \nu) &\leq \int \|k(x, \cdot)\|_{\mathcal{H}} \mu(dx) + \int \|k(x, \cdot)\|_{\mathcal{H}} \nu(dx) \\ &\leq 2 \sup_{x \in \mathbb{R}^d} \sqrt{k(x, x)} < \infty \end{aligned}$$

where we used the fact that  $k$  is bounded.

The map  $\mu \mapsto \int k(x, \cdot) \mu(dx)$  which embeds  $\mu$  onto the RKHS  $\mathcal{H}$  is called *kernel mean embedding*. It follows from Proposition 2.21 that MMD is an IPM, meaning that it is indeed a metric, if and only if the kernel mean embedding is *injective*. Kernels that ensure this property are called *characteristic* and one such example is the Gaussian kernel  $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ . To see this, observe that, up to normalizing constants, the kernel mean embedding is a convolution of  $\mu$  with a Gaussian measure: for any  $y \in \mathbb{R}^d$ , it holds

$$\int k(x, y) \mu(dx) = \int e^{-\frac{\|x-y\|^2}{2\sigma^2}} \mu(dx) = (\sigma\sqrt{2\pi})^d (\mu \star \mathcal{N}(0, \sigma^2 I))(y).$$

Injectivity of the convolution with a Gaussian distribution can be seen readily using characteristic functions. Indeed, the characteristic function of  $\mu \star \mathcal{N}(0, \sigma^2 I)$  is given by

$$\phi_{\mu \star \mathcal{N}(0, \sigma^2 I)}(\cdot) = \phi_{\mu}(\cdot) \phi_{\mathcal{N}(0, \sigma^2 I)}(\cdot) = \phi_{\mu}(\cdot) e^{-\frac{\sigma^2 \|\cdot\|^2}{2}}.$$

Hence, since the characteristic function  $e^{-\frac{\sigma^2 \|\cdot\|^2}{2}}$  of the Gaussian is everywhere positive, we get that

$$\mu \star \mathcal{N}(0, \sigma^2 I) = \nu \star \mathcal{N}(0, \sigma^2 I)$$

if and only if  $\mu = \nu$ .

Clearly, the above argument generalizes to translation-invariant kernels that are of the form  $k(x, y) = K(x - y)$  for some bounded positive definite function  $K : \mathbb{R}^d \rightarrow \mathbb{R}$  and whose Fourier transform is everywhere positive<sup>5</sup>. This includes for example the Laplace kernel  $k(x, y) = e^{-\|x-y\|}$  as well as other examples; see [MFSS17, Table 3.1].

The representation of MMD given by (2.19) gives an easy way to estimate MMD from data. For example, assume that  $X_1, \dots, X_m$  are i.i.d. from  $\mu$  and  $Y_1, \dots, Y_n$  are i.i.d. from  $\mu$ . Denote by  $\mu_m$  and  $\nu_n$  the corresponding empirical distributions. Then,

$$\begin{aligned} \text{MMD}^2(\mu_m, \nu_n) &= \frac{1}{m^2} \sum_{i, i'=1}^m k(X_i, X_{i'}) + \frac{1}{n^2} \sum_{j, j'=1}^n k(Y_j, Y_{j'}) \\ &\quad - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(X_i, Y_j). \end{aligned}$$

A natural question is whether this gives a good estimator of  $\text{MMD}^2(\mu, \nu)$ . Using the triangle inequality, it is sufficient to control  $\text{MMD}(\mu_m, \mu)$ . While MMD is an IPM, the closed-form representation of Proposition 2.21 allows us to bypass the use of empirical process theory to control this quantity.

**Theorem 2.22.** *Let  $k$  be a characteristic kernel such that  $k(x, x) \leq 1$  for any  $x \in \mathbb{R}^d$ . Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations from a distribution  $\mu$  on  $\mathbb{R}^d$  and define the empirical measure*

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

*Then*

$$\mathbb{E}_\mu[\text{MMD}(\mu_n, \mu)] \leq \frac{1}{\sqrt{n}}.$$

*Proof.* It follows from Proposition 2.21 that

$$\mathbb{E}[\text{MMD}^2(\mu_n, \mu)] = \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \{k(X_i, \cdot) - \mathbb{E}k(X_i, \cdot)\} \right\|_{\mathcal{H}}^2$$

---

<sup>5</sup> Note that Bochner's theorem implies that positive definite kernels have a *non-negative* Fourier transform, so this is a stronger requirement.

$$\begin{aligned}
&= \frac{1}{n} \mathbb{E} \|k(X_1, \cdot) - \mathbb{E} k(X_1, \cdot)\|_{\mathcal{H}}^2 \\
&= \frac{1}{n} (\mathbb{E} \|k(X_1, \cdot)\|_{\mathcal{H}}^2 - \|\mathbb{E} k(X_1, \cdot)\|_{\mathcal{H}}^2) \\
&\leq \frac{1}{n} \mathbb{E} \|k(X_1, \cdot)\|_{\mathcal{H}}^2.
\end{aligned}$$

Next, observe that

$$\mathbb{E} \|k(X_1, \cdot)\|_{\mathcal{H}}^2 = \mathbb{E}[k(X_1, X_1)] \leq 1.$$

The claim follows from Jensen's inequality.  $\square$

We see that unlike Wasserstein distances, MMD does not suffer from the curse of dimensionality. This is certainly a desirable feature, but it may also be interpreted from a more cautious perspective. Indeed, while MMD does define a metric, it is less sensitive to deviations between probability measures and tends to make them small. This is why  $\mu_n$ , which according to the 1-Wasserstein distance is quite far from  $\mu$ , appears to be quite close to  $\mu$  from the perspective of MMD.

### 2.8.3 Smoothed Wasserstein distances

We see from Proposition 2.21 that when  $k$  is the Gaussian kernel,  $\text{MMD}(\mu, \nu)$  is a Hilbert space norm involving the densities  $\mu \star \mathcal{N}(0, \sigma^2 I_d)$  and  $\nu \star \mathcal{N}(0, \sigma^2 I_d)$ . We could very well measure this distance between probability measures using other distances, in particular, using Wasserstein distances.

**Definition 2.23.** Fix  $p \geq 1$ . The smoothed  $p$ -Wasserstein distance between two probability measures  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  is defined by

$$W_p^{(\sigma)}(\mu, \nu) := W_p(\mu \star \mathcal{N}(0, \sigma^2 I), \nu \star \mathcal{N}(0, \sigma^2 I)).$$

The idea of computing the Wasserstein distance between smoothed versions of the measures was already used as an analytical tool in the proof of Corollary 2.12; here, we consider it as a notion of distance in its own right.

It follows readily from this definition that the smoothed Wasserstein distance is indeed a distance. Compared to MMD, which embeds distributions in a Hilbert space, the geometry induced on distributions by the smoothed Wasserstein distance is much closer to the original

Wasserstein distance. Like MMD, however, smoothed Wasserstein distances enjoy faster statistical rates of convergence. For simplicity, we focus here on the case  $p = 1$ , but parametric rates have been established for  $p = 2$  as well.

**Theorem 2.24.** *Fix  $\sigma > 0$ . Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations from a distribution  $\mu$  on  $[-1, 1]^d$  and define the empirical measure*

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

Then

$$\mathbb{E}_\mu[W_1^{(\sigma)}(\mu_n, \mu)] \lesssim \frac{1}{\sqrt{n}},$$

where the implicit constant depends on both  $\sigma^2$  and  $d$ .

Before turning to the proof, we note that the constant factor in this bound scales exponentially in the dimension. This poor scaling in  $d$  is, in fact, unavoidable and reflects the fundamental statistical difficulty of estimating the Wasserstein distance.

*Proof.* Denote by  $f$  the density of  $\mu \star \mathcal{N}(0, \sigma^2 I)$  and by  $f_n$  the density of  $\mu_n \star \mathcal{N}(0, \sigma^2 I)$ . Write  $\varphi(z) := (2\pi\sigma^2)^{-d/2} \exp(-\frac{1}{2\sigma^2}\|z\|^2)$  for the density of  $\mathcal{N}(0, \sigma^2 I)$ . Theorem 1.6 implies

$$\begin{aligned} \mathbb{E}W_1^{(\sigma)}(\mu_n, \mu) &\leq \mathbb{E} \int \|z\| |f_n(z) - f(z)| dz \\ &= \int \|z\| \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n \varphi(z - X_i) - \mathbb{E}\varphi(z - X_1) \right| dz \\ &\leq \frac{1}{\sqrt{n}} \int \|z\| (\mathbb{E}(\varphi(z - X_1) - \mathbb{E}\varphi(z - X_1))^2)^{1/2} dz \\ &\leq \frac{1}{\sqrt{n}} \int \|z\| (\mathbb{E}\varphi(z - X_1)^2)^{1/2} dz. \end{aligned}$$

It suffices to show that the integral is bounded. If  $\|z\| \leq 2\sqrt{d}$ , then we can use the crude bound  $(\mathbb{E}\varphi(z - X_1)^2)^{1/2} \leq (2\pi\sigma^2)^{-d/2}$ . If  $\|z\| > 2\sqrt{d}$ , then  $\|z - X_1\| \geq \|z\| - \|X_1\| \geq \|z/2\|$  almost surely, which yields  $(\mathbb{E}\varphi(z - X_1)^2)^{1/2} \leq \varphi(z/2)$ . We obtain

$$\begin{aligned} \mathbb{E}W_1^{(\sigma)}(\mu_n, \mu) &\leq \frac{(2\pi\sigma^2)^{-d/2}}{\sqrt{n}} \int_{\|z\| \leq 2\sqrt{d}} \|z\| dz + \frac{1}{\sqrt{n}} \int \|z\| \varphi(z/2) dz \\ &\lesssim n^{-1/2}, \end{aligned}$$

as claimed.  $\square$

### 2.8.4 Sliced Wasserstein distances

Finally, we close this chapter with yet another method to avoid the curse of dimensionality, this time based on considering the Wasserstein distance between one-dimensional projections.

Formally, let  $\mathbb{S}^{d-1}$  denote the unit sphere in  $\mathbb{R}^d$  and for  $\theta \in \mathbb{S}^{d-1}$  let  $\Pi^\theta : \mathbb{R}^d \rightarrow \mathbb{R}$  be the projection  $\Pi^\theta(x) := \langle \theta, x \rangle$ . Define the *sliced Wasserstein distance* between  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  to be the quantity

$$\text{SW}_p(\mu, \nu) := \left( \int W_p^p(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu) \sigma(d\theta) \right)^{1/p}, \quad (2.20)$$

where  $\sigma$  is the uniform measure on  $\mathbb{S}^{d-1}$ .

The idea of considering one-dimensional projections is rooted in applications to imaging and tomography, for which various integral transforms have been introduced. In particular, the *Radon transform* of a measure  $\mu$  is defined to be the collection of one-dimensional projections  $(\Pi_\#^\theta \mu)_{\theta \in \mathbb{S}^{d-1}}$ . It is a classical fact, known as the *Cramér–Wold theorem*, that the Radon transform of  $\mu$  completely characterizes  $\mu$ , justifying its use in defining a metric over probability measures. Let us start by checking that the axioms of a metric space are indeed satisfied.

**Theorem 2.25.** *For every  $p \geq 1$ ,  $\text{SW}_p$  defines a metric over  $\mathcal{P}_p(\mathbb{R}^d)$ .*

*Proof.* Symmetry and non-negativity follow from the corresponding facts about the Wasserstein distance. For  $\Theta \sim \sigma$ , the triangle inequality is verified via

$$\begin{aligned} \text{SW}_p(\mu, \nu) &= (\mathbb{E} W_p^p(\Pi_\#^\Theta \mu, \Pi_\#^\Theta \nu))^{1/p} \\ &\leq (\mathbb{E} W_p^p(\Pi_\#^\Theta \mu, \Pi_\#^\Theta \rho))^{1/p} + (\mathbb{E} W_p^p(\Pi_\#^\Theta \rho, \Pi_\#^\Theta \nu))^{1/p} \\ &= \text{SW}_p(\mu, \rho) + \text{SW}_p(\rho, \nu). \end{aligned}$$

Finally, we must check that  $\text{SW}_p(\mu, \nu) = 0$  implies  $\mu = \nu$ . Certainly,  $\text{SW}_p(\mu, \nu) = 0$  implies that  $W_p(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu) = 0$  for *almost every*  $\theta \in \mathbb{S}^{d-1}$ , which implies  $\Pi_\#^\theta \mu = \Pi_\#^\theta \nu$ . To finish, we would like to upgrade “almost every” to “every” to apply the Cramér–Wold device.

To do so, we prove a Lipschitz continuity property of the mapping  $\theta \mapsto \Pi_\#^\theta \mu$ . For  $\theta' \in \mathbb{S}^{d-1}$  and  $X \sim \mu$ ,

$$W_p(\Pi_\#^\theta \mu, \Pi_\#^{\theta'} \mu) \leq (\mathbb{E} [|\langle \theta - \theta', X \rangle|^p])^{1/p} \leq (\mathbb{E} \|X\|^p)^{1/p} \|\theta - \theta'\|.$$

Together with the  $W_p$  triangle inequality, it shows that

$$\begin{aligned} |W_p(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu) - W_p(\Pi_\#^{\theta'} \mu, \Pi_\#^{\theta'} \nu)| \\ \leq W_p(\Pi_\#^\theta \mu, \Pi_\#^{\theta'} \mu) + W_p(\Pi_\#^\theta \nu, \Pi_\#^{\theta'} \nu) \lesssim \|\theta - \theta'\|. \end{aligned}$$

Therefore,  $\theta \mapsto W_p(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu)$  is continuous, and  $\text{SW}_p(\mu, \nu) = 0$  implies that this quantity vanishes for every  $\theta \in \mathbb{S}^{d-1}$ .  $\square$

We can now prove that the sliced Wasserstein distance can be estimated at a parametric rate.

**Proposition 2.26.** *Suppose that  $\mu, \nu \in \mathcal{P}(B_1)$ , where  $B_1$  is the unit ball in  $\mathbb{R}^d$ , and let  $\mu_n, \nu_n$  denote the corresponding empirical measures formed from i.i.d. samples  $X_1, \dots, X_n \sim \mu$  and  $Y_1, \dots, Y_n \sim \nu$ . Then,*

$$\mathbb{E}\text{SW}_1(\mu_n, \mu) \lesssim n^{-1/2}.$$

Also,

$$\mathbb{E}|\text{SW}_1(\mu_n, \nu_n) - \text{SW}_1(\mu, \nu)| \lesssim n^{-1/2}.$$

*Proof.* The second inequality follows from the first by the triangle inequality. To establish the first, we can note that  $\langle \theta, X_1 \rangle, \dots, \langle \theta, X_n \rangle$  is an i.i.d. sample from  $\Pi_\#^\theta \mu$ , and  $\Pi_\#^\theta \mu_n$  is the corresponding empirical measure. Hence, from the one-dimensional rate in Proposition 2.5,  $\mathbb{E}W_1(\Pi_\#^\theta \mu_n, \Pi_\#^\theta \mu) \lesssim n^{-1/2}$ . Then, average over  $\theta$ .  $\square$

Although we have motivated the sliced Wasserstein distance for its statistical benefits, fortuitously it also comes with substantial computational ones. Indeed, computation of  $\text{SW}_p$  boils down to one-dimensional optimal transport, which for discrete measures can be solved via sorting; see Exercise 5.

## 2.9 Discussion

**§2.1.** The Wasserstein law of large numbers is discussed in more detail in [Dud02, Chapter 11]. The slow rate of convergence is a manifestation of the fact that  $W_1$  convergence automatically implies convergence of

---

<sup>6</sup> An alternative argument proceeds as follows:  $\Pi_\#^\theta \mu = \Pi_\#^\theta \nu$  for almost every  $\theta$  implies that the characteristic functions of  $\mu, \nu$  are equal almost everywhere. But characteristic functions are uniformly continuous.

all Lipschitz text functions (and, for  $p > 1$ , convergence of all higher moments, as Proposition 1.5 shows); it is therefore not surprising that a large number of samples is needed to obtain such strong control. The implications of the Wasserstein law of large numbers for optimal quantization and  $k$ -means clustering were first developed in a statistical context in [Pol82].

**§2.2.** The usefulness of the dyadic partitioning argument for controlling the Wasserstein distances was first highlighted by [BLG14]; see also [DSS13, FG15] extensions to the unbounded setting. A further discussion of the history of this approach appears in [NWB19], from which this version of the argument was taken. This argument can easily be extended to show that the rate of convergence depends on the intrinsic dimension of the measure  $\mu$  rather than the ambient dimension.

The dyadic partitioning argument also applies to the  $p > 1$  case, and shows that

$$\mathbb{E}W_p(\mu_n, \mu) \lesssim_p \sqrt{d} \cdot \begin{cases} n^{-1/2p}, & \text{if } d < 2p, \\ (\log n)^{1/p}/n^{1/2p}, & \text{if } d = 2p, \\ n^{-1/d}, & \text{if } d > 2p. \end{cases} \quad (2.21)$$

This rate is essentially sharp, apart from the logarithmic factor in the  $d = 2p$  case. On the other hand, the dual bounds we present in this chapter do *not* easily extend to  $p > 1$  since the dual formulation of  $W_p(\mu_n, \mu)$  does not give rise to an empirical process when  $p > 1$ .

The triangle inequality implies that rates of convergence of  $W_p(\mu_n, \nu)$  to  $W_p(\mu, \nu)$  can be derived from the corresponding rates of convergence of  $W_p(\mu_n, \mu)$ ; however, these rates can fail to be sharp. To give one example, [CRL<sup>+</sup>20] showed that

$$\mathbb{E}|W_2^2(\mu, \nu_n) - W_2^2(\mu, \nu)| \lesssim \begin{cases} n^{-1/2}, & \text{if } d < 4, \\ (\log n)/n^{1/2}, & \text{if } d = 4, \\ n^{-2/d}, & \text{if } d > 4. \end{cases} \quad (2.22)$$

Note that when  $\mu \neq \nu$ , this bound is stronger than what could be deduced from (2.21). A similar phenomenon exists for other  $W_p$  distances [MNW24].

More strikingly, the rate at which  $W_p(\mu_n, \nu)$  converges to  $W_p(\mu, \nu)$  can be shown to depend on the *smaller* of the intrinsic dimensions of  $\mu$  and  $\nu$ ; see [HSM24]. In particular, if  $\nu$  is supported on a finite number of points (sometimes known as the *semi-discrete* optimal transport

problem), then  $W_p(\mu_n, \nu)$  converges to  $W_p(\mu, \nu)$  at a rate that does not suffer from the curse of dimensionality. This fact cannot be deduced from bounds on  $W_p(\mu_n, \mu)$  alone.

**§2.3.** Chaining is an idea that goes back implicitly to Kolmogorov. In the form of Proposition 2.6, it is known as Dudley’s entropy integral [Dud67]. For some of the many references on chaining and its applications, see [Dud99, vH14, Ver18, Wai19, Tal21, vdVW23].

**§2.4.** As mentioned, the result of this section is due to [AKT84] and the argument here is taken from [BL21]. The idea of using Fourier transforms to bound matching costs is originally due to [CS91], and was developed extensively by Talagrand [Tal21]. More broadly, there is a large literature on so-called matching problems, e.g., [Led17, LZ21, Tal21], and recently techniques from partial differential equations have been used to derive very sophisticated results when  $d = 2$ , see, e.g., [AST19].

**§2.5.** Applications of Wasserstein distances to testing can be found in [dBCAMRR99, HMS21, GDGSCN23, NWKB23]. Goodness-of-fit and two-sample testing problems with Wasserstein distances were studied in [BNNR11]; for each problem, they showed that the optimal separation is of order  $n^{-c/d}$  for some constant  $c > 1$ .

It was shown in [NWR22] that, up to logarithmic factors, a separation of  $n^{-1/d}$  is necessary for a “robust” version of goodness-of-fit problem described above, with hypotheses

$$H^0 : W_1(\mu, \mu^0) < \varepsilon \quad \text{vs.} \quad H_1 : W_1(\mu, \mu^0) > 2\varepsilon.$$

This lower bound can be used to obtain nearly sharp minimax lower bounds for estimating the Wasserstein distance.

**§2.6.** The lower bound in Theorem 2.14 is due to [Dud69]. The minimax lower bound in Theorem 2.15 was first proved by [SP18]. For expositions of minimax lower bound techniques, see [Tsy09, RH17, Wai19].

**§2.7.** Minimax estimation of smooth densities in the Wasserstein distance was studied in [SUL<sup>+</sup>18, Lia21] for  $W_1$ , and in [NWB22] for  $W_p$ ,  $p > 1$ . The case of  $p > 1$  evinces different behavior from the  $p = 1$  case: [NWB22] showed that the rate in Theorem 2.18 is achievable for  $p > 1$  only under the additional assumption that the density of  $\mu$  is bounded below; without this assumption, rates of estimation are strictly worse. The results for the  $p > 1$  case are confined to densities lying in Besov classes; extending the arguments of [NWB22] to other classes of densities is an open question. A version of this problem on manifolds has been studied in [Div22].

Non-parametric density estimation is itself a classical topic in statistics, albeit usually studied in other distance metrics [Tsy09].

**§2.8.** It is worth mentioning that IPMs have received significant attention in the context of Generative Adversarial Networks (GANs), and in particular, Wasserstein GANs [ACB17] where  $\mathcal{F}$  is chosen to be a family of deep neural networks. For statistical analyses of IPMs and GANs, see, e.g. [USP19, Lia21].

Maximum mean discrepancy was first developed in [BGR<sup>+</sup>06], and is now the subject of a large literature, see, [MFSS17]. The rate given in Theorem 2.22 is folklore. A notable special case of MMD is the class of energy distances, for which we recommend [Ger24, Subsection 1.2.4] for an introduction and references.

The favorable statistical properties of the smoothed Wasserstein distances were first recorded in [GGNWP20]. The simple proof of Theorem 2.24 is taken from [NW18].

Sliced Wasserstein distances were introduced in [RPDB12]. They arose as a device to understand the “iterative distribution transfer” algorithm [PKD07]; see the PhD thesis of Bonnotte [Bon13] for history. For further discussion, consult [San15, Section 5.5.4], and see [NDC<sup>+</sup>20] for generalizations with similar metric and statistical properties. Extensions of Proposition 2.26 appear in [MBW22]. The sharp condition for the fast rate of estimation of sliced Wasserstein distances to hold was obtained in [BL19]. Other results in this vein, including distributional limits, can be found in [MBW22, NWR22, OI22, XNW22, XH22, PS23, GKRS24].

## 2.10 Exercises

1. This exercise shows that the  $\sqrt{d}$  factor in Proposition 2.1 cannot be improved.
  - a) Show that there exists a positive universal constant  $c$  such that for any  $n \geq 1$ , the Lebesgue measure of a ball in  $\mathbb{R}^d$  with radius  $c\sqrt{d}n^{-1/d}$  is at most  $(2n)^{-1}$ . (Hint: recall that the unit ball in  $\mathbb{R}^d$  has Lebesgue measure  $\pi^{d/2}/\Gamma(\frac{d}{2} + 1)$ .)
  - b) Let  $\mu$  be the uniform measure on  $[0, 1]^d$ , and let  $\tilde{\mu}_n$  be any measure supported on  $n$  points  $x_1, \dots, x_n$ . If we denote by  $B(x_i, \epsilon)$  a ball of radius  $\epsilon$  around  $x_i$ , show that  $\mu(\bigcup_{i=1}^n B(x_i, \epsilon)) \leq \frac{1}{2}$  if  $\epsilon = c\sqrt{d}n^{-1/d}$ , where  $c$  is the constant from part (a).
  - c) Conclude that if  $\gamma \in \Gamma(\mu, \tilde{\mu}_n)$ , then  $\int \|x - y\| d\gamma(x, y) \geq \frac{1}{2} \cdot c\sqrt{d}n^{-1/d}$ .

2. Adapt the proof of Proposition 2.5 to establish (2.21).
3. Recall the bounded differences inequality (e.g., [BLM13, Theorem 6.2]). Use it to prove a concentration inequality for  $W_1(\mu_n, \mu)$  around its expectation, where  $X_1, \dots, X_n$  are i.i.d. from a distribution  $\mu$  supported on a ball of radius  $R$  and  $\mu_n$  is the empirical measure  $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ .
4. Prove that  $\text{SW}_p \leq W_p$  for any  $p \geq 1$ . Is this inequality tight? Similarly, show that  $W_1^{(\sigma)} \leq W_1$ . Is this inequality tight?
5. We consider the computational aspects of the sliced Wasserstein distance, defined in Subsection 2.8.4.
  - a) Show that if  $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$  ( $p \geq 1$ ), and  $\mu$  and  $\nu$  are each uniformly distributed on  $n$  points, then  $W_p(\mu, \nu)$  can be computed in  $O(n \log n)$  time (where we treat arithmetic and comparison operations as constant time). Assume that the measures  $\mu, \nu$  are given as (unordered) lists of points.
  - b) Now suppose that  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  are each uniformly distributed on  $n$  points, presented as lists of points in  $\mathbb{R}^d$ . Argue that we can compute  $W_p(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu)$  in  $O(dn + n \log n)$  time.
  - c) This is still insufficient for algorithmic purposes, since computing  $\text{SW}_1(\mu, \nu)$  exactly requires computing  $W_1(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu)$  for uncountably many values of  $\theta$  and integrating. Argue instead that if  $\mu, \nu \in \mathcal{P}(B_1)$  and we draw  $m$  i.i.d. points  $\theta_1, \dots, \theta_m$  from the uniform measure  $\sigma$  on  $\mathbb{S}^{d-1}$ , then the Monte Carlo average

$$\widehat{\text{SW}}_1(\mu, \nu) := \frac{1}{m} \sum_{i=1}^m W_1(\Pi_\#^{\theta_i} \mu, \Pi_\#^{\theta_i} \nu)$$

approximates  $\text{SW}_1(\mu, \nu)$  to an additive error of size  $O(m^{-1/2})$ .

6. Theorem 1.7 implies that if  $\mu \in \mathcal{P}_1(\mathbb{R})$ , then

$$W_1(\mu_n, \mu) = \int_{-\infty}^{\infty} |F_{\mu_n}(t) - F_\mu(t)| dt,$$

where  $F_{\mu_n}$  and  $F_\mu$  are the cumulative distribution functions of  $\mu_n$  and  $\mu$ , respectively. Use this fact to show Proposition 2.1 for  $d = 1$  directly. (Hint:  $\mathbb{E}|F_{\mu_n}(t) - F_\mu(t)|^2 = F_\mu(t)(1 - F_\mu(t))$ .)

7. The minimax lower bound proved in Theorem 2.15 is suboptimal when  $d = 1$ . This exercise proves an optimal bound based on testing between two hypotheses (Theorem 2.2 in [Tsy09]). To use this approach, it suffices to construct two measures  $\mu^{(0)}$  and  $\mu^{(1)}$  with support in  $[0, 1]$  such that

$$W_1(\mu^{(0)}, \mu^{(1)}) \geq 2r_n, \\ \mathsf{KL}(\mu^{(1)} \parallel \mu^{(0)}) \leq \frac{1}{2n}.$$

The existence of such measures implies that for any estimator  $\tilde{\mu}_n$  based on  $n$  i.i.d. observations, there exists  $j \in \{0, 1\}$  such that  $\mathbb{E}_{\mu^{(j)}}[W_1(\tilde{\mu}_n, \mu^{(j)})] \geq \frac{r_n}{4}$ .

- a) Fix  $\varepsilon \in (0, 1/2)$ , and consider  $\mu^{(0)} = (\frac{1}{2} + \varepsilon)\delta_0 + (\frac{1}{2} - \varepsilon)\delta_1$  and  $\mu^{(1)} = (\frac{1}{2} - \varepsilon)\delta_0 + (\frac{1}{2} + \varepsilon)\delta_1$ . Show that  $W_1(\mu^{(0)}, \mu^{(1)}) = 2\varepsilon$ .
- b) Show that this pair of measures satisfies  $\mathsf{KL}(\mu^{(1)}, \mu^{(0)}) \leq \frac{16\varepsilon^2}{1-4\varepsilon^2}$ .
- c) Conclude that there exists a positive universal constant  $c$  such that for any estimator  $\tilde{\mu}_n$  there exists  $j \in \{0, 1\}$  such that  $\mathbb{E}_{\mu^{(j)}}[W_1(\tilde{\mu}_n, \mu^{(j)})] \geq cn^{-1/2}$ .

8. The regularization strategies discussed in Section 2.8 can be applied more broadly. For example, given probability measures  $\mu, \nu$ , define the following quantity and call it “smoothed  $L_2$ ”:

$$d_\sigma^2(\mu, \nu) := \int (\mu \star \mathcal{N}(0, \sigma^2 I) - \nu \star \mathcal{N}(0, \sigma^2 I))^2.$$

Show that it can be estimated at a parametric rate: if  $\mu_n$  denotes the empirical measure formed from  $n$  i.i.d. samples from  $\mu$ , then

$$\mathbb{E}d_\sigma^2(\mu_n, \mu) \leq \frac{1}{(2\pi\sigma^2)^{d/2} n}.$$

9. For  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ , the “max-sliced Wasserstein distance” [DHS<sup>+</sup>19, KNS<sup>+</sup>19]<sup>7</sup> between them is

$$\mathsf{MSW}_p(\mu, \nu) := \max_{\theta \in \mathbb{S}} W_p(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu).$$

The goal of this exercise is to show that  $\mathsf{MSW}_1$  can be estimated at the parametric rate.

- a) Let  $\mu \in \mathcal{P}(B_1)$ , and let  $\mu_n$  be the empirical measure corresponding to  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mu$ . Show that

$$\mathsf{MSW}_1(\mu_n, \mu) = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\},$$

where  $\mathcal{F}$  is the class of functions of the form  $x \mapsto h(\theta^\top x)$  where  $\theta \in \mathbb{S}^{d-1}$  and  $h$  is a 1-Lipschitz function on  $[-1, 1]$  satisfying  $h(0) = 0$ .

---

<sup>7</sup> Also known as the “low-dimensional Wasserstein distance” [NWR22] or the “subspace robust Wasserstein distance” [PC19a].

b) Prove that

$$\log N(\varepsilon, \mathcal{F}) \lesssim 1/\varepsilon + d \log(1 + 1/\varepsilon).$$

*Hint:* Consult Exercise 2 in Chapter 3.

c) Using Proposition 2.6, conclude that

$$\mathbb{E} \text{MSW}_1(\mu_n, \mu) \lesssim \sqrt{d/n}.$$

In fact, a more sophisticated argument shows that the correct rate is dimension free [Boe24].

# 3

---

## Estimation of transport maps

Thus far, the statistical questions we have investigated center around the estimation of optimal transport distances (and their variants), but the gamut of diverse applications of optimal transport (to name but a few: data fusion [CFTC16] adaptation/transfer learning [CFTR17], and computational biology [SST<sup>+</sup>19, BSG<sup>+</sup>23]), it is the optimal transport *map* which is the object of primary interest. In this chapter, we address the question of estimating this map on the basis of finitely many samples.

### 3.1 Problem formulation

Recall from Brenier's theorem that if  $\mu$  has a density,

$$\begin{aligned} W_2^2(\mu, \nu) &= \min_{\gamma \in \Gamma_{\mu, \nu}} \int \|x - y\|^2 \gamma(dx, dy) \\ &= \min_{T: T_{\#}\mu = \nu} \int \|x - T(x)\|^2 \mu(dx). \end{aligned}$$

Moreover, the optimal transport map takes the form  $T = \nabla\varphi$ , where  $\varphi$  is convex. We can also write this as  $(X, \nabla\varphi(X)) \sim \gamma$ , or  $\gamma(dx, dy) = \mu(dx) \delta_{T(x)}(dy)$ .

The statistical question under investigation is formulated as follows. Given samples  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mu$  and  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \nu$ , how can we estimate the optimal transport map  $T$  from  $\mu$  to  $\nu$  via an estimator  $\hat{T}$  constructed on the basis of the samples?

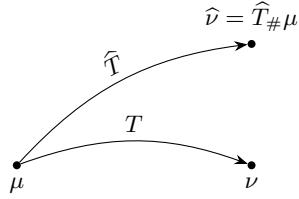
We take as our measure of performance the integrated error

$$\int \|\hat{T}(x) - T(x)\|^2 \mu(dx).$$

The  $L^2$  integrated error is a natural measure of distance that is commonly employed in non-parametric statistics. In this context, however, it takes on an additional interpretation of controlling the Wasserstein distance between the pushforwards of  $\mu$  under the two maps. More precisely, by definition we have  $\nu = T_\# \mu$ . If we define the measure  $\widehat{\nu}$  via  $\widehat{\nu} = \widehat{T}_\# \mu$ , then  $(\widehat{T}(X), T(X))$  for  $X \sim \mu$  is a (suboptimal) coupling of  $\widehat{\nu}$  and  $\nu$ , and hence

$$W_2^2(\widehat{\nu}, \nu) \leq \mathbb{E} \|\widehat{T}(X) - T(X)\|^2 = \int \|\widehat{T}(x) - T(x)\|^2 \mu(dx).$$

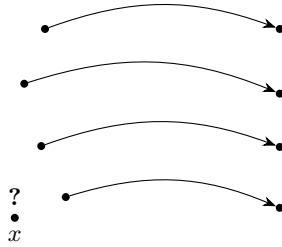
See Figure 3.1 for an illustration.



**Fig. 3.1.** The  $L^2$  error between the transport maps controls the  $W_2^2$  distance between  $\widehat{\nu}$  and  $\nu$ .

A first approach to estimation might be to compute the optimal coupling between the empirical measures  $\mu_n$  and  $\nu_n$ , i.e., solve  $\min_{\gamma \in \Gamma_{\mu_n, \nu_n}} \int \|x - y\|^2 \gamma(dx, dy)$ , but we rapidly recognize an untenable hole in this naïve plan. Namely, even if the optimal transport plan  $\gamma_n$  is induced by a transport map  $T_n$ , so that  $\gamma_n(dx, dy) = \mu_n(dx) \delta_{T_n(x)}(dy)$ , the mapping  $T_n$  is only well-defined on the sample  $\{X_1, \dots, X_n\}$  and it is not clear how to extend it in a principled manner to a mapping over all of  $\mathbb{R}^d$  (Figure 3.2). To remedy this, several approaches have been proposed in the literature aimed at building an interpolation  $\widehat{T}_n$  of  $T_n$  to out-of-sample points. For example, we can take  $\widehat{T}_n(x)$  to equal  $T_n(X_i)$ , where  $X_i$  is the closest sample point to  $x$ . This is a 1-nearest neighbor estimator and it can be shown to be minimax optimal without further smoothness assumptions [MBNW21]; see Exercise 1 for the one-dimensional case. Such an approach, however, cannot take advantage of additional regularity of  $\mu$  and  $\nu$  and we do not pursue it any further here.

Instead, in the next section, we devise an estimator based on the semidual formulation of optimal transport. A benefit of this estimation strategy is that it can be used to flexibly incorporate additional



**Fig. 3.2.** How do we interpolate the empirical optimal transport map at the out-of-sample point  $x$ ?

assumptions—e.g., smoothness—on the population-level transport map  $T$ . Adopting sufficiently strong assumptions gives rise to map estimators that avoid the curse of dimensionality.

### 3.2 The semidual problem and its stability

We recall the semidual problem: if  $\nabla\varphi$  is the optimal transport map, then  $\varphi$  minimizes

$$\mathcal{S}(\phi) := \int \phi \, d\mu + \int \phi^* \, d\nu$$

where  $\phi^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - \phi(x)\}$  is the convex conjugate of  $\phi$ . Crucially, the semidual problem readily lends itself to replacing the population measures  $\mu, \nu$  with their empirical counterparts  $\mu_n, \nu_n$ , leading to a natural estimator for  $\varphi$ : namely, we set

$$\widehat{\varphi} = \arg \min_{\phi \in \mathcal{F}} \mathcal{S}_n(\phi) := \arg \min_{\phi \in \mathcal{F}} \left\{ \int \phi \, d\mu_n + \int \phi^* \, d\nu_n \right\} \quad (3.1)$$

where  $\mathcal{F}$  is a suitable class of functions to be chosen later. We then obtain an estimator for the optimal transport map by setting  $\widehat{T} = \nabla \widehat{\varphi}$ . Through (3.1), we have placed the problem of transport map estimation within the well-studied framework of empirical risk minimization (ERM) which is a cornerstone of modern statistical theory—see, e.g., [Wai19] for a modern overview of these techniques. Akin to many other estimators defined via ERM, it is unclear whether the estimator  $\widehat{T}$  can be computed efficiently; however, our focus here is on the statistical, rather than computational, aspects of transport map estimation.

Through the statistician's lens, the uniqueness assertion in Brenier's theorem ensures that the optimal transport map  $T$  is identifiable, and

hence our statistical question is well-posed. In other words, if  $\mathcal{S}(\phi) = \mathcal{S}(\varphi)$  then  $\nabla\phi = \nabla\varphi$ ,  $\mu$ -a.s. However, in order to obtain rates of estimation, this qualitative assertion needs to be upgraded into a stability statement, which is given as the following theorem.

**Theorem 3.1.** *Assume that  $\phi$  is strongly convex and smooth,*

$$\frac{1}{2} I \preceq \nabla^2\phi \preceq 2I.$$

*Then,*

$$\frac{1}{4} \|\nabla\phi - \nabla\varphi\|_{L^2(\mu)}^2 \leq \mathcal{S}(\phi) - \mathcal{S}(\varphi) \leq \|\nabla\phi - \nabla\varphi\|_{L^2(\mu)}^2. \quad (3.2)$$

Before proving Theorem 3.1, we first describe how the stability result feeds into the overall statistical analysis. The proof is prototypical of analysis of ERM estimators. By definition,  $\widehat{\varphi}$  minimizes  $\mathcal{S}_n$ . Applying the machinery of empirical process theory, we control the fluctuations of the random functional  $\mathcal{S}_n$  from its mean  $\mathcal{S}$ , thereby concluding that  $\mathcal{S}(\widehat{\varphi}) - \mathcal{S}(\varphi)$  is small. The first inequality in (3.2) then implies that the estimation error  $\|\widehat{T} - T\|_{L^2(\mu)}^2$  is small.

Actually, to obtain faster rates of estimation, we improve upon this argument by incorporating another ingredient: the *fixed-point* or *localization* technique. Briefly, the estimation rates depend on a uniform bound on the deviations of  $\mathcal{S}_n$  from  $\mathcal{S}$  over a set of functions that contains the estimator  $\widehat{\varphi}$ . Once we know through the stability inequality (3.2) that  $\widehat{\varphi}$  lies close to  $\varphi$ , we can repeat the argument but restricting to a smaller class of functions, thereby improving our estimation rates further. Seeking the fixed point of this iterative process in which we refine our bounds by localizing the estimator  $\widehat{\varphi}$ , we arrive at our final rates of estimation.

We now turn towards the proof of Theorem 3.1. We repeatedly use the Fenchel–Young inequality (Theorem A.6), as well as the fact that  $\alpha$ -convexity of  $f$  is equivalent to  $\alpha^{-1}$ -smoothness of  $f^*$  (Lemma A.9).

*Proof of Theorem 3.1.* Since  $(\nabla\varphi)_\# \mu = \nu$ ,

$$\mathcal{S}(\phi) = \int \phi(x) \mu(dx) + \int \phi^*(y) \nu(dy) = \int (\phi(x) + \phi^*(\nabla\varphi(x))) \mu(dx).$$

By strong convexity of  $\phi^*$ ,

$$\begin{aligned}\phi^*(\nabla\varphi(x)) &\geq \phi^*(\nabla\phi(x)) + \underbrace{\langle \nabla\phi^*(\nabla\phi(x)), \nabla\varphi(x) - \nabla\phi(x) \rangle}_{=x} \\ &\quad + \frac{1}{4} \|\nabla\varphi(x) - \nabla\phi(x)\|^2\end{aligned}$$

hence

$$\begin{aligned}\phi(x) + \phi^*(\nabla\varphi(x)) &\geq \underbrace{\phi(x) + \phi^*(\nabla\phi(x))}_{=\langle x, \nabla\phi(x) \rangle} + \langle x, \nabla\varphi(x) - \nabla\phi(x) \rangle \\ &\quad + \frac{1}{4} \|\nabla\varphi(x) - \nabla\phi(x)\|^2 \\ &= \langle x, \nabla\varphi(x) \rangle + \frac{1}{4} \|\nabla\varphi(x) - \nabla\phi(x)\|^2.\end{aligned}$$

However,

$$\mathcal{S}(\varphi) = \int (\varphi(x) + \varphi^*(\nabla\varphi(x))) \mu(dx) = \int \langle x, \nabla\varphi(x) \rangle \mu(dx).$$

Therefore, we obtain

$$\mathcal{S}(\phi) \geq \mathcal{S}(\varphi) + \frac{1}{4} \|\nabla\varphi - \nabla\phi\|_{L^2(\mu)}^2.$$

Similarly, by smoothness,

$$\begin{aligned}\phi^*(\nabla\varphi(x)) &\leq \phi^*(\nabla\phi(x)) + \underbrace{\langle \nabla\phi^*(\nabla\phi(x)), \nabla\varphi(x) - \nabla\phi(x) \rangle}_{=x} \\ &\quad + \|\nabla\varphi(x) - \nabla\phi(x)\|^2\end{aligned}$$

hence

$$\begin{aligned}\phi(x) + \phi^*(\nabla\varphi(x)) &\leq \phi(x) + \phi^*(\nabla\phi(x)) + \langle x, \nabla\varphi(x) \rangle - \langle x, \nabla\phi(x) \rangle \\ &\quad + \|\nabla\varphi(x) - \nabla\phi(x)\|^2 \\ &= \langle x, \nabla\phi(x) \rangle + \varphi(x) + \varphi^*(\nabla\varphi(x)) - \langle x, \nabla\phi(x) \rangle \\ &\quad + \|\nabla\varphi(x) - \nabla\phi(x)\|^2\end{aligned}$$

and the result follows from integration.  $\square$

Note that we have proved something even stronger: for

$$\mathcal{S}(\phi) = \int \underbrace{(\phi(x) + \phi^*(\nabla\varphi(x)))}_{s_\phi(x)} \mu(dx)$$

we have the pointwise bounds

$$\frac{1}{4} \|\nabla\varphi(x) - \nabla\phi(x)\|^2 \leq s_\phi(x) - s_\varphi(x) \leq \|\nabla\varphi(x) - \nabla\phi(x)\|^2.$$

### 3.3 A special case: affine transport maps

As a sanity check, we first show that the semidual estimation technique is reasonable for a very simple problem. Consider the one-sample setting, where  $\mu = \mathcal{N}(0, I)$  is known, and we obtain samples from  $\nu = (\nabla\varphi)_{\#}\mu$  for some  $\varphi \in \mathcal{F}$ , where  $\mathcal{F}$  consists of all convex quadratic functions  $x \mapsto \frac{1}{2}x^T Ax + b^T x$  with  $A \succeq 0$ . If  $\varphi$  is of this form, then the transport map  $\nabla\varphi$  is the affine map  $Ax + b$ , and  $\nu = \mathcal{N}(b, A^2)$ .

In this setting, it is natural to estimate the transport map by first computing the empirical mean  $\hat{m}$  and covariance  $\hat{\Sigma}$  of  $\nu$ , and setting

$$\hat{T}(x) = \hat{\Sigma}^{1/2}x + \hat{m}. \quad (3.3)$$

This estimator is studied in more generality in [FLF20] by leveraging techniques to derive rates of estimation for covariance matrices.

The next result shows that  $\hat{T}$  defined in (3.3) is precisely the estimator computed by minimizing the empirical semidual functional.

**Proposition 3.2.** *Let  $\mathcal{F}$  be the set of all convex quadratic functions on  $\mathbb{R}^d$ . Let  $\mu = \mathcal{N}(0, I)$ , and write  $\nu_n$  for an empirical measure consisting of i.i.d. samples from a probability measure  $\nu$ . If*

$$\hat{\varphi} = \arg \min_{\phi \in \mathcal{F}} \left\{ \int \phi \, d\mu + \int \phi^* \, d\nu_n \right\},$$

then

$$\nabla\hat{\varphi}(x) = \hat{\Sigma}^{1/2}x + \hat{m},$$

where  $\hat{m}$  and  $\hat{\Sigma}$  are the mean and covariance of  $\nu_n$ , respectively.

*Proof.* By definition,  $\hat{\varphi} = \frac{1}{2}x^T \hat{A}x + \hat{b}^T x$ , where  $(\hat{A}, \hat{b})$  solve

$$\min_{A \succeq 0, b \in \mathbb{R}^d} \left[ \int \left( \frac{1}{2}x^T Ax + b^T x \right) \mu(dx) + \int \left( \frac{1}{2}y^T Ay + b^T y \right)^* \nu_n(dy) \right].$$

By Lemma A.13, the convex conjugate of a quadratic is also a quadratic, so the second integral only depends on moments of  $\nu_n$  of order at most 2. We can therefore replace the integration over  $\nu_n$  with any other measure that matches the first two moments, in particular  $\mathcal{N}(\hat{m}, \hat{\Sigma})$ . Then, since the function class contains all Kantorovich potentials between Gaussians (see Example 1.19), it follows that the minimizer is the one which corresponds to the optimal transport from  $\mathcal{N}(0, I)$  to  $\mathcal{N}(\hat{m}, \hat{\Sigma})$ .  $\square$

In the setting of Proposition 3.2, it is easy to analyze the performance of  $\widehat{T}$  directly, since it is defined explicitly in terms of the sample mean and covariance. However, for more general families  $\mathcal{F}$ , we typically cannot solve the semidual problem explicitly, and we need to use more abstract arguments to analyze  $\widehat{\varphi}$ .

### 3.4 Obtaining the slow rate

In this and the following section, we focus on estimating maps arising from potentials that lie in a suitable class  $\Phi$  whose covering numbers—in the sense of Section 2.3—are suitably bounded. The size of these covering numbers directly affects the quality of the estimator obtained by minimizing the empirical semidual, as in (3.1).

To begin our analysis, we make several technical assumptions on the measures  $\mu$  and  $\nu$  and the family  $\Phi$ .

**Assumption 3.3.** *There exists  $\varphi \in \Phi$  such that  $\nu = (\nabla \varphi)_{\#} \mu$ , where  $\mu$ ,  $\nu$ , and  $\Phi$  satisfy:*

- *The supports of  $\mu$  and  $\nu$  lie in  $\Omega = B_1(0)$ .*
- *The set  $\Phi$  is bounded in  $L^\infty$  on  $\Omega$ , i.e.,  $\sup_{\phi \in \Phi} \|\phi\|_{L^\infty(\Omega)} \lesssim 1$ .*
- *The potentials satisfy  $\phi(0) = 0$  and  $\phi(x) = +\infty$  if  $x \notin \Omega$ .*
- *The potentials are lower-semicontinuous everywhere, and smooth and strongly convex on  $\Omega$ :  $\frac{1}{2}I \preceq \nabla^2 \phi(x) \preceq 2I$  if  $\|x\| < 1$ .*

The first and second of these assumptions can be weakened under suitably strong moment assumptions, but we do not pursue this avenue here. The third is without loss of generality: since subtracting a constant from  $\phi$  does not affect the semidual objective or the gradient  $\nabla \phi$ , we can always assume that  $\phi(0) = 0$ , and since the supports of  $\mu$  and  $\nu$  lie in  $B_1(0)$ , we may define  $\phi$  to be infinity outside of this set without affecting the semidual problem. The fact that  $\phi = +\infty$  identically outside of  $\Omega$  simplifies several arguments involving the conjugate function, since it implies that

$$\phi^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - \phi(x)\} = \sup_{x \in \Omega} \{\langle x, y \rangle - \phi(x)\}$$

for all  $y \in \mathbb{R}^d$  and  $\phi \in \Phi$ . The fourth assumption is the most important, because it guarantees the stability of the semidual problem via Theorem 3.1.

With these assumptions in hand, we can carry out the first step of the analysis.

**Lemma 3.4.** *Adopt Assumption 3.3, and assume that  $\nu = \nabla \varphi \# \mu$  for some  $\varphi \in \Phi$ . Let  $\widehat{\varphi}$  be given by*

$$\widehat{\varphi} = \arg \min_{\phi \in \Phi} \mathcal{S}_n(\phi).$$

*Then*

$$\|\nabla \widehat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 \lesssim \sup_{\phi \in \Phi} \{ |(\mu_n - \mu)(\phi)| + |(\nu_n - \nu)(\phi^*)| \}. \quad (3.4)$$

*Proof.* The proof is an application of a standard argument in empirical risk minimization. Theorem 3.1 implies

$$\begin{aligned} \|\nabla \widehat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 &\lesssim \mathcal{S}(\widehat{\varphi}) - \mathcal{S}(\varphi) \\ &= \mathcal{S}(\widehat{\varphi}) - \mathcal{S}_n(\widehat{\varphi}) + \mathcal{S}_n(\widehat{\varphi}) - \mathcal{S}_n(\varphi) + \mathcal{S}_n(\varphi) - \mathcal{S}(\varphi) \\ &\leq 2 \sup_{\phi \in \Phi} |\mathcal{S}_n(\phi) - \mathcal{S}(\phi)|, \end{aligned}$$

where the last inequality uses that  $\mathcal{S}_n(\widehat{\varphi}) - \mathcal{S}_n(\varphi) \leq 0$  by definition of  $\widehat{\varphi}$ . Expanding the definitions of  $\mathcal{S}$  and  $\mathcal{S}_n$  yields the claim.  $\square$

To bound the right side of (3.4), we employ the chaining technique of Proposition 2.6. For simplicity, we focus on the case where the class of functions is small enough that the covering numbers satisfy

$$\log N(\varepsilon, \Phi) \lesssim \varepsilon^{-\gamma} \log(1 + \varepsilon^{-1}), \quad \gamma \in [0, 1) \quad (3.5)$$

for all sufficiently small  $\varepsilon$ .

A paradigmatic example of such classes are parametric classes, where the set  $\Phi$  is finite dimensional, that is, where  $\Phi = \{\phi_\theta\}_{\theta \in \Theta}$  is indexed by a parameter  $\theta \in \Theta \subseteq \mathbb{R}^M$ . Indeed, in this case, we have the following bound.

**Lemma 3.5.** *Assume that  $\Phi = \{\phi_\theta\}_{\theta \in \Theta}$ , where  $\Theta \subseteq \mathbb{R}^M$  is bounded, and the potentials satisfy  $\|\phi_\theta - \phi_{\theta'}\|_{L^\infty(\Omega)} \lesssim \|\theta - \theta'\|$ . Then there exists a positive constant  $C$  such that the covering numbers of  $\Phi$  satisfy  $\log N(\varepsilon, \Phi) = 0$  if  $\varepsilon \geq C$  and*

$$\log N(\varepsilon, \Phi) \lesssim \log(1 + \varepsilon^{-1})$$

*otherwise.*

*Proof.* By assumption,  $\Theta \subseteq B_R(0)$  for some  $R > 0$ , so  $\|\phi - \psi\|_{L^\infty(\Omega)} \lesssim R$  for any  $\phi, \psi \in \Phi$ . Therefore, if  $\varepsilon$  is larger than a sufficiently large constant, any element of  $\Phi$  constitutes a one-element  $\varepsilon$ -cover of  $\Phi$ .

For any  $\delta > 0$ , Exercise 2 shows that there exists  $\theta_1, \dots, \theta_N$  with  $N \leq (1 + 2R\delta^{-1})^M$  such that  $\bigcup_{i=1}^N B_\delta(\theta_i) \supseteq \Theta$ . Then  $\phi_{\theta_1}, \dots, \phi_{\theta_N}$  is an  $O(\delta)$ -cover of  $\Phi$ . Indeed, for any  $\theta \in \Theta$ , we may choose  $i \in [N]$  such that  $\|\phi_\theta - \phi_{\theta_i}\|_{L^\infty(\Omega)} \lesssim \|\theta - \theta_i\| \leq \delta$ . Taking  $\delta = c\varepsilon$  for a sufficiently small positive constant  $c$  yields the claim.  $\square$

To give some examples of parametric classes,  $\{\phi_\theta\}_{\theta \in \Theta}$  could be a set of convex quadratic functions, as in Section 3.3, or it could consist of linear combinations of a fixed dictionary  $\{\phi_1, \dots, \phi_M\}$ , with

$$\phi_\theta = \sum_{i=1}^M \theta_i \phi_i.$$

Note that it is common in non-parametric statistics to choose the dictionary carefully to balance approximation and estimation errors, but we do not delve into such questions here in order to focus on the core statistical content.

More generally, condition (3.5) allows for infinite-dimensional function classes which are nevertheless not “too large”. By contrast, it excludes classes whose complexity grows with the ambient dimension, such as the class of Lipschitz functions studied in Lemma 2.7.

What is the optimal rate of estimating  $T = \nabla \varphi$  under this assumption? If  $\Phi$  is a parametric class, we expect the minimax rate to be  $n^{-1}$ —in particular, we expect that the map estimation problem avoids the curse of dimensionality. Indeed, rates avoiding the curse of dimensionality are achievable whenever a bound such as (3.5) is satisfied.

Lemma 3.4 involves both the potential  $\phi$  and its conjugate  $\phi^*$ . Unfortunately, even if the set  $\Phi$  has a simple form, the set  $\Phi^* = \{\phi^* : \phi \in \Phi\}$  may defy easy description. However, we make the following crucial observation: the covering numbers of  $\Phi$  control those of  $\Phi^*$ .

**Lemma 3.6.** *For any  $\varepsilon > 0$ ,*

$$N(\varepsilon, \Phi^*) \leq N(\varepsilon, \Phi).$$

*Proof.* The result follows from the fact that the conjugation operation is a contraction in  $L^\infty$ . Indeed, given any pair of functions  $\phi, \psi \in \Phi$ ,

$$\begin{aligned}
|\phi^*(y) - \psi^*(y)| &= \left| \sup_{x \in \Omega} \{\langle x, y \rangle - \phi(x)\} - \sup_{x' \in \Omega} \{\langle x', y \rangle - \psi(x')\} \right| \\
&\leq \sup_{x \in \Omega} |\phi(x) - \psi(x)| = \|\phi - \psi\|_{L^\infty(\Omega)}.
\end{aligned}$$

In particular, if  $\phi_1, \dots, \phi_N$  is an  $\varepsilon$ -net for  $\Phi$ , then  $\phi_1^*, \dots, \phi_N^*$  is an  $\varepsilon$ -net for  $\Phi^*$ .  $\square$

We can now prove our first convergence rate for map estimation.

**Theorem 3.7.** *Adopt Assumption 3.3 and assume (3.5) holds. The semidual estimator  $\hat{\varphi}$  satisfies the bound*

$$\mathbb{E} \|\nabla \hat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 \lesssim n^{-1/2}.$$

*Proof.* Lemma 3.4 implies

$$\mathbb{E} \|\nabla \hat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 \lesssim \mathbb{E} \sup_{\phi \in \Phi} |(\mu_n - \mu)(\phi)| + \mathbb{E} \sup_{\phi \in \Phi} |(\nu_n - \nu)(\phi^*)|.$$

By Assumption 3.3, there exists a positive constant  $R$  such that  $\|\phi\|_{L^\infty(\Omega)} \leq R$  and  $\|\phi^*\| \leq R$  for all  $\phi \in \Phi$ . Applying Proposition 2.6 with  $\tau = 0$  yields

$$\mathbb{E} \|\nabla \hat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 \lesssim \frac{1}{\sqrt{n}} \int_0^R (\sqrt{\log N(\varepsilon, \Phi)} + \sqrt{\log N(\varepsilon, \Phi^*)}) \, d\varepsilon.$$

Applying (3.5) and Lemma 3.6, we obtain

$$\mathbb{E} \|\nabla \hat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 \lesssim \frac{1}{\sqrt{n}} \int_0^R \varepsilon^{-\gamma/2} \sqrt{\log(1 + \varepsilon^{-1})} \, d\varepsilon \lesssim n^{-1/2},$$

as desired.  $\square$

As anticipated, the parametric assumption on the class  $\Phi$  translates to a rate of convergence that avoids the curse of dimensionality. However, the “slow rate”  $n^{-1/2}$  is not quite what we hoped to prove. To obtain the “fast rate”  $n^{-1}$ , we need to *localize* and exploiting this localization step requires imposing additional assumptions on  $\mu$ .

### 3.5 The fixed point argument

As mentioned above, the chaining argument in Theorem 3.7 fails to give the correct rate of convergence because it is based on bounding the deviations of  $\mathcal{S}_n$  from  $\mathcal{S}$  uniformly over the set  $\Phi$ . However, Theorem 3.7 shows that  $\widehat{\varphi}$  is close to  $\varphi$  when  $n$  is large, which suggests that it is not necessary to bound the deviations of  $\mathcal{S}_n$  from  $\mathcal{S}$  uniformly over the set  $\Phi$ , but only over that subset near  $\varphi$ .

The argument below, due to van de Geer, formalizes this process in the context of map estimation. The main idea is to apply the reasoning of Lemma 3.4 not to  $\widehat{\varphi}$  but to a convex combination of  $\widehat{\varphi}$  and  $\varphi$  itself. For simplicity, to apply this argument, we make one more assumption on  $\Phi$ .

**Assumption 3.8.** *The set  $\Phi$  is convex, that is, if  $\phi, \psi \in \Phi$ , then  $(1 - \lambda)\phi + \lambda\psi \in \Phi$  for all  $\lambda \in [0, 1]$ .*

Define

$$\varphi_\varepsilon = (1 - \lambda)\varphi + \lambda\widehat{\varphi}, \quad \lambda = \frac{\varepsilon}{\varepsilon + \|\nabla\widehat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}.$$

Then,

$$\begin{aligned} \|\nabla\varphi_\varepsilon - \nabla\varphi\|_{L^2(\mu)} &= \lambda \|\nabla\widehat{\varphi} - \nabla\varphi\|_{L^2(\mu)} \\ &= \varepsilon \left( \frac{\|\nabla\widehat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}{\varepsilon + \|\nabla\widehat{\varphi} - \nabla\varphi\|_{L^2(\mu)}} \right) \leq \varepsilon. \end{aligned}$$

Moreover:

$$\begin{aligned} \|\nabla\varphi_\varepsilon - \nabla\varphi\|_{L^2(\mu)} \leq \frac{\varepsilon}{2} &\iff \frac{\varepsilon \|\nabla\widehat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}{\varepsilon + \|\nabla\widehat{\varphi} - \nabla\varphi\|_{L^2(\mu)}} \leq \frac{\varepsilon}{2} \\ &\iff \|\nabla\widehat{\varphi} - \nabla\varphi\|_{L^2(\mu)} \leq \varepsilon, \end{aligned}$$

so in considering  $\|\nabla\varphi_\varepsilon - \nabla\varphi\|_{L^2(\mu)}$  instead of  $\|\nabla\widehat{\varphi} - \nabla\varphi\|_{L^2(\mu)}$  we only lose a factor of 2. Finally, Assumption 3.8 guarantees that  $\varphi_\varepsilon \in \Phi$ , since it is a convex combination of elements of  $\Phi$ .

Also, note that for  $\lambda \in [0, 1]$ , pointwise we have  $((1 - \lambda)\phi_0 + \lambda\phi_1)^* \leq (1 - \lambda)\phi_0^* + \lambda\phi_1^*$ , from which it follows that  $\mathcal{S}_n$  is a convex functional. If we set  $\overline{\mathcal{S}} = \mathcal{S} - \mathcal{S}(\varphi)$  and  $\overline{\mathcal{S}}_n = \mathcal{S}_n - \mathcal{S}_n(\varphi)$ , then by convexity,

$$\overline{\mathcal{S}}_n(\varphi_\varepsilon) \leq (1 - \lambda) \underbrace{\overline{\mathcal{S}}_n(\varphi)}_{=0} + \lambda \underbrace{\overline{\mathcal{S}}_n(\widehat{\varphi})}_{\leq 0} \leq 0$$

by the definition of  $\widehat{\varphi}$ . Therefore, by Theorem 3.1, if  $\varphi_\varepsilon$  is  $\frac{1}{2}$ -strongly convex and 2-smooth,

$$\begin{aligned} \frac{1}{4} \|\nabla \varphi_\varepsilon - \nabla \varphi\|_{L^2(\mu)}^2 &\leq \bar{\mathcal{S}}(\varphi_\varepsilon) \leq (\bar{\mathcal{S}} - \bar{\mathcal{S}}_n)(\varphi_\varepsilon) \\ &= (\mathcal{S} - \mathcal{S}_n)(\varphi_\varepsilon) - (\mathcal{S} - \mathcal{S}_n)(\varphi) \\ &\leq \sup_{\phi \in \Phi_\varepsilon} (\mathcal{S} - \mathcal{S}_n)(\phi) - (\mathcal{S} - \mathcal{S}_n)(\varphi) \\ &\leq \sup_{\phi \in \Phi_\varepsilon} \{ |(\mu_n - \mu)(\phi - \varphi)| + |(\nu_n - \nu)(\phi^* - \varphi^*)| \}, \end{aligned}$$

where  $\Phi_\varepsilon = \{\phi \in \Phi : \|\nabla \phi - \nabla \varphi\|_{L^2(\mu)} \leq \varepsilon\}$ .

If the right side of the above inequality is less than  $\varepsilon^2/16$ , then  $\|\nabla \varphi_\varepsilon - \nabla \varphi\|_{L^2(\mu)} \leq \varepsilon/2$ , which in turn implies that  $\|\nabla \widehat{\varphi} - \nabla \varphi\|_{L^2(\mu)} \leq \varepsilon$ . We therefore can control the risk of  $\widehat{\varphi}$  if we can find an  $\varepsilon$  for which the supremum of the empirical process over  $\mathcal{F}_\varepsilon$  is of order  $\varepsilon^2$ . We formalize the above considerations in the following proposition.

**Proposition 3.9.** *Adopt Assumptions 3.3 and 3.8. For  $\varepsilon > 0$ , let*

$$r(\varepsilon) = \sup_{\phi \in \Phi_\varepsilon} \{ |(\mu_n - \mu)(\phi - \varphi)| + |(\nu_n - \nu)(\phi^* - \varphi^*)| \}.$$

*Then on the event  $\{r(\varepsilon) \leq \varepsilon^2/16\}$ , the semidual estimator  $\widehat{\varphi}$  satisfies  $\|\nabla \widehat{\varphi} - \nabla \varphi\|_{L^2(\mu)} \leq \varepsilon$ .*

*In particular, if  $\varepsilon_n$  is a deterministic quantity such that  $r(\varepsilon_n) \leq \varepsilon_n^2/16$  with high probability, then the risk of  $\widehat{\varphi}$  is bounded by  $\varepsilon_n$  with high probability.*

Comparing Lemma 3.4 with Proposition 3.9 shows that we have replaced the task of bounding the deviations of an empirical process uniformly over  $\Phi$  by the task of bounding them over the smaller set  $\Phi_\varepsilon$ .

### 3.6 Obtaining the fast rate

In order to exploit the fact that we now seek to bound the empirical process only over  $\Phi_\varepsilon$ , we need to formalize the notion that  $\Phi_\varepsilon$  is much smaller than  $\Phi$ . A complicating factor is that the chaining technique given in Proposition 2.6 measures the “size” of  $\Phi$  by its  $\varepsilon$ -covering numbers, which are defined in terms of  $L^\infty$  covers. By contrast, the restriction of  $\Phi$  to  $\Phi_\varepsilon$  is based on the additional restriction on the  $L^2$

norm of the gradients of  $\phi$ . We therefore need a version of the chaining bound which is able to exploit the size of a function class with respect to both  $L^\infty$  and  $L^2$ .

The following modified chaining bound addresses this deficit.

**Proposition 3.10** ([vdVW23, Theorem 2.14.21]). *Let  $P$  be a probability measure on a set  $\Omega \subseteq \mathbb{R}^d$ . Let  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P$ . If  $\mathcal{F}$  is a set of real-valued functions such that  $\|f\|_{L^2(P)} \leq \sigma$  and  $\|f\|_{L^\infty(\Omega)} \leq R$  for all  $f \in \mathcal{F}$ , then*

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\} &\lesssim \frac{1}{\sqrt{n}} \int_0^\sigma \sqrt{\log N(\varepsilon, \mathcal{F})} \, d\varepsilon \\ &\quad + \frac{1}{n} \int_0^R \log N(\varepsilon, \mathcal{F}) \, d\varepsilon. \end{aligned} \quad (3.6)$$

Note that in the first term in (3.6), the upper limit of the integral is  $\sigma$  rather than  $R$ . The second integral incurs a worse dependence on the covering number, but appears with a prefactor of  $\frac{1}{n}$  rather than  $\frac{1}{\sqrt{n}}$ . We may therefore hope that when  $n$  is large enough, the first term dominates. If the  $L^2$  size of  $\mathcal{F}$  is small, as captured by  $\sigma$ , then the first term may be substantially smaller than the bound obtained by applying Proposition 2.6 directly.

Proposition 3.10 also comes with a tail bound, showing that the quantity on the right-hand side of (3.6) also bounds the empirical process with high probability.

**Proposition 3.11.** *Let  $J_n(\mathcal{F})$  denote the right side of (3.6). Under the same assumptions as Proposition 3.10, there exists a positive universal constant  $C$  such that for any  $t \geq 0$ ,*

$$\mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \{f(X_i) - \mathbb{E}f(X_i)\} \geq C \left( J_n(\mathcal{F}) + \sigma \sqrt{\frac{t}{n}} + R \frac{t}{n} \right) \right) \leq \exp(-t).$$

Proposition 3.10 requires us to control the  $L^2$  norm of the elements of our function class; however,  $\Phi_\varepsilon$  is defined using the  $L^2$  norms of the gradients of the elements of  $\Phi$ . We therefore adopt the final assumption on  $\mu$ , which allows us to move back and forth between these notions.

**Definition 3.12.** *A measure  $P$  satisfies a Poincaré inequality (with constant  $C$ ) if for all  $f \in L^2(P)$ ,*

$$\int \left( f - \int f \, dP \right)^2 \, dP \leq C \int \|\nabla f\|^2 \, dP.$$

**Assumption 3.13.** *The measure  $\mu$  satisfies a Poincaré inequality.*

The Poincaré inequality is a quantitative form of the statement that the support of  $\mu$  is connected. Indeed, a Poincaré inequality holds for any measure having a density bounded away from zero and infinity on a bounded Lipschitz domain.

Under this new assumption, we obtain  $L^2$  bounds on  $\Phi_\varepsilon$  and  $\Phi_\varepsilon^*$ .

**Proposition 3.14.** *If Assumptions 3.3 and 3.13 hold, then*

$$\begin{aligned}\|\phi - \varphi - \mu(\phi - \varphi)\|_{L^2(\mu)} &\lesssim \varepsilon \\ \|\phi^* - \varphi^* - \nu(\phi^* - \varphi^*)\|_{L^2(\nu)} &\lesssim \varepsilon\end{aligned}$$

for all  $\phi \in \Phi_\varepsilon$ .

*Proof.* The first bound follows directly from the Poincaré inequality: for  $\phi \in \Phi_\varepsilon$ ,

$$\|\phi - \varphi - \mu(\phi - \varphi)\|_{L^2(\mu)}^2 \leq C \|\nabla(\phi - \varphi)\|_{L^2(\mu)}^2 \lesssim \varepsilon^2.$$

To prove the second bound, we first use the strong convexity of  $\phi$  and  $\varphi$ . Consider the functional  $\mathcal{T}$  defined by

$$\mathcal{T}(\psi) := \int \psi \, d\nu + \int \psi^* \, d\mu.$$

That is,  $\mathcal{T}$  is the semidual functional obtained by exchanging the roles of  $\mu$  and  $\nu$ . Since  $(\phi^*)^* = \phi$  for all convex and lower semicontinuous  $\phi$ , we have that  $\mathcal{S}(\phi) = \mathcal{T}(\phi^*)$  for all such  $\phi$ . In particular, the minimizer of  $\mathcal{T}$  is  $\varphi^*$ , and Theorem 3.1 implies that

$$\frac{1}{4} \|\nabla\phi^* - \nabla\varphi^*\|_{L^2(\nu)}^2 \leq \mathcal{T}(\phi^*) - \mathcal{T}(\varphi^*) = \mathcal{S}(\phi) - \mathcal{S}(\varphi) \leq \|\nabla\phi - \nabla\varphi\|_{L^2(\mu)}^2.$$

Therefore  $\|\nabla\phi^* - \nabla\varphi^*\|_{L^2(\nu)}^2 \lesssim \varepsilon^2$  for all  $\phi \in \Phi_\varepsilon$ .

To obtain the bound, all that is left is to show that  $\nu$  also satisfies a Poincaré inequality, since we can then conclude as in the proof of the first inequality. To see this, we use the fact that  $\nu = (\nabla\varphi)_\# \mu$ . The fact that  $\varphi$  is smooth means that for any  $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ,

$$\|\nabla(f \circ \nabla\varphi)(x)\| = \|\nabla^2\varphi(x) \nabla f(\nabla\varphi(x))\| \leq 2 \|\nabla f(\nabla\varphi(x))\|.$$

The Poincaré inequality for  $\mu$  implies that for any  $f \in L^2(\nu)$ ,

$$\begin{aligned}
\int \left( f - \int f \, d\nu \right)^2 d\nu &= \int \left( f \circ \nabla \varphi - \int f \circ \nabla \varphi \, d\mu \right)^2 d\mu \\
&\leq C \int \|\nabla(f \circ \nabla \varphi)\|^2 d\mu \\
&\leq 4C \int \|\nabla f(\nabla \varphi(x))\|^2 d\mu \\
&= 4C \int \|\nabla f\|^2 d\nu.
\end{aligned}$$

Therefore  $\nu$  also satisfies a Poincaré inequality, with constant  $4C$ . Hence we may conclude as in the first case.  $\square$

We are finally ready to prove the desired rate. Note that in the finite-dimensional setting, when Lemma 3.5 holds, this theorem shows that the map can be estimated at nearly the parametric rate.

**Theorem 3.15.** *Adopt Assumptions 3.3, 3.8, and 3.13. If (3.5) holds, then the semidual estimator  $\widehat{\varphi}$  satisfies*

$$\|\nabla \widehat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 \lesssim \left( \frac{\log n}{n} \right)^{\frac{2}{2+\gamma}} + \frac{t+1}{n} \quad (3.7)$$

with probability at least  $1 - e^{-t}$ . In particular,

$$\mathbb{E} \|\nabla \widehat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 \lesssim \left( \frac{\log n}{n} \right)^{\frac{2}{2+\gamma}}.$$

*Proof.* By Proposition 3.9, it suffices to show that  $r(\varepsilon_n) \leq \varepsilon_n^2/16$  with probability at least  $1 - e^{-t}$  for  $\varepsilon_n \asymp \left( \frac{\log n}{n} \right)^{\frac{1}{2+\gamma}} + \sqrt{\frac{t+1}{n}}$ . Let us first bound  $\sup_{\phi \in \Phi_\varepsilon} |(\mu_n - \mu)(\phi - \varphi)|$ . Since  $(\mu_n - \mu)h = 0$  if  $h$  is a constant function, we have

$$\sup_{\phi \in \Phi_\varepsilon} |(\mu_n - \mu)(\phi - \varphi)| = \sup_{\phi \in \Phi_\varepsilon} |(\mu_n - \mu)(\phi - \varphi - \mu(\phi - \varphi))|.$$

We can apply Proposition 3.10, Proposition 3.11, and Proposition 3.14 along with the fact that  $\phi - \varphi - \mu(\phi - \varphi)$  is bounded to obtain that there exists a constant  $C_2$  such that

$$\begin{aligned}
&\sup_{\phi \in \Phi_\varepsilon} |(\mu_n - \mu)(\phi - \varphi - \mu(\phi - \varphi))| \\
&\lesssim \frac{1}{\sqrt{n}} \int_0^{C_2 \varepsilon} \delta^{-\gamma/2} \sqrt{\log(1 + \delta^{-1})} \, d\delta
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{n} \int_0^{C_2} \delta^{-\gamma} \log(1 + \delta^{-1}) d\delta + \varepsilon \sqrt{\frac{t}{n}} + \frac{t}{n} \\
& \lesssim \varepsilon^{1-\gamma/2} \sqrt{\frac{\log(1 + \varepsilon^{-1})}{n}} + \varepsilon \sqrt{\frac{t}{n}} + \frac{t+1}{n}
\end{aligned}$$

with probability at least  $1 - e^{-t}$ . Therefore, taking

$$\varepsilon_n = C \left( \left( \frac{\log n}{n} \right)^{\frac{1}{2+\gamma}} + \sqrt{\frac{t+1}{n}} \right) \quad (3.8)$$

for a sufficiently large constant  $C$ , we can ensure that

$$\sup_{\phi \in \Phi_{\varepsilon_n}} |(\mu_n - \mu)(\phi - \varphi - \mu(\phi - \varphi))| \leq \varepsilon_n^2 / 32$$

with probability at least  $1 - e^{-t}/2$ . An analogous argument yields

$$\sup_{\phi \in \Phi_{\varepsilon_n}} |(\nu_n - \nu)(\phi^* - \varphi^* - \nu(\phi^* - \varphi^*))| \leq \varepsilon_n^2 / 32$$

with the same probability. By a union bound, we obtain that  $r(\varepsilon_n) \leq \varepsilon_n^2 / 16$  for  $\varepsilon_n$  as in (3.8) as claimed.

The second bound following from integrating the tail.  $\square$

### 3.7 Discussion

**§3.1.** The empirical “plug-in” approach based on nearest neighbors was developed as a simple alternative to the semidual approach in [MBNWW21, DGS21, GS22]. Although the nearest neighbors estimator does not adapt to the smoothness of  $\mu$  and  $\nu$ , one recovers minimax rates via the optimal transport map between density estimators [MBNWW21], and even central limit theorems [MBNWW23]. However, compared to the plug-in approach, the semidual approach developed here is overall more flexible and can be combined with other tools such as kernel SoS [VMB<sup>+</sup>24].

**§3.2.** The semidual approach to map estimation was introduced in the paper [HR21], which also proved the semidual stability estimates and minimax lower bounds. That paper showed that, if the map between  $\mu$  and  $\nu$  is assumed to be  $s$ -smooth (i.e., to possess  $s$  bounded derivatives), then a suitable semidual estimator achieves the minimax-optimal rate

$$\mathbb{E} \|\nabla \hat{\varphi} - \nabla \varphi\|_{L^2(\mu)}^2 \lesssim n^{-\frac{2\alpha}{2\alpha-2+d}} (\log n)^2 + \frac{1}{n}. \quad (3.9)$$

This approach was then explored in great generality in [DNWP22], and the arguments in that paper are closely related to those in this chapter. However, the tools we describe here are not strong enough to prove (3.9), since the class of  $s$ -smooth functions does not satisfy (3.5). More information about how to obtain (3.9) along the lines of the arguments we have presented in this chapter can be found in [DNWP22, Section 4.4].

The alternative semidual stability estimates in Exercises 3 and 5 are taken from [MBNWW21].

**§3.3.** Estimating the transport map between Gaussians was given as an example in [DNWP22] in which the semidual approach yields parametric rates; see the paper for other function classes of interest.

**§3.4.** Standard references for empirical risk minimization (or M-estimation) include [vdV98, Wai19]. The “slow rate” is characteristic of M-estimation problems in the absence of strong convexity; the Poincaré inequality assumption adopted in §3.6 can be viewed as the appropriate strong convexity condition for the semi-dual functional  $\mathcal{S}$ .

**§3.5.** The one-shot localization we use is due to van de Geer [vdG87, vdG02] and provides an alternative to the usual localization arguments (e.g., [Wai19, Chapter 14]).

**§3.6.** The improved chaining bound of Proposition 3.10 is obtained by the “generic chaining” technique developed by Talagrand [Tal96]. This technique is at the heart of the study of Gaussian processes, see [Tal21]. The tail bound in Proposition 3.11 follows from a more general result for generic chaining bounds [vdVW23, Theorem 2.2.19].

The Poincaré inequality is a standard tool in high-dimensional probability, see [BLM13, BGL14, vH14]. It is an open question whether the rates presented in this chapter are achievable without making this assumption.

### 3.8 Exercises

1. Let  $\mu, \nu \in \mathcal{P}([0, 1])$  and let  $X_1, \dots, X_n \sim \mu$ ,  $Y_1, \dots, Y_n \sim \nu$  be i.i.d. samples independent from each other. Assume that  $\mu, \nu$  have differentiable CDFs  $F_\mu, F_\nu$  respectively, such that

$$0 < c \leq F'_\mu, F'_\nu \leq C < \infty \quad \text{on } [0, 1].$$

This is equivalent to  $\mu, \nu$  having densities on  $[0, 1]$  which are bounded away from zero and infinity.

Let us show that the following estimator  $\hat{T}_n$  obeys a parametric rate of convergence. Let  $X_{(1)} < \dots < X_{(n)}$ ,  $Y_{(1)} < \dots < Y_{(n)}$  denote the samples in sorted order, and given  $x \in [0, 1]$  let  $X_{(i)}$  denote the largest sample from  $\mu$  such that  $X_{(i)} \leq x$ . We then set  $\hat{T}_n(x) := Y_{(i)}$  (if no such  $X_{(i)}$  exists, then output  $\hat{T}_n(x) := 0$ ). This estimator can be viewed as a piecewise constant interpolation of the empirical optimal coupling, or as a 1-nearest neighbor estimator. For simplicity, we fix  $x \in [0, 1]$  and prove

$$\mathbb{E}[|\hat{T}_n(x) - T(x)|^2] \lesssim 1/n$$

where  $T$  is the true optimal transport map  $F_\nu^{-1} \circ F_\mu$  from  $\mu$  to  $\nu$ , although it is a straightforward exercise to extend the results of this problem to the integrated risk  $\mathbb{E} \int |\hat{T}_n - T|^2 d\mu$ .

- a) Let  $N_x := \sum_{i=1}^n \mathbb{1}\{X_i \leq x\}$  and  $N'_y := \sum_{i=1}^n \mathbb{1}\{Y_i \leq y\}$ . Argue that if  $N'_y < N_x$ , then  $\hat{T}_n(x) \geq y$ ; if  $N'_y > N_x$ , then  $\hat{T}_n(x) \leq y$ .
- b) Using the Dvoretzky–Kiefer–Wolfowitz inequality, argue that for any  $\delta \in (0, 1)$ , the following hold simultaneously with probability at least  $1 - \delta$ :

$$|N_x - nF_\mu(x)| \lesssim \sqrt{n \log(1/\delta)}$$

and

$$|N'_y - nF_\nu(y)| \lesssim \sqrt{n \log(1/\delta)} \quad \text{for all } y \in [0, 1].$$

- c) Use the previous two parts to conclude.
- 2. This exercise shows that the ball  $B_R(0)$  in  $\mathbb{R}^d$  can be covered by  $(1 + 2R\delta^{-1})^d$  balls of radius  $\delta$ .
  - a) Argue by rescaling that it suffices to consider the case  $R = 1$ .
  - b) Let  $\mathcal{X} = \{x_1, \dots, x_N\}$  be any set of elements of  $B_1(0)$  such that  $\|x_i - x_j\| > \delta$  for all  $i \neq j$ . Such a set is called a  $\delta$ -packing of  $B_1(0)$ . Show that if  $\mathcal{X}$  is a  $\delta$ -packing of  $B_1(0)$ , then the sets  $\{B_{\frac{\delta}{2}}(x)\}_{x \in \mathcal{X}}$  are disjoint subsets of  $B_{1+\frac{\delta}{2}}(0)$ . Conclude that  $|\mathcal{X}| \leq (1 + 2\delta^{-1})^d$ .
  - c) Suppose that  $\mathcal{X}$  is a maximal  $\delta$ -packing of  $B_R(0)$ , i.e., there does not exist a strict superset of  $\mathcal{X}$  which is also a  $\delta$ -packing. Argue (via the contrapositive) that  $B_R(0) \subseteq \bigcup_{x \in \mathcal{X}} B_\delta(x)$ .
  - d) Use the previous two parts to conclude.
- 3. Let  $\mu, \nu$  be probability measures over  $\mathbb{R}^d$  and let  $\nabla\varphi$  denote the optimal transport map from  $\mu$  to  $\nu$ . Assume that  $\nabla\varphi$  is  $L$ -Lipschitz.

In this exercise, we establish the following estimate: for any other convex function  $\phi$ , if  $\hat{\nu} := (\nabla\phi)_{\#}\mu$ , then

$$\begin{aligned} & \|\nabla\phi - \nabla\varphi\|_{L^2(\mu)}^2 \\ & \leq L \left( W_2^2(\mu, \hat{\nu}) - W_2^2(\mu, \nu) - \int (\|\cdot\|^2 - 2\varphi^*) d(\hat{\nu} - \nu) \right). \end{aligned}$$

*Hint:* First, argue that by strong convexity of  $\varphi^*$ , it holds that

$$\begin{aligned} & \frac{1}{2L} \|\nabla\phi - \nabla\varphi\|_{L^2(\mu)}^2 \\ & \leq \int \varphi^* d(\hat{\nu} - \nu) - \int \langle x, \nabla\phi(x) - \nabla\varphi(x) \rangle \mu(dx). \end{aligned}$$

Then, expand out the quantity  $W_2^2(\mu, \hat{\nu}) - W_2^2(\mu, \nu)$  and substitute this into the above inequality.

4. We now use the stability estimate from the previous exercise to deduce rates for map estimation in the one-sample setting. Let  $\mu, \nu$  be probability measures with densities supported on the ball  $B_1(0)$  of radius 1 and assume that the optimal transport map  $\nabla\varphi$  from  $\mu$  to  $\nu$  is Lipschitz. Assume that we have access to  $\mu$ , and to  $n$  i.i.d. samples from  $\nu$ . Let  $\nabla\hat{\varphi}$  denote the optimal transport map from  $\mu$  to the empirical measure  $\nu_n$ . Using (2.22), prove that

$$\mathbb{E} \|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}^2 \lesssim \begin{cases} n^{-1/2}, & d < 4, \\ n^{-1/2} \log n, & d = 4, \\ n^{-2/d}, & d > 4. \end{cases}$$

5. Starting with Exercise 3, assume additionally that  $\varphi$  is  $\ell$ -strongly convex. Let  $\gamma$  denote the optimal coupling between  $\nu$  and  $\hat{\nu}$ , and let  $(Y, \hat{Y}) \sim \gamma$ . Then,  $(\nabla\varphi^*(Y), \hat{Y})$  is a (suboptimal) coupling between  $\mu$  and  $\hat{\nu}$ , hence  $W_2^2(\mu, \hat{\nu}) \leq \mathbb{E} \|\nabla\varphi^*(Y) - \hat{Y}\|^2$ . Expand this out and use the smoothness of  $\varphi^*$  to deduce the stronger stability estimate

$$\|\nabla\phi - \nabla\varphi\|_{L^2(\mu)} \leq \sqrt{L/\ell} W_2(\nu, \hat{\nu}).$$



---

## Entropic optimal transport

Entropic regularization is one of the most active research areas in modern optimal transport. As a regularization technique, it technically falls under the scope of Section 2.8. Indeed, we show in this chapter that it yields parametric rates, like many of the other regularization approaches we have discussed. But entropic optimal transport is, in fact, much more.

Since the groundbreaking work of Cuturi [Cut13], it has been primarily used as a computational device that enables fast computation of optimal transport distances using the Sinkhorn algorithm. However, our focus remains statistical and we refer the reader to the excellent manuscript [PC19b] of Gabriel Peyré and Marco Cuturi for more details on the computational benefits of entropic regularization.

The basic principle of entropic optimal transport is to modify the definition of optimal transport to include a penalization term based on the entropy of the coupling, that is, to consider the optimization problem

$$\inf_{\gamma \in \Gamma_{\mu,\nu}} \left\{ \int \|x - y\|^2 \gamma(dx, dy) - \varepsilon \text{Ent}(\gamma) \right\}, \quad (4.1)$$

where  $\text{Ent}(\gamma)$  denotes the differential entropy  $\int \gamma(x) \log \frac{1}{\gamma(x)} dx$  for an absolutely continuous probability measure  $\gamma$ . In fact, Cuturi originally considered a discrete version of this problem, where  $\mu$  and  $\nu$  are finitely supported and the coupling  $\gamma$  can therefore be identified with a matrix. He considered the problem

$$\inf_{\gamma \in \Gamma_{\mu,\nu}} \left\{ \sum_{i,j} \|x_i - y_j\|^2 \gamma_{i,j} - \varepsilon H(\gamma) \right\}, \quad (4.2)$$

where  $H(\gamma)$  denotes the Shannon entropy  $\sum_{i,j} \gamma_{i,j} \log \frac{1}{\gamma_{i,j}}$ .

The role of the penalty terms in both (4.1) and (4.2) is to encourage the coupling to be *more spread out* than the solution to the unregularized optimal transport problem. Informally, the entropy of a measure increases when its mass is more evenly spread. Indeed, Exercise 1 shows that uniform distributions (over a subset of  $\mathbb{R}^d$  in continuous case, or over a finite set in the discrete case) have the maximum possible entropy. The entropic penalty biases the solutions to (4.1) and (4.2) towards that extreme.

To put (4.1) and (4.2) on a common footing, we introduce the KL divergence between probability measures:

$$\text{KL}(P \parallel Q) = \begin{cases} \int \frac{dP}{dQ} \log \frac{dP}{dQ} dQ & \text{if } P \ll Q, \\ +\infty & \text{otherwise.} \end{cases}$$

Exercise 2 shows that both (4.1) and (4.2) are equivalent to

$$\inf_{\gamma \in \Gamma_{\mu, \nu}} \left\{ \int \|x - y\|^2 \gamma(dx, dy) + \varepsilon \text{KL}(\gamma \parallel \mu \otimes \nu) \right\}, \quad (4.3)$$

in the sense of yielding the same optimal  $\gamma$ , and we take (4.3) as the main definition of entropic OT.

In the next section, we give a non-rigorous motivation for this regularization approach from the perspective of convex duality. We analyze the resulting dual problem in Section 4.2.

## 4.1 Derivation of entropic optimal transport

In this section, we attempt to motivate the definition of entropic OT from basic optimization and duality principles. For simplicity, we assume for now that we work on a compact set  $\Omega \subseteq \mathbb{R}^d$ . Given  $\mu, \nu \in \mathcal{P}(\Omega)$ , recall from Theorem 1.14 that  $W_2^2(\mu, \nu)$  can be written

$$W_2^2(\mu, \nu) = \sup_{\substack{f, g \in C_b(\Omega) \\ f(x) + g(y) \leq \|x - y\|^2}} \left\{ \int f d\mu + \int g d\nu \right\}.$$

Formally, we can rewrite this as an unconstrained maximization problem by introducing a penalization term that enforces the constraint. Indeed, if we define

$$\iota(f, g) = \begin{cases} 0 & \text{if } f(x) + g(y) \leq \|x - y\|^2 \text{ } \mu \otimes \nu\text{-a.e.}, \\ +\infty & \text{otherwise,} \end{cases}$$

then we obtain

$$W_2^2(\mu, \nu) = \sup_{f,g \in C_b(\Omega)} \left\{ \int f \, d\mu + \int g \, d\nu - \iota(f, g) \right\}.$$

This is a concave maximization problem, so it is (in principle) benign; however, from a computational and statistical perspective, the fact that  $\iota$  fails to be continuous, much less smooth, is a source of difficulty. To remedy this, we can consider a relaxed version of this problem obtained by replacing  $\iota$  by a smoothed version.<sup>1</sup> Define

$$\iota^\varepsilon(f, g) = \varepsilon \iint \left( e^{(f(x)+g(y)-\|x-y\|^2)/\varepsilon} - 1 \right) \mu(dx) \nu(dy).$$

The function  $\iota^\varepsilon$  is convex and continuous on the space  $C_b(\Omega)$  of bounded, continuous functions on  $\Omega$ . Moreover, it is easy to see that as  $\varepsilon \searrow 0$ , we recover the original hard constraint.

**Lemma 4.1.** *For any measurable  $f, g$ ,*

$$\lim_{\varepsilon \searrow 0} \iota^\varepsilon(f, g) = \iota(f, g).$$

*Proof.* Suppose first that  $\iota(f, g) = 0$ , so that  $f(x) + g(y) \leq \|x - y\|^2$   $\mu \otimes \nu$ -almost everywhere. Then the integral

$$\iint \left( e^{(f(x)+g(y)-\|x-y\|^2)/\varepsilon} - 1 \right) \mu(dx) \nu(dy)$$

is bounded as  $\varepsilon \searrow 0$ , and hence  $\iota^\varepsilon(f, g) \rightarrow 0$ .

On the other hand, if  $\iota(f, g) = +\infty$ , then there exists  $\delta > 0$  and a set  $U$  of positive  $\mu \otimes \nu$  measure such that  $e^{(f(x)+g(y)-\|x-y\|^2)/\varepsilon} \geq e^{\delta/\varepsilon}$  for all  $(x, y) \in U$ . We obtain

$$\begin{aligned} \varepsilon \iint \left( e^{(f(x)+g(y)-\|x-y\|^2)/\varepsilon} - 1 \right) \mu(dx) \nu(dy) &\geq \varepsilon e^{\delta/\varepsilon} (\mu \otimes \nu)(U) - \varepsilon \\ &\rightarrow \infty. \end{aligned}$$

This concludes the proof.  $\square$

---

<sup>1</sup> The smoothing we employ is reminiscent of the “softmax” function in machine learning.

We are therefore led to consider the following “ $\varepsilon$ -smoothed” dual version of the  $W_2^2$  distance:

$$\sup_{f,g \in C_b(\Omega)} \left\{ \int f \, d\mu + \int g \, d\nu - \iota^\varepsilon(f, g) \right\} \quad (\varepsilon\text{-D-}W_2^2)$$

Now that we have derived a relaxation of the dual problem, we can ask what this corresponds to in the primal problem. It turns out that the relaxation we have proposed in the dual corresponds to an *entropic penalty* in the primal problem.

To obtain this connection, let us define a version of the Kullback–Leibler divergence over the space  $\mathcal{M}_+(\Omega)$  of all positive (not necessarily probability) Borel measures on  $\Omega$ :

$$\mathsf{KL}(P \parallel Q) = \begin{cases} \int \left( \frac{dP}{dQ} \log \frac{dP}{dQ} - \frac{dP}{dQ} + 1 \right) dQ & \text{if } P \ll Q, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that the integrand is non-negative, so that the integral is always well defined, and  $\mathsf{KL}$  is always non-negative. When  $P$  and  $Q$  are probability measures, the terms  $-\frac{dP}{dQ} + 1$  cancel out and we obtain the usual definition.

The importance of this definition is that the convex conjugate of the functional  $\mathsf{KL}(\cdot \parallel Q)$  (see Appendix A.1) is precisely the exponential term appearing in  $\iota^\varepsilon$ . This fact is a variant of what is commonly known as the *Gibbs variational principle*.

**Proposition 4.2.** *For any bounded measurable function  $h$ ,*

$$\sup_{P \in \mathcal{M}_+(\Omega)} \left\{ \int h \, dP - \mathsf{KL}(P \parallel Q) \right\} = \int (\exp h - 1) \, dQ.$$

Moreover the supremum is achieved at  $P_h$  satisfying  $\frac{dP_h}{dQ} = \exp h$ .

*Proof.* We show that, for any Borel measure  $P$ , the difference

$$\Delta := \int (\exp h - 1) \, dQ - \int h \, dP + \mathsf{KL}(P \parallel Q)$$

is non-negative, and equals 0 when  $P = P_h$ . We may assume without loss of generality that  $\mathsf{KL}(P \parallel Q) < +\infty$ , since otherwise the claim is vacuous. Expanding the definition of  $\mathsf{KL}(P \parallel Q)$ , we obtain

$$\begin{aligned}\Delta &= \int \left( e^h - 1 - h \frac{dP}{dQ} + \frac{dP}{dQ} \log \frac{dP}{dQ} - \frac{dP}{dQ} + 1 \right) dQ \\ &= \int \left( \frac{dP}{dQ} \log \left( e^{-h} \frac{dP}{dQ} \right) - \frac{dP}{dQ} + e^h \right) dQ. \quad (4.4)\end{aligned}$$

By change of measure, we have

$$\frac{dP}{dQ} = \frac{dP_h}{dQ} \frac{dP}{dP_h} = e^h \frac{dP}{dP_h}.$$

Therefore (4.4) can be written

$$\Delta = \int e^h \left( \frac{dP}{dP_h} \log \frac{dP}{dP_h} - \frac{dP}{dP_h} + 1 \right) dQ = \mathsf{KL}(P \parallel P_h).$$

Since  $\mathsf{KL}(P \parallel P_h) \geq 0$  and  $\mathsf{KL}(P_h \parallel P_h) = 0$ , this proves the claim.  $\square$

We can therefore rewrite  $(\varepsilon\text{-D-W}_2^2)$  as

$$\begin{aligned}&\sup_{f,g \in C_b(\Omega)} \left\{ \int f d\mu + \int g d\nu \right\} \\ &- \varepsilon \sup_{\gamma \in \mathcal{M}_+(\Omega)} \left\{ \int \frac{f(x) + g(y) - \|x - y\|^2}{\varepsilon} \gamma(dx, dy) - \mathsf{KL}(\gamma \parallel \mu \otimes \nu) \right\},\end{aligned}$$

and, rearranging,

$$\begin{aligned}&\sup_{f,g \in C_b(\Omega)} \inf_{\gamma \in \mathcal{M}_+(\Omega)} \left\{ \int \|x - y\|^2 \gamma(dx, dy) + \varepsilon \mathsf{KL}(\gamma \parallel \mu \otimes \nu) \right. \\ &\quad \left. + \int f d\mu + \int g d\nu - \int f \oplus g d\gamma \right\},\end{aligned}$$

where we define  $(f \oplus g)(x, y) = f(x) + g(y)$ .

As in Subsection 1.5.1, we can swap the inf and sup to get an upper bound on the value of  $(\varepsilon\text{-D-W}_2^2)$ :

$$\begin{aligned}&\inf_{\gamma \in \mathcal{M}(\Omega)} \left\{ \int \|x - y\|^2 \gamma(dx, dy) + \varepsilon \mathsf{KL}(\gamma \parallel \mu \otimes \nu) \right. \\ &\quad \left. + \sup_{f,g \in C_b(\Omega)} \left\{ \int f d\mu + \int g d\nu - \int f \oplus g d\gamma \right\} \right\}. \quad (4.5)\end{aligned}$$

We have already seen that

$$\sup_{f,g \in C_b(\Omega)} \left\{ \int f d\mu + \int g d\nu - \int f \oplus g d\gamma \right\} = \begin{cases} 0, & \text{if } \gamma \in \Gamma_{\mu,\nu}, \\ \infty, & \text{otherwise.} \end{cases}$$

Therefore, (4.5) is equivalent to

$$\inf_{\gamma \in \Gamma_{\mu, \nu}} \left\{ \int \|x - y\|^2 \gamma(dx, dy) + \varepsilon \text{KL}(\gamma \parallel \mu \otimes \nu) \right\}. \quad (\varepsilon\text{-W}_2^2)$$

This is the *primal* version of the entropic OT problem, and it is the version that is usually taken as the definition of entropic regularization. This choice of regularization is usually justified *a posteriori* by the existence of Sinkhorn's algorithm (see Section 4.2), but as we have seen it also arises naturally from a simple relaxation of the dual Kantorovich problem. The argument in this section establishes a form of weak duality, showing that the value of  $(\varepsilon\text{-W}_2^2)$  is lower bounded by the value of  $(\varepsilon\text{-D-W}_2^2)$ . The next section establishes a tight connection between the primal and dual problems, both in terms of their optimal value and their optimal solutions. This connection has been heavily exploited in the statistical analysis of entropic OT.

## 4.2 Duality

In this section, we show that the values of the primal problem  $(\varepsilon\text{-W}_2^2)$  and dual problem  $(\varepsilon\text{-D-W}_2^2)$  actually agree, and that an optimal solution to one problem can be extracted from the optimal solution to the other.

**Proposition 4.3.** *Let  $f^*, g^*$  solve  $(\varepsilon\text{-D-W}_2^2)$ . Then*

$$\gamma^*(dx, dy) = \exp\left(\frac{f^*(x) + g^*(y) - \|x - y\|^2}{\varepsilon}\right) \mu(dx) \nu(dy) \quad (4.6)$$

*is the unique solution to  $(\varepsilon\text{-W}_2^2)$ , and*

$$\begin{aligned} & \int \|x - y\|^2 \gamma^*(dx, dy) + \varepsilon \text{KL}(\gamma^* \parallel \mu \otimes \nu) \\ &= \int f^* d\mu + \int g^* d\nu - \varepsilon(f^*, g^*) = \int f^* d\mu + \int g^* d\nu. \end{aligned}$$

*Proof.* It suffices to show that  $\gamma^*$  is a solution to  $(\varepsilon\text{-W}_2^2)$ , since uniqueness follows immediately from the strict convexity of  $\text{KL}(\gamma \parallel \mu \otimes \nu)$ .

We need to show that  $\gamma^* \in \Gamma_{\mu, \nu}$ . Clearly  $\gamma^*$  is positive, so it suffices to show that it has the correct marginals. To that end, we need to verify that for any Borel set  $A$ ,

$$\nu(A) = \int_A \int \exp\left(\frac{f^*(x) + g^*(y) - \|x - y\|^2}{\varepsilon}\right) \mu(dx) \nu(dy).$$

Equivalently, we need to show that

$$\int \exp\left(\frac{f^*(x) + g^*(y) - \|x - y\|^2}{\varepsilon}\right) \mu(dx) = 1 \quad \nu\text{-a.e.} \quad (4.7)$$

Let us define the function

$$\bar{g}(y) = -\varepsilon \log \int \exp\left(\frac{f^*(x) - \|x - y\|^2}{\varepsilon}\right) \mu(dx). \quad (4.8)$$

We show that  $\bar{g} = g^*$ ,  $\nu$ -almost everywhere. Since

$$\int \exp\left(\frac{f^*(x) + \bar{g}(y) - \|x - y\|^2}{\varepsilon}\right) \mu(dx) = 1 \quad \forall y \in \mathbb{R}^d \quad (4.9)$$

holds by definition, this establishes that (4.7) holds as well.

Let  $\bar{\gamma}(dx, dy) = \exp((f^*(x) + \bar{g}(y) - \|x - y\|^2)/\varepsilon) \mu(dx) \nu(dy)$ . Then,

$$\begin{aligned} 0 \leq \mathsf{KL}(\bar{\gamma} \parallel \gamma^*) &= \int \left( \frac{d\bar{\gamma}}{d\gamma^*} \log \frac{d\bar{\gamma}}{d\gamma^*} - \frac{d\bar{\gamma}}{d\gamma^*} + 1 \right) d\bar{\gamma} \\ &= \int \frac{\bar{g}(y) - g^*(y)}{\varepsilon} \bar{\gamma}(dx, dy) - \int d\bar{\gamma} + \int d\gamma^* \\ &= \frac{1}{\varepsilon} \left( \int (\bar{g} - g^*) d\nu - \iota^\varepsilon(f^*, \bar{g}) + \iota^\varepsilon(f^*, g^*) \right), \end{aligned}$$

where the last step uses that the second marginal of  $\bar{\gamma}$  is  $\nu$ , by (4.9). We obtain that

$$\begin{aligned} \varepsilon \mathsf{KL}(\bar{\gamma} \parallel \gamma^*) &= \int \bar{g} d\nu + \int f^* d\mu - \iota^\varepsilon(f^*, \bar{g}) \\ &\quad - \left( \int g^* d\nu + \int f^* d\mu - \iota^\varepsilon(f^*, g^*) \right) \\ &\leq 0, \end{aligned}$$

since  $(f^*, g^*)$  are optimal for  $(\varepsilon\text{-D-W}_2^2)$ .

Therefore  $\mathsf{KL}(\bar{\gamma} \parallel \gamma^*) = 0$ , so  $\bar{\gamma} = \gamma^*$ , and  $\bar{g} = g^*$ ,  $\nu$ -almost everywhere. This establishes that the second marginal of  $\gamma^*$  is  $\nu$ , and an analogous argument shows that the first marginal of  $\gamma^*$  is  $\mu$ . We obtain that  $\gamma^*$  is feasible in  $(\varepsilon\text{-W}_2^2)$ .

To conclude, we compute the value that  $\gamma^*$  achieves in the primal problem:

$$\int \|x - y\|^2 \gamma^*(dx, dy) + \varepsilon \mathsf{KL}(\gamma^* \parallel \mu \otimes \nu)$$

$$\begin{aligned}
&= \int \|x - y\|^2 \gamma^*(dx, dy) \\
&\quad + \int (f^*(x) + g^*(y) - \|x - y\|^2) \gamma^*(dx, dy) \\
&= \int (f^*(x) + g^*(y)) \gamma^*(dx, dy) \\
&= \int f^* d\mu + \int g^* d\nu - \iota^\varepsilon(f^*, g^*),
\end{aligned}$$

where the last step uses that  $\gamma^* \in \Gamma_{\mu, \nu}$  and that  $\iota^\varepsilon(f^*, g^*) = 0$  since  $\gamma^*$  is a probability measure. Therefore, the primal objective evaluated at  $\gamma^*$  and the dual objective evaluated at  $(f^*, g^*)$  have the same value, and weak duality (see Section 4.1) shows that  $\gamma^*$  and  $(f^*, g^*)$  are an optimal pair of primal/dual solutions.  $\square$

Proposition 4.3 deserves several remarks. First, note that the hypothesis of the proposition is the existence of optimal solutions to  $(\varepsilon\text{-D-W}_2^2)$ . We do not justify the existence of such solutions here, but it can be shown that if  $\mu$  and  $\nu$  are compactly supported, then there exist  $f^*, g^* \in C_b(\Omega)$ . More generally, if  $\mu$  and  $\nu$  have finite second moments, then optima exist in  $L^1(\mu)$  and  $L^1(\nu)$ , respectively, and Proposition 4.3 continues to hold.

The proof of Proposition 4.3 actually shows that if  $f, g$  are such that  $\gamma(dx, dy) = \exp((f(x) + g(y) - \|x - y\|^2)/\varepsilon) \mu(dx) \nu(dy)$  is a valid coupling between  $\mu$  and  $\nu$ , then  $\gamma$  is optimal for  $(\varepsilon\text{-W}_2^2)$  and  $f, g$  are optimal for  $(\varepsilon\text{-D-W}_2^2)$ . This fact can be viewed as an entropic variant of the complementary slackness condition  $\bar{f}(x) + \bar{g}(y) = \|x - y\|^2$ ,  $\bar{\gamma}$ -almost everywhere, which holds for the optimal solutions of  $(\text{W}_2^2)$  and  $(\text{D-W}_2^2)$ . (See Theorem 1.14.) We can therefore conclude that  $f^*$  and  $g^*$  are optimal for  $(\varepsilon\text{-D-W}_2^2)$  if and only if they satisfy the marginal constraint (4.7) and the analogous constraint for the other marginal:

$$\int \exp\left(\frac{f^*(x) + g^*(y) - \|x - y\|^2}{\varepsilon}\right) \nu(dy) = 1 \quad \mu\text{-a.e.} \quad (4.10)$$

The marginal constraints (4.7) and (4.10), sometimes known as the *Schrödinger system*, are at the core of the theory of entropic OT. Even though these equations *a priori* only hold  $\nu$ - and  $\mu$ -almost everywhere, respectively, the construction in (4.8) shows that we can construct canonical extensions of  $f^*$  and  $g^*$  so that the marginal constraints hold everywhere in  $\mathbb{R}^d$ . Moreover, the dominated convergence theorem

shows that if  $\mu$  and  $\nu$  are compactly supported, then these extensions are continuous (even  $C^\infty$ ) functions on  $\mathbb{R}^d$ . In what follows, we may therefore always assume that  $f^*$  and  $g^*$  are defined everywhere on  $\mathbb{R}^d$ , and that (4.7) and (4.10) hold for all  $y$  and  $x \in \mathbb{R}^d$ , respectively.

Finally, we note that Proposition 4.3 is the basis for the celebrated *Sinkhorn algorithm* for entropic optimal transport. This algorithm is defined by initializing  $f_0 \equiv 0$ , and for  $t \geq 1$  performing the updates

$$g_t(y) = -\varepsilon \log \int \exp\left(\frac{f_{t-1}(x) - \|x - y\|^2}{\varepsilon}\right) \mu(dx), \quad (4.11)$$

$$f_t(y) = -\varepsilon \log \int \exp\left(\frac{g_t(y) - \|x - y\|^2}{\varepsilon}\right) \nu(dy). \quad (4.12)$$

Proposition 4.3 shows that a fixed point of this algorithm yields an optimal solution to  $(\varepsilon\text{-D-W}_2^2)$ , and therefore an optimal solution to  $(\varepsilon\text{-W}_2^2)$ .

### 4.3 Statistical rates for dual solutions

In this and the following section, we consider the statistical behavior of empirical versions of the entropic OT problem. In contrast to the results of Chapter 2, the rates of convergence (as a function of  $n$ ) no longer suffer from the curse of dimensionality. However, the price to pay for this improvement is a steep dependence on  $1/\varepsilon$ .

The strategy for proving statistical bounds is to analyze the dual problem  $(\varepsilon\text{-D-W}_2^2)$ . We then transfer these bounds to the primal solution using the connection between primal and dual solutions given by Proposition 4.3.

Let us denote by  $S(\mu, \nu)$  the value of the primal problem  $(\varepsilon\text{-W}_2^2)$ . Given i.i.d. samples from  $\mu$  and  $\nu$ , we are chiefly interested in estimating two quantities:

- The cost  $S(\mu, \nu)$ ,
- The *entropic map* or *entropic regression function*

$$b^*(x) = \mathbb{E}_{(X, Y) \sim \gamma^*}[Y \mid X = x].$$

Estimating the first quantity is the entropic analogue of the question we considered in Chapter 2. Estimating the second quantity is the entropic analogue of the map estimation task described in Chapter 3. Indeed,  $b^*$  is a projection of  $\gamma^*$ , in the sense of  $L^2$ , onto the space of maps; however,

we stress that  $b^*$  is not a valid transport map between  $\mu$  and  $\nu$ , since  $(b^*)_\# \mu \neq \nu$ .

As in Chapter 2, we analyze *plug-in estimators* for these quantities:  $S(\mu_n, \nu_n)$  for the cost, and  $\hat{b}(x) = \mathbb{E}_{(X,Y) \sim \hat{\gamma}}[Y \mid X = x]$ , where  $\hat{\gamma}$  is the optimal solution to the empirical entropic OT problem between  $\mu_n$  and  $\nu_n$ .<sup>2</sup>

As emphasized above, our main tool to analyze these quantities is the duality relationship established in Proposition 4.3. Denote by  $C_b^\oplus$  the subspace of  $C_b(\Omega \times \Omega)$  consisting of functions of the form  $f \oplus g$  for  $f, g \in C_b(\Omega)$ . The dual problem  $(\varepsilon\text{-D-W}_2^2)$  depends on  $f$  and  $g$  only through their sum  $h = f \oplus g \in C_b^\oplus$ . In particular, if  $(f, g)$  is a dual solution, then so is  $(f + \lambda, g - \lambda)$  for any  $\lambda \in \mathbb{R}$ . Inspired by this fact, let us define the *dual functional*  $\Phi : C_b^\oplus \rightarrow \mathbb{R}$  given by

$$h \mapsto \int (h - \varepsilon(e^{(h-c)/\varepsilon} - 1)) d(\mu \otimes \nu), \quad (4.13)$$

where  $c(x, y) = \|x - y\|^2$  is the squared Euclidean cost. The dual problem can then be written succinctly as  $\sup_{h \in C_b^\oplus} \Phi(h)$ .

Suppose we wish to compare  $S(\mu, \nu)$  to  $S(\mu_n, \nu_n)$ , where, as in Chapter 2,  $\mu_n$  and  $\nu_n$  denote empirical measures corresponding to i.i.d. samples from  $\mu$  and  $\nu$ . We can define an empirical version of the dual functional by

$$\hat{\Phi}(h) = \int (h - \varepsilon(e^{(h-c)/\varepsilon} - 1)) d(\mu_n \otimes \nu_n).$$

Then

$$S(\mu_n, \nu_n) - S(\mu, \nu) = \sup_{h \in C_b^\oplus} \hat{\Phi}(h) - \sup_{h \in C_b^\oplus} \Phi(h) = \hat{\Phi}(\hat{h}) - \Phi(h^*),$$

where  $\hat{h}$  and  $h^*$  are maximizers of  $\hat{\Phi}$  and  $\Phi$ , respectively.

Exercise 7 sketches a direct approach to obtain an upper bound on  $\hat{\Phi}(\hat{h}) - \Phi(h^*)$  based on empirical process theory, analogous to the one developed in Section 2.3 for the unregularized optimal transport problem. However, we pursue a different path, which leverages the strong concavity of the dual functional.

---

<sup>2</sup> Though the formulas for the entropic maps  $b^*$  and  $b_n$  define them as elements of  $L^1(\mu)$  and  $L^1(\mu_n)$ , respectively, the canonical extensions described in Section 4.2 can be used to define continuous versions of  $b^*$  and  $b_n$ .

We begin with a non-rigorous sketch of the argument. Strong concavity of the functional  $\widehat{\Phi}$  should imply there exists  $\delta > 0$  such that

$$\widehat{\Phi}(\hat{h}) \leq \widehat{\Phi}(h^*) + \langle \nabla \widehat{\Phi}(h^*), \hat{h} - h^* \rangle_{L^2(\mu_n \otimes \nu_n)} - \frac{\delta}{2} \|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)}^2.$$

While it is possible to define a suitable notion of gradient for  $\nabla \widehat{\Phi}$ , it is sufficient for our purposes to interpret the above inner product as a directional (Gâteaux) derivative. In contrast to the empirical process theory approach, this inequality implies that we can obtain a bound by controlling  $\widehat{\Phi}$  and  $\nabla \widehat{\Phi}$  at the *fixed* function  $h^*$  rather than the random function  $\hat{h}$ . In particular, there is no need to “sup-out”  $\hat{h}$  which allows us to circumvent the use of empirical process theory.

We make the following assumption.

**Assumption 4.4.** *The supports of  $\mu$  and  $\nu$  lie in  $\Omega \subseteq B_{1/2}(0)$ .*

In particular, under Assumption 4.4,  $\text{diam}(\Omega) \leq 1$ . This assumption implies simple *a priori* bounds on  $\hat{h}$  and  $h^*$ .

**Proposition 4.5.** *Under Assumption 4.4, it holds that*

$$\|\hat{h}\|_{L^\infty}, \|h^*\|_{L^\infty} \leq 2.$$

*Proof.* We first prove the claim for  $h^*$ . Recall that  $h^* = f^* \oplus g^*$  for  $f^*, g^* \in C_b(\Omega)$  and that thanks to canonical extensions, we may assume that the marginal constraints (4.7) and (4.10) hold for all  $x, y \in \Omega$ . Since  $c(x, y) \leq 1$  for all  $x, y \in \Omega$ , we get

$$\begin{aligned} 1 &= \int e^{(f^* \oplus g^* - c)/\varepsilon} \mu(dx) \geq e^{(g^*(y) - 1)/\varepsilon} \int e^{f^*(x)/\varepsilon} \mu(dx), \quad \forall y \in \Omega, \\ 1 &= \int e^{(f^* \oplus g^* - c)/\varepsilon} \nu(dy) \geq e^{(f^*(x) - 1)/\varepsilon} \int e^{g^*(y)/\varepsilon} \nu(dy), \quad \forall x \in \Omega. \end{aligned}$$

Multiplying these two inequalities yields

$$e^{(h^* - 2)/\varepsilon} \int e^{h^*/\varepsilon} d(\mu \otimes \nu) \leq 1.$$

Next note that by Jensen’s inequality, we get

$$\int e^{h^*/\varepsilon} d(\mu \otimes \nu) \geq e^{S(\mu, \nu)/\varepsilon} \geq 1,$$

where we used Proposition 4.3. From the above two displays, we get that  $h^* \leq 2$  for all  $x, y \in \Omega$ .

Next, since  $c \geq 0$  on  $\Omega \times \Omega$ , we get from the same argument that

$$1 \leq e^{h^*/\varepsilon} \int e^{h^*/\varepsilon} d(\mu \otimes \nu) \leq e^{(h^*+2)/\varepsilon},$$

where we used the bound  $h^* \leq 2$  that we just proved. Hence  $h^* \geq -2$  for all  $x, y \in \Omega$ .

Since the only fact that was used about  $h^*$  is that it maximizes the dual functional  $\Phi$  corresponding to measures whose supports lie in  $\Omega$ , the claim also holds for  $\hat{h}$  when replacing  $(\mu, \nu)$  with  $(\mu_n, \nu_n)$  in Proposition 4.3. Again, canonical extensions play a crucial role here.  $\square$

We require some fundamental differentiability and concavity properties of the empirical dual functional. If we let  $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$  and  $\nu_n = \frac{1}{n} \sum_{j=1}^n \delta_{Y_j}$ , for  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mu$  and  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \nu$ , then we can write  $\hat{\Phi}$  explicitly as

$$\hat{\Phi}(h) = \frac{1}{n^2} \sum_{i,j=1}^n \left( h(X_i, Y_j) - \varepsilon \left( e^{(h(X_i, Y_j) - \|X_i - Y_j\|^2)/\varepsilon} - 1 \right) \right). \quad (4.14)$$

Rather than appealing to functional analysis to study differentiability of the functional  $\hat{\Phi}$ , it is sufficient to study the function  $\varphi$  defined on  $[0, 1]$  by

$$\varphi(t) = \hat{\Phi}(h_t), \quad \text{where } h_t := (1-t)\hat{h} + th^*. \quad (4.15)$$

In particular, it is twice differentiable with derivatives given by

$$\varphi'(t) = \frac{1}{n^2} \sum_{i,j=1}^n \left( (h^*(X_i, Y_j) - \hat{h}(X_i, Y_j)) (1 - e^{(h_t(X_i, Y_j) - \|X_i - Y_j\|^2)/\varepsilon}) \right), \quad (4.16)$$

and

$$\begin{aligned} \varphi''(t) &= -\frac{1}{\varepsilon n^2} \sum_{i,j=1}^n \left( (h^*(X_i, Y_j) - \hat{h}(X_i, Y_j))^2 e^{(h_t(X_i, Y_j) - \|X_i - Y_j\|^2)/\varepsilon} \right) \\ &\leq -\frac{e^{-3/\varepsilon}}{\varepsilon} \|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)}^2, \end{aligned} \quad (4.17)$$

where we used Proposition 4.5 and Assumption 4.4 in the above inequality. We readily get that  $\varphi$  is strongly concave on  $[0, 1]$ .

The expression (4.16) reveals that the derivative of  $\varphi$  is an  $L^2(\mu_n \otimes \nu_n)$  inner product with a function in the space  $C_b^\oplus$ . This inner product can be well understood using the Hoeffding (a.k.a. Efron–Stein, a.k.a. ANOVA) decomposition [Hoe48].

**Definition 4.6.** Let  $X, Y$  be two independent random variables with distributions  $P$  and  $Q$  respectively. Given  $k \in L^2(P \otimes Q)$ , the Hoeffding decomposition of  $k(X, Y)$  in  $L^2(P \otimes Q)$  is given by

$$k(X, Y) = \bar{k}_1(X) + \bar{k}_2(Y) + \bar{\bar{k}} + \mathbf{r}(X, Y)$$

where

$$\begin{aligned}\bar{\bar{k}} &= \mathbb{E}[k(X, Y)] \in \mathbb{R}, \\ \bar{k}_1(x) &= \mathbb{E}[k(X, Y) \mid X = x] - \bar{\bar{k}}, \\ \bar{k}_2(y) &= \mathbb{E}[k(X, Y) \mid Y = y] - \bar{\bar{k}},\end{aligned}$$

and

$$\mathbf{r}(X, Y) = k(X, Y) - \bar{k}_1(X) - \bar{k}_2(Y) - \bar{\bar{k}}.$$

It is easy to check (exercise!) that the Hoeffding decomposition is orthogonal in  $L^2(P \otimes Q)$ . In fact, the same calculations reveal the relevance of this decomposition to our problem: for any  $h = f \oplus g \in C_b^\oplus$ , it holds

$$\begin{aligned}\langle h, k \rangle_{L^2(P \otimes Q)} &= \langle f, \bar{k}_1 \rangle_{L^2(P)} + \langle g, \bar{k}_2 \rangle_{L^2(Q)} + \langle h, \bar{\bar{k}} \rangle_{L^2(P \otimes Q)} \\ &= \langle h, \bar{k}_1 + \bar{k}_2 + \bar{\bar{k}} \rangle_{L^2(P \otimes Q)}.\end{aligned}$$

Using Cauchy–Schwarz and orthogonality of the Hoeffding decomposition, we get

$$\begin{aligned}\langle h, k \rangle_{L^2(P \otimes Q)}^2 &\leq \|h\|_{L^2(P \otimes Q)}^2 \|\bar{k}_1 + \bar{k}_2 + \bar{\bar{k}}\|_{L^2(P \otimes Q)}^2 \\ &= \|h\|_{L^2(P \otimes Q)}^2 (\|\bar{k}_1\|_{L^2(P \otimes Q)}^2 + \|\bar{k}_2\|_{L^2(P \otimes Q)}^2 + \bar{\bar{k}}^2).\end{aligned}\tag{4.18}$$

Using orthogonality again implies

$$\begin{aligned}\mathbb{E}[(\mathbb{E}[k(X, Y) \mid X])^2] &= \|\bar{k}_1 + \bar{\bar{k}}\|_{L^2(P \otimes Q)}^2 \\ &= \|\bar{k}_1\|_{L^2(P \otimes Q)}^2 + \bar{\bar{k}}^2\end{aligned}$$

and similarly

$$\mathbb{E}[(\mathbb{E}[k(X, Y) \mid Y])^2] = \|\bar{k}_2\|_{L^2(P \otimes Q)}^2 + \bar{\bar{k}}^2.$$

These two identities together with (4.18) yield

$$\frac{\langle h, k \rangle_{L^2(P \otimes Q)}^2}{\|h\|_{L^2(P \otimes Q)}^2} \leq \mathbb{E}[\mathbb{E}[k(X, Y) | X]^2] + \mathbb{E}[\mathbb{E}[k(X, Y) | Y]^2] - \bar{k}^2.$$

Applying this result with  $P = \mu_n$ ,  $Q = \nu_n$  we get the following lemma.

**Lemma 4.7.** *For any  $k \in L^2(\mu_n \otimes \nu_n)$  and any  $h \in C_b^\oplus$ , we have*

$$\langle h, k \rangle_{L^2(\mu_n \otimes \nu_n)}^2 \leq \|h\|_{L^2(\mu_n \otimes \nu_n)}^2 (\|\nu_n(k)\|_{L^2(\mu_n)}^2 + \|\mu_n(k)\|_{L^2(\nu_n)}^2),$$

where

$$\mu_n(k)(y) = \frac{1}{n} \sum_{i=1}^n k(X_i, y), \quad \nu_n(k)(x) = \frac{1}{n} \sum_{j=1}^n k(x, Y_j).$$

We are now in a position to obtain an important lemma showing that  $\hat{h}$  is a good estimator of  $h^*$ . In turn, rates of convergence for the cost and the entropic map follow from this lemma.

**Lemma 4.8.** *Let Assumption 4.4 hold. Then*

$$\mathbb{E}\|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)}^2 \leq \frac{2\varepsilon^2 e^{10/\varepsilon}}{n}.$$

*Proof.* Since  $\varphi$  is strongly concave, using respectively (4.17), the optimality condition  $\varphi'(0) = 0$ , and (4.16), we get

$$\begin{aligned} \frac{e^{-3/\varepsilon}}{\varepsilon} \|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)}^2 &\leq \varphi'(0) - \varphi'(1) = -\varphi'(1) \\ &= \frac{1}{n^2} \sum_{i,j=1}^n \left( (h^*(X_i, Y_j) - \hat{h}(X_i, Y_j)) (e^{(h^*(X_i, Y_j) - \|X_i - Y_j\|^2)/\varepsilon} - 1) \right) \\ &= \langle h^* - \hat{h}, p^* - 1 \rangle_{L^2(\mu_n \otimes \nu_n)}, \end{aligned} \tag{4.19}$$

where

$$p^*(x, y) = e^{(h^*(x, y) - \|x - y\|^2)/\varepsilon}.$$

Since  $h^* - \hat{h} \in C_b^\oplus$ , Lemma 4.7 implies

$$\langle h^* - \hat{h}, p^* - 1 \rangle_{L^2(\mu_n \otimes \nu_n)} \leq \|h^* - \hat{h}\|_{L^2(\mu_n \otimes \nu_n)} \delta_n,$$

where

$$\delta_n = (\|\nu_n(p^* - 1)\|_{L^2(\mu_n)}^2 + \|\mu_n(p^* - 1)\|_{L^2(\nu_n)}^2)^{1/2}.$$

Combining this with (4.19) yields

$$\|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)}^2 \leq \varepsilon^2 e^{6/\varepsilon} \delta_n^2. \quad (4.20)$$

Recall from (4.7) that

$$\mathbb{E}[p^*(X_i, Y_j) \mid Y_j] = 1,$$

so that

$$\begin{aligned} \mathbb{E}\|\mu_n(p^* - 1)\|_{L^2(\nu_n)}^2 &= \frac{1}{n} \mathbb{E} \sum_{j=1}^n \left( \frac{1}{n} \sum_{i=1}^n p^*(X_i, Y_j) - \mathbb{E}[p^*(X_i, Y_j) \mid Y_j] \right)^2 \\ &= \frac{1}{n} \mathbb{E} \text{var}(p^*(X_1, Y_1) \mid Y_1) \\ &\leq \frac{\mathbb{E}[p^*(X_1, Y_1)^2]}{n} \leq \frac{e^{4/\varepsilon}}{n}. \end{aligned}$$

An analogous bound holds for  $\mathbb{E}\|\nu_n(p^* - 1)\|_{L^2(\mu_n)}^2$ , which implies that

$$\mathbb{E}\delta_n^2 \leq \frac{2e^{4/\varepsilon}}{n}, \quad (4.21)$$

proving the claim.  $\square$

#### 4.4 Statistical rates for primal solutions

Lemma 4.8 shows that solutions to the dual problem ( $\varepsilon$ -D-W<sub>2</sub><sup>2</sup>) converge at the parametric rate. In this section, we use this result to give bounds for the primal problem as well.

We now turn to our first quantity of interest, the cost  $S(\mu, \nu)$ . The following result shows that the mean squared error and bias of the estimator  $S(\mu_n, \nu_n)$  are both of order  $n^{-1}$ , albeit with constants that scale exponentially in  $1/\varepsilon$ . Strikingly, the  $n^{-1}$  rate we have obtained for the variance and bias is characteristic of *parametric* estimation problems, despite the non-parametric setting. It is of course to be contrasted with the slow, non-parametric rates of Chapter 2.

**Theorem 4.9.** *If  $\mu$  and  $\nu$  satisfy Assumption 4.4, then the mean squared error and bias of  $S(\mu_n, \nu_n)$  satisfy*

$$\mathbb{E}(S(\mu_n, \nu_n) - S(\mu, \nu))^2 \lesssim \frac{1}{n},$$

$$|\mathbb{E}S(\mu_n, \nu_n) - S(\mu, \nu)| \lesssim \frac{1}{n}, \quad (4.22)$$

where the implicit constants depends exponentially on  $1/\varepsilon$ .

*Proof.* We begin with establishing (4.22) by studying the bias. Jensen's inequality implies

$$\mathbb{E}S(\mu_n, \nu_n) = \mathbb{E} \sup_{h \in C_b^\oplus} \widehat{\Phi}(h) \geq \sup_{h \in C_b^\oplus} \Phi(h) = S(\mu, \nu).$$

Hence

$$\begin{aligned} 0 \leq b_n &:= \mathbb{E}S(\mu_n, \nu_n) - S(\mu, \nu) \\ &= \mathbb{E}[\widehat{\Phi}(\hat{h}) - \widehat{\Phi}(h^*)] \\ &= \mathbb{E}[\varphi(0) - \varphi(1)] \leq -\mathbb{E}\varphi'(1), \end{aligned}$$

where we recall that the concave function  $\varphi$  is defined in (4.15). It follows from (4.16) that  $-\varphi'(1)$  is given by

$$\begin{aligned} \frac{1}{n^2} \sum_{i,j=1}^n &\left( (h^*(X_i, Y_j) - \hat{h}(X_i, Y_j)) (e^{(h^*(X_i, Y_j) - \|X_i - Y_j\|^2)/\varepsilon} - 1) \right) \\ &\leq \|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)} \delta_n, \end{aligned}$$

where  $\delta_n$  is defined as in the proof of Lemma 4.8 and we have applied Lemma 4.7. By the Cauchy–Schwarz inequality,

$$\begin{aligned} 0 \leq b_n &\leq \sqrt{\mathbb{E}\|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)}^2} \mathbb{E}\delta_n^2 \\ &\leq \frac{\sqrt{2}\varepsilon e^{5/\varepsilon}}{\sqrt{n}} \frac{\sqrt{2}e^{2/\varepsilon}}{\sqrt{n}} = \frac{2\varepsilon e^{7/\varepsilon}}{n}, \end{aligned}$$

where we used Lemma 4.8 and (4.21).

To prove the bound on the mean squared error, we first use a bias–variance decomposition to write

$$\begin{aligned} \mathbb{E}(S(\mu_n, \nu_n) - S(\mu, \nu))^2 &= \text{var}(S(\mu_n, \nu_n)) + |\mathbb{E}S(\mu_n, \nu_n) - S(\mu, \nu)|^2 \\ &= \text{var}(S(\mu_n, \nu_n)) + O(n^{-2}). \end{aligned}$$

It therefore suffices to show that the variance of  $S(\mu_n, \nu_n)$  is  $O(n^{-1})$ . For this purpose, we employ the *Efron–Stein inequality*. More precisely, [BLM13, Corollary 3.2] is sufficient for our purposes. It says that

if  $f = f(Z_1, \dots, Z_m)$  is a function of independent random variables that satisfies the bounded differences inequality:

$$|f(z_1, \dots, z_m) - f(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_m)| \leq 2c \quad (4.23)$$

for all  $z_1, \dots, z_m, z'_i$  and all  $i \in [m]$ , then  $\text{var}(f) \leq c^2 m$ .

Let us view  $S(\mu_n, \nu_n)$  as a function of the  $m = 2n$  independent random variables  $X_1, \dots, X_n, Y_1, \dots, Y_n$ . Fix  $(X_2, \dots, X_n) = (x_2, \dots, x_n)$  and  $(Y_1, \dots, Y_n) = (y_1, \dots, y_n)$ , and view the dual functional  $\widehat{\Phi} = \widehat{\Phi}_{x_1}$  as a function of the value of  $X_1 = x_1$  alone. Then for any  $x_1 \in \Omega$ , under Assumption 4.4, Proposition 4.5 implies that the maximizer of the dual functional  $\widehat{\Phi}_{x_1}$  over  $C_b^\oplus$  is achieved at an  $h$  satisfying  $\|h\|_\infty \leq 2$ . Therefore

$$\begin{aligned} \left| \sup_{h \in C_b^\oplus} \widehat{\Phi}_{x_1}(h) - \sup_{h' \in C_b^\oplus} \widehat{\Phi}_{x'_1}(h') \right| &\leq \sup_{h \in C_b^\oplus, \|h\|_\infty \leq 2} |\widehat{\Phi}_{x_1}(h) - \widehat{\Phi}_{x'_1}(h)| \\ &\leq \frac{2\varepsilon e^{2/\varepsilon} + 4}{n} =: 2c, \end{aligned}$$

where we have used the fact that each term in (4.14) is bounded. Repeating this argument for  $X_2, \dots, X_n, Y_1, \dots, Y_n$ , we obtain that  $S(\mu_n, \nu_n)$  satisfies (4.23) with

$$c = \frac{\varepsilon e^{2/\varepsilon} + 2}{n} \leq \frac{2\varepsilon e^{2/\varepsilon}}{n},$$

since  $\varepsilon e^{2/\varepsilon} > 2$  for all  $\varepsilon > 0$ .

Applying the Efron–Stein inequality, we obtain

$$\text{var}(S(\mu_n, \nu_n)) \leq 2c^2 n \leq \frac{8\varepsilon^2 e^{4/\varepsilon}}{n},$$

as claimed.  $\square$

We now conclude with an analogous sample complexity result for the entropic map  $b^*$ .

Recall from (4.6) that the density of  $\gamma^*$  with respect to  $\mu \otimes \nu$  is

$$p^* = e^{(h^* - c)/\varepsilon}.$$

Similarly, let  $\hat{p} = e^{(\hat{h} - c)/\varepsilon}$  denote the density of  $\hat{\gamma}$  with respect to  $\mu_n \otimes \nu_n$ . Note that thanks to canonical extensions, these two functions may be defined on the whole space.

In the sequel, we use the fact that these densities are uniformly bounded. Indeed, from Proposition 4.5 and Assumption 4.4,

$$\|\hat{p}\|_{L^\infty}, \|p^*\|_{L^\infty} \leq e^{2/\varepsilon}. \quad (4.24)$$

With these definitions, we have the identities

$$\begin{aligned} b^*(x) &= \int y p^*(x, y) \nu(dy), \\ \hat{b}(x) &= \int y \hat{p}(x, y) \nu_n(dy). \end{aligned}$$

The following bound holds.

**Theorem 4.10.** *Adopt Assumption 4.4. The empirical entropic map satisfies*

$$\mathbb{E} \|b^* - \hat{b}\|_{L^2(\mu_n)}^2 \lesssim \frac{1}{n},$$

where the implicit constant depends exponentially on  $1/\varepsilon$ .

*Proof.* Fix  $x \in \Omega$ . Young's inequality and Jensen's inequality imply

$$\begin{aligned} &\|b^*(x) - \hat{b}(x)\|^2 \\ &\leq 2 \left\| \int y p^*(x, y) (\nu - \nu_n)(dy) \right\|^2 + 2 \left\| \int y (p^* - \hat{p})(x, y) \nu_n(dy) \right\|^2 \\ &\leq 2 \left\| \int y p^*(x, y) (\nu - \nu_n)(dy) \right\|^2 + 2 \int \|y\|^2 |(p^* - \hat{p})(x, y)|^2 \nu_n(dy) \\ &\leq 2 \left\| \int y p^*(x, y) (\nu - \nu_n)(dy) \right\|^2 + 2 \|p^*(x, \cdot) - \hat{p}(x, \cdot)\|_{L^2(\nu_n)}^2, \end{aligned}$$

where in the last inequality we use the fact that  $\|y\| \leq 1$  on the support of  $\nu_n$ , by Assumption 4.4. We therefore obtain

$$\begin{aligned} &\|b^* - \hat{b}\|_{L^2(\mu_n)}^2 \\ &\leq \frac{2}{n} \sum_{i=1}^n \left\| \int y p^*(X_i, y) (\nu - \nu_n)(dy) \right\|^2 + 2 \|p^* - \hat{p}\|_{L^2(\mu_n \otimes \nu_n)}^2. \end{aligned}$$

To control the first term, observe that

$$\begin{aligned} &\mathbb{E} \left[ \left\| \int y p^*(X_i, y) (\nu - \nu_n)(dy) \right\|^2 \mid X_i \right] \\ &= \frac{1}{n} \mathbb{E} \left[ \left\| Y_1 p^*(X_i, Y_1) - \mathbb{E}[Y_1 p^*(X_i, Y_1)] \right\|^2 \mid X_i \right] \leq \frac{e^{4/\varepsilon}}{n}, \end{aligned}$$

where we used Assumption 4.4 and (4.24).

To control the second term, we use the fact that the exponential function  $e^x$  is  $e^M$ -Lipschitz on  $(-\infty, M]$  for any  $M$ . Hence, using Assumption 4.4 and Proposition 4.5, we get also that

$$|p^*(x, y) - \hat{p}(x, y)| \leq e^{2/\varepsilon} |h^*(x, y) - \hat{h}(x, y)| \quad \forall x, y \in \Omega.$$

Therefore

$$\mathbb{E}\|p^* - \hat{p}\|_{L^2(\mu_n \otimes \nu_n)}^2 \leq e^{4/\varepsilon} \mathbb{E}\|h^* - \hat{h}\|_{L^2(\mu_n \otimes \nu_n)}^2 \leq \frac{\varepsilon^2 e^{14/\varepsilon}}{n},$$

by Lemma 4.8. We have proved that

$$\mathbb{E}\|b^* - \hat{b}\|_{L^2(\mu_n)}^2 \leq \frac{2e^{4/\varepsilon}}{n} + \frac{2\varepsilon^2 e^{14/\varepsilon}}{n} \lesssim \frac{1}{n}.$$

□

The preceding theorem gives a bound on the empirical entropic map in expected  $L^2(\mu_n)$  norm. At the price of a larger constant factor, it is also possible to obtain a bound in  $L^2(\mu)$ .

**Theorem 4.11.** *Adopt Assumption 4.4. The empirical entropic map satisfies*

$$\mathbb{E}\|b^* - \hat{b}\|_{L^2(\mu)}^2 \lesssim \frac{1}{n},$$

where the implicit constant depends exponentially on  $1/\varepsilon$ .

*Proof.* As in the proof of Theorem 4.10, we have the pointwise bound

$$\begin{aligned} & \|b^*(x) - \hat{b}(x)\|^2 \\ & \lesssim \left\| \int y p^*(x, y) (\nu - \nu_n)(dy) \right\|^2 + \|h^*(x, \cdot) - \hat{h}(x, \cdot)\|_{L^2(\nu_n)}^2. \end{aligned}$$

Integrating with respect to  $\mu$  and taking expectation, we obtain

$$\mathbb{E}\|b^* - \hat{b}\|_{L^2(\mu)}^2 \lesssim \frac{1}{n} + \mathbb{E}\|h^* - \hat{h}\|_{L^2(\mu \otimes \nu_n)}^2.$$

It is thus sufficient to establish that

$$\mathbb{E}\|h^* - \hat{h}\|_{L^2(\mu \otimes \nu_n)}^2 \lesssim \frac{1}{n}. \quad (4.25)$$

To that end, we use the fact that the logarithm and exponential functions are locally Lipschitz, with Lipschitz constant depending on the magnitude of the arguments. In particular, we recall the elementary inequalities

$$e^{\min\{a,b\}} |a - b| \leq |e^a - e^b| \leq e^{\max\{a,b\}} |a - b|,$$

which also imply the bound  $|\log a - \log b| \leq \frac{|a-b|}{\min\{a,b\}}$  for  $a, b > 0$ .

Recall that

$$\hat{h}(x, y) = \hat{f}(x) + \hat{g}(y) = -\varepsilon \log \int e^{(\hat{g}(y) - \|x-y\|^2)/\varepsilon} \nu_n(dy) + \hat{g}(y),$$

and

$$\begin{aligned} h^*(x, y) &= f^*(x) + g^*(y) = -\varepsilon \log \int e^{(g^*(y) - \|x-y\|^2)/\varepsilon} \nu(dy) + g^*(y) \\ &= -\varepsilon \log \int e^{(g^*(y) - \|x-y\|^2)/\varepsilon} \nu_n(dy) + g^*(y) + \Delta(x), \end{aligned}$$

where

$$\Delta(x) = \varepsilon \log \int e^{(g^*(y) - \|x-y\|^2)/\varepsilon} \nu_n(dy) - \varepsilon \log \int e^{(g^*(y) - \|x-y\|^2)/\varepsilon} \nu(dy).$$

Moreover, we may assume that  $\int \hat{g} d\nu_n = \int g^* d\nu_n$  without loss of generality since dual solutions are defined up to an additive constant. Using the Lipschitz properties listed above together with Assumption 4.4 and Proposition 4.5, we get

$$\begin{aligned} |\hat{h}(x, y) - h^*(x, y)| &\lesssim \int |\hat{g} - g^*| d\nu_n + |\hat{g}(y) - g^*(y)| \\ &\quad + \left| \int e^{(g^*(y) - \|x-y\|^2)/\varepsilon} (\nu_n - \nu)(dy) \right|. \end{aligned}$$

Using Jensen's inequality and a trivial variance bound for the average of independent and bounded random variables, we finally obtain

$$\mathbb{E} \|\hat{h} - h^*\|_{L^2(\mu \otimes \nu_n)}^2 \lesssim \mathbb{E} \|\hat{g} - g^*\|_{L^2(\nu_n)}^2 + \frac{1}{n}.$$

Finally, note that

$$\|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)}^2 = \int [(\hat{f} - f^*) \oplus (\hat{g} - g^*)]^2 d(\mu_n \otimes \nu_n)$$

$$\begin{aligned}
&= \|\hat{f} - f^*\|_{L^2(\mu_n)}^2 + \|\hat{g} - g^*\|_{L^2(\nu_n)}^2 \\
&\quad + 2 \int (\hat{f} - f^*) d\mu_n \int (\hat{g} - g^*) d\nu_n \\
&\geq \|\hat{g} - g^*\|_{L^2(\nu_n)}^2,
\end{aligned}$$

where in the last inequality we used the fact that  $\int \hat{g} d\nu_n = \int g^* d\nu_n$ .

Hence we have proved that

$$\mathbb{E}\|\hat{h} - h^*\|_{L^2(\mu \otimes \nu_n)}^2 \lesssim \mathbb{E}\|\hat{h} - h^*\|_{L^2(\mu_n \otimes \nu_n)}^2 + \frac{1}{n}.$$

Together with Lemma 4.8, it completes the proof of (4.25), and hence of the theorem.  $\square$

The conclusion of this section is quite striking: non-parametric quantities can be estimated at a parametric rate. An inspection of the proofs of these results indicates that strong convexity is key to achieve such a result. In retrospect it is not surprising that the empirical risk minimizer of a strongly convex functional should enjoy such dimension-free rates. Indeed, stochastic gradient descent on such an objective does (see, e.g., [KNS16, Theorem 4]). This phenomenon is not new: it is known for specific losses such as the ones employed in Chapter 8 and was observed in [EHL18] for example.

## 4.5 Discussion

**§4.1.** Entropic optimal transport was first popularized for computational purposes in [Cut13]. See [PC19b] for an introduction to computational optimal transport which nicely complements our treatment of statistical optimal transport. Entropic optimal transport between Gaussians (Exercise 4) was computed in [JMP20, MGM22].

Besides statistical applications, entropic optimal transport has also been used to establish mathematical results for unregularized optimal transport [FGP20, GLRT20, CP23].

**§4.2.** The convergence of Sinkhorn's algorithm is discussed in many places, see [KLRS08, ANWR17, DGK18, DBTHD21, Lég21, AFKL22, GN22, BB23, GNCD23, CDV24].

**§4.3.** The proofs in this section are based on [RS22]. Note that the bounds obtained here are *dimension-free*, but scale exponentially w.r.t.

$1/\varepsilon$ . The sample complexity of entropic optimal transport was first established by [GCB<sup>+</sup>19], who proved bounds that scaled as  $e^{O(1/\varepsilon)}\varepsilon^{-O(d)}$ . Later, [MNW19] observed that it was possible to slightly modify their argument to remove the exponential factor. These bounds can be better in low dimension, but provide poor control when the dimension is large.

Recent works have also focused on obtaining bounds which instead scale as  $\varepsilon^{-O(d^*)}$ , where the parameter  $d^*$  denotes the intrinsic dimensionality of the measures (in fact, the minimum intrinsic dimension among  $\mu$  and  $\nu$ ). See Exercise 8 and [Str23] for an approach close to the one taken here, and [GH23] for an empirical process argument.

The techniques used in this section can be used not only to prove sample complexity bounds, but also to obtain distributional limits [MNW19, dBSLNW23, GSLNW24]. Such bounds were originally obtained in the discrete case by [BCP19a, KTM20].

One notable quirk about entropic optimal transport is that in general,  $S(\mu, \mu) > 0$  due to the presence of the entropic term in the objective. In light of this, Genevay et al. [GPC18] proposed the “debiased” quantity  $D(\mu, \nu) := S(\mu, \nu) - \frac{1}{2}(S(\mu, \mu) + S(\nu, \nu))$ , called the *Sinkhorn divergence* between  $\mu$  and  $\nu$ . It can be shown that the Sinkhorn divergence is convex in each of its variables, non-negative, and vanishes if and only if  $\mu = \nu$  [FSV<sup>+</sup>19]. Like entropic optimal transport, the Sinkhorn divergence can be estimated at a parametric rate [GCB<sup>+</sup>19, dBSLNW23]. The Sinkhorn divergence has been advocated as tool for estimating Wasserstein distances [CRL<sup>+</sup>20], although there are caveats when using it for map estimation [PCNW22].

**§4.4.** Theorem 4.11 provides a rate for estimating the entropic map  $b^*$ , but combined with an approximation result quantifying the distance between  $b^*$  and the true optimal transport map  $T$ , one can use  $\hat{b}$  as a computationally efficient estimator for  $T$  (c.f. [PNW22]). Although it is not minimax in general, it is in the semi-discrete case [PDNW23].

As the alternative nomenclature “entropic regression function” indicates, the entropic map also solves a regression problem with respect to the entropic coupling  $\gamma^*$ ; indeed,  $b^* = \arg \min_{f: \mathbb{R}^d \rightarrow \mathbb{R}^d} \mathbb{E}_{\gamma^*} \|Y - f(X)\|^2$ . Analyzing this regression problem when the minimization is taken over a smaller class of candidate regression functions rather than all maps from  $\mathbb{R}^d \rightarrow \mathbb{R}^d$  is an open problem.

## 4.6 Exercises

1. a) Let  $\Omega$  be a compact subset of  $\mathbb{R}^d$  with positive Lebesgue measure. Show that the uniform measure on  $\Omega$  has the largest differential entropy of any probability measure on  $\Omega$ .  
b) Let  $m$  be a positive integer. Show that the uniform measure on  $[m]$  has the largest Shannon entropy of any probability measure on  $[m]$ .
2. a) Show that if  $\mu$ ,  $\nu$ , and  $\gamma$  are absolutely continuous, and  $\gamma \in \Gamma_{\mu,\nu}$ , then  $\text{KL}(\gamma \parallel \mu \otimes \nu) = \text{Ent}(\mu) + \text{Ent}(\nu) - \text{Ent}(\gamma)$ . Conclude that if  $\mu$  and  $\nu$  are absolutely continuous, then the optimization problem (4.1) is equivalent to (4.3).  
b) Show by an analogous calculation that if  $\mu$  and  $\nu$  are discrete, then (4.2) is equivalent to (4.3).
3. Let  $\gamma_\varepsilon$  denote the entropic optimal transport plan between  $\mu$  and  $\nu$ , with corresponding potentials  $f_\varepsilon$ ,  $g_\varepsilon$ . Define  $\varphi_\varepsilon := \frac{1}{2} \|\cdot\|^2 - f_\varepsilon$ . Compute the derivatives of  $\varphi_\varepsilon$  and conclude that

$$\nabla \varphi_\varepsilon(x) = \mathbb{E}_{\gamma_\varepsilon}[Y \mid X = x], \quad (4.26)$$

$$\nabla^2 \varphi_\varepsilon(x) = \varepsilon^{-1} \text{cov}_{\gamma_\varepsilon}(Y \mid X = x). \quad (4.27)$$

In particular, since we expect that  $\varphi_\varepsilon$  converges to the unregularized Brenier potential  $\varphi$  as  $\varepsilon \rightarrow 0$  (proven rigorously in [NW22]), and  $\varphi_\varepsilon$  is convex by (4.27), this gives another explanation for Brenier's Theorem 1.16.

4. Compute the entropic optimal transport solution (i.e., the potentials, the plan, the cost) between two Gaussians. *Hint:* as you might expect, the entropic potentials are quadratic functions.
5. In this exercise, we present another view on the Sinkhorn iterations (4.11), (4.12). Consider the joint distributions

$$\begin{aligned} \gamma_{t-\frac{1}{2}}(dx, dy) &\propto \exp((f_{t-1}(x) + g_t(y) - \|x - y\|^2)/\varepsilon) \mu(dx) \nu(dy), \\ \gamma_t(dx, dy) &\propto \exp((f_t(x) + g_t(y) - \|x - y\|^2)/\varepsilon) \mu(dx) \nu(dy). \end{aligned}$$

Here, we take  $\gamma_0(dx, dy) \propto \exp(-\|x - y\|^2/\varepsilon) \mu(dx) \nu(dy)$ . Show that the Sinkhorn updates correspond to iteratively “fixing the marginals”; i.e.,  $\gamma_{t-\frac{1}{2}}$  is obtained from  $\gamma_{t-1}$  by keeping the conditional distribution of  $X \mid Y$  fixed but setting the  $Y$ -marginal to  $\nu$ , and  $\gamma_t$  is obtained from  $\gamma_{t-\frac{1}{2}}$  by keeping the conditional distribution of  $Y \mid X$  fixed but setting the  $X$ -marginal to  $\mu$ .

6. In the discrete setting where  $\mu, \nu$  are finitely supported on  $\{x_1, \dots, x_m\}$  and  $\{y_1, \dots, y_n\}$  respectively, Sinkhorn's algorithm shows that there are positive *scalings*  $\kappa \in \mathbb{R}_+^m, \lambda \in \mathbb{R}_+^n$  of the rows and columns of the matrix  $M$  with entries  $M_{i,j} = \exp(-\|x_i - y_j\|^2/\varepsilon)$ , such that the scaled matrix  $\text{diag}(\kappa) M \text{diag}(\lambda)$  has marginals  $\mu$  and  $\nu$  respectively; see [PC19b] for details.

As a special case, suppose that  $\mu, \nu$  are *uniformly* distributed and  $m = n$ . Then, the scaled matrix  $\tilde{M} := \text{diag}(\kappa) M \text{diag}(\lambda)$  has marginals  $\mu$  and  $\nu$  if and only if  $n\tilde{M}$  is doubly stochastic, i.e., it belongs to the Birkhoff polytope (1.2). In this case, prove the existence of these scalings for any matrix  $M$  with positive entries by considering the KL minimization problem

$$\underset{\gamma \in \text{Birk}}{\text{minimize}} \quad \sum_{i,j=1}^n \left( \gamma_{i,j} \log \frac{\gamma_{i,j}}{M_{i,j}} - \gamma_{i,j} + M_{i,j} \right).$$

Namely, show that a solution to this problem exists, and using Lagrange multipliers, show that  $\gamma$  is of the form  $\text{diag}(\kappa) M \text{diag}(\lambda)$  for positive scalings  $\kappa, \lambda$ .

7. The strong convexity arguments in Section 4.3 are designed to avoid the use of empirical process theory. However, this exercise shows how to use empirical process theory to prove sample complexity bounds using techniques analogous to those in Section 2.3. Unlike the approach in Section 4.3, these bounds depend polynomially on  $1/\varepsilon$ , but with exponent scaling with  $d$ . For simplicity, we focus on the *one-sample problem*, and prove bounds on the quantity  $S(\mu_n, \nu) - S(\mu, \nu)$ .

a) Let  $f$  and  $g$  be solutions to  $(\varepsilon\text{-D-W}_2)$  for any pair of measures supported on  $\Omega = B_{1/2}(0)$ . Let  $s$  be a positive integer. Arguing as in Exercise 3, show that there exists a positive constant  $C_s$  such that for all multi-indices  $\alpha = (\alpha_1, \dots, \alpha_d)$  with  $|\alpha| = s$ , we have the bound

$$\sup_{x \in \Omega} |\partial_\alpha f(x)| \leq C_s \varepsilon^{1-s}.$$

Argue that we can assume that  $f(0) = 0$  without loss of generality, and thereby obtain the bound  $\sup_{x \in \Omega} |f(x)| \leq C_0$  for some positive constant  $C_0$ .

b) For  $L > 0$ , define

$$\mathcal{F}_s(L) := \left\{ f : \Omega \rightarrow \mathbb{R}^d : \sum_{k=0}^s \sum_{\alpha:|\alpha|=k} \|\partial_\alpha f\|_{L^\infty(\Omega)} \leq L \right\}.$$

Fix a positive integer  $s$ . Argue that there exists a constant  $C = C(s)$  such that for  $L = C(1 + \varepsilon^{1-s})$ , we have

$$|S(\mu_n, \nu) - S(\mu, \nu)| \leq \sup_{f \in \mathcal{F}_s(L)} \left| \int f \, d\mu_n - \int f \, d\mu \right|.$$

*Hint:* let  $f^*$  and  $g^*$  be solutions to  $(\varepsilon\text{-D-}\mathcal{W}_2^2)$  for  $\mu$  and  $\nu$ , and let  $\hat{f}$  and  $\hat{g}$  be the solutions for  $\mu_n$  and  $\nu$ . Argue that  $h^* = f^* \oplus g^*$  and  $\hat{h} = \hat{f} \oplus \hat{g}$  satisfy

$$S(\mu_n, \nu) - S(\mu, \nu) = \widehat{\Phi}(\hat{h}) - \Phi(h^*) \leq \int \hat{f} \, d\mu_n - \int \hat{f} \, d\mu,$$

and analogously

$$S(\mu, \nu) - S(\mu_n, \nu) \leq \int f^* \, d\mu - \int f^* \, d\mu_n.$$

Then apply part (a).

c) It can be shown that  $\log N(\delta, \mathcal{F}_s(L)) \lesssim (L/\delta)^{d/s}$ , and moreover that this bound holds also for fractional  $s$ , where  $\mathcal{F}_s$  is interpreted as a suitable Hölder space. Taking  $s = d/2+1$ , use Proposition 2.6 to conclude

$$\mathbb{E}|S(\mu_n, \nu) - S(\mu, \nu)| \lesssim (1 + \varepsilon^{-d/2}) n^{-1/2}.$$

8. The statistical results in Section 4.3 rely on pointwise bounds on the density  $p^*$ . Here, we show a different way to control  $\|p^*\|_{L^2(\mu \otimes \nu)}$ , which provides an entry point into [Str23].

- Argue that the dual potentials  $f^*, g^*$  are  $O(1)$ -Lipschitz, and that  $\log p^*$  is  $O(\varepsilon^{-1})$ -Lipschitz. (Here, we are still working over a bounded domain.)
- Prove that for all  $\delta > 0$ ,  $\int \nu(B(z, \delta))^{-1} \nu(dz) \leq N(\delta/4, \text{supp } \nu)$ , where  $N(\delta/4, \text{supp } \nu)$  is the covering number of  $\text{supp } \nu$  at scale  $\delta/4$ . (Let  $z_1, \dots, z_K \in \text{supp } \nu$  be a  $\delta/2$ -covering of  $\text{supp } \nu$  with  $K \leq N(\delta/4, \text{supp } \nu)$ . Bound the integral by summing over the integrals over  $B(z_k, \delta/2)$  for  $k = 1, \dots, K$ .)
- Using the fact that

$$1 = p^*(x, y) \int \frac{p^*(x, y')}{p^*(x, y)} \nu(dy') \geq p^*(x, y) \int_{B(y, r)} \frac{p^*(x, y')}{p^*(x, y)} \nu(dy')$$

and the log-Lipschitz property from (a), show that  $p^*(x, y) \lesssim \nu(B(y, r))^{-1}$ , where  $r \asymp \varepsilon$ .

d) Combining this with the estimate in (b), prove that  $\|p^*\|_{L^2(\mu \otimes \nu)} \lesssim \sqrt{N(r', \text{supp } \nu)}$  where  $r' \asymp \varepsilon$ . Explain why this implies that if  $\text{supp } \mu$  is  $d_\mu$ -dimensional and  $\text{supp } \nu$  is  $d_\nu$ -dimensional, then  $\|p^*\|_{L^2(\mu \otimes \nu)} \lesssim \varepsilon^{-(d_\mu \wedge d_\nu)/2}$ .

## Wasserstein gradient flows: theory

We have seen in Proposition 1.3 that  $\mathcal{P}_2(\mathbb{R}^d)$ , once endowed with the  $W_2$  distance, has the structure of a metric space. It has in fact a much richer geometric structure, as it resembles a Riemannian manifold. Consequently, we can bring to bear the calculation rules of Riemannian geometry, known in this context as *Otto calculus*, on the design and interpretation of algorithms over the space of probability measures.

The identification of  $\mathcal{P}_2(\mathbb{R}^d)$  with a Riemannian manifold is purely “formal” (that is, heuristic). For instance,  $\mathcal{P}_2(\mathbb{R}^d)$  is not locally homeomorphic to a Hilbert space. However, the Riemannian view of  $\mathcal{P}_2(\mathbb{R}^d)$  is nevertheless a powerful tool for understanding the geometric properties of the Wasserstein space.

To elucidate this Riemannian viewpoint, we work with absolutely continuous measures in this chapter, for which the Riemannian formalism can be put on a more rigorous footing. Our main goal in constructing this formalism is to define interesting dynamics on the Wasserstein space, given by *gradient flows*. Having defined these dynamics, we shall see that they often make sense even for non-absolutely continuous measures—in particular, they give rise to well-defined dynamics for discrete measures, viewed as particle systems. Once derived, these non-trivial dynamics can be studied directly for discrete measures without the need for making rigorous sense of the Riemannian calculations in the discrete case. The reader interested in seeing a fully rigorous derivation of gradient flows for general measures should consult the seminal monograph of Ambrosio, Gigli, and Savaré [AGS08], or [San17] for a quick overview.

## 5.1 Metric derivative and the continuity equation

The utility of optimal transport lies in its endowment of the space of probability measures with a geometric structure which respects that of the underlying space. For example, we saw that the mapping  $x \mapsto \delta_x$  is an isometric embedding of  $(\mathbb{R}^d, \|\cdot\|)$  into  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ . Accordingly, as we now seek to understand dynamics on  $\mathcal{P}_2(\mathbb{R}^d)$ , our approach is to “lift” the corresponding dynamics of particles on  $\mathbb{R}^d$ .

The general way to prescribe dynamics on  $\mathbb{R}^d$  using differential calculus is via ordinary differential equations (ODEs). Namely, given a time-dependent family of vector fields  $(v_t)_{t \geq 0}$ , consider the ODE

$$\dot{X}_t = v_t(X_t). \quad (5.1)$$

Under standard assumptions on  $(v_t)_{t \geq 0}$ ,<sup>1</sup> there is a unique solution to the ODE for any given initial condition  $X_0$ . Suppose now that  $X_0$  is drawn *randomly* from a measure  $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ , and similarly let  $\mu_t$  denote the law of  $X_t$  for all  $t \geq 0$ . We think of the curve of measures  $(\mu_t)_{t \geq 0}$  as describing the evolution of a *collection* of particles. Then, the dynamics of  $(\mu_t)_{t \geq 0}$  is described by a partial differential equation (PDE), known as the *continuity equation*.

**Proposition 5.1 (Continuity equation).** *Suppose that  $X_0 \sim \mu_0$ , and that  $(X_t)_{t \geq 0}$  evolves according to the dynamics (5.1), which we assume is well-posed. Let  $\mu_t$  denote the law of  $X_t$  for all  $t \geq 0$ . Then,  $(\mu_t)_{t \geq 0}$  satisfies the following equation in the weak sense,*

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0, \quad (5.2)$$

i.e., for all compactly supported and smooth test functions  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ , it holds that

$$\partial_t \int \varphi \, d\mu_t = \int \langle \nabla \varphi, v_t \rangle \, d\mu_t. \quad (5.3)$$

The equation (5.3), when written in probabilistic language, reads  $\partial_t \mathbb{E} \varphi(X_t) = \mathbb{E} \langle \nabla \varphi(X_t), v_t(X_t) \rangle$ , and it simply follows from (5.1) and the chain rule. The real content of the proposition actually lies in (5.2): when  $\mu_t$  admits a smooth density w.r.t. Lebesgue measure, which by an

---

<sup>1</sup> For example, if the vector fields are Lipschitz uniformly in time, then well-posedness follows from the Cauchy–Lipschitz theorem.

abuse of notation we denote also by  $\mu_t$ , then integration by parts yields the equation

$$\int \varphi \partial_t \mu_t = \partial_t \int \varphi d\mu_t = \int \langle \nabla \varphi, v_t \rangle \mu_t = - \int \varphi \operatorname{div}(\mu_t v_t).$$

Since this equality is supposed to hold for all suitable test functions  $\varphi$ , it follows that (5.2) holds pointwise. To summarize, we see that  $(\mu_t)_{t \geq 0}$  solves the PDE (5.2), at least when  $\mu_t$  admits a smooth density for all  $t \geq 0$ . In general, it is more convenient to make statements that hold for curves  $(\mu_t)_{t \geq 0}$  without knowing in advance the regularity of  $\mu_t$ , in which case (5.2) should be interpreted to hold in the weak sense (5.3). However, for the sake of developing the framework of Otto calculus unencumbered by technical distractions, from now on we ignore such issues of regularity and pretend that we are working with curves of smooth densities. See [AGS08] for a more rigorous treatment.

The equations (5.1) and (5.2) provide us with dual perspectives on the same dynamics; in the field of fluid dynamics, these perspectives are known as *Lagrangian* and *Eulerian* respectively. The Lagrangian perspective describes the evolution of individual particle trajectories, whereas the Eulerian perspective tracks the evolution of aggregate quantities through the notions of mass density  $\mu_t$  and velocity field  $v_t$ .

To foreshadow the development of geometry over  $\mathcal{P}_2(\mathbb{R}^d)$ , let us first examine how to develop geometry over  $\mathbb{R}^d$ ; for now, we refer to concepts from Riemannian geometry loosely, but we return to the subject in Section 5.2. For a single particle trajectory  $t \mapsto X_t$  evolving according to (5.1), the kinetic energy at time  $t$  (assuming the particle has unit mass) is  $\|\dot{X}_t\|^2 = \|v_t(X_t)\|^2$ . The total energy of the curve over the time interval  $[0, 1]$  is  $\int_0^1 \|v_t(X_t)\|^2 dt$ , and if we minimize this energy over all curves  $(X_t)_{t \in [0, 1]}$  evolving according to (5.1) with endpoints fixed at  $X_0$  and  $X_1$ , we obtain the constant-speed geodesic  $t \mapsto X_t := (1 - t)X_0 + tX_1$ . Geometrically, we think of  $v_t(X_t)$  as the *tangent vector* to the curve  $(X_t)_{t \in [0, 1]}$  at time  $t$ , and we measure its length using the Euclidean norm  $\|\cdot\|$ .

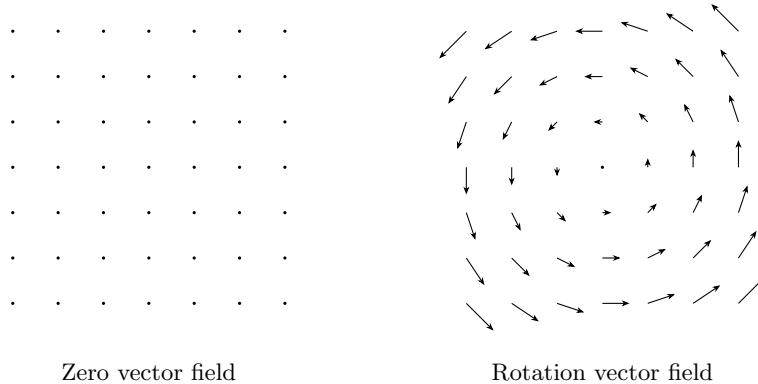
We now try to lift this picture to  $\mathcal{P}_2(\mathbb{R}^d)$ . For a curve  $(\mu_t)_{t \geq 0}$  evolving according to (5.2), it is natural to think of the velocity vector field  $v_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$  as an abstract “tangent vector” to  $(\mu_t)_{t \geq 0}$  at time  $t$ , and to measure its squared “length” via the kinetic energy<sup>2</sup>

---

<sup>2</sup> Since  $\mu_t$  plays the role of a mass density, then  $\|v_t\|^2 \mu_t$  is the kinetic energy density, and integrating this over all of space yields the kinetic energy.

$$\|v_t\|_{\mu_t}^2 := \int \|v_t\|^2 d\mu_t. \quad (5.4)$$

However, we immediately arrive at an obstacle in doing so: given a curve  $(\mu_t)_{t \geq 0}$ , there is not a *unique* choice of vector fields  $(v_t)_{t \geq 0}$  for which (5.2) holds, and hence it is unclear what vector field  $v_t$  to choose as our tangent vector. Indeed, if we start with any family of vector fields  $(v_t)_{t \geq 0}$  for which (5.2) holds, and if  $w_t$  satisfies  $\operatorname{div}(\mu_t w_t) = 0$  for all  $t \geq 0$ , then  $(v_t + w_t)_{t \geq 0}$  is another family of vector fields for which (5.2) holds by linearity of the divergence operator. To see a concrete example of non-uniqueness, suppose that  $\mu_t$  is the standard Gaussian distribution on  $\mathbb{R}^2$  for all  $t \geq 0$ . Then, one natural choice of vector fields is to take  $v_t = 0$  for all  $t \geq 0$ ; however, due to the rotational invariance of the standard Gaussian, another choice is to choose vector fields inducing a rotation (see Figure 5.1).



**Fig. 5.1.** Two vector fields which preserve the standard Gaussian on  $\mathbb{R}^2$ .

We see that the vector field on the right of Figure 5.1 induces extraneous motion for the particles and is therefore not the most parsimonious explanation for the dynamics of  $(\mu_t)_{t \geq 0}$ . To resolve the ambiguity in the choice of vector fields, we can elect to declare as our tangent vector the vector field which minimizes the kinetic energy (5.4) while still explaining the dynamics of  $(\mu_t)_{t \geq 0}$ . As discussed below, the minimization of kinetic energy falls naturally in line with the philosophy of “optimal” transport of mass.

To further motivate this choice, we introduce the notion of the metric derivative of a curve  $(x_t)_{t \geq 0}$  in a metric space  $(S, d)$ . Although in general we cannot make sense of the notion of a tangent vector (or

the “derivative”) of a curve in a general metric space, it turns out that we can make sense of its *speed*.

**Definition 5.2 (Metric derivative).** *Let  $(S, d)$  be a metric space and let  $(x_t)_{t \geq 0}$  be a curve in  $S$ . Then, the metric derivative of the curve at time  $t$  is given by*

$$|\dot{x}|(t) := \lim_{s \rightarrow t, s \neq t} \frac{d(x_s, x_t)}{|s - t|},$$

*provided that the limit exists.*

The next theorem shows that our selection principle produces a well-defined choice of tangent vector  $v_t$  which can be characterized in any one of three ways: (1) as the vector field  $v_t$  with minimal kinetic energy subject to the constraint (5.2); (2) as the unique choice of vector field solving (5.2) with length  $\|v_t\|_{\mu_t}$  equal to the metric derivative  $|\dot{\mu}|(t)$ ; (3) as a limit of Brenier maps.

However, let us first introduce the concept of the *flow map* associated with the ODE (5.1) (or equivalently, with the family of vector fields  $(v_t)_{t \geq 0}$ ). The map  $F_{0,t} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is defined as the map which, given  $X_0 \in \mathbb{R}^d$ , outputs the solution  $X_t$  to the ODE (5.1) at time  $t$  when started at  $X_0$ . In an analogous manner, we can define the flow map  $F_{s,t} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  for any pair of times  $0 \leq s \leq t$ . The significance of this definition is that it shifts our attention away from thinking about the ODE as describing a single trajectory, and instead views the effect of the ODE as a deformation of the entire space  $\mathbb{R}^d$ .

Given a pair of probability measures  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  such that  $\mu$  admits a density, we write  $T_{\mu \rightarrow \nu}$  for the Brenier map from  $\mu$  to  $\nu$ , and denote by  $|\dot{\mu}|$  the metric derivative of a curve in  $\mathcal{P}_2(\mathbb{R}^d)$  with respect to the metric  $d = W_2$ .

**Theorem 5.3.** *Let  $(\mu_t)_{t \geq 0}$  be a regular<sup>3</sup> curve of probability measures. Then, for every family of vector fields  $(v_t)_{t \geq 0}$  for which (5.2) holds, we have  $|\dot{\mu}|(t) \leq \|v_t\|_{\mu_t}$  for all  $t \geq 0$ .*

*Conversely, there exists a unique family  $(v_t)_{t \geq 0}$  such that (5.2) holds and for which  $|\dot{\mu}|(t) = \|v_t\|_{\mu_t}$  for every  $t \geq 0$ . This family is such that*

---

<sup>3</sup> Here, “regular” can be taken to mean that  $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$  and admits a density, and that the metric derivative  $|\dot{\mu}|(t)$  exists for all  $t \geq 0$ . The qualifier “for all  $t \geq 0$ ” in the assumptions and conclusions can be replaced by “for almost every  $t \geq 0$ ”, and the assumed existence of a density can also be relaxed.

$$v_t = \lim_{h \searrow 0} \frac{T_{\mu_t \rightarrow \mu_{t+h}} - \text{id}}{h} \quad \text{in } L^2(\mu_t), \quad (5.5)$$

where  $\text{id}$  is the identity function.

*Proof sketch.* We start with the first statement. Let  $(X_t)_{t \geq 0}$  be the curve of random variables with  $X_0 \sim \mu_0$  solving  $\dot{X}_t = v_t(X_t)$ . Since the continuity equation (5.2) holds by assumption, then  $X_t \sim \mu_t$  for all  $t \geq 0$ , and in particular,  $\mu_{t+h} = (F_{t,t+h})_\# \mu_t$ . We can upper bound  $W_2(\mu_t, \mu_{t+h})$  using this suboptimal coupling:

$$\frac{W_2^2(\mu_t, \mu_{t+h})}{h^2} \leq \mathbb{E} \left[ \frac{\|F_{t,t+h}(X_t) - X_t\|^2}{h^2} \right].$$

Observe that  $F_{t,t+h}(X_t) = X_t + h v_t(X_t) + o(h)$  so that letting  $h \rightarrow 0$ , we obtain  $|\dot{\mu}|(t) \leq \sqrt{\mathbb{E}[\|v_t(X_t)\|^2]} = \|v_t\|_{\mu_t}$ .

Note that the inequality above arises from the use of a suboptimal coupling. Intuitively, if we take  $X_t$  and  $X_{t+h}$  to be optimally coupled and thereby define  $v_t$  according to (5.5), then we ought to obtain an equality. This is in fact the case but we omit the proof.

Finally, by combining the two statements, we obtain that if  $(v_t)_{t \geq 0}$  is such that (5.2) holds and  $|\dot{\mu}|(t) = \|v_t\|_{\mu_t}$ , then

$$0 \in \arg \min_{w_t: \mathbb{R}^d \rightarrow \mathbb{R}^d} \|v_t + w_t\|_{\mu_t} \quad \text{s.t.} \quad \text{div}(\mu_t w_t) = 0.$$

Moreover, this is a strictly convex problem, so the minimizer is unique. We obtain that if  $(v_t)_{t \geq 0}$  and  $(u_t)_{t \geq 0}$  are two vector fields satisfying the continuity equation and  $\|v_t\|_{\mu_t} = \|u_t\|_{\mu_t} = |\dot{\mu}|(t)$ , then  $w_t = u_t - v_t = 0$ , as claimed.  $\square$

In the above theorem, we used the fact that  $\mu_t$  admits a density in order to write (5.5), i.e., to assert that the optimal transport map exists. In order to facilitate the discussion, we restrict to this class of measures from now on, although we return to the subject of particle methods at the end of the chapter.

**Definition 5.4.**  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  is the class of probability measures over  $\mathbb{R}^d$  with finite second moment and which are absolutely continuous (i.e., admit a density w.r.t. Lebesgue measure).

## 5.2 Elements of Riemannian geometry

Before proceeding further, we provide a brief and informal exposition to the concepts from Riemannian geometry that we need.

A *manifold*  $\mathcal{M}$  is a set which is locally homeomorphic to a Euclidean space  $\mathbb{R}^d$ . At each point  $p \in \mathcal{M}$ , we can associate a *tangent space*  $T_p\mathcal{M}$ , which is a  $d$ -dimensional vector space and represents all possible velocity vectors of curves passing through  $p$ . A *Riemannian metric* is a choice of an inner product  $\mathfrak{g}_p$  on each tangent space  $T_p\mathcal{M}$ , and once endowed with a Riemannian metric, the manifold is then called a *Riemannian manifold*. To emphasize the Hilbertian structure, we usually simply write  $\langle \cdot, \cdot \rangle_p$  for the metric  $\mathfrak{g}_p$ . Usually, one imposes additional smoothness assumptions for these objects in order to properly build up a theory of differential calculus, but here we focus on introducing the basic language without delving into details. In the case of the Wasserstein space, note that we have already identified a natural norm for a “tangent vector” (velocity vector field)  $v_t$  at  $\mu_t$ —the  $L^2$  norm,  $\sqrt{\mathfrak{g}_{\mu_t}(v_t, v_t)} = \|v_t\|_{\mu_t}$ —indicating the possibility of identifying further Wasserstein analogues of Riemannian theory.

The next important concept is that of a geodesic. For a curve  $(p_t)_{t \in [0,1]}$  in  $\mathcal{M}$ , let  $\dot{p}_t \in T_{p_t}\mathcal{M}$  denote the tangent vector at time  $t$ . Given  $p_0, p_1 \in \mathcal{M}$ , *geodesics* or shortest paths<sup>4</sup> between  $p_0$  and  $p_1$  are obtained by solving either of the following variational problems,

$$\min_{(p_t)_{t \in [0,1]}} \int \|\dot{p}_t\|_{p_t} dt \quad \text{or} \quad \min_{(p_t)_{t \in [0,1]}} \int \|\dot{p}_t\|_{p_t}^2 dt$$

over curves  $(p_t)_{t \in [0,1]}$  joining  $p_0$  to  $p_1$ . In the first problem, the objective functional is the *arc length* of the curve; in the second problem, the objective functional is called the *energy*. The second variational problem is technically more convenient. Indeed, the arc length is invariant under reparametrization (i.e., replacing  $(p_t)_{t \in [0,1]}$  by  $(p_{f(t)})_{t \in [0,1]}$  for any continuous and strictly increasing function  $f : [0, 1] \rightarrow [0, 1]$ ), so solutions to the first variational problem can only be unique up to reparametrization. In contrast, the second variational problem singles out a specific parametrization of the optimal curves, namely, curves with *constant speed* (i.e.,  $t \mapsto \|\dot{p}_t\|_{p_t}$  is constant). Henceforth, we only

---

<sup>4</sup> In Riemannian geometry, it is more customary to define geodesics to only be *locally* length-minimizing, but here we always use the word “geodesic” to refer to shortest paths.

consider constant-speed geodesics and therefore omit the adjective “constant-speed”.

The values of the variational problems are  $\mathbf{d}(p_0, p_1)$  and  $\mathbf{d}^2(p_0, p_1)$  respectively, where  $\mathbf{d}$  defines a metric (in the sense of metric spaces) on  $\mathcal{M}$  induced by the Riemannian metric  $\langle \cdot, \cdot \rangle$ .

The *exponential map*<sup>5</sup> is a mapping  $\exp_p : T_p \mathcal{M} \rightarrow \mathcal{M}$  which maps a tangent vector  $v$  to  $p_1$ , where  $(p_t)_{t \in [0,1]}$  is the constant-speed geodesic such that  $p_0 = p$  and  $\dot{p}_0 = v$ . The inverse map is called the *logarithmic map*  $\log_p : \mathcal{M} \rightarrow T_p \mathcal{M}$ , which maps  $q \mapsto \exp_p^{-1}(q)$ . Actually, in general, the exponential map may not be defined over all of  $T_p \mathcal{M}$  because a geodesic, once extended too far, may no longer remain a shortest path between its endpoints; think, for instance, of extending the geodesic from the north pole to the south pole of the sphere.

With the idea of a geodesic in hand, we can then define the concepts of convexity, gradients, and gradient flows, which form the building blocks of optimization over curved spaces. We say that a set  $C \subseteq \mathbb{R}^d$  is *convex* if for all  $p_0, p_1 \in C$  and all  $t \in [0, 1]$ , it holds that  $(1 - t)p_0 + t p_1 \in C$ . In this definition,  $t \mapsto (1 - t)p_0 + t p_1$  is the Euclidean geodesic joining  $p_0$  to  $p_1$ . We can generalize this definition to Riemannian manifolds: we say that  $C \subseteq \mathcal{M}$  is *geodesically convex* if for all  $p_0, p_1 \in C$ , all geodesics joining  $p_0$  to  $p_1$  also lie in  $C$ .

We can also define convexity for functions: given  $\alpha \in \mathbb{R}$ , a function  $f : \mathcal{M} \rightarrow \mathbb{R}$  is called  $\alpha$ -*geodesically convex* if

$$f(p_t) \leq (1 - t)f(p_0) + t f(p_1) - \frac{\alpha t(1 - t)}{2} \mathbf{d}^2(p_0, p_1) \quad (5.6)$$

for all  $t \in [0, 1]$  and all geodesics  $(p_t)_{t \in [0,1]}$ . Equivalently, we have the first-order condition

$$f(q) \geq f(p) + \langle \nabla f(p), \log_p(q) \rangle_p + \frac{\alpha}{2} \mathbf{d}^2(p, q) \quad \forall p, q \in \mathcal{M}$$

where  $\nabla f$ , the Riemannian gradient, is defined so that for all curves  $(p_t)_{t \geq 0}$ ,  $\nabla f(p_t) \in T_{p_t} \mathcal{M}$  satisfies  $\partial_t f(p_t) = \langle \nabla f(p_t), \dot{p}_t \rangle_{p_t}$ . Equivalently, we also have the second-order condition

$$\nabla^2 f(p)[v, v] \geq \alpha \|v\|_p^2 \quad \forall p \in \mathcal{M}, \forall v \in T_p \mathcal{M},$$

---

<sup>5</sup> The name is motivated by a classical example of a manifold, the set of orthogonal matrices, in which the tangent space at the identity matrix is the set of anti-symmetric matrices and the exponential map  $\exp_I(A) = \exp(A)$  coincides with the matrix exponential.

where  $\nabla^2 f$ , the Riemannian Hessian, can be defined via  $\nabla^2 f(p)[v, v] := \partial_t^2 f(p_t)|_{t=0}$ , where  $(p_t)_{t \in [0,1]}$  is the geodesic with  $p_0 = p$  and  $\dot{p}_0 = v$ .

In the next section, we return to  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  which, despite not being a *bona fide* Riemannian manifold, carries enough structure to apply calculation rules from Riemannian geometry (and indeed, the formidable book [AGS08] is devoted to the task of placing this endeavor on rigorous footing). It leads to a toolbox, known as *Otto calculus* after Felix Otto, for the study of gradient flows over the space of probability measures.

### 5.3 The Riemannian structure of Wasserstein space

We are now in a position to define a formal Riemannian structure over  $\mathcal{P}_2(\mathbb{R}^d)$ . Recall from Brenier's theorem (Theorem 1.16) that optimal transport maps for the quadratic cost are gradients of convex functions. From (5.5), it follows that optimal velocity vector fields lie in the  $L^2$  closure of the space of gradients of functions (which are not necessarily convex, since we have subtracted the identity map).

**Definition 5.5.** *Let  $\mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ . We define the tangent space to  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  at  $\mu$  to be*

$$T_\mu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) := \overline{\{\nabla \psi \mid \psi : \mathbb{R}^d \rightarrow \mathbb{R} \text{ compactly supported, smooth}\}}^{L^2(\mu)}$$

where  $\overline{\{\cdot\}}^{L^2(\mu)}$  denotes the  $L^2(\mu)$  closure. We endow  $T_\mu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  with the  $L^2(\mu)$  inner product,

$$\langle \nabla \psi_1, \nabla \psi_2 \rangle_\mu := \int \langle \nabla \psi_1, \nabla \psi_2 \rangle \, d\mu.$$

One can show that requiring  $v_t$  to be the gradient of a function,  $v_t = \nabla \psi_t$ , in fact furnishes a fourth characterization of the optimal vector field  $v_t$  in Theorem 5.3, thus justifying Definition 5.5, but we do not prove this here.

*Remark 5.6.* We pause to describe a common alternative convention: instead of defining the tangent vector at  $\mu_t$  to be the driving velocity vector field  $\nabla \psi$ , we could take it to be the ordinary time derivative  $\partial_t \mu_t$  which is given by the continuity equation:  $\partial_t \mu_t = -\text{div}(\mu_t \nabla \psi) =: \chi$ . In this case, the tangent space becomes the space of signed measures with zero total mass, and the metric becomes  $\langle \chi, \chi' \rangle_\mu = \int \langle \nabla \psi, \nabla \psi' \rangle \, d\mu$ ,

where  $\psi, \psi'$  solve the equations  $\chi = -\operatorname{div}(\mu \nabla \psi)$ ,  $\chi' = -\operatorname{div}(\mu \nabla \psi')$ . This just amounts to a change of notation:  $\nabla \psi \mapsto \chi$  is an isometry between our convention and the alternative convention.

Although our logical development thus far has strongly hinted at a connection between this Riemannian structure and the theory of optimal transport, we have not yet stated any result to this effect. The following theorem computes the constant-speed geodesics in the metric defined above. In the language of Section 5.2, it asserts that the metric induced by the Riemannian structure we defined over  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  is indeed the Wasserstein distance.

**Theorem 5.7 (Benamou–Brenier).** *Let  $\mu_0, \mu_1 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ . Then,*

$$W_2^2(\mu_0, \mu_1) = \inf \left\{ \int_0^1 \|v_t\|_{\mu_t}^2 dt \mid (\mu_t, v_t)_{t \in [0,1]} \text{ solves (5.2)} \right\}. \quad (5.7)$$

*The optimal curve  $(\mu_t)_{t \in [0,1]}$  is unique and is described by  $X_t \sim \mu_t$ , where  $X_t = (1-t)X_0 + tX_1$  and  $(X_0, X_1) \sim \bar{\gamma} \in \Gamma_{\mu_0, \mu_1}$  with  $\bar{\gamma}$  being an optimal coupling.*

*Proof.* Let  $(\mu_t, v_t)_{t \in [0,1]}$  solve (5.2), and let  $\dot{X}_t = v_t(X_t)$  with  $X_0 \sim \mu_0$ . Then, it holds that

$$\begin{aligned} W_2^2(\mu_0, \mu_1) &\leq \mathbb{E}[\|X_0 - X_1\|^2] = \mathbb{E}\left[\left\|\int_0^1 \dot{X}_t dt\right\|^2\right] \leq \int_0^1 \mathbb{E}[\|\dot{X}_t\|^2] dt \\ &= \int_0^1 \|v_t\|_{\mu_t}^2 dt. \end{aligned}$$

To study the equality case, note that in the above calculations we employed two inequalities. The first inequality is an equality if and only if  $(X_0, X_1)$  are optimally coupled. The second inequality is an equality if and only if  $t \mapsto \dot{X}_t$  is constant, which forces  $\dot{X}_t = X_1 - X_0$  for all  $t \in [0, 1]$ .

It therefore suffices to show that there exists a family of vector fields  $(v_t)_{t \geq 0}$  such that  $X_1 - X_0 = v_t((1-t)X_0 + tX_1)$  for all  $t \in [0, 1]$ . Since  $(X_0, X_1)$  are optimally coupled, there exists a convex function  $\varphi$  such that  $\nabla \varphi(X_0) = X_1$ . This implies that  $(1-t)X_0 + tX_1 = \nabla \varphi_t(X_0)$ , where  $\varphi_t(x) := (1-t)\frac{\|x\|^2}{2} + t\varphi(x)$ . In particular,  $\varphi_t$  is strongly convex for all  $t \in [0, 1]$ , so  $\nabla \varphi_t$  is invertible. We may therefore define  $v_t := (\nabla \varphi - \text{id}) \circ \nabla \varphi_t^{-1}$ , which satisfies  $v_t((1-t)X_0 + tX_1) = v_t(\nabla \varphi_t(X_0)) = \nabla \varphi(X_0) - X_0 = X_1 - X_0$ , as claimed.  $\square$

Note that we can also write  $\mu_t = [(1-t)\text{id} + tT]_\# \mu_0$ , where  $T$  is the optimal transport map from  $\mu_0$  to  $\mu_1$ . Hence, we formulate the following definition.

**Definition 5.8.** *Let  $\mu_0, \mu_1 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  and let  $T$  denote the optimal transport map from  $\mu_0$  to  $\mu_1$ . The constant-speed geodesic joining  $\mu_0$  to  $\mu_1$  is the curve  $(\mu_t)_{t \in [0,1]}$ , where*

$$\mu_t = [(1-t)\text{id} + tT]_\# \mu_0. \quad (5.8)$$

*This curve is known as the displacement interpolation, McCann's interpolation, or simply the Wasserstein geodesic joining  $\mu_0$  to  $\mu_1$ .*

From (5.8), we can identify  $\log_\mu(\nu) = T_{\mu \rightarrow \nu} - \text{id}$ , and hence  $\exp_\mu(\nabla\psi) = (\text{id} + \nabla\psi)_\# \mu$ . Note that the exponential map is not well-defined if  $\nabla^2\psi$  has an eigenvalue smaller than  $-1$ , since then  $\text{id} + \nabla\psi$  is not the gradient of a convex function and thus not an optimal transport map. This reflects our earlier discussion that the exponential map is not necessarily defined on the full tangent space of a Riemannian manifold.<sup>6</sup>

## 5.4 Otto calculus

The next step is to identify the Wasserstein gradient, which, in turn, allow us to define Wasserstein gradient flows. After obtaining criteria for functionals to be geodesically convex, we can then obtain rates of convergence thereof.

It turns out that the Wasserstein gradient can be expressed in terms of the first variation.

**Definition 5.9 (First variation).** *Let  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  be a functional. The first variation of  $\mathcal{F}$  at  $\mu$ , denoted  $\delta\mathcal{F}(\mu) : \mathbb{R}^d \rightarrow \mathbb{R}$ , is the function defined by*

$$\lim_{\varepsilon \searrow 0} \frac{\mathcal{F}(\mu + \varepsilon\chi) - \mathcal{F}(\mu)}{\varepsilon} = \int \delta\mathcal{F}(\mu) \, d\chi, \quad (5.9)$$

*for all signed measures  $\chi$  such that  $\mu + \varepsilon\chi \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  for all sufficiently small  $\varepsilon$ .*

---

<sup>6</sup> However, the domain of the exponential map for a Riemannian manifold always contains a neighborhood of the origin, whereas this is not true for the Wasserstein space. This is one of the reasons why the Wasserstein space is not truly a Riemannian manifold, even an infinite-dimensional one.

If  $(\mu_t)_{t \geq 0}$  is a curve of densities, then we can write the linear approximation  $\mu_{t+\varepsilon} \approx \mu_t + \varepsilon \partial_t \mu_t$  for  $\varepsilon$  small, where  $\partial_t \mu_t$  denotes the usual time derivative. We can take  $\chi = \partial_t \mu_t$ , in which case (5.9) reads  $\partial_t \mathcal{F}(\mu_t) = \int \delta \mathcal{F}(\mu_t) \partial_t \mu_t$ . Note also that the first variation is only defined up to an additive constant, since the perturbations  $\chi$  always satisfy  $\int d\chi = 0$ .

We can now define Wasserstein gradients. By definition, given a curve of measures  $(\mu_t)_{t \geq 0}$  with tangent vectors  $(v_t)_{t \geq 0}$ , the gradient of a functional  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  is the element of  $T_{\mu_t} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  such that  $\partial_t \mathcal{F}(\mu_t) = \langle \nabla \mathcal{F}(\mu_t), v_t \rangle_{\mu_t}$ . The following proposition shows that this definition can be written directly in terms of the first variation.

**Proposition 5.10.** *Let  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  be a functional with first variation  $\delta \mathcal{F}$ . Then, the Wasserstein gradient of  $\mathcal{F}$  is the vector field  $\nabla \mathcal{F}(\mu) : \mathbb{R}^d \rightarrow \mathbb{R}^d$  defined by*

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

where  $\nabla$  on the right-hand side denotes the usual Euclidean gradient.

*Proof.* Let  $(\mu_t)_{t \geq 0}$  be a curve of measures with tangent vectors  $(v_t)_{t \geq 0}$ . The fact that  $v_t$  is the tangent vector at time  $t$  means that it solves the continuity equation (5.2). From the above discussion of the first variation,

$$\begin{aligned} \partial_t \mathcal{F}(\mu_t) &= \int \delta \mathcal{F}(\mu_t) \partial_t \mu_t = - \int \delta \mathcal{F}(\mu_t) \operatorname{div}(\mu_t v_t) \\ &= \int \langle \nabla \delta \mathcal{F}(\mu_t), v_t \rangle d\mu_t = \langle \nabla \delta \mathcal{F}(\mu_t), v_t \rangle_{\mu_t}. \end{aligned}$$

Moreover, since  $\nabla \delta \mathcal{F}(\mu_t)$  is the gradient of a function, from Definition 5.5 we have  $\nabla \delta \mathcal{F}(\mu_t) \in T_{\mu_t} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ . From this, we conclude that  $\nabla \delta \mathcal{F}(\mu_t)$  is indeed the Wasserstein gradient of  $\mathcal{F}$  at  $\mu_t$ .  $\square$

To compute the Wasserstein gradient, we therefore have to compute the first variation and then take its gradient. We illustrate this on three canonical examples of functionals over  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ .

*Example 5.11 (Potential energy).* Let  $\mathcal{F}(\mu) := \int V d\mu$  for some (potential) function  $V : \mathbb{R}^d \rightarrow \mathbb{R}$ . Then,  $\partial_t \mathcal{F}(\mu_t) = \int V \partial_t \mu_t$  and we can identify  $\delta \mathcal{F}(\mu) = V$ . Thus,  $\nabla \mathcal{F}(\mu) = \nabla V$ .

*Example 5.12 (Internal energy).* Let  $\mathcal{F}(\mu) := \int U(\mu(x)) dx$  for some function  $U : \mathbb{R}_+ \rightarrow \mathbb{R}$ . For example,  $U(x) = x \log x$  gives rise to the entropy<sup>7</sup> functional. Then,  $\partial_t \mathcal{F}(\mu_t) = \int U'(\mu_t) \partial_t \mu_t$ , so we can identify  $\delta \mathcal{F}(\mu) = U' \circ \mu$  and therefore  $\nabla \mathcal{F}(\mu) = \nabla(U' \circ \mu)$ . In the case of entropy,  $\delta \mathcal{F}(\mu) = \log \mu + 1$ , and  $\nabla \mathcal{F}(\mu) = \nabla \log \mu$ .

*Example 5.13 (Interaction energy).* Take a symmetric kernel  $K : \mathbb{R}^d \rightarrow \mathbb{R}$ , i.e.,  $K(-z) = K(z)$ . Set  $\mathcal{F}(\mu) := \frac{1}{2} \iint K(x - y) \mu(dx) \mu(dy)$ . For example, we could consider a Gaussian kernel  $K(x) = \exp(-\frac{\|x\|^2}{2\sigma^2})$ . Then,  $\partial_t \mathcal{F}(\mu_t) = \iint K(x - y) \mu_t(dy) \partial_t \mu_t(dx)$ , so we can identify  $\delta \mathcal{F}(\mu) = \int K(\cdot - y) \mu(dy)$ , and  $\nabla \mathcal{F}(\mu) = \int \nabla K(\cdot - y) \mu(dy)$ .

We can now define the Wasserstein gradient flow of a functional  $\mathcal{F}$  over  $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ . The gradient flow is a curve of measures  $(\mu_t)_{t \geq 0}$  such that the tangent vector to the curve at time  $t$  equals  $-\nabla \mathcal{F}(\mu_t)$ . Recalling that the tangent vectors governs the evolution of  $(\mu_t)_{t \geq 0}$  through the continuity equation (5.2), we arrive at the following definition.

**Definition 5.14 (Wasserstein gradient flow).** Let  $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \rightarrow \mathbb{R}$  be a functional. Then,  $(\mu_t)_{t \geq 0}$  is called the Wasserstein gradient flow of  $\mathcal{F}$  if it solves the PDE

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla \mathcal{F}(\mu_t)).$$

As is well-understood in optimization, gradient flows are natural dynamics for minimizing the objective functional  $\mathcal{F}$  because, as discussed shortly, they always reduce the value of the objective. Wasserstein gradient flows therefore constitute a principled approach for designing dynamics over the space of probability measures aimed at minimizing some criterion, a task which is ubiquitous in applied mathematics, statistics, and beyond; see Chapter 6.

A quick calculation using the definition of the Wasserstein gradient flow  $(\mu_t)_{t \geq 0}$  of  $\mathcal{F}$  yields

$$\partial_t \mathcal{F}(\mu_t) = \langle \nabla \mathcal{F}(\mu_t), v_t \rangle_{\mu_t} = -\|\nabla \mathcal{F}(\mu_t)\|_{\mu_t}^2 \quad (5.10)$$

where  $v_t = -\nabla \mathcal{F}(\mu_t)$  is the tangent vector at time  $t$ . This equality, which states that the objective functional is dissipated at a rate equal to the squared norm of the gradient, is a generic fact about gradient

---

<sup>7</sup> This is the negative of the thermodynamic entropy.

flows. In particular, if  $\mathcal{F}$  is bounded below, it implies that any limit point of the gradient flow must be a stationary point of  $\mathcal{F}$ .

However, we can say more once we have a quantitative lower bound on the rate of dissipation. The simplest such condition is the *Polyak–Lojasiewicz (PL) inequality*.

**Definition 5.15 (Polyak–Lojasiewicz (PL) inequality).** *We say that  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  satisfies a PL inequality with constant  $\alpha > 0$  if for all  $\mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ ,*

$$\|\nabla \mathcal{F}(\mu)\|_\mu^2 \geq 2\alpha (\mathcal{F}(\mu) - \inf \mathcal{F}).$$

The PL inequality over  $\mathbb{R}^d$  is discussed in Appendix A.2; the above definition adapts this concept to the present setting. From (5.10), the PL inequality yields

$$\partial_t(\mathcal{F}(\mu_t) - \inf \mathcal{F}) \leq -2\alpha (\mathcal{F}(\mu_t) - \inf \mathcal{F}).$$

Let  $\phi(t) := \mathcal{F}(\mu_t) - \inf \mathcal{F}$ , so that  $\dot{\phi}(t) \leq -2\alpha\phi(t)$ . If this inequality were an equality, then we could solve the differential equation to obtain  $\phi(t) = \phi(0) \exp(-2\alpha t)$ . In general, when we have a differential *inequality*, we can bound  $\phi$  by the solution to the differential equation; this is formalized as Grönwall's inequality.

**Lemma 5.16 (Grönwall's inequality).** *Let  $c \in \mathbb{R}$ . Let  $\phi : [0, T] \rightarrow \mathbb{R}$  be differentiable, satisfying  $\dot{\phi}(t) \leq c\phi(t)$  for all  $t \in [0, T]$ . Then,*

$$\phi(t) \leq \phi(0) \exp(ct) \quad \forall t \in [0, T].$$

*Proof.* It holds that

$$\partial_t[\exp(-ct) \phi(t)] = \exp(-ct) [-c\phi(t) + \dot{\phi}(t)] \leq 0.$$

This implies  $\exp(-ct) \phi(t) \leq \exp(-c \cdot 0) \phi(0) = \phi(0)$ .  $\square$

On the other hand, applying the same argument as in Lemma A.11, one can show that  $\mathcal{F}$  satisfies the PL inequality with constant  $\alpha$  as soon as  $\mathcal{F}$  is  $\alpha$ -geodesically convex. We deduce the following useful corollary.

**Corollary 5.17.** *Let  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  be  $\alpha$ -geodesically convex. Then, along the Wasserstein gradient flow  $(\mu_t)_{t \geq 0}$  for  $\mathcal{F}$ , it holds*

$$\mathcal{F}(\mu_t) - \inf \mathcal{F} \leq e^{-2\alpha t} (\mathcal{F}(\mu_0) - \inf \mathcal{F}).$$

## 5.5 Bures–Wasserstein

It is illuminating to specialize the concepts in the previous section to the submanifold of Wasserstein space consisting of Gaussian measures.

**Definition 5.18.** *The Bures–Wasserstein space  $\text{BW}(\mathbb{R}^d)$  is the space of non-degenerate Gaussians on  $\mathbb{R}^d$ , equipped with the Wasserstein metric.*

Concretely, since Gaussians are parameterized by the mean and covariance matrix, we can think of  $\text{BW}(\mathbb{R}^d) \cong \mathbb{R}^d \times \mathbf{S}_{++}^d$ , where  $\mathbf{S}_{++}^d$  is the cone of symmetric positive definite  $d \times d$  matrices. Recall from Example 1.19 that for any  $\mu_0, \mu_1 \in \text{BW}(\mathbb{R}^d)$ , the optimal transport map  $T$  from  $\mu_0$  to  $\mu_1$  is an affine map, and the Wasserstein geodesic joining  $\mu_0$  to  $\mu_1$  is

$$\mu_t = \underbrace{[(1-t) \text{id} + t T] \# \mu_0}_{\text{affine}}, \quad t \in [0, 1].$$

Since the pushforward of a non-degenerate Gaussian by a non-singular affine map is also a non-degenerate Gaussian, the Wasserstein geodesic from  $\mu_0$  to  $\mu_1$  lies entirely in  $\text{BW}(\mathbb{R}^d)$ , or in other words:

**Proposition 5.19.**  $\text{BW}(\mathbb{R}^d) \subseteq \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  is geodesically convex.

Recall that a functional  $\mathcal{F}$  on a Riemannian manifold  $\mathcal{M}$  is  $\alpha$ -geodesically convex if the mapping  $[0, 1] \rightarrow \mathcal{M}$ ,  $t \mapsto \mathcal{F}(p_t)$  is  $\alpha$ -convex for all geodesics  $(p_t)_{t \in [0, 1]}$  on  $\mathcal{M}$ . The geodesic convexity of  $\text{BW}(\mathbb{R}^d)$  means that the intrinsic geodesics of  $\text{BW}(\mathbb{R}^d)$  coincide with the Wasserstein geodesics, which immediately furnishes the following corollary.

**Corollary 5.20.** *Let  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  be an  $\alpha$ -geodesically convex functional. Then,  $\mathcal{F}$  is also  $\alpha$ -geodesically convex when viewed as a functional over  $\text{BW}(\mathbb{R}^d)$ .*

We make use of this fact in Subsection 6.1.2.

The Riemannian structure of  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  descends to  $\text{BW}(\mathbb{R}^d)$  and endows the Bures–Wasserstein space with the structure of a bona fide finite-dimensional Riemannian manifold. The tangent space at  $\mu \in \text{BW}(\mathbb{R}^d)$  is

$$\begin{aligned} T_\mu \text{BW}(\mathbb{R}^d) &= \{ \lambda (T_{\mu \rightarrow \nu} - \text{id}) \mid \lambda > 0, \nu \in \text{BW}(\mathbb{R}^d) \} \\ &= \{ x \mapsto Sx + a \mid a \in \mathbb{R}^d, S \in \mathbf{S}^d \}, \end{aligned}$$

where  $\mathbf{S}^d$  is the space of symmetric  $d \times d$  matrices. Actually, it is convenient to reparametrize the tangent space as

$$T_\mu \text{BW}(\mathbb{R}^d) = \{x \mapsto S(x - m) + a \mid a \in \mathbb{R}^d, S \in \mathbf{S}^d\},$$

where  $m = \mathbb{E}_{X \sim \mu}[X]$ .

By definition, the Riemannian structure induced on  $\text{BW}(\mathbb{R}^d)$  is the restriction of the inner product on  $T_\mu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  to the subspace  $T_\mu \text{BW}(\mathbb{R}^d) \subseteq T_\mu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ , i.e., the  $L^2(\mu)$  inner product. Using this, we could compute the BW gradient of a functional  $\mathcal{F}$  from scratch. However, since we have already computed the Wasserstein gradient of  $\mathcal{F}$  at  $\mu$  to be the vector field  $\nabla \mathcal{F}(\mu) = \nabla \delta \mathcal{F}(\mu)$  (Proposition 5.10), a more expedient approach is to now compute the orthogonal projection of  $\nabla \mathcal{F}(\mu)$  onto  $T_\mu \text{BW}(\mathbb{R}^d)$ .

**Theorem 5.21.** *Let  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  be a functional with first variation  $\delta \mathcal{F}(\mu)$  at  $\mu$ . Then, the Bures–Wasserstein gradient of  $\mathcal{F}$  at  $\mu \in \text{BW}(\mathbb{R}^d)$  is the affine mapping*

$$x \mapsto \left( \int \nabla^2 \delta \mathcal{F}(\mu) \, d\mu \right) (x - m) + \int \nabla \delta \mathcal{F}(\mu) \, d\mu,$$

where  $m = \int x \, \mu(dx)$  is the mean of  $\mu$ .

*Proof.* Recall that  $\nabla \mathcal{F}(\mu) = \nabla \delta \mathcal{F}(\mu)$  (Proposition 5.10). The BW gradient at  $\mu$  is the orthogonal projection of  $\nabla \delta \mathcal{F}(\mu)$  onto  $T_\mu \text{BW}(\mathbb{R}^d)$ ; by definition, this is the element  $\nabla_{\text{BW}} \mathcal{F}(\mu) \in T_\mu \text{BW}(\mathbb{R}^d)$  which satisfies

$$\langle \nabla_{\text{BW}} \mathcal{F}(\mu), v \rangle_\mu = \langle \nabla \delta \mathcal{F}(\mu), v \rangle_\mu \quad \forall v \in T_\mu \text{BW}(\mathbb{R}^d). \quad (5.11)$$

We can write out this condition more explicitly. Since  $\nabla_{\text{BW}} \mathcal{F}(\mu), v \in T_\mu \text{BW}(\mathbb{R}^d)$ , they are of the form

$$\begin{aligned} \nabla_{\text{BW}} \mathcal{F}(\mu) &= S(\cdot - m) + a, \\ v &= \tilde{S}(\cdot - m) + \tilde{a}. \end{aligned}$$

On one hand,

$$\begin{aligned} \langle \nabla_{\text{BW}} \mathcal{F}(\mu), v \rangle_\mu &= \mathbb{E}_{X \sim \mu} \langle S(X - m) + a, \tilde{S}(X - m) + \tilde{a} \rangle \\ &= \langle S, \Sigma \tilde{S} \rangle + \langle a, \tilde{a} \rangle, \end{aligned} \quad (5.12)$$

where  $\Sigma$  is the covariance matrix of  $\mu$ . On the other hand,

$$\begin{aligned}\langle \nabla \delta \mathcal{F}(\mu), v \rangle_\mu &= \mathbb{E}_{X \sim \mu} \langle [\nabla \delta \mathcal{F}(\mu)](X), \tilde{S} (X - m) + \tilde{a} \rangle \\ &= \langle \mathbb{E}_{X \sim \mu} [\nabla \delta \mathcal{F}(\mu)(X) (X - m)^\top], \tilde{S} \rangle \quad (5.13)\end{aligned}$$

$$+ \langle \mathbb{E}_{X \sim \mu} \nabla \delta \mathcal{F}(\mu)(X), \tilde{a} \rangle. \quad (5.14)$$

Also, integration by parts yields

$$\begin{aligned}\mathbb{E}_{X \sim \mu} [\nabla \delta \mathcal{F}(\mu)(X) (X - m)^\top] &= \int \nabla \delta \mathcal{F}(\mu) (\Sigma \Sigma^{-1} (\cdot - m))^\top d\mu \\ &= - \int \nabla \delta \mathcal{F}(\mu) (\nabla \log \mu)^\top d\mu \Sigma \\ &= - \int \nabla \delta \mathcal{F}(\mu) (\nabla \mu)^\top \Sigma \\ &= \int \nabla^2 \delta \mathcal{F}(\mu) d\mu \Sigma.\end{aligned}$$

Hence,

$$\begin{aligned}\langle \mathbb{E}_{X \sim \mu} [\nabla \delta \mathcal{F}(\mu)(X) (X - m)^\top], \tilde{S} \rangle &= \left\langle \int \nabla^2 \delta \mathcal{F}(\mu) d\mu \Sigma, \tilde{S} \right\rangle \\ &= \left\langle \Sigma \int \nabla^2 \delta \mathcal{F}(\mu) d\mu, \tilde{S} \right\rangle \\ &= \left\langle \int \nabla^2 \delta \mathcal{F}(\mu) d\mu, \Sigma \tilde{S} \right\rangle. \quad (5.15)\end{aligned}$$

Since (5.11) is supposed to hold for all  $\tilde{a} \in \mathbb{R}^d$  and all  $\tilde{S} \in \mathbf{S}^d$ , by comparing (5.12), (5.13), (5.14), and (5.15), we can identify

$$S = \int \nabla^2 \delta \mathcal{F}(\mu) d\mu \quad \text{and} \quad a = \int \nabla \delta \mathcal{F}(\mu) d\mu.$$

This completes the derivation.  $\square$

Once we have identified the BW gradient, we can use the Lagrangian interpretation of the continuity equation to implement the gradient flow via the dynamics

$$\begin{aligned}\dot{X}_t &= -\nabla_{\text{BW}} \mathcal{F}(\mu_t)(X_t) \\ &= -\left( \int \nabla^2 \delta \mathcal{F}(\mu_t) d\mu_t \right) (X_t - m_t) - \int \nabla \delta \mathcal{F}(\mu_t) d\mu_t,\end{aligned}$$

where  $X_t \sim \mu_t$  and we denote the mean and covariance of  $\mu_t$  by  $m_t$  and  $\Sigma_t$  respectively. However, since  $\mu_t$  is a Gaussian for each  $t \geq 0$ ,

it is expedient to instead track  $\mu_t$  *exactly* through the mean  $m_t$  and covariance  $\Sigma_t$ . They follow the dynamics:

$$\dot{m}_t = \mathbb{E}\dot{X}_t = - \int \nabla \delta \mathcal{F}(\mu_t) d\mu_t,$$

and

$$\begin{aligned} \dot{\Sigma}_t &= \partial_t \mathbb{E}[(X_t - m_t)(X_t - m_t)^\top] \\ &= \mathbb{E}[\dot{X}_t (X_t - m_t)^\top] + \mathbb{E}[(X_t - m_t) \dot{X}_t^\top] \\ &= -\mathbb{E}\left[\left(\left(\int \nabla^2 \delta \mathcal{F}(\mu_t) d\mu_t\right)(X_t - m_t)\right.\right. \\ &\quad \left.\left. + \int \nabla \delta \mathcal{F}(\mu_t) d\mu_t\right)(X_t - m_t)^\top\right] + \dots \\ &= -\left(\int \nabla^2 \delta \mathcal{F}(\mu_t) d\mu_t\right) \mathbb{E}[(X_t - m_t)(X_t - m_t)^\top] + \dots \\ &= -\left(\int \nabla^2 \delta \mathcal{F}(\mu_t) d\mu_t\right) \Sigma_t - \Sigma_t \left(\int \nabla^2 \delta \mathcal{F}(\mu_t) d\mu_t\right), \end{aligned}$$

where above,  $A + \dots$  is shorthand for the expression  $A + A^\top$ . Finally, we have arrived at an explicit system of equations.

**Theorem 5.22.** *The BW gradient flow of the functional  $\mathcal{F}$  is the curve  $(\mu_t = \mathcal{N}(m_t, \Sigma_t))_{t \geq 0}$ , where*

$$\begin{aligned} \dot{m}_t &= -\mathbb{E}\nabla \delta \mathcal{F}(\mu_t)(X_t), \\ \dot{\Sigma}_t &= -\mathbb{E}\nabla^2 \delta \mathcal{F}(\mu_t)(X_t) \Sigma_t - \Sigma_t \mathbb{E}\nabla^2 \delta \mathcal{F}(\mu_t)(X_t), \end{aligned} \tag{5.16}$$

and  $X_t \sim \mu_t$ .

## 5.6 Gaussian mixtures

Building on top of the ideas introduced in Section 5.5, we now consider gradient flows over the space of Gaussian mixtures, which is a far richer space. In fact, as explained below, *any* measure over  $\mathbb{R}^d$  can be viewed as a Gaussian mixture when viewed through the right lens.

Before doing so, we first note that simply constraining the Wasserstein gradient flow to lie on the space of Gaussian mixtures, similarly to how we proceeded in Section 5.5, does not work. The problem is that we cannot explicitly compute the optimal transport map between

two Gaussian mixtures, even infinitesimally, and so we cannot identify the tangent space—unless each Gaussian mixture has one component, or one dimension. Nevertheless, following [CGT19, DD20], there is a natural geometric structure we can consider: Wasserstein over Bures–Wasserstein.

Gaussian mixtures are typically introduced as distributions of the form  $\sum_{k=1}^K w_k \mathcal{N}(m_k, \Sigma_k)$  for *mixing weights*  $w_k \geq 0$ ,  $\sum_{k=1}^K w_k = 1$ , but we can define a Gaussian mixture more broadly as a measure of the form  $\int \mathcal{N}(m, \Sigma) \nu(dm, d\Sigma)$ . The finite Gaussian mixture above corresponds to a discrete *mixing measure*  $\nu$ :  $\nu = \sum_{k=1}^K w_k \delta_{(m_k, \Sigma_k)}$ . This new, broader definition of a Gaussian mixture, is nearly useless, since any measure  $\mu$  can be represented thus:  $\mu = \int \delta_x \mu(dx)$ , where  $\delta_x = \mathcal{N}(x, 0)$  is a degenerate Gaussian. Also, the representation of a Gaussian mixture by a mixing measure  $\nu$  is “overparametrized”, i.e.,  $\nu$  is highly non-unique: consider the equality  $\mathcal{N}(0, I) = \int \mathcal{N}(x, \tau I) \nu(dx)$  where  $\nu = \mathcal{N}(0, (1 - \tau)I)$ , valid for any  $\tau \in [0, 1]$ . Nevertheless, the utility of this perspective is that it leads to a natural interpretation: *a mixing measure for a Gaussian mixture is simply a probability measure over the Bures–Wasserstein space*. Let us see how this leads to the definition of a geometric structure.

*The Bures–Wasserstein space is a Riemannian manifold.* As noted earlier, the space  $\text{BW}(\mathbb{R}^d)$  is isometric to the manifold  $\mathbb{R}^d \times \mathbf{S}_{++}^d$  equipped with a certain Riemannian metric. Hereafter we consider the metric arising from Otto calculus but any metric, including the Euclidean one could be used here.

*We can consider the space of probability measures (with finite second moment) over any metric space, and endow it with the 2-Wasserstein distance.* Indeed, recall from Section 1.1 that the optimal transport problem can be defined with more general costs, so we can take the squared distance function over the metric space as our cost.

*When the metric space in question is a Riemannian manifold, the results from Sections 5.1–5.4 continue to hold with appropriate modifications.* We do not justify this in detail here, but we invite the reader to revisit these sections with a fresh perspective. For example, the ODE (5.1) still makes sense, keeping in mind that a vector field  $v$  on a manifold  $\mathcal{M}$  is an assignment  $x \mapsto v(x)$  of a tangent vector  $v(x) \in T_x \mathcal{M}$  at each point  $x \in \mathcal{M}$ . The continuity equation still makes sense in its weak form (5.3), where  $\nabla$  now refers to the Riemannian gradient, and even (5.2) makes sense if we interpret  $\mu_t$  as a density with respect to the

volume measure, etc. In fact, this geometric setting is the source of some of the deepest developments in optimal transport theory; see [Vil09b].

Crucially, for our purposes, the formula for the Wasserstein gradient given in Proposition 5.10 still holds, where we again interpret  $\nabla$  as the Riemannian gradient.

Putting this discussion together, we can derive the Wasserstein gradient flow which lives in the space  $(\mathcal{P}_2(\text{BW}(\mathbb{R}^d)), W_2)$ .

**Theorem 5.23.** *Given  $\nu \in \mathcal{P}_2(\text{BW}(\mathbb{R}^d))$ , let  $G_\nu$  denote the corresponding Gaussian mixture  $G_\nu = \int \mathcal{N}(m, \Sigma) \nu(dm, d\Sigma)$ . Let  $\mathcal{F}$  be a functional over  $\mathcal{P}_2(\mathbb{R}^d)$ , and let  $\mathcal{G}$  be the corresponding functional over  $\mathcal{P}_2(\text{BW}(\mathbb{R}^d))$  given by  $\nu \mapsto \mathcal{F}(G_\nu)$ . Then, the Wasserstein gradient flow of  $\mathcal{G}$  is the curve  $(\nu_t)_{t \geq 0}$  described as follows:  $\nu_t = \text{law}(m_t, \Sigma_t)$ , where*

$$\begin{aligned}\dot{m}_t &= -\mathbb{E}\nabla\delta\mathcal{F}(G_{\nu_t})(X_t), \\ \dot{\Sigma}_t &= -\mathbb{E}\nabla^2\delta\mathcal{F}(G_{\nu_t})(X_t) \Sigma_t - \Sigma_t \mathbb{E}\nabla^2\delta\mathcal{F}(G_{\nu_t})(X_t),\end{aligned}$$

and  $X_t \sim \mathcal{N}(m_t, \Sigma_t)$ .

*Proof.* The first variation of  $\mathcal{G}$  is computed as follows. If  $(\nu_t)_{t \in \mathbb{R}}$  is a curve in  $\mathcal{P}_2(\text{BW}(\mathbb{R}^d))$ , then  $\partial_t G_{\nu_t} = \int \mathcal{N}(m, \Sigma) \partial_t \nu_t(dm, d\Sigma)$ . Hence,

$$\begin{aligned}\partial_t \mathcal{G}(\nu_t) &= \partial_t \mathcal{F}(G_{\nu_t}) = \int \delta\mathcal{F}(G_{\nu_t}) \partial_t G_{\nu_t} \\ &= \iint \delta\mathcal{F}(G_{\nu_t}) d\mathcal{N}(m, \Sigma) \partial_t \nu_t(dm, d\Sigma).\end{aligned}$$

This implies that the first variation is

$$\delta\mathcal{G}(\nu) : (m, \Sigma) \mapsto \int \delta\mathcal{F}(G_\nu) d\mathcal{N}(m, \Sigma).$$

Based on our identification of  $\text{BW}(\mathbb{R}^d)$  with  $\mathbb{R}^d \times \mathbf{S}_{++}^d$ , the Riemannian gradient of  $\delta\mathcal{G}(\nu)$ , evaluated at  $(m, \Sigma)$ , is the same as the Bures–Wasserstein gradient of  $\mu \mapsto \int \delta\mathcal{F}(G_\mu) d\mu$ , evaluated at  $\mu = \mathcal{N}(m, \Sigma)$ , and we computed the latter in Theorem 5.21. Therefore, the result follows from Theorem 5.22.  $\square$

## 5.7 Wasserstein–Fisher–Rao

We now describe a variation of the Wasserstein geometry that is often useful in applications as illustrated in Chapter 6. This variation, known

as the *Wasserstein–Fisher–Rao* (WFR) or *Hellinger–Kantorovich* distance, was originally proposed and studied as a model of *unbalanced optimal transport*, that is, optimal transport between positive measures not necessarily containing the same total mass. A notable example is the cellular trajectory reconstruction application mentioned in the Preface, in which measures are used to represent snapshots of the cell population at different times, and for which the total mass indeed changes due to the birth and death of individual cells.

From the static perspective, WFR defines a variant of the optimal transport problem, and its various properties such as the cost, metric properties, duality, etc. can all be investigated in a similar vein as we did in Chapter 1. We refer to the references [LMS16, KMV16, CPSV18, LMS18] for detailed investigations in this direction. In this section, we follow [LCB<sup>+</sup>22, Appendix H] and focus on the Riemannian structure of the resulting metric space for the purpose of deriving gradient flows.

### *Fisher–Rao*

Before turning toward Wasserstein–Fisher–Rao, we first describe one of its key components: the *Fisher–Rao* metric. This is a metric over the space  $\mathcal{M}_+(\mathbb{R}^d)$  of positive measures over  $\mathbb{R}^d$ , defined via

$$d_{\text{FR}}^2(\mu_0, \mu_1) := \int (\sqrt{\mu_0} - \sqrt{\mu_1})^2,$$

where as usual we identify measures with their Lebesgue densities, assuming that they exist.<sup>8</sup> When  $\mu_0, \mu_1$  are probability measures, then  $d_{\text{FR}}$  coincides with the statistician’s *Hellinger distance*.<sup>9</sup>

Geometrically,  $d_{\text{FR}}$  is the metric over (say) non-negative densities obtained by demanding that  $\mu \mapsto \sqrt{\mu}$  be an isometry into the Hilbert space  $L^2(\mathbb{R}^d)$ . Therefore, the geometry of  $(\mathcal{M}_+(\mathbb{R}^d), d_{\text{FR}})$  is flat, and we can obtain a Riemannian structure through the isometry. Namely, if  $\dot{\mu}$  denotes the derivative in time of a curve of densities, then the derivative of the square root is  $\dot{\sqrt{\mu}} = \dot{\mu}/(2\sqrt{\mu})$ . If we measure the “length” of the latter in the  $L^2(\mathbb{R}^d)$  norm, we arrive at the induced Riemannian metric

$$g_\mu(\dot{\mu}, \dot{\mu}) := \|\dot{\sqrt{\mu}}\|_{L^2(\mathbb{R}^d)}^2 = \int \frac{\dot{\mu}^2}{4\mu}$$

---

<sup>8</sup> When the densities do not exist, we can define the distance via  $d_{\text{FR}}^2(\mu_0, \mu_1) := \int (\sqrt{\frac{d\mu_0}{d\lambda}} - \sqrt{\frac{d\mu_1}{d\lambda}})^2 d\lambda$  with respect to any common dominating measure  $\lambda$ .

<sup>9</sup> This explains the competing naming conventions for the WFR metric; note that  $\{\text{Wasserstein, Fisher–Rao}\} \cong \{\text{Hellinger, Kantorovich}\}$ .

over the tangent space  $T_\mu \mathcal{M}_+(\mathbb{R}^d)$  of functions  $\dot{\mu} : \mathbb{R}^d \rightarrow \mathbb{R}$ .

This geometry is set over the space of all positive measures, including measures with differing amounts of mass, and indeed this geometry proves useful for modeling physical situations in which change of mass naturally occurs. A canonical example is when  $\mu$  represents the concentration of a chemical substance, and the concentration changes over time due to chemical reactions. This is modelled by the *reaction equation*  $\partial_t \mu_t = \alpha_t \mu_t$ , where  $\alpha_t : \mathbb{R}^d \rightarrow \mathbb{R}$  dictates the rate of reaction at each point in space. Note that in this notation,  $\alpha = \dot{\mu}/\mu$ , which amounts to a reparametrization of the tangent space. In other words, we can equivalently think of the tangent space as consisting of functions  $\alpha : \mathbb{R}^d \rightarrow \mathbb{R}$  equipped with the metric

$$\tilde{\mathfrak{g}}_\mu(\alpha, \alpha) := \mathfrak{g}_\mu(\alpha\mu, \alpha\mu) = \frac{1}{4} \int \alpha^2 d\mu. \quad (5.17)$$

Going forward, we adopt this as our definition of the metric, and hence we write  $\|\alpha\|_\mu^2 := \tilde{\mathfrak{g}}_\mu(\alpha, \alpha)$ .

We can draw comparisons with the definition of the Wasserstein geometry: at each measure  $\mu$ , the “tangent space”  $T_\mu \mathcal{M}_+(\mathbb{R}^d)$  at  $\mu$  is now defined to be the space of all functions  $\alpha : \mathbb{R}^d \rightarrow \mathbb{R}$ , equipped with the metric (5.17), and the continuity equation is replaced by the reaction equation  $\partial_t \mu_t = \alpha_t \mu_t$ . Compared to the Wasserstein metric, the Fisher–Rao metric is based on an entirely different intuition: rather than *transportation* of mass, the reaction equation now describes spontaneous *creation and destruction* of mass.

Despite motivating the Fisher–Rao geometry for problems involving change of mass, we may also wish to apply it to problems in which we want to maintain a flow on the space of probability measures. To do so, we consider the induced geometry over  $\mathcal{P}(\mathbb{R}^d)$ . The equation  $\partial_t \mu_t = \alpha_t \mu_t$  preserves the total mass if and only if  $\int \alpha_t d\mu_t = 0$  for all  $t \geq 0$ , so we restrict the tangent space to  $T_\mu \mathcal{P}(\mathbb{R}^d) = \{\alpha : \mathbb{R}^d \rightarrow \mathbb{R} \mid \int \alpha d\mu = 0\}$ , equipped with the metric (5.17).<sup>10</sup> The preservation of mass ensures that any mass that is destroyed is also instantly created elsewhere. To adhere to the lexicon of transport, this phenomenon is sometimes referred to as *teleportation*; however, it should be noted that it merely corresponds to a reweighting.

<sup>10</sup> A discrete analogy: endow the space of probability measures on  $\{1, \dots, d\}$  (i.e., the simplex in  $\mathbb{R}^d$ ) with a geometry via an isometry  $p \mapsto \sqrt{p}$ , where  $\sqrt{p}$  is an element of the unit sphere  $\mathbb{S}^{d-1}$ . See Exercise 13.

What about gradient flows? Given a functional  $\mathcal{F} : \mathcal{M}_+(\mathbb{R}^d) \rightarrow \mathbb{R}$  or  $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}$ , the gradient by definition satisfies  $\partial_t \mathcal{F}(\mu_t) = \langle \nabla_{\text{FR}} \mathcal{F}(\mu_t), \alpha_t \rangle_{\mu_t}$  along every curve  $\partial_t \mu_t = \alpha_t \mu_t$ . By unpacking the definitions, one checks (see Exercise 14) that

$$\nabla_{\text{FR}} \mathcal{F}(\mu) = \delta \mathcal{F}(\mu) \quad \text{or} \quad \nabla_{\text{FR}} \mathcal{F}(\mu) = \delta \mathcal{F}(\mu) - \int \delta \mathcal{F}(\mu) \, d\mu \quad (5.18)$$

depending on whether we are working over  $\mathcal{M}_+(\mathbb{R}^d)$  or  $\mathcal{P}(\mathbb{R}^d)$  respectively; here,  $\delta \mathcal{F}$  denotes the first variation of  $\mathcal{F}$  (Definition 5.9). To disambiguate the two cases and to emphasize the original motivation of the WFR metric from unbalanced optimal transport, we refer to the former case as *unbalanced Fisher–Rao* and the latter as simply *Fisher–Rao*. The Fisher–Rao gradient flow of  $\mathcal{F}$  follows the tangent vector  $-\nabla_{\text{FR}} \mathcal{F}(\mu_t)$  at time  $t$  and is given by

$$\partial_t \mu_t = -\nabla_{\text{FR}} \mathcal{F}(\mu_t) \mu_t. \quad (5.19)$$

### Wasserstein–Fisher–Rao

We now combine both the Wasserstein and Fisher–Rao geometries into a hybrid geometry that incorporates both mass transport and creation/destruction (a.k.a. reweighting, a.k.a. teleportation). The idea is to simply consider the continuity equation with reaction,  $\partial_t \mu_t + \text{div}(\mu_t v_t) = \alpha_t \mu_t$ . The governing equation is parameterized by a function  $\alpha$  and a vector field  $v$ , which together form a tangent vector. It is then natural to consider the metric<sup>11</sup>

$$\|(\alpha, v)\|_{\mu}^2 = \int (\alpha^2 + \|v\|^2) \, d\mu. \quad (5.20)$$

However, similarly to our discussion in Section 5.1, this does not uniquely define a tangent vector because there is too much freedom to choose the pair  $(\alpha, v)$  while maintaining the same evolution of measures  $(\mu_t)_{t \geq 0}$ . It can be shown that the *optimal* pair  $(\alpha, v)$ , in the sense of minimizing the norm (5.20) (c.f. the discussion in Section 5.1) can be characterized as follows:  $\alpha = \psi$  and  $v = \nabla \psi$  for some function  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ . Hence, we can define the WFR tangent space at  $\mu$  to be

$$T_{\mu} \mathcal{M}_+(\mathbb{R}^d) = \overline{\{(\psi, \nabla \psi) \mid \psi : \mathbb{R}^d \rightarrow \mathbb{R} \text{ compact supp., smooth}\}}^{L^2(\mu)}$$

<sup>11</sup> Strictly speaking, to add the Wasserstein and Fisher–Rao geometries, we should add a factor of  $\frac{1}{4}$  in front of the  $\alpha^2$  term, and this is indeed the convention adopted in some works. We omit this factor for parsimony.

equipped with the norm

$$\|(\psi, \nabla\psi)\|_{\mu}^2 := \int (\psi^2 + \|\nabla\psi\|^2) d\mu. \quad (5.21)$$

This has the pleasing interpretation of “completing” the Wasserstein metric  $\|\nabla\psi\|_{L^2(\mu)}^2$  to the full Sobolev norm of  $\psi$ . Note also that the governing equation becomes

$$\partial_t \mu_t + \operatorname{div}(\mu_t \nabla\psi_t) = \psi_t \mu_t. \quad (5.22)$$

As before, we can also restrict to the space of probability measures  $\mathcal{P}(\mathbb{R}^d)$ , in which case we restrict to pairs  $(\psi, \nabla\psi)$  such that  $\int \psi d\mu = 0$ , endowed with the same metric (5.21). We refer to WFR over the full space  $\mathcal{M}_+(\mathbb{R}^d)$  as *unbalanced WFR*, henceforth reserving the use of WFR for the restriction to  $\mathcal{P}(\mathbb{R}^d)$ .

The following theorem computes the WFR gradient.

**Theorem 5.24.** *Let  $\mathcal{F}$  be a functional over  $\mathcal{M}_+(\mathbb{R}^d)$  or  $\mathcal{P}(\mathbb{R}^d)$ . Then, unbalanced WFR gradient of  $\mathcal{F}$ , denoted  $\nabla_{\text{FR}} \mathcal{F}$ , is given by*

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

and the WFR gradient by

$$\nabla_{\text{FR}} \mathcal{F}(\mu) = \left( \delta \mathcal{F}(\mu) - \int \delta \mathcal{F}(\mu) d\mu, \nabla \delta \mathcal{F}(\mu) \right).$$

*Proof.* Let  $(\mu_t)_{t \geq 0}$  satisfy (5.22). Then, by integration by parts,

$$\partial_t \mathcal{F}(\mu_t) = \int \langle \nabla \delta \mathcal{F}(\mu_t), \nabla \psi_t \rangle d\mu_t + \int \delta \mathcal{F}(\mu_t) \psi_t d\mu_t.$$

In the unbalanced case, we can identify this as

$$\langle (\delta \mathcal{F}(\mu), \nabla \delta \mathcal{F}(\mu)), (\psi_t, \nabla \psi_t) \rangle_{\mu_t}$$

according to the definition of the metric (5.21). In the balanced case, we have  $\int \psi_t d\mu_t = 0$ , and since the WFR gradient is by definition an element of the tangent space its first component must also have mean zero, so the claim follows.  $\square$

The WFR gradient flow is therefore given by

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla \delta \mathcal{F}(\mu_t)) - \left( \delta \mathcal{F}(\mu_t) - \int \delta \mathcal{F}(\mu_t) d\mu_t \right) \mu_t. \quad (5.23)$$

## 5.8 Mean-field particle systems

We conclude this chapter by describing Wasserstein and WFR gradient flows from a particle systems perspective. These arise naturally when these gradient flows are initialized at finite measures. Indeed, a key observation is that since both gradient flows can be implemented using ordinary differential equations, if  $\mu_0$  is a finite measure,  $\mu_t$  remains a finite measure at all times  $t$  along these gradient flows.

### 5.8.1 Particle Wasserstein gradient flow

To illustrate this point, let  $\mathcal{F}$  be a function over  $\mathcal{P}(\mathbb{R}^d)$  and recall from Definition 5.14 that the Wasserstein gradient flow of  $\mathcal{F}$  is the continuity equation associated with the ODE

$$\dot{X}_t = -\nabla \mathcal{F}(\mu_t)(X_t), \quad (5.24)$$

where  $\mu_t$  denotes the law of  $X_t$ . In particular, we only need to describe these dynamics on the support of  $\mu_t$ .

Assume now that the Wasserstein gradient flow is initialized at

$$\mu_0 := \frac{1}{N} \sum_{j=1}^N \delta_{X_0^j},$$

for a given collection of points  $X_0^1, \dots, X_0^N \in \mathbb{R}^d$ . We get that

$$\mu_t := \frac{1}{N} \sum_{j=1}^N \delta_{X_t^j},$$

where for  $i \in [N]$ ,

$$\dot{X}_t^i = -\nabla \mathcal{F}(\mu_t)(X_t^i). \quad (5.25)$$

These dynamics describe an interacting particle system where particles  $(X_t^1, \dots, X_t^N)$  are subject to dynamics of the form

$$\dot{X}_t^i = V_t^i(X_t^1, \dots, X_t^N), \quad i \in [N]. \quad (5.26)$$

Note that in the case of Wasserstein gradient flows, we further have that:

- (a) each particle  $X_t^i$  interacts with the others only through the effect of their distribution  $\mu_t$ , and

(b) these interactions have the same form for all the particles.

Slightly overloading notation, this means that the general dynamics in (5.26) simplify to

$$\dot{X}_t^i = V_t^i(X_t^1, \dots, X_t^N) \stackrel{(a)}{=} V_t^i(X_t^i, \mu_t) \stackrel{(b)}{=} V_t(X_t^i, \mu_t), \quad i \in [N].$$

These two properties are precisely captured by (5.24): the first one is obvious and the second one is manifest due to the absence of a superscript  $i$ , which indicates that each particle is subject to the same vector field. Such a system is said to exhibit *mean-field interactions*. Both the Wasserstein and WFR gradient flows are of this form.

Mean-field interaction systems are convenient because it is strictly equivalent to describe the dynamics of each particle and that of their distribution. The latter takes the form of a PDE given by the continuity equation (5.2).

Since the Wasserstein gradient flow only moves particles, the weights in  $\mu_0$  do not change over time: if the Wasserstein gradient flow is initialized at

$$\mu_0 := \sum_{j=1}^N w_0^j \delta_{X_0^j}, \quad (5.27)$$

where  $w_0^j \geq 0$ ,  $j \in [N]$  and  $\sum_{j=1}^N w_0^j = 1$ , then

$$\mu_t := \sum_{j=1}^N w_0^j \delta_{X_t^j},$$

where  $X_t^1, \dots, X_t^N$  evolve according to (5.25). To also impose dynamics on the weights, we employ instead a WFR gradient flow.

### 5.8.2 Particle WFR gradient flow

Recall that tangent vector fields for the Wasserstein space are displacement maps of the form  $\nabla\psi$ . The Wasserstein–Fisher–Rao (WFR) geometry reinterprets the tangent space by replacing the governing continuity equation (5.2) with the reaction-transport equation (5.22). In particular, it offers the possibility of traversing the space of probability measures, say from initial distribution to target distribution, more efficiently by reweighting particles rather than having to move them across the space in a continuous fashion. When initialized at a finite

measure of the form (5.27), this effect manifests itself in the form of time-varying weights:

$$\mu_t := \sum_{j=1}^N w_t^j \delta_{X_t^j},$$

where  $w_t^j \geq 0$ ,  $j \in [N]$  and  $\sum_{j=1}^N w_t^j = 1$ .

The particle updates follow the Wasserstein geometry, and the weight updates follow the Fisher–Rao geometry: for  $i \in [N]$ ,

$$\begin{aligned} \dot{X}_t^i &= -\nabla \delta \mathcal{F}(\mu_t)(X_t^i), \\ \dot{w}_t^i &= -\left( \delta \mathcal{F}(\mu_t)(X_t^i) - \int \delta \mathcal{F}(\mu_t) d\mu_t \right) w_t^i. \end{aligned} \quad (5.28)$$

### 5.8.3 Gaussian particles

Following Section 5.6, we can also take a finite Gaussian mixture with mixing measure

$$\nu_t = \frac{1}{K} \sum_{k=1}^K \delta_{(m^k, \Sigma^k)}$$

that evolves according to the Wasserstein gradient flow for the functional  $\nu \mapsto \mathcal{G}(\nu) = \mathcal{F}(G_\nu)$ . By Theorem 5.23, this flow takes the following form: for each  $k \in [K]$ ,

$$\begin{aligned} \dot{m}_t^k &= -\mathbb{E} \nabla \delta \mathcal{F}(G_{\nu_t})(X_t^k), \\ \dot{\Sigma}_t^k &= -\mathbb{E} \nabla^2 \delta \mathcal{F}(G_{\nu_t})(X_t^k) \Sigma_t^k - \Sigma_t^k \mathbb{E} \nabla^2 \delta \mathcal{F}(G_{\nu_t})(X_t^k), \end{aligned} \quad (5.29)$$

where  $X_t^k \sim \mathcal{N}(m_t^k, \Sigma_t^k)$ . Note that this is an interacting system of “particles”  $(m_t^k, \Sigma_t^k)$ , but each particle corresponds to a Gaussian component  $\mathcal{N}(m_t^k, \Sigma_t^k)$ , and the collection thereof to the Gaussian mixture  $\nu_t$ . We therefore refer to  $\mathcal{N}(m_t^k, \Sigma_t^k)$  as a *Gaussian particle*.

We emphasize that these dynamics do *not* implement the Wasserstein gradient flow for  $\mathcal{F}$ . Nevertheless, these dynamics are perfectly valid for minimizing  $\mathcal{F}$  over the space of  $K$ -component Gaussian mixtures.

Recall that in Section 5.6, we equipped the space of probability measures over  $\text{BW}(\mathbb{R}^d)$ —i.e., the space of mixing measures—with the Wasserstein geometry. But we could have equally well considered equipping this space with the WFR geometry. The corresponding particle dynamics evolves the finite Gaussian mixture

$$\nu_t = \sum_{k=1}^K w_t^k \delta_{(m_t^k, \Sigma_t^k)}$$

with changing weights, governed by

$$\begin{aligned}\dot{m}_t^k &= -\mathbb{E} \nabla \delta \mathcal{F}(G_{\nu_t})(X_t^k), \\ \dot{\Sigma}_t^k &= -\mathbb{E} \nabla^2 \delta \mathcal{F}(G_{\nu_t})(X_t^k) \Sigma_t^k - \Sigma_t^k \mathbb{E} \nabla^2 \delta \mathcal{F}(G_{\nu_t})(X_t^k), \\ \dot{w}_t^k &= -\left(\mathbb{E} \delta \mathcal{F}(G_{\nu_t})(X_t^k) - \frac{1}{K} \sum_{k'=1}^K \mathbb{E} \delta \mathcal{F}(G_{\nu_t})(X_t^{k'})\right) w_t^k,\end{aligned}$$

where  $X_t^k \sim \mathcal{N}(m_t^k, \Sigma_t^k)$ .

#### 5.8.4 Implementation strategies for gradient flows

For both the Wasserstein gradient flow and the WFR gradient flow, one needs to compute the Wasserstein gradient  $\nabla \mathcal{F}(\mu_t) = \nabla \delta \mathcal{F}(\mu_t)$  on the support of  $\mu_t$ . When  $\mu_t$  is a discrete measure, this quantity may not be well-defined. This is the case for example when  $\mathcal{F}$  is the entropy functional which is itself not defined on discrete measures, let alone its Wasserstein gradient. (Note, however, that it may be well-defined when we use Gaussian particles.)

In practice, the particle implementations discussed here typically needs to be combined with other tricks (e.g., “kernelization” as in Subsection 6.1.4). These implementation strategies are described in the next chapter.

## 5.9 Discussion

**§5.1.** Detailed treatments of the metric derivative and the continuity equation can be found in [AGS08, Vil09b, San15]. In particular, a rigorous version of Theorem 5.3 can be found in [AGS08, Chapter 8] or [San15, Chapter 5].

**§5.2.** There are many excellent textbooks covering Riemannian geometry, e.g., [dC92].

**§5.3.** The Benamou–Brenier formula is often called the “dynamical” formulation of optimal transport (as opposed to Chapter 1, which describes the “static” picture). There is also a dynamical version of the dual problem, in which the dual potentials evolve according to

the *Hamilton–Jacobi equation*; see [Vil03, Section 8.1]. The dynamical version of entropic optimal transport, introduced in Chapter 4, is closely tied to the well-known *Schrödinger bridge problem* [Léo14, CGP21].

**§5.4.** The formal calculation rules described in this section were first laid out by Otto [Ott01], although some of the ideas were already anticipated in the earlier work of [Laf88].

The tangent space at a measure  $\mu$ , together with its metric, can be viewed as a linearization of the geometric structure at  $\mu$ . This gives rise to “linearized optimal transport” which has formed the basis for numerous applications [WSB<sup>+</sup>13, BKR14, KR15, SC15, KTOR16, BGKL17, PT18, CCCC20].

Besides gradient flows, there have also been proposals for adaptations of other optimization algorithms to the Wasserstein space, e.g., [CLZ20, WL20, WL22, Tan23, CLTW24].

**§5.5.** The Bures–Wasserstein geometry is named after Donald Bures [Bur69], who introduced this metric over the PSD cone in his work on quantum information theory. BW geometry is further explored in [Mod17, BJL19, HMJG21, vO22]. The connection with the Burer–Monteiro factorization, as described in Exercise 12, has been explored in the context of low-rank matrix recovery [LGT22, MLGR23].

**§5.6.** As mentioned in the main text, the geometry described in this section was first considered in [CGT19, DD20], although the gradient flow equations were obtained in [LCB<sup>+</sup>22].

**§5.7.** The Fisher–Rao geometry is well-studied in information geometry [AN00, AJLS17].

**§5.8.** In some sources, such as [LMS16], WFR gradient flows are written in terms of the square root of the weights, i.e., in terms of  $r := \sqrt{w}$ . This convention is motivated by the fact that the FR distance corresponds to the Euclidean distance between the square roots of the weights, and the WFR distance can therefore be interpreted as a coupling cost on the space of  $(r, x)$  pairs equipped with a “cone” metric (c.f. Subsection 7.3.3). Since we do not cover this perspective here, we adopt the more straightforward parametrization in terms of  $w$ .

The use of Gaussian particles was first advocated in [LCB<sup>+</sup>22].

## 5.10 Exercises

In the following exercises, you may use the following formula for the Wasserstein gradient of the squared Wasserstein distance:

$$[\mathbb{W}W_2^2(\cdot, \nu)](\mu) = 2(\text{id} - T_{\mu \rightarrow \nu}). \quad (5.30)$$

We do not give the full proof of (5.30) here, but it is straightforward to establish the upper bound (Exercise 3).

1. Let  $X_0 \sim \mu_0$ , where  $\mu_0$  is the standard Gaussian over  $\mathbb{R}^2$ . For  $t \geq 0$ , let  $X_t = R_t X_0$  where  $R_t$  is a rotation by  $\theta(t)$  radians. Compute the vector field  $v_t$  such that  $\dot{X}_t = v_t(X_t)$  and show that  $\text{div}(\mu_0 v_t) = 0$  for all  $t \geq 0$  (and hence that  $X_t \sim \mu_0$  for all  $t \geq 0$ ).
2. Show that the set of product measures over  $\mathbb{R}^d$  is a geodesically convex subset of  $\mathcal{P}_2(\mathbb{R}^d)$ .
3. Suppose that  $(X_t)_{t \in \mathbb{R}}$  follows the ODE  $\dot{X}_t = v_t(X_t)$ , so that  $\mu_t = \text{law}(X_t)$  evolves according to the continuity equation  $\partial_t \mu_t + \text{div}(\mu_t v_t) = 0$ , and suppose that  $\mu_0 = \mu$ . Prove that

$$\limsup_{h \searrow 0} \frac{W_2^2(\mu_h, \nu) - W_2^2(\mu_0, \nu)}{h} \leq 2 \langle \text{id} - T_{\mu \rightarrow \nu}, v_0 \rangle_{\mu}.$$

*Hint:* Let  $X_0 \sim \mu$  and  $Y \sim \nu$  be optimally coupled.

4. Let  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R} \cup \{\infty\}$  be a geodesically convex functional which is minimized at  $\pi$ . Let  $(\mu_t)_{t \geq 0}$  denote the Wasserstein gradient flow for  $\mathcal{F}$ . By differentiating  $t \mapsto 2t \mathcal{F}(\mu_t) + W_2^2(\mu_t, \pi)$ , prove

$$\mathcal{F}(\mu_t) - \inf \mathcal{F} \leq \frac{W_2^2(\mu_0, \pi)}{2t}.$$

5. Let  $\mathcal{F} : \mathcal{P}_{2,\text{ac}} \rightarrow \mathbb{R} \cup \{\infty\}$  be a functional. We saw that  $\alpha$ -strong convexity of  $\mathcal{F}$  implies the Polyak–Lojasiewicz (PL) inequality

$$\|\mathbb{W}\mathcal{F}(\mu)\|_{\mu}^2 \geq 2\alpha(\mathcal{F}(\mu) - \inf \mathcal{F}), \quad \forall \mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d). \quad (5.31)$$

- a) Show that (5.31) implies the quadratic growth inequality

$$\mathcal{F}(\mu) - \inf \mathcal{F} \geq \frac{\alpha}{2} W_2^2(\mu, \pi), \quad \forall \mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d), \quad (5.32)$$

where  $\pi$  is the minimizer of  $\mathcal{F}$ . This is known as the *Otto–Villani theorem* after [OV00].

*Hint:* Differentiate  $t \mapsto \sqrt{\frac{\alpha}{2}} W_2(\mu_t, \mu_0) + \sqrt{\mathcal{F}(\mu_t) - \inf \mathcal{F}}$  along the Wasserstein gradient flow of  $\mathcal{F}$ . You may assume that the gradient flow converges to  $\pi$ , which is a consequence of (5.32) if  $\mathcal{F}$  is uniquely minimized.

b) In general, (5.32) does not imply (5.31). However, prove that when  $\mathcal{F}$  is geodesically convex, then (5.32) implies (5.31) but with  $\alpha$  replaced by  $\alpha/4$ .

6. Suppose that instead of the PL inequality (5.31), we instead have the inequality

$$\|\nabla \mathcal{F}(\mu)\|_\mu^p \geq c(\mathcal{F}(\mu) - \inf \mathcal{F}), \quad \forall \mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d),$$

for some power  $0 < p < 2$ . Show that the Wasserstein gradient flow dissipates  $\mathcal{F}$  at a polynomial rate:  $\mathcal{F}(\mu_t) - \inf \mathcal{F} = O(1/t^{\frac{2}{p}-1})$ . What happens in the case  $p > 2$ ?

7. Consider  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  which is the operator norm of the second moment matrix:

$$\mathcal{F}(\mu) := \left\| \int x x^\top \mu(dx) \right\|_{\text{op}}.$$

Prove that  $\mathcal{F}$  is geodesically convex.

8. a) Compute the Wasserstein gradient of the chi-squared divergence  $\chi^2(\cdot \parallel \pi)$  at  $\mu$ . Recall that  $\chi^2(\mu \parallel \pi) := \int \frac{d\mu}{d\pi} d\mu - 1$ . Also, write down the equation for the Wasserstein gradient flow of the chi-squared divergence.

b) Prove that when  $\pi$  is log-concave, then  $\chi^2(\cdot \parallel \pi)$  is geodesically convex.

9. We say that a functional  $\mathcal{F} : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R} \cup \{\infty\}$  is  $\alpha$ -convex along *generalized geodesics* if for all triples  $\mu_0, \mu_1, \nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ , if we define the generalized geodesic joining  $\mu_0$  to  $\mu_1$  with base  $\nu$  via

$$\mu_t^\nu := [(1-t)T_{\mu_0 \rightarrow \nu} + tT_{\mu_1 \rightarrow \nu}]_\# \nu,$$

then it holds:

$$\mathcal{F}(\mu_t^\nu) \leq (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) - \frac{\alpha t(1-t)}{2} W_2^2(\mu_0, \mu_1).$$

a) Explain why, if  $\mathcal{F}$  is  $\alpha$ -convex along generalized geodesics, then it is  $\alpha$ -strongly convex. Also, explain why being  $\alpha$ -convex along generalized geodesics is equivalent to the mapping  $\mathcal{F} \circ \exp_\nu$  being  $\alpha$ -strongly convex on the tangent space  $T_\nu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ .

b) Show that for  $\pi \propto \exp(-V)$  where  $V$  is  $\alpha$ -strongly convex, then  $\text{KL}(\cdot \parallel \pi)$  is  $\alpha$ -convex along generalized geodesics (this strengthens the convexity result of Corollary 6.4).

c) Show that for any  $\mu_0, \mu_1, \nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ , there *exists* at least one generalized geodesic joining  $\mu_0$  to  $\mu_1$ , along which  $\frac{1}{2}W_2^2(\cdot, \nu)$  is 1-strongly convex.

*Remark:* Generalized geodesics play an important role in studying the geometry of the Wasserstein space. The result of the third part of this question was used to show existence of the minimizing movements scheme, which in turn is used to rigorously construct Wasserstein gradient flows; see [AGS08] for further reading.

10. Compute the Wasserstein geodesic joining two Gaussians.
11. Use (5.30) to show that if  $(\mu_t)_{t \geq 0}$  is the Wasserstein gradient flow of  $\frac{1}{2}W_2^2(\cdot, \nu)$ , then  $t \mapsto \mu_{1-\exp(-t)}$  is the constant-speed Wasserstein geodesic joining  $\mu_0$  to  $\nu$ .
12. Let  $\mathcal{F}$  be a functional over  $\mathcal{P}_2(\mathbb{R}^d)$ , and consider the functional  $(m, U) \mapsto F(m, U) := \mathcal{F}(\mathcal{N}(m, UU^\top))$ . Show that the *Euclidean gradient flow* for  $F$  over  $\mathbb{R}^d \times \mathbb{R}^{d \times d}$  yields the same dynamics (up to rescaling time) as the Bures–Wasserstein gradient flow (5.16) where  $\Sigma = UU^\top$ . Similarly, show that the Euclidean gradient flow of  $(m^1, \dots, m^K, U^1, \dots, U^K) \mapsto \mathcal{F}(\frac{1}{K} \sum_{k=1}^K \mathcal{N}(m^k, U^k(U^k)^\top))$  recovers the Gaussian mixture flow (5.29) (up to rescaling time).

*Remark:* The parametrization  $\Sigma = UU^\top$  is often referred to as the *Burer–Monteiro parametrization*, especially when it is used to constrain  $\Sigma$  to have low rank [BM03, BM05].

13. Let  $p$  be an element in the interior of the simplex, i.e., a strictly positive probability distribution over the finite alphabet  $\{1, \dots, d\}$ . Consider the isometry  $f : p \mapsto \sqrt{p}$  that maps  $p$  to an element of the sphere:  $\sqrt{p} \in \mathbb{S}^{d-1}$ . Show that under this isometry, a tangent vector  $\dot{p}$  on the simplex is mapped to  $v = \dot{p}/(2\sqrt{p})$  and conclude that  $\dot{p}$  is tangent to the simplex (i.e.,  $\sum_{i \in [d]} \dot{p}_i = 0$ ) if and only if  $v$  is tangent to the sphere (i.e.,  $\sqrt{p} \perp v$ ).
14. Verify the expressions (5.18) for the (unbalanced) Fisher–Rao gradient of a functional.
15. Compute the FR and WFR gradients of the functionals listed in Examples 5.11, 5.12, and 5.13.
16. Show that the FR gradient flow (5.19) and the WFR gradient flow (5.23) maintain the property that  $\mu_t$  is a probability measure for all  $t \geq 0$ .
17. Show that (5.28) indeed follows the WFR gradient flow (5.23).

---

## Wasserstein gradient flows: applications

In the previous chapter, we developed a Riemannian structure on the space  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$  in order to define Wasserstein and WFR gradient flows. In this chapter, we use these gradient flows as optimization algorithms over the space of probability measures for various tasks arising in statistics and machine learning. Each task corresponds to choosing a specific functional  $\mathcal{F}$  over this space. In particular, akin to the notion of convexity in classical optimization (see, e.g., [Bub15]), the notion of geodesic convexity is instrumental in deriving rates of convergence.

### 6.1 Variational inference

As our first application of gradient flow theory, we consider a rich source of optimization problems over the space of measures arising from the burgeoning field of *variational inference* (VI) [JGJS99, WJ08, BKM17]. In VI, we posit access to a probability measure  $\pi$  over  $\mathbb{R}^d$  via an expression for its density, and our goal is to perform inference. A typical example arises when  $\pi$  is the posterior distribution from a Bayesian inference problem, in which case VI is also known as *variational Bayes*, and it has gradually emerged as an appealing computational counterpoint to traditional Markov chain Monte Carlo (MCMC) methods, which we study in Section 6.2.

The idea of VI is to approximate  $\pi$  with an element of a simpler class  $\mathcal{Q}$  of probability measures by solving the optimization problem

$$q_\star = \arg \min_{q \in \mathcal{Q}} \mathsf{KL}(q \parallel \pi). \quad (\text{VI})$$

Although a plethora of variants have been proposed which replace the KL divergence with other objectives<sup>1</sup>, the one we present here is particularly popular in practice. This is because typically we do not have access directly to  $\pi$  but rather to an unnormalized density  $\tilde{\pi}$ , and the unknown normalization constant  $Z = \int \tilde{\pi}$  does not affect the optimization objective in (VI).

This modelling choice also has fortuitous consequences for the convexity of the VI problem over the Wasserstein space, leading to the development of flow-based algorithms.

### 6.1.1 Convexity of the VI problem

In order to apply the gradient flow machinery to (VI), we are inexorably led to our next undertaking: studying the geodesic convexity of the KL divergence over the Wasserstein space.

Henceforth, we always assume that  $\pi$  admits a density of the form  $\pi \propto \exp(-V)$ , where  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  is called the potential function. The first observation is that the KL divergence decomposes into a sum of two functionals:

$$\mathcal{F}(\mu) := \text{KL}(\mu \parallel \pi) = \underbrace{\int \mu \log \frac{\mu}{\pi}}_{=: \mathcal{V}(\mu)} + \underbrace{\int \mu \log \mu}_{=: \mathcal{H}(\mu)} + \text{const.} \quad (6.1)$$

where  $\mathcal{V}$  is the *potential energy*,  $\mathcal{H}$  is the *entropy*, and “const.” denotes an additive constant that does not depend on  $\mu$  (and hence is irrelevant for studying properties of the gradient flow).

In Examples 5.11 and 5.12, we have already computed the Wasserstein gradients  $\nabla \mathcal{V}(\mu) = \nabla V$  and  $\nabla \mathcal{H}(\mu) = \nabla \log \mu$ . Adding these together, we obtain  $\nabla \mathcal{F}(\mu) = \nabla \log \mu + \nabla V = \nabla \log(\mu/\pi)$ . From Definition 5.14, the Wasserstein gradient flow of  $\mathcal{F}$  solves

$$\partial_t \mu_t = \text{div} \left( \mu_t \nabla \log \frac{\mu_t}{\pi} \right). \quad (6.2)$$

Leveraging (6.1), we can study the convexity of the two functionals  $\mathcal{V}$  and  $\mathcal{H}$  separately. The potential energy is straightforward.

**Theorem 6.1.** *Suppose that  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\alpha$ -convex on  $\mathbb{R}^d$ . Then, the corresponding potential energy functional  $\mathcal{V}$  defined by  $\mathcal{V}(\mu) := \int V \, d\mu$  is  $\alpha$ -geodesically convex on  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ .*

<sup>1</sup> Including the KL divergence with the order of arguments swapped, which is closer to the statistician’s concept of maximum likelihood.

*Proof.* We use the second-order condition from Section 5.2. Namely, let  $X_t \sim \mu_t$  for  $t \in [0, 1]$ , where  $(\mu_t)_{t \in [0,1]}$  is a Wasserstein geodesic; thus,  $X_t = (1-t)X_0 + tT(X_0)$  where  $T$  is the optimal transport map from  $\mu_0$  to  $\mu_1$ . We compute

$$\begin{aligned}\mathbb{W}^2\mathcal{V}(\mu_0)[T - \text{id}, T - \text{id}] &= \partial_t^2\mathcal{V}(\mu_t)|_{t=0} = \partial_t^2\mathbb{E}V(X_t)|_{t=0} \\ &= \mathbb{E}\langle T(X_0) - X_0, \nabla^2V(X_0)(T(X_0) - X_0) \rangle \\ &\geq \alpha \mathbb{E}[\|T(X_0) - X_0\|^2] \\ &= \alpha \|T - \text{id}\|_{\mu_0}^2,\end{aligned}$$

where we used the assumption  $\nabla^2V \succeq \alpha I$ .  $\square$

Next, we show that the entropy  $\mathcal{H}$  is geodesically convex. For this, we invoke the change of variables formula.

**Lemma 6.2 (Change of variables).** *Let  $\mu$  be a density on  $\mathbb{R}^d$ , let  $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$  be a diffeomorphism, and let  $\nu := T_\#\mu$ . Then,  $\nu$  has density given by*

$$\nu(T(x)) = \frac{\mu(x)}{|\det \nabla T(x)|}.$$

Recall the mnemonic for memorizing this rule: under the change of variables  $y = T(x)$ , one has  $dy = |\det \nabla T(x)| dx$ , since the Jacobian determinant  $|\det \nabla T(x)|$  measures the volume distortion of the map  $T$ . The pushforward satisfies, by definition,  $\int \varphi \circ T d\mu = \int \varphi d\nu$  for all test functions  $\varphi$ . We can write this as

$$\begin{aligned}\int \varphi(T(x)) \mu(x) dx &= \int \varphi(y) \nu(y) dy \\ &= \int \varphi(T(x)) \nu(T(x)) |\det \nabla T(x)| dx\end{aligned}$$

and Lemma 6.2 follows. In applications, we do not always know that optimal transport maps are diffeomorphisms, but nevertheless a variant of Lemma 6.2 still holds, and we refer to [Vil03, Theorem 4.8] for the technical details.

The change of variables formula furnishes the quickest proof of geodesic convexity of  $\mathcal{H}$ .

**Theorem 6.3.** *The entropy functional  $\mathcal{H}$ , given by  $\mathcal{H}(\mu) := \int \mu \log \mu$ , is geodesically convex on  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ .*

*Proof.* Again let  $(\mu_t)_{t \in [0,1]}$  be a Wasserstein geodesic and let  $T_t := (1-t) \text{id} + t T$ , so that  $\mu_t = (T_t)_\# \mu_0$ . Then, by Lemma 6.2 (which we apply blithely despite not knowing that  $T_t$  is a diffeomorphism),

$$\begin{aligned} \mathcal{H}(\mu_t) &= \int (\log \mu_t) \, d\mu_t = \int \log(\mu_t \circ T_t) \, d\mu_0 = \int \log \frac{\mu_0}{\det \nabla T_t} \, d\mu_0 \\ &= \mathcal{H}(\mu_0) - \int \log \det \nabla T_t \, d\mu_0. \end{aligned}$$

It is a standard exercise to show that  $-\log \det$  is convex over the positive definite cone, and  $t \mapsto \nabla T_t$  is affine; therefore, the composition  $t \mapsto -\log \det \nabla T_t$  is convex. In turn, it shows that  $t \mapsto \mathcal{H}(\mu_t)$  is convex, which is what we wanted to show.  $\square$

**Corollary 6.4.** *Let  $\pi \propto \exp(-V)$  be a density, where  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\alpha$ -convex. Then, the functional  $\mathcal{F} := \text{KL}(\cdot \parallel \pi)$  is  $\alpha$ -geodesically convex on  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ .*

Distributions  $\pi \propto \exp(-V)$  for which  $V$  is strongly convex are known as *strongly log-concave distributions*. Therefore, we have shown that the KL divergence w.r.t. a (strongly) log-concave measure is (strongly) geodesically convex.

For (VI), our goal is to minimize the KL divergence over a subset  $\mathcal{Q} \subseteq \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ . If  $\mathcal{Q}$  is geodesically convex (see Section 5.2), then we immediately obtain the following corollary.

**Corollary 6.5.** *Let  $\pi \propto \exp(-V)$  be a density on  $\mathbb{R}^d$ , where  $V$  is  $\alpha$ -convex. Let  $\mathcal{Q} \subseteq \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  be geodesically convex. Then,  $\text{KL}(\cdot \parallel \pi)$  is  $\alpha$ -geodesically convex over  $\mathcal{Q}$ .*

*In particular, the solution  $q_*$  to (VI) is unique.*

Recall from Lemma A.11 and Section 5.4 that  $\alpha$ -convexity implies a PL inequality, which in turn implies rapid convergence for the gradient flow:

**Corollary 6.6.** *Let  $\pi \propto \exp(-V)$  be a density on  $\mathbb{R}^d$ , where  $V$  is  $\alpha$ -convex. Let  $\mathcal{Q} \subseteq \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  be geodesically convex. Then, the Wasserstein gradient flow  $(q_t)_{t \geq 0}$  of  $\text{KL}(\cdot \parallel \pi)$  constrained to lie in  $\mathcal{Q}$  satisfies*

$$\text{KL}(q_t \parallel \pi) - \text{KL}(q_* \parallel \pi) \leq e^{-2\alpha t} \{ \text{KL}(q_0 \parallel \pi) - \text{KL}(q_* \parallel \pi) \}.$$

In the sequel, our aim is show how the constrained Wasserstein gradient flow can be implemented in several important cases.

### 6.1.2 Gaussian VI

In this section, we study the problem of *Gaussian VI*, in which the variational family  $\mathcal{Q}$  consists of all non-degenerate Gaussian measures over  $\mathbb{R}^d$ . This family is simple yet abundantly motivated: if we can approximate  $\pi \propto \exp(-V)$  by a Gaussian  $\mathcal{N}(m, \Sigma)$ , then the parameters  $(m, \Sigma)$  of the Gaussian are a reasonable guess for the mean and covariance matrix of  $\pi$ , which already suffice to construct credible regions. The *Laplace approximation* takes  $m = \theta^*$  and  $\Sigma = [\nabla^2 V(\theta^*)]^{-1}$  where  $\theta^* = \arg \min V$  is the mode of  $\pi$ . When  $\pi$  is a Bayesian posterior, the validity of this approximation can be justified in the large-sample limit by the Bernstein–von Mises theorem.

To go beyond the Laplace approximation, we can ask for the *optimal* Gaussian approximation, which is formulated as the VI problem

$$q_* = \arg \min_{q \in \text{BW}(\mathbb{R}^d)} \text{KL}(q \parallel \pi) \quad (\text{GVI})$$

where  $\text{BW}(\mathbb{R}^d)$  is the Bures–Wasserstein space introduced in Section 5.5. Note that if we had considered the KL divergence with the arguments swapped,  $q \mapsto \text{KL}(\pi \parallel q)$ , then the optimal solution is the one that matches the mean and covariance of  $\pi$ , which defeats the purpose of VI since they are precisely the parameters we are trying to compute.

Following [LCB<sup>+</sup>22], our approach to solve (GVI) is to follow the Wasserstein gradient flow constrained to lie in the Bures–Wasserstein space, see Section 5.5. By Corollary 5.20,  $\text{BW}(\mathbb{R}^d)$  is geodesically convex, and hence the guarantee of Corollary 6.6 applies. It remains to derive the form of the BW gradient flow using Theorem 5.21 in order to arrive at an implementable algorithm.

For the KL divergence  $\mathcal{F} = \text{KL}(\cdot \parallel \pi)$ ,  $\nabla \delta \mathcal{F}(q) = \nabla V + \nabla \log q$  and  $\nabla^2 \delta \mathcal{F}(q) = \nabla^2 V + \nabla^2 \log q$ . Hence,

$$\begin{aligned} \nabla_{\text{BW}} \mathcal{F}(q)(x) &= \left( \int (\nabla^2 V + \nabla^2 \log q) \, dq \right) (x - m_q) \\ &\quad + \int \nabla V \, dq + \underbrace{\int \nabla \log q \, dq}_{=0} \\ &= \left( \int \nabla^2 V \, dq - \Sigma_q^{-1} \right) (x - m_q) + \int \nabla V \, dq. \end{aligned}$$

By setting the BW gradient equal to zero, we also deduce the first-order stationarity conditions, which are both necessary and sufficient by convexity.

**Proposition 6.7.** *Suppose that  $\pi \propto \exp(-V)$ , where  $V$  is  $\alpha$ -convex for some  $\alpha > 0$ . Then, the unique minimizer  $q_\star$  in (GVI) is characterized by the conditions*

$$\int \nabla V \, dq_\star = 0 \quad \text{and} \quad \int \nabla^2 V \, dq_\star = \Sigma_{q_\star}^{-1}.$$

We can now write down the BW gradient flow using Theorem 5.22. Note that there is a slight simplification since the covariance matrix  $\Sigma_t$  cancels with the Hessian of the first variation of the entropy.

**Theorem 6.8.** *The BW gradient flow of the functional  $\text{KL}(\cdot \parallel \pi)$ , where  $\pi \propto \exp(-V)$ , is the curve  $(q_t = \mathcal{N}(m_t, \Sigma_t))_{t \geq 0}$ , where*

$$\begin{aligned} \dot{m}_t &= -\mathbb{E} \nabla V(X_t), \\ \dot{\Sigma}_t &= -\mathbb{E} \nabla^2 V(X_t) \Sigma_t - \Sigma_t \mathbb{E} \nabla^2 V(X_t) + 2I, \end{aligned} \tag{6.3}$$

and  $X_t \sim q_t$ .

To implement the gradient flow, the system of ODEs (6.3) can be discretized in time. At each iteration  $t$ , since we keep track of the mean  $m_t$  and covariance  $\Sigma_t$ , the expectations  $\mathbb{E} \nabla V(X_t)$  and  $\mathbb{E} \nabla^2 V(X_t)$  can be approximated via Monte Carlo averages by drawing samples from  $q_t = \mathcal{N}(m_t, \Sigma_t)$ , or via quadrature rules. Furthermore, [DBCS23] observed that the splitting (6.1) of the KL divergence naturally suggests a proximal gradient method for (GVI).

We mention two appealing features of the gradient flow perspective. First, it comes with principled guarantees: by mimicking optimization proofs over the Bures–Wasserstein space, the papers [LCB<sup>+</sup>22, DBCS23] translate Corollary 6.6 into non-asymptotic convergence rates for the stochastic gradient-based implementations of (6.3). Second, it readily leads to an extension to variational inference over the richer class of mixtures of Gaussians by applying either of the Gaussian particle methods from Subsection 5.8.3; see [LCB<sup>+</sup>22] for details.

### 6.1.3 Mean-field VI

Another important variational family is the class  $\mathcal{Q} = \mathcal{P}_{2,\text{ac}}(\mathbb{R})^{\otimes d}$  of product measures over  $\mathbb{R}^d$ , in which case the problem (VI) is known as *mean-field VI*:

$$q_\star = \arg \min_{q \in \mathcal{P}_{2,\text{ac}}(\mathbb{R})^{\otimes d}} \text{KL}(q \parallel \pi). \tag{MFVI}$$

This form of VI has its roots in product measure approximations of spin systems from statistical physics and can be motivated statistically by the desire to compute integrals of separable test functions  $\phi(x) = \sum_{i=1}^d \phi_i(x_i)$  against the posterior.

Since the family of product measures is geodesically convex (Exercise 2 in Chapter 5), Corollary 6.6 once again shows that the constrained Wasserstein gradient flow converges rapidly. One can also show that for a functional  $\mathcal{F}$ , the component of the Wasserstein gradient  $\nabla \mathcal{F}(\mu)$  which is tangential to the space of product measures takes the form

$$x \mapsto \sum_{i=1}^d \left( \int \nabla \mathcal{F}(\mu)(x) \mu(dx_1, \dots, dx_{i-1}, dx_{i+1}, \dots, dx_d) \right) e_i, \quad (6.4)$$

where  $e_i$  is the  $i$ -th standard basis vector; see Exercise 2.

Note that the particle approach of Subsection 5.8.1 does not apply because the Wasserstein gradient of the KL divergence is not defined for discrete measures. One way to circumvent this issue is via stochastic dynamics, see Subsection 6.3. In this section, we instead describe the approach of [JCP24] which, similarly to the previous subsection, is based on finite-dimensional parameterization. Key to this approach is that for mean-field VI, the measure  $q_*$  is not intrinsically high-dimensional due to the product structure, which allows for efficient parameterization.

The idea is to parameterize an element  $q \in \mathcal{P}_{2,\text{ac}}(\mathbb{R})^{\otimes d}$  via the Brenier map  $T_{\rho \rightarrow q}$  from the standard Gaussian measure  $\rho$ . Since both  $\rho$  and  $q$  are product measures, one sees that the transport map is separable, i.e., it is of the form  $T_{\rho \rightarrow q}(x) = (T_1(x_1), \dots, T_d(x_d))$  for some *univariate* (and increasing, by Theorem 1.14) maps  $T_i : \mathbb{R} \rightarrow \mathbb{R}$ . However, the class of such maps is still infinite-dimensional and needs to be further restricted for implementation purposes.

To do so, we take a finite family  $\mathcal{M}$  of optimal transport maps—called the *dictionary*—and take as our eventual family of maps the set  $\text{cone}(\mathcal{M})$  of all conic combinations of elements of  $\mathcal{M}$ :

$$\text{cone}(\mathcal{M}) = \left\{ \sum_{T \in \mathcal{M}} \lambda_T T \mid \lambda \in \mathbb{R}_+^{\mathcal{M}} \right\}.$$

We denote  $T^\lambda := \sum_{T \in \mathcal{M}} \lambda_T T$  and  $\rho^\lambda := (T^\lambda)_\# \rho$ . This set is now finite-dimensional, and in fact is parameterized by the positive orthant  $\mathbb{R}_+^{\mathcal{M}}$ . Therefore, we can now optimize the functional  $\lambda \mapsto \text{KL}(\rho^\lambda \parallel \pi)$  over  $\mathbb{R}_+^{\mathcal{M}}$ . Note that we have replaced our original variational family  $\mathcal{P}_{2,\text{ac}}(\mathbb{R})^{\otimes d}$

with the smaller set  $\text{cone}(\mathcal{M})\# \rho$ , but the hope is that for an appropriate choice of  $\mathcal{M}$ , the family  $\text{cone}(\mathcal{M})\# \rho$  is expressive enough to approximately capture all of  $\mathcal{P}_{2,\text{ac}}(\mathbb{R})^{\otimes d}$ . In [JCP24], this is indeed shown to be the case, e.g., when  $\mathcal{M}$  consists of increasing and piecewise linear functions which act on a single coordinate, and that the total size of  $\mathcal{M}$  is polynomially bounded in the problem parameters.

We have the following lemma, whose proof we leave as Exercise 4.

**Lemma 6.9.** *Assume that  $\mathcal{M}$  consists of maps  $T$  which are separable, in the sense that  $T(x) = (T_1(x_1), \dots, T_d(x_d))$  for increasing univariate maps  $T_1, \dots, T_d : \mathbb{R} \rightarrow \mathbb{R}$ . Then,  $\text{cone}(\mathcal{M})\# \rho$  is geodesically convex.*

Moreover, the map  $(\mathbb{R}_+^{\mathcal{M}}, \|\cdot\|_Q) \rightarrow (\text{cone}(\mathcal{M})\# \rho, W_2)$ ,  $\lambda \mapsto \rho^\lambda$  is an isometry, where  $\|x\|_Q^2 := \langle x, Q x \rangle$  and  $Q$  is the  $|\mathcal{M}| \times |\mathcal{M}|$  matrix with entries

$$Q_{T,T'} := \langle T, T' \rangle_\rho, \quad T, T' \in \mathcal{M}. \quad (6.5)$$

The first statement of Lemma 6.9 shows that under strong log-concavity for  $\pi$ , the problem of minimizing  $\text{KL}(\cdot \parallel \pi)$  over  $\text{cone}(\mathcal{M})\# \rho$  is a strongly convex problem in the Wasserstein geometry, and in particular, that Corollary 6.6 applies. The second statement shows that implementing the Wasserstein gradient flow in this case amounts to implementing a *Euclidean* gradient flow up to preconditioning by  $Q^{-1}$ . Indeed, the isometry implies that the Wasserstein gradient flow of the KL divergence over  $\text{cone}(\mathcal{M})\# \rho$  is equivalent to the gradient flow of  $\lambda \mapsto \text{KL}(\rho^\lambda \parallel \pi)$  over  $(\mathbb{R}_+^{\mathcal{M}}, \|\cdot\|_Q)$ , and the gradient operator in the  $\|\cdot\|_Q$  norm is simply the Euclidean gradient operator premultiplied by the matrix  $Q^{-1}$ . In particular, we can write down the resulting algorithm explicitly.

**Theorem 6.10.** *The Wasserstein gradient flow of  $\text{KL}(\cdot \parallel \pi)$  restricted to  $\text{cone}(\mathcal{M})\# \rho$  is given by  $(\rho^{\lambda(t)})_{t \geq 0}$ , where*

$$\dot{\lambda}_T = - \sum_{T' \in \mathcal{M}} (Q^{-1})_{T,T'} \int [\langle \nabla V \circ T^\lambda, T' \rangle - \langle (\nabla T^\lambda)^{-1}, \nabla T' \rangle] d\rho$$

and the matrix  $Q$  is given in (6.5).

*Proof.* Lemma 6.9 implies that the Wasserstein gradient flow is given by  $\dot{\lambda}_t = -Q^{-1} \nabla_\lambda \text{KL}(\rho^{\lambda_t} \parallel \pi)$ . We compute

$$\partial_{\lambda_T} \mathcal{V}(\rho^\lambda) = \partial_{\lambda_T} \int V \circ T^\lambda d\rho = \int \langle \nabla V \circ T^\lambda, T \rangle d\rho$$

and

$$\begin{aligned}\partial_{\lambda_T} \mathcal{H}(\rho^\lambda) &= \partial_{\lambda_T} \int \rho^\lambda \log \rho^\lambda \\ &= -\partial_{\lambda_T} \int \log \det \nabla T^\lambda \, d\rho \\ &= - \int \langle (\nabla T^\lambda)^{-1}, \nabla T \rangle \, d\rho,\end{aligned}$$

where we used respectively Lemma 6.2 and a classical result of matrix calculus to compute the gradient of the log det function.  $\square$

As in the previous subsection, the expectations can be estimated using Monte Carlo averages.

#### 6.1.4 Stein variational gradient descent

We now ask whether we can minimize the KL divergence over the family  $\mathcal{Q}$  of empirical measures—measures of the form  $\frac{1}{N} \sum_{i=1}^N \delta_{x^i}$ . Unlike the preceding two subsections, this class is arbitrarily expressive: as  $N \rightarrow \infty$ , we can always find a sequence of empirical measures that converges weakly to  $\pi$ , as a consequence of the law of large numbers; see Chapter 2.

Recall from Section 5.8 that the Wasserstein gradient flow of a functional  $\mathcal{F}$  over the full Wasserstein space can be represented via

$$\dot{X}_t = -\nabla \mathcal{F}(\mu_t)(X_t), \quad X_t \sim \mu_t, \quad (6.6)$$

provided that  $\nabla \mathcal{F}(\mu_t)$  makes sense. Note also that unlike the previous two subsections, we do not have to do anything special to ensure that  $\mu_t$  remains an empirical measure for all  $t \geq 0$ : the gradient flow (6.6) *automatically* preserves the space of empirical measures.

If we specialize this to  $\mathcal{F} = \text{KL}(\cdot \parallel \pi)$ , then (6.2) leads to

$$\dot{X}_t = -\nabla \log \frac{\mu_t}{\pi}(X_t), \quad X_t \sim \mu_t. \quad (\text{WGF})$$

Unfortunately, the expression  $\nabla \log(\mu_t/\pi)$  does *not* make sense when  $\mu_t$  is an empirical measure.

The next idea that springs to mind is to replace  $\mu_t$  in (WGF) with a smoothed version via a kernel density estimator (KDE). More precisely, let us initialize  $N$  particles  $X_0^1, \dots, X_0^N$ , and let  $k : \mathbb{R}^d \rightarrow \mathbb{R}_+$  be a symmetric kernel with  $\int k = 1$ . At time  $t$ , we replace  $\mu_t$  by the KDE  $\hat{\mu}_t = \frac{1}{N} \sum_{i=1}^N k(\cdot - X_t^i)$ , which leads to an interacting system of particles:

$$\dot{X}_t^i = -\nabla V(X_t^i) - \left[ \nabla \log \frac{1}{N} \sum_{j=1}^N k(\cdot - X_t^j) \right] (X_t^i), \quad i \in [N].$$

In the mean-field limit  $N \rightarrow \infty$ , we expect that  $\frac{1}{N} \sum_{j=1}^N k(\cdot - X_t^j) \rightarrow \int k(\cdot - y) \mu_t(dy) = k \star \mu_t$ , where  $\mu_t = \text{law}(X_t)$ , leading to the dynamics

$$\dot{X}_t = -\nabla V(X_t) - \nabla \log(k \star \mu_t)(X_t)$$

or equivalently

$$\partial_t \mu_t = \text{div}(\mu_t (\nabla V + \nabla \log(k \star \mu_t))).$$

However, these dynamics do not necessarily converge to the target  $\pi \propto \exp(-V)$ . This issue can be fixed by introducing a *bandwidth* parameter to the kernel  $k$  which tends to zero as  $N \rightarrow \infty$ , but there is an alternative which stems from the RKHS literature (recall the discussion in Subsection 2.8.2) which we describe next. The *Stein variational gradient descent* (SVGD) algorithm, due to [LW16], manages to use a *fixed* kernel  $k$  but still admits  $\pi$  as a stationary solution.

We define the integral operator

$$\mathcal{K}_\mu : f \mapsto \int f(y) k(\cdot - y) \mu(dy),$$

and we follow the dynamics

$$\partial_t \mu_t = \text{div}\left(\mu_t \mathcal{K}_{\mu_t} \nabla \log \frac{\mu_t}{\pi}\right), \quad (\text{SVGD})$$

where the integral operator acts on vector fields coordinate-wise. Clearly these dynamics leave  $\pi$  stationary. Let us calculate the effect of the integral operator above. First,

$$\mathcal{K}_\mu \nabla \log \frac{1}{\pi} = \mathcal{K}_\mu \nabla V = \int \nabla V(y) k(\cdot - y) \mu(dy). \quad (6.7)$$

For the other term, we use integration by parts:

$$\begin{aligned} \mathcal{K}_\mu \nabla \log \mu &= \int \frac{\nabla \mu(y)}{\mu(y)} k(\cdot - y) \mu(dy) = \int \nabla \mu(y) k(\cdot - y) dy \\ &= \int \nabla k(\cdot - y) \mu(dy). \end{aligned} \quad (6.8)$$

Both (6.7) and (6.8) are expectations w.r.t.  $\mu$ , so they can be replaced by empirical averages over  $N$  particles. This leads to the algorithm

$$\dot{X}_t^i = -\frac{1}{N} \sum_{j=1}^N [k(X_t^i - X_t^j) \nabla V(X_t^j) + \nabla k(X_t^i - X_t^j)].$$

Although SVGD has been an active subject of research, many theoretical questions regarding its convergence remain open.

Recall that we motivated SVGD with the idea of approximating  $\pi$  by an empirical measure, noting that the class of empirical measures over  $N$  atoms is arbitrarily expressive as  $N \rightarrow \infty$ . But if our ultimate goal is fidelity with respect to  $\pi$ , we may as well ask if we can directly output samples from  $\pi$  itself. In the next section, we discuss the problem of sampling via MCMC methods, which can be viewed as *stochastic* implementations of the Wasserstein gradient flow.

## 6.2 Sampling

One of the most compelling applications of the theory of Wasserstein gradient flows is to provide a geometric interpretation of the Langevin diffusion, as put forth in the seminal work of Jordan, Kinderlehrer, and Otto [JKO98]. As before, let  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  be a smooth potential with  $\int \exp(-V) < \infty$  and let  $\pi$  denote the probability measure over  $\mathbb{R}^d$  with density  $\pi \propto \exp(-V)$ .

Suppose that we wish to *sample* from the distribution  $\pi$ . In other words, we want to design an algorithm for producing a random variable whose law is close to  $\pi$ . For example,  $\pi$  could be the posterior distribution in a Bayesian inference problem, in which case basic downstream tasks such as constructing credible regions or point estimates are often intractable in non-conjugate models. Nevertheless, we can usually solve these tasks approximately and efficiently, given a subroutine for drawing approximate samples from the posterior. Beyond the application to computational Bayesian statistics, sampling also plays an important role in scientific computing through Monte Carlo integration and for the design of randomized algorithms.

The predominant approach to this problem, dubbed Markov chain Monte Carlo (MCMC), is to design a Markov chain whose unique stationary distribution is, or at least is close to, the target  $\pi$ . When  $\pi$  admits a positive and smooth density, as we assume in this section,

then we can write  $\pi \propto \exp(-V)$  without loss of generality (with  $V = \log(1/\pi) + \text{const.}$ ). In this case, a canonical MCMC algorithm is obtained by discretizing the *Langevin diffusion*, which is the solution to the stochastic differential equation (SDE)

$$dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t,$$

where  $(B_t)_{t \geq 0}$  is a standard Brownian motion. As soon as  $\nabla V$  is, e.g., Lipschitz continuous, there is a unique strong solution to this SDE for any prescribed initial condition, and its stationary distribution is  $\pi$ .

In the next section, we show that when we track the evolution of the marginal law  $\mu_t := \text{law}(X_t)$  of the Langevin diffusion, then  $(\mu_t)_{t \geq 0}$  follows the Wasserstein gradient flow of the KL divergence  $\text{KL}(\cdot \parallel \pi)$ . More broadly, this story is the starting point of a fruitful literature which has blossomed in recent years on an optimization perspective (i.e., the application of optimization algorithms such as gradient flows) on the problem of sampling.

### 6.2.1 The Langevin diffusion as a Wasserstein gradient flow

To spoil the surprise, the fundamental reason why (6.2) admits a stochastic implementation is because we can rewrite

$$\text{div}(\mu_t \nabla \log \mu_t) = \text{div}\left(\mu_t \frac{\nabla \mu_t}{\mu_t}\right) = \text{div}(\nabla \mu_t) = \Delta \mu_t$$

where  $\Delta f = \sum_{i=1}^d \partial_i^2 f$  is the *Laplacian* of  $f$ . On the other hand, second-order parabolic PDEs—the heat equation  $\partial_t \mu_t = \Delta \mu_t$  being the most fundamental example—classically describe the evolution in law of stochastic differential equations driven by Brownian motion. The rest of this subsection aims to make this connection precise.

Using the computation above, we rewrite the Wasserstein gradient flow of the KL divergence, given in (6.2), as

$$\partial_t \mu_t = \Delta \mu_t + \text{div}(\mu_t \nabla V). \quad (6.9)$$

We next compute the marginal evolution of the Langevin diffusion in order to compare with (6.9). The usual method for doing so is to use Itô's formula from stochastic calculus, but we instead proceed more informally. First, let us condition on  $X_0 = x_0$ , and write the Langevin diffusion in integral form.

$$X_t = x_0 - \int_0^t \nabla V(X_s) \, ds + \sqrt{2} B_t.$$

Recall also that  $B_t \sim \mathcal{N}(0, tI)$ . In particular,  $\|\int_0^t \nabla V(X_s) \, ds\| = O_{\mathbb{P}}(t)$  and  $\|B_t\| = O_{\mathbb{P}}(t^{1/2})$  for small  $t$ , so the Brownian motion term dominates and  $\|X_t - x_0\| = O_{\mathbb{P}}(t^{1/2})$ . By duality, to calculate the evolution of  $\mu_t$ , it suffices to compute the evolution of the expectation of any test function  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ . A Taylor expansion yields

$$\begin{aligned} \varphi(X_t) &= \varphi\left(x_0 - \int_0^t \nabla V(X_s) \, ds + \sqrt{2} B_t\right) \\ &= \varphi(x_0) + \left\langle \nabla \varphi(x_0), - \int_0^t \nabla V(X_s) \, ds + \sqrt{2} B_t \right\rangle \\ &\quad + \frac{1}{2} \left\langle \nabla^2 \varphi(x_0), \left( - \int_0^t \nabla V(X_s) \, ds + \sqrt{2} B_t \right)^{\otimes 2} \right\rangle + O_{\mathbb{P}}(t^{3/2}). \end{aligned}$$

The first-order term equals

$$-t \langle \nabla \varphi(x_0), \nabla V(x_0) \rangle + \sqrt{2} \langle \nabla \varphi(x_0), B_t \rangle + O_{\mathbb{P}}(t^{3/2}).$$

The second-order term equals

$$\langle \nabla^2 \varphi(x_0) B_t, B_t \rangle + O_{\mathbb{P}}(t^{3/2}).$$

Therefore,

$$\begin{aligned} \varphi(X_t) &= \varphi(x_0) - t \langle \nabla \varphi(x_0), \nabla V(x_0) \rangle + \sqrt{2} \langle \nabla \varphi(x_0), B_t \rangle \\ &\quad + \langle \nabla^2 \varphi(x_0) B_t, B_t \rangle + O_{\mathbb{P}}(t^{3/2}). \end{aligned}$$

Taking expectations and using  $\mathbb{E}[B_t] = 0$ ,  $\mathbb{E}[B_t B_t^T] = tI$ ,

$$\begin{aligned} \mathbb{E}\varphi(X_t) &= \varphi(x_0) + t (\text{tr } \nabla^2 \varphi(x_0) - \langle \nabla \varphi(x_0), \nabla V(x_0) \rangle) + O_{\mathbb{P}}(t^{3/2}) \\ &= \varphi(x_0) + t (\Delta \varphi(x_0) - \langle \nabla \varphi(x_0), \nabla V(x_0) \rangle) + O_{\mathbb{P}}(t^{3/2}). \end{aligned}$$

Subtracting  $\varphi(x_0)$ , dividing by  $t$ , and letting  $t \searrow 0$ ,

$$\partial_t \mathbb{E}\varphi(X_t) \Big|_{t=0} = \Delta \varphi(x_0) - \langle \nabla \varphi(x_0), \nabla V(x_0) \rangle. \quad (6.10)$$

In the language of Markov semigroup theory, we have computed the *generator* of the Langevin diffusion to be the second-order differential operator  $\mathcal{L}$ , defined by  $\mathcal{L}\varphi := \Delta \varphi - \langle \nabla \varphi, \nabla V \rangle$ .

More generally, by first conditioning on the value of  $X_t$  and using the Markov property and (6.10), it holds that

$$\partial_t \mathbb{E} \varphi(X_t) = \mathbb{E} \mathcal{L} \varphi(X_t).$$

Expressed in terms of the marginal law  $\mu_t$ , it reads

$$\int \varphi \partial_t \mu_t = \int (\Delta \varphi - \langle \nabla \varphi, \nabla V \rangle) d\mu_t.$$

In order to identify an equation for  $\partial_t \mu_t$ , we must compute the adjoint (w.r.t. Lebesgue measure) of  $\mathcal{L}$ . This is accomplished through integration by parts, which shows that the right-hand side equals

$$\int \varphi (\Delta \mu_t + \operatorname{div}(\mu_t \nabla V)).$$

We have established the following theorem.

**Theorem 6.11 (Fokker–Planck equation).** *The marginal law  $\mu_t := \operatorname{law}(X_t)$  of the Langevin diffusion with potential  $V$  is given by the solution to the Fokker–Planck equation*

$$\partial_t \mu_t = \Delta \mu_t + \operatorname{div}(\mu_t \nabla V).$$

Comparing with (6.9), it yields:

**Corollary 6.12.** *The marginal law of the Langevin diffusion with potential  $V$  is the Wasserstein gradient flow of  $\operatorname{KL}(\cdot \parallel \pi)$ , where  $\pi$  has density proportional to  $\exp(-V)$ .*

As a special case when  $V = 0$ , we also obtain the following corollary.

**Corollary 6.13.** *If  $(\mu_t)_{t \geq 0}$  is the marginal law of a (rescaled) Brownian motion  $(\sqrt{2} B_t)_{t \geq 0}$ , then  $(\mu_t)_{t \geq 0}$  solves the heat equation  $\partial_t \mu_t = \Delta \mu_t$ , and it is the Wasserstein gradient flow of the entropy functional  $\mathcal{H}$ .*

We showed in Corollary 6.6 that the strong log-concavity of  $\pi$  implies rapid convergence of the Wasserstein gradient flow of the KL divergence. Therefore, we immediately obtain the following elegant convergence result for the Langevin diffusion.

**Corollary 6.14.** *Let  $\pi$  be an  $\alpha$ -strongly log-concave measure, and let  $(\mu_t)_{t \geq 0}$  denote the marginal law of the Langevin diffusion with stationary distribution  $\pi$ . Then,*

$$\operatorname{KL}(\mu_t \parallel \pi) \leq e^{-2\alpha t} \operatorname{KL}(\mu_0 \parallel \pi).$$

To conclude this section, we take stock of the situation at hand. Recall that the Wasserstein gradient flow of the KL divergence can be implemented via the deterministic evolution (WGF). On the other hand, in this subsection, we started with the Langevin diffusion, which is a stochastic evolution:

$$dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t. \quad (\text{LD})$$

We showed in Theorem 6.11 that the marginal law  $\mu_t := \text{law}(X_t)$  evolves according to the Fokker–Planck equation

$$\partial_t \mu_t = \Delta \mu_t + \text{div}(\mu_t \nabla V), \quad (\text{FP})$$

and moreover that this evolution coincides with the Wasserstein gradient flow of  $\text{KL}(\cdot \parallel \pi)$ . Note that (FP) is strictly coarser than (LD) because (LD) also includes information about correlations between different time points of the stochastic process.

Ultimately, we have the equivalence (FP)  $\Leftrightarrow$  (LD)  $\Leftrightarrow$  (WGF) in the sense that they correspond to the same curve on the space of probability measures—clearly at the particle level, they are different—but these three differing perspectives provide new avenues for algorithm design and theoretical study.

### 6.2.2 Sampling as optimization

The gradient flow perspective on the Langevin diffusion is the starting point of a flourishing literature on an optimization perspective on sampling. In this subsection, we study the basic properties of KL divergence minimization as an optimization problem, inspired by the treatment in [Wib18]. Then, in the next subsection, we provide a glimpse of the recent impact of this perspective on the theory of log-concave sampling. We refer to the monograph [Che24] for a detailed exposition.

Let  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  be a potential which is  $\alpha$ -convex,  $\alpha > 0$ , and recall that our goal is to output a sample from the target  $\pi \propto \exp(-V)$ . Corollary 6.14 then ensures that the continuous-time Langevin diffusion converges rapidly to  $\pi$ , but in order to obtain algorithmic guarantees we must *discretize* the process, and for this we also impose the dual assumption of smoothness,  $\nabla^2 V \preceq \beta I$ , to ensure stability.

Error estimates for discretizations of the Langevin diffusion are by now well-established. Under our assumptions of strong convexity

and smoothness of  $V$ , non-asymptotic convergence guarantees can be established for the Euler–Maruyama scheme

$$X_{k+1} = X_k - h \nabla V(X_k) + \mathcal{N}(0, 2hI), \quad k = 0, 1, 2, \dots, \quad (6.11)$$

which parallels the complexity theory for optimization [Nes18]. Unlike the situation in optimization, however, the discretization (6.11) is asymptotically biased: the stationary distribution of the Markov chain (6.11) does not equal the target  $\pi$ . This leads to slower rates of convergence, as the step size  $h$  must be chosen small to mitigate the bias. We now explain how an optimization perspective sheds light on the source of asymptotic bias and suggests a proximal scheme for removing it.

As in (6.1), we write the KL divergence as the sum of the potential energy and the entropy,

$$\text{KL}(\mu \parallel \pi) = \int V \, d\mu + \int \mu \log \mu + \text{const.} = \mathcal{V}(\mu) + \mathcal{H}(\mu) + \text{const.}$$

The problem of sampling from  $\pi$  is cast as the minimization of this objective functional over  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ , and we have already begun studying its properties. Namely:

- The potential energy is strongly convex and smooth,  $\alpha \leq \mathbb{W}^2 \mathcal{V} \leq \beta$ . We proved the lower bound in Theorem 6.1, and the upper bound follows by a similar computation.
- The entropy is convex,  $0 \leq \mathbb{W}^2 \mathcal{H}$ . We proved this as Theorem 6.3. However, the entropy is *non-smooth*.<sup>2</sup>

The situation at hand is one that is commonly encountered in optimization, known as *composite optimization*: minimize the sum  $f + g$ , where  $f$  is (strongly) convex and smooth, and  $g$  is convex but non-smooth. The prototypical example is the  $\ell_1$ -penalized least squares objective (or LASSO),  $\theta \mapsto \|y - X\theta\|^2 + \lambda \|\theta\|_1$ . In optimization theory, the canonical algorithm designed for such problems is the *proximal gradient* method

$$x_{k+1} = \text{prox}_{hg}(x_k - h \nabla f(x_k)), \quad (6.12)$$

where

$$\text{prox}_{hg}(y) := \arg \min_{x \in \mathbb{R}^d} \left\{ hg(x) + \frac{1}{2} \|y - x\|^2 \right\}. \quad (6.13)$$

---

<sup>2</sup> One can in fact show that  $\mathbb{W}^2 \mathcal{H}(\mu)[v, v] = \int \|\nabla v - I\|_{\text{HS}}^2 \, d\mu$  and there is no constant  $C > 0$  such that  $\mathbb{W}^2 \mathcal{H}(\mu)[v, v] \leq C \|v\|_\mu^2$  for all  $v \in T_\mu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ .

When  $g$  has a simple structure, such as the  $\ell_1$  norm, then the proximal mapping sometimes admits a closed-form solution.

The proximal gradient algorithm is unbiased, meaning that its only fixed points are minimizers of  $f + g$ . Moreover, one can show that the iteration (6.12) converges at the same rate that gradient descent would for a convex and smooth objective, despite the non-smoothness of  $g$ .

More broadly, we have introduced two discretization schemes: gradient descent, and the proximal step (6.13). Each has its relative merits. Whereas gradient descent is cheaper to implement (especially for functions which do not have a simple structure like  $\|\cdot\|_1$ ), the proximal scheme converges even without smoothness. Therefore, we are motivated to apply one discretization method to  $f$ , and the other to  $g$ ; this is known as a *splitting scheme*. Not all splitting schemes are unbiased, however, and the combination of gradient descent and the proximal map is an especially auspicious match.<sup>3</sup>

With these principles from optimization in mind, let us now consider the situation for sampling. The discretization (6.11) can be viewed as the splitting scheme

$$\begin{aligned} X_{k+1/2} &= X_k - h \nabla V(X_k), \\ X_{k+1} &= X_{k+1/2} + \mathcal{N}(0, 2hI). \end{aligned}$$

The two steps correspond, respectively, to *gradient descent*<sup>4</sup> for  $\mathcal{V}$  and the *gradient flow* of  $\mathcal{H}$ ; the latter statement is Corollary 6.13. The combination of gradient descent and gradient flow does not produce an unbiased splitting scheme.

The intuition from optimization suggests to replace the gradient flow for  $\mathcal{H}$  with the proximal map for  $\mathcal{H}$ . Generalizing the definition (6.13) to the Wasserstein space, we arrive at

$$\text{prox}_{h\mathcal{H}}(\mu) := \arg \min_{\nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)} \left\{ h\mathcal{H}(\nu) + \frac{1}{2} W_2^2(\mu, \nu) \right\}. \quad (6.14)$$

(In fact, the proximal operator on the Wasserstein space was the device through which [JKO98] first made precise the Wasserstein gradient flow interpretation of the Langevin diffusion in Subsection 6.2.1.) The resulting proximal gradient algorithm on the Wasserstein space can indeed

<sup>3</sup> In numerical analysis, these two discretizations are so-called “adjoints” to each other, see [Wib18, Appendix B].

<sup>4</sup> Check that if  $\mu_k = \text{law}(X_k)$ ,  $\mu_{k+1/2} = \text{law}(X_{k+1/2})$ , and  $h \leq 1/\beta$ , then  $\mu_{k+1/2} = \exp_{\mu_k}(-h \nabla V(\mu_k))$ , which is the Riemannian analogue of gradient descent.

be shown to converge rapidly to  $\pi$  [SKL20]. However, the proximal map (6.14) is in general intractable, so the proximal gradient algorithm is merely wishful thinking.

Actually, it is not just  $\text{prox}_{h\mathcal{H}}$  that is intractable; gradient *descent* on  $\mathcal{H}$  is also impractical because it requires knowing the entire probability density (this is the motivation for our discussion of SVGD in Section 6.1.4). It seems that the only operation we can reasonably implement for minimizing  $\mathcal{H}$  is the gradient *flow*, due to the fortuitous link with Brownian motion. This can be considered both as a blessing and curse. The blessing is that the routine we can implement—gradient flow—succeeds despite the non-smoothness of  $\mathcal{H}$ , which explains why sampling is possible at all. The curse, however, is that the gradient flow is “mismatched” with our discretization for  $\mathcal{V}$ , and hence the discretization (6.11) incurs asymptotic bias.

This is perhaps representative of the subject as a whole: although optimization theory suggests a huge number of algorithmic paradigms which we hope to port over to the world of sampling, the execution of these ideas requires care. In the next subsection, we survey a number of examples in which this philosophy has been successfully carried out, including a surprising “proximal” algorithm for sampling.

### 6.2.3 Some recent developments

#### Algorithms

Open any modern book on convex optimization to find a formidable arsenal of methods: coordinate descent, gradient descent, interior point, mirror descent, Newton’s method, Nesterov’s fast gradient method, proximal gradient, stochastic gradient descent, etc. In recent years, a substantial research effort has been devoted to developing sampling analogues of all of these methods and more, of which we describe only a select few.

Our first example is the aforementioned proximal algorithm for sampling. Since it is easier to motivate *a posteriori*, we begin by defining the algorithm. Augment the target distribution  $\pi$  to form a distribution  $\boldsymbol{\pi}$  over  $\mathbb{R}^d \times \mathbb{R}^d$  with density given by

$$\boldsymbol{\pi}(x, y) \propto \exp\left(-V(x) - \frac{1}{2h} \|y - x\|^2\right).$$

The *proximal sampler*, introduced in [LST21], applies Gibbs sampling to the new target  $\boldsymbol{\pi}$ . Explicitly, repeat the following steps:

1. Given  $X = x$ , resample  $Y \sim \pi^{Y|X=x} = \mathcal{N}(x, hI)$ .
2. Given  $Y = y$ , resample  $X \sim \pi^{X|Y=y}$ , where the conditional distribution  $\pi^{X|Y=y}$ , called the *restricted Gaussian oracle* (RGO), has density  $\pi^{X|Y=y}(x) \propto \exp(-V(x) - \frac{1}{2h} \|y - x\|^2)$ .

A few simple properties can be verified immediately. First, the  $X$ -marginal of  $\pi$  is the original target  $\pi \propto \exp(-V)$ , so it suffices to sample from the augmented target. Second, the proximal sampler is unbiased—its stationary distribution is  $\pi$ —because Gibbs sampling is so. As stated, however, it is still an idealized algorithm, since it is unclear how to implement step two. But notice the following analogy: if minimizing  $V$  (optimization) corresponds to sampling from  $\pi \propto \exp(-V)$  (sampling), then computing the proximal map (6.13) for  $V$  corresponds to sampling from  $\pi^{X|Y=y}$ ; it is in this sense that the proximal sampler resembles a “proximal” algorithm for sampling.

But perhaps the most convincing justification is based on the adage “if it looks like a duck...”: the convergence analyses in [LST21, CCSW22] show that the convergence rates for the proximal sampler *exactly* replicate the rates for the proximal point method from convex optimization. Through careful implementations of the RGO, the proximal sampler has played an essential role in extending sampling guarantees to both non-log-concave and non-log-smooth settings [FYC23, AC24].

Our next example is the adaptation of mirror descent. Recall that the *mirror descent* algorithm for minimizing  $V$ , originally introduced in [NY83] for optimization w.r.t. non-Euclidean norms, starts by choosing a strictly convex function (the “mirror map”)  $\phi : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  and iterating

$$\nabla\phi(x_{k+1}) = \nabla\phi(x_k) - h \nabla V(x_k), \quad k = 0, 1, 2, \dots.$$

It turns out that the “mirror” analogue of (LD) is the so-called *mirror Langevin diffusion* [ZPFP20, CLGL<sup>+</sup>20b]:

$$Y_t := \nabla\phi(X_t), \quad dY_t = -\nabla V(X_t) dt + \sqrt{2 \nabla^2 \phi(X_t)} dB_t.$$

The mirror Langevin diffusion can also be suitably interpreted as a Wasserstein “mirror” flow of the KL divergence.<sup>5</sup> An important special case is obtained when  $V$  is strictly convex and we take the mirror

---

<sup>5</sup> More specifically, it is the gradient flow of  $\text{KL}(\cdot \parallel \pi)$  with respect to the Wasserstein geometry induced by the Hessian metric induced by  $\phi$ .

map  $\phi = V$ , leading to the *Newton–Langevin diffusion*—the sampling analogue of Newton’s method:

$$Y_t := \nabla V(X_t), \quad dY_t = -Y_t dt + \sqrt{2 \nabla^2 V(X_t)} dB_t.$$

The Newton–Langevin diffusion inherits some appealing properties of Newton’s method, such as its affine invariance. However, discretization of the Newton–Langevin diffusion (or of the more general mirror Langevin diffusion) is currently less well-understood than for (LD).

Finally, our last example is the adaptation of Nesterov’s *accelerated gradient descent* [Nes83], which is an optimal first-order method for convex smooth minimization. The corresponding SDE system, called the *underdamped* (or *kinetic*) *Langevin diffusion*, dates back at least to [Kol34] and is given by

$$dX_t = P_t dt, \quad dP_t = -\nabla V(X_t) dt - \gamma P_t dt + \sqrt{2\gamma} dB_t.$$

Here,  $P$  represents a momentum variable, and  $\gamma > 0$  is the “friction” parameter. This diffusion has already formed the basis for numerous state-of-the-art guarantees for log-concave sampling. But in the world of optimization, Nesterov’s method is best known for improving the complexity of strongly convex and smooth minimization to  $\tilde{O}(\sqrt{\kappa})$ , where  $\kappa$  is the “condition number”. This remarkable result, which saves a factor of  $\sqrt{\kappa}$  over the basic rate for gradient descent, has been dubbed the *acceleration* phenomenon, and it remains an intriguing open question to establish such a phenomenon for sampling.

### Complexity

Since the work of [NY83], a major goal of optimization research has been to precisely characterize the minimax complexity of optimization over various function classes and oracle models. With the advent of this mindset to MCMC, it became natural to do the same for sampling, starting with the non-asymptotic upper bounds in works such as [Dal17, DM17, CB18, DMM19]. The similarities are striking: both fields consider similar choices for the function classes (e.g., strongly convex and smooth functions) and for the oracle models. This connection also motivated the search for oracle *lower bounds* which could certify the optimality of our existing algorithms. This has proven to be a challenging problem, with some modest progress made recently.

Another example of the transfer of ideas is the development of a sampling analogue of “approximate first-order stationarity” which

provides an alternative approach to the quantitative study of sampling in general non-log-concave settings [BCE<sup>+</sup>22].

Despite rapid progress in this direction, there are still many fundamental unresolved questions regarding the complexity of log-concave sampling. We refer to the monograph [Che24] for an introduction to this active field and for further references.

### 6.3 Interacting particle systems

The SDE systems we encountered in Section 6.2 all involve evolving a single particle (possibly over an expanded state space) at a time. More generally, we can consider an *interacting system* of particles, either deterministic or stochastic. This is highly relevant because as we discussed in Subsection 5.8.1, Wasserstein gradient flows initialized at discrete measures can be implemented using mean-field interacting particle systems. Indeed, recall that if we initialize the Wasserstein gradient flow (6.6) at

$$\mu_0^N = \frac{1}{N} \sum_{j=1}^N \delta_{X_0^j},$$

then  $\mu_t$  is given by the empirical measure

$$\mu_t^N = \frac{1}{N} \sum_{j=1}^N \delta_{X_t^j},$$

where

$$\dot{X}_t^i = -\nabla \mathcal{F}\left(\frac{1}{N} \sum_{j=1}^N \delta_{X_t^j}\right), \quad i \in [N]. \quad (6.15)$$

This is true whenever the Wasserstein gradient is defined at the discrete measure  $\mu_t^N$ , and in this case, it is generally expected (and can be rigorously established under assumptions on  $\mathcal{F}$ ) that as  $N \rightarrow \infty$  and  $\mu_0^N \rightarrow \mu_0 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ , the dynamics (6.15) converges to (6.6) initialized at  $\mu_0$ . Using additional tools, one may also show that if  $\mu_0 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  then  $\mu_t \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  for all  $t \geq 0$ .

### 6.3.1 McKean–Vlasov equations

In this subsection, we consider other examples of interacting systems arising from Wasserstein gradient flows. Recall that in Section 5.4, we gave three fundamental examples of functionals over the Wasserstein space: potential energy, internal energy, and interaction energy. What if we consider the sum of the three?

$$\mathcal{F}(\mu) := \int V \, d\mu + \iint W(x - y) \mu(dx) \mu(dy) + \frac{\sigma^2}{2} \int \mu \log \mu, \quad (6.16)$$

where  $W$  is even. By using the trick at the beginning of Subsection 6.2.1 and by computing Wasserstein gradients, convince yourself of the following theorem.

**Theorem 6.15.** *The Wasserstein gradient flow  $(\mu_t)_{t \geq 0}$  of (6.16) can be described as follows:  $\mu_t = \text{law}(X_t)$ , where  $(X_t)_{t \geq 0}$  solves the SDE*

$$dX_t = -\nabla V(X_t) dt - \int \nabla W(X_t - y) \mu_t(dy) dt + \sigma dB_t. \quad (6.17)$$

Note that the coefficients of the SDE system (6.17) depend on the law of the process. Such systems are called *McKean–Vlasov processes* [McK66]. To approximate (6.17) by an interacting particle system, we replace the integral over  $\mu_t$  with an average over particles:

$$dX_t^i = -\nabla V(X_t^i) dt - \frac{1}{N-1} \sum_{j \in [N] \setminus i} \nabla W(X_t^i - X_t^j) dt + \sigma dB_t^i, \quad (6.18)$$

where  $\{B_t^i\}_{i \in [N]}$  is a collection of independent Brownian motions. A natural question that arises is to quantify how close the finite-particle system (6.18) is to its mean-field limit (6.17). This problem is addressed by the mathematical theory of *propagation of chaos* [Szn91].

More generally, suppose that we have the general entropically-regularized functional

$$\mathcal{F}(\mu) := \mathcal{F}_0(\mu) + \frac{\sigma^2}{2} \int \mu \log \mu. \quad (6.19)$$

One can show that the Wasserstein gradient flow  $(\mu_t)_{t \geq 0}$  of (6.19) can be described as the marginal law  $\mu_t = \text{law}(X_t)$  of the SDE system

$$dX_t = -\nabla \mathcal{F}_0(\mu_t)(X_t) dt + \sigma dB_t.$$

This system is known as the *mean-field* (or *interacting*) *Langevin dynamics* [CRW23, SNW23]. The corresponding finite-particle system,

$$dX_t^i = -\nabla \mathcal{F}_0\left(\frac{1}{N} \sum_{j=1}^N \delta_{X_t^j}\right)(X_t^i) dt + \sigma dB_t^i, \quad i \in [N],$$

is actually the Langevin diffusion corresponding to the target

$$\hat{\pi}_N(x^1, \dots, x^N) \propto \exp\left(-\frac{2N}{\sigma^2} \mathcal{F}_0\left(\frac{1}{N} \sum_{j=1}^N \delta_{x_t^j}\right)\right).$$

The complexity of sampling from the minimizers of the functionals (6.16) and (6.19) was studied in [KZC<sup>+</sup>24].

We conclude this section with an application to the mean-field VI problem introduced in Subsection 6.1.3. By writing down the stochastic implementation of the Wasserstein gradient flow therein, [Lac23] arrived at the McKean–Vlasov SDE

$$dX_t = - \int \nabla_1 W(X_t, y) \mu_t(dy) dt + \sqrt{2} dB_t,$$

where  $\mu_t = \text{law}(X_t)$ ,  $\nabla_1$  denotes the gradient taken with respect to the first argument, and

$$W(x, y) := \sum_{i=1}^d V(y_1, \dots, y_{i-1}, x_i, y_{i+1}, \dots, y_d).$$

When this SDE is initialized at  $X_0 \sim \mu_0$  which is a product measure,  $\mu_t$  remains a product measure for all  $t \geq 0$  so that  $(\mu_t)_{t \geq 0}$  is indeed the constrained Wasserstein gradient flow.

The corresponding interacting particle system is given by

$$dX_t^j = -\frac{1}{N-1} \sum_{j' \in [N] \setminus i} \nabla_1 W(X_t^j, X_t^{j'}) dt + \sqrt{2} dB_t^j, \quad j \in [N].$$

One then expects that if we take  $N$  sufficiently large and run the particle system, then the law of (say) the first particle  $X_t^1$  will be close to the mean-field minimizer  $q_*$ .

### 6.3.2 Birth-death sampling

We return to the sampling problem from Section 6.2. Instead of following the Wasserstein gradient flow of the KL divergence, what if we follow the WFR gradient flow introduced in Section 5.7? The fundamental difference between these approaches is that the Langevin diffusion is a *local* algorithm and hence struggles to jump between well-separated modes. This manifests itself in the convergence rate in Corollary 6.14, which depends on the log-Sobolev constant of the target  $\pi$ . On the other hand, the “teleportation” effect of the Fisher–Rao component gives rise to a universal exponential convergence rate [DEP23].

The main challenge is to implement the flow. Recall that a particle implementation for the WFR gradient flow was presented in Subsection 5.8.2, but it does not apply to the choice of functional  $\mathcal{F} = \text{KL}(\cdot \parallel \pi)$  since it assumed there that the first variation  $\delta\mathcal{F}$  can be evaluated at an empirical measure. An implementation was provided in [LLN19b] under the name of “birth-death” sampling, which we now describe.

The implementation is based on an interacting system of  $N$  particles, which at time  $t$  are denoted  $X_t^1, \dots, X_t^N$ . The approach of Subsection 5.8.2 would associate with each particle  $X_t^i$  a weight  $w_t^i$  which evolves via  $\dot{w}_t^i = -\alpha_t(X_t^i) w_t^i$ , where  $\alpha_t(x) := \delta\mathcal{F}(\mu_t)(x) - \int \delta\mathcal{F}(\mu_t) d\mu_t$ . Here,  $\alpha_t(x)$  represents an exponential rate of decay/growth (according to  $\alpha_t(x) > 0$  or  $\alpha_t(x) < 0$ ) of the density at  $x$ . We instead replace the use of weights with a procedure that “kills” or “duplicates” the particle after a random wait time. More precisely, associate with each particle  $X_t^i$  an independent clock which rings in the next instantaneous time interval  $[t, t + dt]$  with probability  $\alpha_t(X_t^i) dt$ . Whenever one of these clock rings—corresponding to, say,  $X_t^i$ —we either remove  $X_t^i$  from the system (if  $\alpha_t(X_t^i) > 0$ ) or we duplicate  $X_t^i$  (if  $\alpha_t(X_t^i) < 0$ ). To keep the total number of particles constant, in the former (resp. latter) case we randomly duplicate (resp. kill) one of the other particles.

The birth-death process implements the Fisher–Rao component of the WFR gradient flow. To implement the Wasserstein component, we stipulate that each particle evolves independently according to a Langevin diffusion between birth-death events.

The preceding discussion is ambiguous: what is the measure  $\mu_t$ ? Ideally it should be the marginal law of  $X_t^i$  (which is independent of  $i$  due to exchangeability), but we do not have access to this marginal law for implementation purposes. The methodology from the previous subsection suggests to replace  $\mu_t$  by the empirical measure  $\mu_t^N :=$

$\frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}$  over the particles, but for the KL divergence the first variation is not well-defined at such a measure. Therefore, as suggested in [LLN19b], we resort to a kernel density estimator, replacing  $\mu_t$  with  $k \star \mu_t^N$  for an appropriate kernel function  $k : \mathbb{R}^d \rightarrow \mathbb{R}$ . This leads to the following algorithm.

1. Associate with each particle  $X_t^i$  an independent clock that rings with instantaneous rate  $\hat{\alpha}_t(X_t^i)$ , where

$$\hat{\alpha}_t(x) := \log \frac{k \star \mu_t^N}{\pi}(x) - \frac{1}{N} \sum_{j=1}^N \log \frac{k \star \mu_t^N}{\pi}(X_t^j).$$

Note that when  $\pi$  is given as an unnormalized density  $\pi \propto \exp(-V)$ , the computation of  $\hat{\alpha}_t$  does not require knowledge of the normalization constant for  $\pi$ .

2. When one of the clock rings, kill or duplicate the corresponding particle  $X_t^i$ , and randomly duplicate or kill another particle, according to  $\hat{\alpha}_t(X_t^i) > 0$  or  $\hat{\alpha}_t(X_t^i) < 0$ .
3. Between the rings of the clocks, each particle evolves according to a Langevin diffusion:

$$dX_t^i = -\nabla V(X_t^i) dt + \sqrt{2} dB_t^i, \quad i \in [N],$$

where  $\{B^i\}_{i \in [N]}$  are i.i.d. Brownian motions.

As shown in [LLN19b], the marginal laws of this process converge to the WFR gradient flow as  $N \rightarrow \infty$  and the bandwidth of the kernel tends to zero appropriately.

## 6.4 Non-parametric maximum likelihood

We have just seen that sampling can be viewed as optimization over the space of probability measures using the perspective originally put forward in [JKO98] and [Wib18]. In this section we study a classical statistical problem that is readily of this nature.

Consider a Gaussian mixture on  $\mathbb{R}^d$  with density given by:

$$G_\rho = \int_{\mathbb{R}^d} \phi(\cdot - y) \rho(dy) = \phi \star \rho,$$

where  $\phi(z) = (2\pi)^{-d/2} \exp(-\|z\|^2/2)$  denotes the density of the standard isotropic Gaussian distribution  $\mathcal{N}(0, I)$  and  $\rho$  is the mixing distribution

of interest. Note that in comparison with the general Gaussian mixtures introduced in Section 5.6, we constrain all the Gaussian components to have identity covariance matrix for simplicity.

Let  $\rho^*$  be an unknown mixing distribution on  $\mathbb{R}^d$ . Given  $n$  independent observations  $X_1, \dots, X_n$  drawn from  $G_{\rho^*}$ , our goal is to estimate  $\rho^*$ . This is a Gaussian deconvolution problem, for which the rates of convergence are known to be very slow [RNW19]. When  $\rho^*$  is assumed to be smooth, a classical approach that leads to optimal rates of convergence uses kernel smoothing [Fan91]. In this section, we explore a different approach called non-parametric maximum likelihood estimation following the paper [YWR24].

The negative log-likelihood for this problem is defined as:

$$\ell_n(\rho) := -\frac{1}{n} \sum_{i=1}^n \log G_\rho(X_i) = -\frac{1}{n} \sum_{i=1}^n \log \phi \star \rho(X_i).$$

The non-parametric maximum likelihood estimator, or NPMLE, is defined as any minimizer of  $\ell_n$ :

$$\hat{\rho} = \arg \min_{\rho \in \mathcal{P}(\mathbb{R}^d)} \ell_n(\rho). \quad (6.20)$$

Before turning to the computational aspects of this problem, we first note a surprising connection with entropic optimal transport, highlighted in [RNW18]. Writing  $\mu_n$  for the empirical measure  $\frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ , it turns out that the NPMLE satisfies

$$\hat{\rho} = \arg \min_{\rho \in \mathcal{P}(\mathbb{R}^d)} S_{2\sigma^2}(\mu_n, \rho), \quad (6.21)$$

where  $S_{2\sigma^2}(\cdot, \cdot)$  is the entropic optimal transport cost with regularization parameter  $\varepsilon = 2\sigma^2$ . In other words, the NPMLE precisely minimizes the entropic OT cost to the data  $\mu_n$ .

This connection arises from duality. A version of the Gibbs variational principle (see Proposition 4.2) tailored to probability measures rather than general positive measures implies that for suitable  $h : \mathbb{R}^d \rightarrow \mathbb{R}$ , it holds

$$\log \int \exp(h) dQ = \sup_{P \in \mathcal{P}(\mathbb{R}^d)} \left\{ \int h dP - \text{KL}(P \parallel Q) \right\}.$$

This result can be found for example as Proposition 1.4.2 in [DE97]. We can use this expression to obtain a variational formulation of  $\log G_\rho(X_i)$ :

$$\begin{aligned}
-\log G_\rho(X_i) &= (2\pi)^{d/2} - \log \int \exp\left(-\frac{\|X_i - y\|^2}{2\sigma^2}\right) \rho(dy) \\
&= (2\pi)^{d/2} + \inf_{P_i \in \mathcal{P}(\mathbb{R}^d)} \int \frac{1}{2\sigma^2} \|X_i - y\|^2 P_i(dy) + \mathsf{KL}(P_i \parallel \rho).
\end{aligned}$$

This shows that the optimization problem in (6.20) is equivalent to

$$\hat{\rho} = \arg \min_{\rho \in \mathcal{P}(\mathbb{R}^d)} \inf_{P_1, \dots, P_n \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{n} \sum_{i=1}^n \left[ \int \|X_i - y\|^2 P_i(dy) + 2\sigma^2 \mathsf{KL}(P_i \parallel \rho) \right].$$

The minimization problem over  $P_1, \dots, P_n$  can be equivalently rewritten as a minimization over measures of the form  $\gamma = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \otimes P_i$ , which are precisely joint measures whose first marginal is  $\mu_n$ . It can be shown that the second marginal can be taken to be  $\rho$ , which leads to the representation in (6.21).

The convex infinite-dimensional optimization problem (6.20) has primarily been studied in the case where  $d = 1$  where it can be shown that the solution is unique and supported on a small number of atoms [Lin83, PW24]. Using these properties, various computational schemes have been proposed by restricting the set of measures to ones with a small support. In higher dimensions, much less is known.

The definition (6.20) of the NPMLE is precisely an optimization over probability measures and in the rest of this section we describe algorithms based on Wasserstein gradient flows to solve it.

We first compute the Wasserstein gradient. To that end, we need the first variation of  $\ell_n$ . Fix  $\varepsilon > 0$  and  $\xi$  be a perturbation such that  $\rho + \varepsilon \xi$  is a probability measure and observe that

$$\begin{aligned}
\ell_n(\rho + \varepsilon \xi) &= -\frac{1}{n} \sum_{i=1}^n \log(\phi \star \rho + \varepsilon \phi \star \xi)(X_i) \\
&= -\frac{1}{n} \sum_{i=1}^n \log(\phi \star \rho)(X_i) - \frac{\varepsilon}{n} \sum_{i=1}^n \frac{\phi \star \xi(X_i)}{\phi \star \rho(X_i)} + O(\varepsilon^2).
\end{aligned}$$

Hence

$$\lim_{\varepsilon \searrow 0} \frac{\ell_n(\rho + \varepsilon \xi) - \ell_n(\rho)}{\varepsilon} = -\frac{1}{n} \sum_{i=1}^n \frac{\phi \star \xi(X_i)}{\phi \star \rho(X_i)} = -\frac{1}{n} \sum_{i=1}^n \frac{\int \phi(\cdot - X_i) d\xi}{\phi \star \rho(X_i)},$$

and we readily identify that the first variation is given by

$$\delta\ell_n(\rho) = -\frac{1}{n} \sum_{i=1}^n \frac{\phi(\cdot - X_i)}{\phi \star \rho(X_i)}.$$

The Wasserstein gradient flow of  $\ell_n$  correspond to the following ODE:

$$\begin{aligned} \dot{\theta}_t &= -\nabla \ell_n(\rho_t)(\theta_t) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\nabla \phi(\theta_t - X_i)}{\phi \star \rho_t(X_i)} \\ &= -\frac{1}{n} \sum_{i=1}^n \frac{(\theta_t - X_i) \phi(\theta_t - X_i)}{\phi \star \rho_t(X_i)}, \end{aligned} \quad (6.22)$$

where  $\rho_t = \text{law}(\theta_t)$ . In light of the continuity equation (5.2), we see that the Wasserstein gradient flow of  $\ell_n$  is the curve described by the PDE:

$$\partial_t \rho_t = \frac{1}{n} \sum_{i=1}^n \text{div} \left( \rho_t \frac{(\cdot - X_i) \phi(\cdot - X_i)}{\phi \star \rho_t(X_i)} \right).$$

Since the velocity field in (6.22) depends on  $\rho_t$ , we use a particle implementation of this gradient flow: given  $N$  particles  $\theta_t^1, \dots, \theta_t^N$  with marginal distribution  $\rho_t$ , we replace  $\rho_t$  with the empirical distribution  $N^{-1} \sum_{j=1}^N \delta_{\theta_t^j}$ . It results in the system of coupled ODEs: for  $j \in [N]$ ,

$$\dot{\theta}_t^j = -\frac{1}{n} \sum_{i=1}^n \frac{(\theta_t^j - X_i) \phi(\theta_t^j - X_i)}{\frac{1}{N} \sum_{k=1}^N \phi(\theta_t^k - X_i)}.$$

Unfortunately, this Wasserstein gradient flow is difficult to analyze. Moreover, time-discretizations of this Wasserstein gradient flow do not perform well in practice. Instead, [YWR24] propose to study the Wasserstein–Fisher–Rao gradient flow of  $\ell_n$ . In this context, the measure  $\rho_t$  is approximated by

$$\sum_{j=1}^N w_t^j \delta_{\theta_t^j}$$

where for any  $j \in [N]$ , we use the following dynamics:

$$\begin{aligned} \dot{\theta}_t^j &= -\frac{1}{n} \sum_{i=1}^n \frac{(\theta_t^j - X_i) \phi(\theta_t^j - X_i)}{\sum_{k=1}^N w_t^k \phi(\theta_t^k - X_i)}, \\ \dot{w}_t^j &= \left[ \frac{1}{n} \sum_{i=1}^n \frac{\phi(\theta_t^j - X_i)}{\sum_{k=1}^N w_t^k \phi(\theta_t^k - X_i)} - 1 \right] w_t^j. \end{aligned}$$

Under some conditions, the convergence guarantees of this system can be established but they are entirely driven by the Fisher–Rao part and the proof largely consists in finding conditions under which the Wasserstein part does not get in the way of convergence.

## 6.5 Mean-field neural networks

A *two-layer*<sup>6</sup> neural network is a parameterized function

$$f(x; \theta) = \frac{1}{m} \sum_{j=1}^m a_j \sigma(\langle w_j, x \rangle + b_j), \quad (6.23)$$

where  $\theta := \{(a_j, w_j, b_j), j \in [m]\}$  represents the *parameters* (or *weights*) of the network, and  $\sigma(\cdot)$  is a non-linearity, e.g., the common ReLU activation  $\sigma(\cdot) = (\cdot)_+ = \max(0, \cdot)$ .

The first layer is the map  $\ell_1 : \mathbb{R}^d \rightarrow \mathbb{R}^m$  defined by

$$\ell_1(x) = (\sigma(\langle w_1, x \rangle + b_1), \dots, \sigma(\langle w_m, x \rangle + b_m))^T =: \sigma(Wx + b). \quad (6.24)$$

It produces an internal *representation* of the vector  $x$  that is more suitable for the subsequent task; e.g. classification or regression. This representation is then passed on to the second and terminal layer  $\ell_2 : \mathbb{R}^m \rightarrow \mathbb{R}$  which collapses the representation  $z := \ell_1(x)$  into a scalar prediction using a linear projection:

$$\ell_2(z) = \frac{1}{m} \sum_{j=1}^m a_j z_j = \frac{1}{m} \langle a, z \rangle.$$

This terminal layer is tailored to a regression task but it is also common to employ terminal layers that are tailored to classification. For binary classification for example, it is desirable to have the output of the neural network lie in the interval  $[0, 1]$  so further process the output  $y = \ell_2 \circ \ell_1(x)$  using

$$\text{logistic}(y) = \frac{e^y}{1 + e^y}.$$

The reader will recognize here the logistic function employed in generalized linear models. In the rest of this section we focus on two-layer

---

<sup>6</sup> In other words, the network has one hidden layer.

neural networks of the form  $\ell_2 \circ \ell_1$  and leave the study of  $\text{logistic} \circ \ell_2 \circ \ell_1$  as an exercise for the reader.

The parametrization (6.23) is called the *mean-field* parametrization, and it lends itself to passing to the limit  $m \rightarrow \infty$ .

Consider a simple regression task in which we have  $n$  data points  $\{(X_i, Y_i), i \in [n]\}$  with  $X_i \in \mathbb{R}^d$  and  $Y_i \in \mathbb{R}$ . To train the neural network, we can minimize the squared error

$$L(\theta) := \sum_{i=1}^n (Y_i - f(X_i; \theta))^2. \quad (6.25)$$

The training dynamics for minimizing the objective (6.25) are complex because the parametrization  $\theta \mapsto f(\cdot; \theta)$  is non-linear and consequently the loss  $L$  is non-convex. One approach to study these dynamics is to lift the optimization problem to one set over the space of probability measures, with the hope that the lifted problem affords simplifications. In doing so, we must preserve the connection with the original dynamics, and this leads naturally to Wasserstein gradient flows.

The lifting is carried out as follows. Let  $\Omega = \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$  denote the space of  $(a, w, b)$  triples, and let  $\mu$  be a measure over  $\Omega$ . Set

$$f(x; \mu) = \int \rho(x; \omega) \mu(d\omega), \quad \rho(x; \omega) := a\sigma(\langle w, x \rangle + b),$$

where  $\omega = (a, w, b)$ . One can check that if we encode parameters  $\theta = \{(a_i, w_i, b_i)\}_{i=1}^m$  via the empirical measure  $\mu_\theta := \frac{1}{m} \sum_{i=1}^m \delta_{(a_i, w_i, b_i)}$ , then  $f(\cdot; \mu_\theta) = f(\cdot; \theta)$ , so this definition indeed generalizes (6.23). However, we can now formulate the problem of optimizing, over the space  $\mathcal{P}_2(\Omega)$  of probability measures over  $\Omega$ , the objective

$$L(\mu) := \sum_{i=1}^n (Y_i - f(X_i; \mu))^2. \quad (6.26)$$

As discussed above, we require that the dynamics of minimizing  $\mu \mapsto L(\mu)$  over  $\mathcal{P}_2(\Omega)$  be compatible with the original dynamics of minimizing  $\theta \mapsto L(\theta)$  over  $\Omega^m$ . Herein lies the utility of the Wasserstein geometry, as it was set up precisely to ensure that dynamics over the base space  $\Omega$  lift gracefully to dynamics over  $\mathcal{P}_2(\Omega)$ . More precisely:

**Proposition 6.16.** *The Wasserstein gradient flow  $(\mu_t)_{t \geq 0}$  of (6.26), when initialized at a measure of the form  $\mu_0 = \mu_\theta$ , is such that  $\mu_t = \mu_{\theta_t}$  for all  $t \geq 0$ , where  $(\theta_t)_{t \geq 0}$  is the (time-rescaled) Euclidean gradient flow of (6.25) initialized at  $\theta$ .*

We can also rewrite the objective (6.26) as follows:

$$\begin{aligned}
L(\mu) &= \sum_{i=1}^n \left( Y_i^2 - 2Y_i \int \rho(X_i; \omega) \mu(d\omega) \right. \\
&\quad \left. + \iint \rho(X_i; \omega) \rho(X_i; \omega') \mu(d\omega) \mu(d\omega') \right) \\
&= \text{const.} - 2 \int \sum_{i=1}^n Y_i \rho(X_i; \omega) \mu(d\omega) \\
&\quad + \iint \sum_{i=1}^n \rho(X_i; \omega) \rho(X_i; \omega') \mu(d\omega) \mu(d\omega') .
\end{aligned}$$

We can recognize the second term as a potential energy (in the sense of Example 5.11) and the third term as an interaction energy (in the sense of Example 5.13), albeit a generalized version in which the interaction is not of the form  $(x, y) \mapsto K(x - y)$ . As expected, the loss  $L$  is not in general geodesically convex.

Since the Wasserstein perspective is essentially a reformulation of the original neural network problem, a skeptic may ask what advantages it brings. The answer is that we can now consider more general initializations than empirical measures (measures of the form  $\mu_\theta$ ), and in particular, a well-known result of Chizat and Bach [CB18] uses this approach to establish global convergence in the mean-field regime, under certain assumptions. Their result requires the initialization to be absolutely continuous, and since such a measure can only be approximated by empirical measures in the limit  $m \rightarrow \infty$ , this corresponds in some sense to “infinitely wide” neural networks. The error incurred for finite  $m$  can be controlled and leads to insights for finite-width networks in various settings [MMN18, ABAM22, ABAM23].

It has also been proposed to add an entropic regularization term to the loss (6.26) and to train the network via the mean-field Langevin dynamics from Subsection 6.3.1 [CRW23, SNW23, TR24].

## 6.6 Transformers

Since their introduction in 2017 in the paper “Attention is all you need” [VSP<sup>+</sup>17], transformers have profoundly transformed practical deep neural networks, most notably in natural language processing

(NLP), but also in computer vision and robotics. Central to this new architecture is the so-called *attention* mechanism, a layer that is markedly different from a perceptron (a.k.a. feed-forward) layer such as the one in (6.23).

Unlike the neural networks that we have seen in the previous sections that are functions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , an attention layer is a sequence-to-sequence map

$$\begin{aligned} g : (\mathbb{R}^d)^N &\rightarrow (\mathbb{R}^d)^N, \\ (x^1, \dots, x^N) &\mapsto (g^1(x^1, \dots, x^N), \dots, g^N(x^1, \dots, x^N)). \end{aligned}$$

More specifically, a input (a sentence in NLP or an image in computer vision) is broken into *tokens*  $x^1, \dots, x^N \in \mathbb{R}^d$  and processed through an attention layer  $g = (g^1, \dots, g^N)$  where for each  $i \in [N]$ ,

$$g^i(x^1, \dots, x^N) = x^i + V \frac{\sum_{j=1}^N x^j e^{\langle Qx^i, Kx^j \rangle}}{\sum_{j=1}^N e^{\langle Qx^i, Kx^j \rangle}},$$

where  $K$ ,  $Q$ , and  $V$  are three  $d \times d$  matrices called *key*, *query*, and *value* respectively.

While practical transformers combine perceptron layers with attention layers—and also normalization layers that are briefly discussed below—we focus here on composing multiple attention layers with the same matrices  $(K, Q, V)$ . This composition results in a iterative scheme where tokens are updated as:

$$x_{t+1}^i = x_t^i + V \frac{\sum_{j=1}^N x_t^j e^{\langle Qx_t^i, Kx_t^j \rangle}}{\sum_{j=1}^N e^{\langle Qx_t^i, Kx_t^j \rangle}}, \quad i \in [N].$$

In turn, taking the same perspective as in neural ODEs [CRBD18], we can view the above iterations as a time discretization of the following dynamical system of interacting particles:

$$\dot{x}_t^i = V \frac{\sum_{j=1}^N x_t^j e^{\langle Qx_t^i, Kx_t^j \rangle}}{\sum_{j=1}^N e^{\langle Qx_t^i, Kx_t^j \rangle}}, \quad i \in [N]. \quad (6.27)$$

The above equation describes a system of  $N$  ordinary differential equations (ODEs), one for each token/particle, that are called *self-attention dynamics* by [GLPR23, GLPR24]. The way these tokens interact is not completely wild: each token evolves according to its own position and

the *empirical distribution* of all the tokens. Indeed, let  $\mu_t$  denote this empirical distribution at time  $t$ :

$$\mu_t = \frac{1}{N} \sum_{i=1}^N \delta_{x_t^i}.$$

We can rewrite (6.27) as

$$\dot{x}_t^i = V \frac{\int y e^{\langle Qx_t^i, Ky \rangle} \mu_t(dy)}{\int e^{\langle Qx_t^i, Ky \rangle} \mu_t(dy)}, \quad i \in [N].$$

It becomes now clear that the tokens all have the same mean-field dynamics so we can drop the index  $i$ :

$$\dot{x}_t = V \frac{\int y e^{\langle Qx_t, Ky \rangle} \mu_t(dy)}{\int e^{\langle Qx_t, Ky \rangle} \mu_t(dy)}, \quad (6.28)$$

where  $\mu_t = \text{law}(x_t)$ . Using the continuity equation, we get that  $\mu_t$  evolves according to the following PDE:

$$\partial_t \mu_t + \text{div} \left( \mu_t V \frac{\int y e^{\langle Q \cdot, Ky \rangle} \mu_t(dy)}{\int e^{\langle Q \cdot, Ky \rangle} \mu_t(dy)} \right) = 0.$$

This perspective on transformers was first put forward in [SABP22] which raised the question of whether this curve could be viewed as a Wasserstein gradient flow. To investigate this question, assume that  $K = Q = V = I$  so that (6.28) becomes:

$$\dot{x}_t = \frac{\int y e^{\langle x_t, y \rangle} \mu_t(dy)}{\int e^{\langle x_t, y \rangle} \mu_t(dy)} = \nabla \left[ \log \int e^{\langle \cdot, y \rangle} \mu_t(dy) \right] (x_t).$$

The form of the velocity field is suggestive and readily begs the question of whether there exists a functional  $\mathcal{F}$  such that its first variation is given by

$$\delta \mathcal{F}(\mu) = \log \int e^{\langle \cdot, y \rangle} \mu_t(dy).$$

Unfortunately, [SABP22] also show that this is not the case due to a lack of symmetry. To overcome this limitation, one may consider instead the *unnormalized self-attention* dynamics introduced in [GLPR24]. These dynamics are of the form

$$\dot{x}_t = \int y e^{\langle x_t, y \rangle} \mu_t(dy) = \nabla \left[ \int e^{\langle \cdot, y \rangle} \mu_t(dy) \right] (x_t).$$

We readily get that these dynamics describe the Wasserstein gradient flow of the interaction energy

$$\mathcal{F}(\mu) := - \iint e^{\langle x, y \rangle} \mu(dx) \mu(dy).$$

Unfortunately, it is easy to see that this functional does not admit a global minimum over the space of probability measure, indeed, for any Dirac delta  $\mu = \delta_x$ , we have  $\mathcal{F}(\delta_x) = -e^{\|x\|^2} \rightarrow -\infty$  as  $x \rightarrow \infty$ . In particular this suggests that tokens undergoing these dynamics will simply diverge to infinity.

In practice however, tokens are restricted to live on the unit sphere  $\mathcal{S}^{d-1}$  of  $\mathbb{R}^d$  using a procedure know as *layer normalization* (or simply “layernorm”). With layernorm, the unnormalized self-attention dynamics then become

$$\dot{x}_t = \mathsf{P}_{x_t} \int y e^{\langle x_t, y \rangle} \mu_t(dy) = \nabla_{x_t} \int e^{\langle x_t, y \rangle} \mu_t(dy),$$

where for any  $x \in \mathcal{S}^{d-1}$ ,  $y \in \mathbb{R}^d$ , we write  $\mathsf{P}_{xy} := y - \langle x, y \rangle x$  for the projection of  $y$  onto the tangent space of the sphere  $\mathcal{S}^{d-1}$  at  $x$  and  $\nabla_x := \mathsf{P}_x \nabla$  denotes the spherical (Riemannian) gradient at  $x \in \mathcal{S}^{d-1}$ . Using the version Otto calculus on Riemannian manifolds alluded to in Section 5.6, we get that these dynamics correspond to a Wasserstein gradient flow of the interaction energy  $\mathcal{F}$  now defined on the sphere. In fact, since for  $x, y \in \mathcal{S}^{d-1}$ , it holds that  $\|x - y\|^2 = 2(1 - \langle x, y \rangle)$ , we can write  $\mathcal{F}$  as

$$\mathcal{F}(\mu) := -e \iint_{\mathcal{S}^{d-1} \times \mathcal{S}^{d-1}} e^{-\frac{\|x-y\|^2}{2}} \mu(dx) \mu(dy). \quad (6.29)$$

It can be readily seen that the maximizers of this functional are precisely Dirac deltas  $\delta_x$  for any  $x \in \mathcal{S}^{d-1}$ . Hence unnormalized self-attention is a Wasserstein gradient flow of a functional minimized at Dirac deltas, which correspond to states where all the tokens are clustered. Unfortunately, this functional is not geodesically convex and admits many stationary points where the Wasserstein gradient vanishes. Using this framework [CRMB24, GLPR24] show that these points are in fact saddle points, guaranteeing asymptotic convergence to a single cluster when dynamics are initialized in a generic position.

## 6.7 Discussion

**§6.1.** Another natural geometric approach to VI is *natural gradient descent*, which is motivated by the parametrization invariance of the Fisher–Rao geometry [Ama98, AN00]. However, it has been challenging to analyze this approach since the VI problem is often non-convex. See [AR15] for an early work on algorithmic guarantees for VI.

Our discussion of Gaussian VI follows [LCB<sup>+</sup>22]. Algorithms for Gaussian VI which are closely related to (6.3) have been proposed and studied in works such as [AR20, Dom20, GFPO21]; here, our emphasis is on the derivation via Otto calculus. There have been many subsequent works on Gaussian VI, both on the computational (e.g., [DBCS23, DGG23, KOW<sup>+</sup>23, BLB24]) and statistical ([KR24]) aspects, as well as applications to bandits [CHD24] and control [LBB23, LBB24].

The potential application of Wasserstein geometry to mean-field VI was noticed by several authors [GLNZ22, Lac23, YY23]. The works [GLNZ22, Lac23] also wrote down interacting SDE implementations of the gradient flow described in Subsection 6.3.1. The wider literature on mean-field VI, which usually focuses on coordinate ascent variational inference (CAVI), is vast and we do not survey it here, but see [AL24, LZ24] for analyses leveraging Otto calculus.

SVGD was introduced in [LW16], and geometric interpretations of SVGD are given in [Liu17, CLGL<sup>+</sup>20a, DNS23]. Convergence theory remains underdeveloped, see, e.g., [LLN19a, KSA<sup>+</sup>20, SSR22, DN23, SM23, PBS24].

Corollary 6.4 can be generalized to measures over Riemannian manifolds, in which case the strong convexity parameter of the KL divergence captures information about the Ricci curvature. The seminal work of Lott and Villani [LV09] and Sturm [Stu06a, Stu06b] leverages this to define a synthetic notion of Ricci curvature lower bounds for measured geodesic spaces (see Chapter 7) that recover classical Ricci curvature lower bounds when specialized to the Riemannian setting. These ideas were later extended to discrete settings, e.g., [Oll10, OV12, Oll13].

Finally note that while the KL divergence plays a preponderant role in variation inference, other distances between measures can be considered for this task; see, e.g., [AKSG19] who use Maximum Mean Discrepancy.

**§6.2.** As mentioned in the main text, the interpretation of the Langevin diffusion as a Wasserstein gradient flow goes back to the seminal work of [JKO98]. Otto calculus was first applied to obtain quantitative results

for the Langevin diffusion, as in Exercise 6 below, in [OV00]; in this context, Exercise 4 from Chapter 5 can be sharpened by a factor of 2 [BGL01, OV01]. See also [CE02] for proofs in this spirit which do not require as much differential structure. For textbook treatments on stochastic calculus, see [Ste01] or [Le 16]. The Wasserstein PL inequality for the KL divergence functional is known as the *log-Sobolev inequality* and it plays a key role in the study of high-dimensional probability and Markov processes. Crucially, although we have presented strong log-concavity as a sufficient condition for the validity of the log-Sobolev inequality, it is not necessary. See [BGL14] for further detail and [OT11, OT13, BB18] for generalizations.

The optimization perspective on sampling dates back to early works such as [DT12]; our discussion largely follows [Wib18]. For an exposition to the modern complexity theory of log-concave sampling and further references, see [Che23, Che24]. Recent works also apply this perspective for parameter estimation [ACG<sup>+</sup>23, CKPJ24].

The proximal sampler has been applied to structured log-concave sampling [LST21], to non-Euclidean [GLL<sup>+</sup>23] and heavy-tailed [HMHBE24] sampling, and to sampling from convex bodies [KVZ24].

As noted in [CLGL<sup>+</sup>20b], the Newton–Langevin diffusion converges to any strictly log-concave target with a universal exponential rate as a consequence of the Brascamp–Lieb inequality [BL76]. There is a sense in which it is an optimal preconditioning of Langevin [CTZ24].

Convergence of the underdamped Langevin diffusion requires heavier machinery than Wasserstein gradient flows and is based on the theory of hypocoercivity, for which the standard reference is [Vil09a].

**§6.3.** Analysis of birth-death sampling was first carried out in [LLN19b] and improved in [LSW23].

**§6.4.** The WFR gradient flow for NPMLE can be adapted to more general mixtures. Usually, asymptotic convergence of the gradient flow to the NPMLE is only established conditionally on convergence to a limit point. In fact, it is shown in [YWR24] that the NPMLE is the only stationary point of the gradient flow initialized at a measure that is absolutely continuous.

**§6.5.** The study of training dynamics of two-layer neural networks from the mean-field perspective was proposed in four independent papers that were released within about a month period in 2018: [MMN18] on April 18, [SS20] and [RVE22] both on May 2, and [CB18] on May 24. It is worth noting that the normalization  $1/m$  in (6.23) is critical for the

mean-field interpretation of the problem. Other works have proposed to use the normalization  $1/\sqrt{m}$  which results in the so called *neural tangent kernel (NTK)* (a.k.a. *lazy training*) regime. In this regime, which will remind the reader of the normalization employed in the central limit theorem, it can be shown that the parameters do not move far away from a random initialization and the neural network can be studied using linear approximation around initialization [JGH18]. For more details on NTK and its relationship with the mean-field regime see [MM23].

**§6.6.** The functional  $\mathcal{F}$  defined in (6.29) has appeared in the literature on optimal configuration. In this line of work, the *maximizers* of this functional are of interest. It is known that  $\mathcal{F}$  is maximized by the uniform distribution on the sphere [Tan17]. Finding maximizers subject to a cardinality constraint on the support of  $\mu$  is directly connected to questions arising in sphere packing; see [CK07].

## 6.8 Exercises

1. Let  $K : \mathbb{R}^d \rightarrow \mathbb{R}$  be a symmetric function on  $\mathbb{R}^d$ , and consider the corresponding interaction energy as defined in Example 5.13:  $\mathcal{F}(\mu) := \frac{1}{2} \iint K(x - y) \mu(dx) \mu(dy)$ .
  - Show that if  $K$  is convex, then  $\mathcal{F}$  is geodesically convex on  $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ . *Hint:* let  $X_t = (1-t)X_0 + tT(X_0)$ , so that  $(\mu_t)_{t \in [0,1]} := (\text{law}(X_t))_{t \in [0,1]}$  is a Wasserstein geodesic, then apply (5.6).
  - Show that  $\mathcal{F}$  is never  $\alpha$ -geodesically convex for any  $\alpha > 0$ . *Hint:* Consider the geodesic  $(\mathcal{N}(tv, I))_{t \in [0,1]}$  for a nonzero vector  $v \in \mathbb{R}^d$ .
2. Show that the tangent space to the space of product measures at  $\mu$  is given by the space of *separable* vector fields:

$$T_\mu \mathcal{P}_{2,\text{ac}}(\mathbb{R})^{\otimes d} = \overline{\{x \mapsto (\psi'_1(x_1), \dots, \psi'_d(x_d)) \mid \psi_1, \dots, \psi_d : \mathbb{R} \rightarrow \mathbb{R}\}}^{L^2(\mu)},$$

where  $\psi_1, \dots, \psi_d$  are smooth and compactly supported. Then, show that the Wasserstein gradient projected to this subspace takes the form (6.4).

3. Compute the gradient of the KL divergence restricted to the space of product measures. From the first-order optimality condition, write down a fixed-point equation for the density of the solution  $q_*$  to (MFVI).

4. Prove Lemma 6.9. *Hint:* Use the separability of the transport maps to reduce to one-dimensional optimal transport, for which we can apply the results of Section 1.3 and Proposition 1.18.
5. Compute the derivative of  $t \mapsto \text{KL}(\mu_t \parallel \pi)$  when  $(\mu_t)_{t \geq 0}$  evolves according to (SVGD).
6. Interpret Exercises 4 and 5 from Chapter 5 for the Langevin diffusion.
7. Consider the Euler–Maruyama scheme (6.11) where the initial distribution is  $\mathcal{N}(m, \Sigma)$  and the target distribution is  $\pi = \mathcal{N}(0, I)$ . Compute the law of  $X_k$  for each  $k \geq 0$ . Use this to compute the stationary distribution  $\hat{\pi}$  of (6.11), and compute the KL divergence  $\text{KL}(\hat{\pi} \parallel \pi)$ . How small should we choose the step size  $h$  if we want to ensure  $\text{KL}(\hat{\pi} \parallel \pi) \leq \varepsilon^2$ ?
8. Let  $\mu = \mathcal{N}(m, \Sigma)$  be a Gaussian measure with  $\Sigma \succ 0$ . Evaluate  $\text{prox}_{h\mathcal{F}}(\mu)$ , where  $\text{prox}_{h\mathcal{F}}$  is the proximal map for the entropy defined in (6.14). This computation is used as the basis for the Gaussian VI algorithm of [DBCS23].
9. For  $\mathcal{F} := \frac{1}{2} W_2^2(\cdot, \nu)$ , compute the iterations  $\mu_{n+1} = \text{prox}_{h\mathcal{F}}(\mu_n)$  of the JKO scheme, where

$$\text{prox}_{h\mathcal{F}}(\mu) := \arg \min_{\mu' \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ h\mathcal{F}(\mu') + \frac{1}{2} W_2^2(\mu, \mu') \right\}.$$

Letting  $h \searrow 0$  while  $nh \rightarrow t$ , show that one recovers the gradient flow from Exercise 11 from Chapter 5.

10. Consider the proximal sampler with initial distribution  $\mathcal{N}(m, \Sigma)$  and target distribution  $\pi = \mathcal{N}(0, I)$ . Compute the law of the  $k$ -th iterate  $X_k$  for each  $k \geq 0$  and estimate the rate of convergence to  $\pi$ .
11. Let  $V(x) = \frac{1}{2} \langle x, Ax \rangle$  and  $W(x) = \frac{\lambda}{2} \|x\|^2$ , where  $A \succ 0$  and  $\lambda \geq 0$ . Compute the stationary distributions  $\pi$  and  $\hat{\pi}_N$  of the McKean–Vlasov SDE (6.17) and the finite-particle system (6.18) respectively.
12. Prove Proposition 6.16. Also, generalize to the case of a two-layer neural network composed with a logistic function when the training data satisfies  $Y_i \in \{0, 1\}$  for each  $i \in [n]$  and we use the cross-entropy loss:  $L(\mu) = -\sum_{i=1}^n \{(1 - Y_i) \log(1 - f(X_i; \mu)) + Y_i \log f(X_i; \mu)\}$ .

## Metric geometry of the Wasserstein space

In the previous two chapters, we studied the space  $\mathcal{W}_2 = (\mathcal{P}_2(\mathbb{R}^d), W_2)$  through the lens of Riemannian geometry. Although such an approach yields considerable geometric insight and can even be treated rigorously (see [AGS08]), it is important to keep in mind that  $\mathcal{W}_2$  is not a *bona fide* Riemannian manifold, and consequently technical issues abound.

Despite what its name suggests, *metric geometry* requires a bit more structure than simply a metric space. Indeed, in the rest of this chapter, we will talk about length/geodesic spaces which have a continuous flavor. In particular it is possible to take derivatives of functions along smooth curves. This primitive differential structure is often sufficient to understand questions about curvature which we will employ to establish rates of convergence for Wasserstein barycenters in the next chapter.

The goal of this chapter is to gather basic material from metric geometry for a general metric space  $(S, \mathsf{d})$  following the classical book [BBI01]; see also [AKP22] for a more advanced coverage. Main concepts (curvature, tangent cone, logarithmic map, etc.) are instantiated to the (2-)Wasserstein space.

### 7.1 Geodesics

We already appealed to an intuitive notion of geodesics in Chapter 5. In this chapter we properly define these objects as length minimizing.

#### 7.1.1 Length and geodesic spaces

Let  $(S, \mathsf{d})$  be a metric space. A *path* in  $S$  is a continuous map  $\omega : I \rightarrow S$  where  $I \subset \mathbb{R}$  is an interval. The *length*  $L(\omega) \in \mathbb{R} \cup \{\infty\}$  of a path

$\omega : I \rightarrow S$  is defined by

$$L(\omega) := \sup \sum_{i=1}^{n-1} d(\omega(t_i), \omega(t_{i+1})), \quad (7.1)$$

where the supremum is taken over all  $n \geq 1$  and all  $n$ -tuples  $t_1 < \dots < t_n$  in  $I$ .

A path is called *rectifiable* if it has finite length.

For any path  $\omega$  and any interval  $J \subset \mathbb{R}$ , we write  $\omega_J$  to denote the restriction of  $\omega$  to  $I \cap J$ . The following lemma holds.

**Lemma 7.1.** *For any rectifiable path  $\omega : I \rightarrow S$ , the function  $t \mapsto \ell(t) = L(\omega_{(-\infty, t]})$  is continuous on  $I$ .*

*Proof.* We prove left continuity, i.e., that for any  $\varepsilon > 0$ , there exists  $\delta > 0$  such if  $t - \delta < t' \leq t$ , we have

$$\ell(t) - \varepsilon \leq \ell(t') \leq \ell(t).$$

By continuity of  $\omega$ , there exists  $\delta_1 > 0$  such that  $t' \in (t - \delta_1, t]$  implies

$$d(\omega(t'), \omega(t)) \leq \frac{\varepsilon}{2}. \quad (7.2)$$

Next, let  $n$  and  $t_1 < \dots < t_n = t$  be such that

$$\ell(t) - \frac{\varepsilon}{2} \leq \sum_{i=1}^{n-1} d(\omega(t_i), \omega(t_{i+1})) \leq \ell(t), \quad (7.3)$$

define  $\delta_2 = \min_{i=1, \dots, n-1} |t_{i+1} - t_i| > 0$ , and let  $\delta = \min(\delta_1, \delta_2) > 0$ . Observe that for any  $t'$  such that  $t \geq t' > t - \delta$  it holds  $t_{n-1} < t' \leq t$  so that

$$\begin{aligned} \ell(t') &\geq \sum_{i=1}^{n-2} d(\omega(t_i), \omega(t_{i+1})) + d(\omega(t_{n-1}), \omega(t')) \\ &\geq \sum_{i=1}^{n-1} d(\omega(t_i), \omega(t_{i+1})) - d(\omega(t'), \omega(t)) \\ &\geq \ell(t) - \varepsilon \end{aligned}$$

where we used the triangle inequality in the second line and (7.2)–(7.3) in the third. This completes the proof of left continuity. Right continuity follows using the same argument.  $\square$

Two paths  $\omega_1 : I_1 \rightarrow S$  and  $\omega_2 : I_2 \rightarrow S$  are *equivalent* if there exists a continuous, non-decreasing, and surjective function  $\varphi : I_1 \rightarrow I_2$  such that  $\omega_1 = \omega_2 \circ \varphi$ . In this case,  $\omega_2$  is a *reparametrization* of  $\omega_1$  (and vice-versa) and it is easy to check that  $L(\omega_1) = L(\omega_2)$ .

Finally a path  $\omega : [a, b] \rightarrow S$  is said to have *constant speed* if for all  $a \leq s \leq t \leq b$ ,

$$L(\omega_{[s,t]}) = \frac{t-s}{b-a} L(\omega). \quad (7.4)$$

**Proposition 7.2.** *Any rectifiable path  $\omega : [a, b] \rightarrow S$  has a constant-speed reparametrization  $\bar{\omega} : [0, 1] \rightarrow S$ .*

*Proof.* Let us first reparametrize  $\omega$  so that it is *never locally constant* meaning that there exists no interval  $[c, d] \subset [a, b]$  such that  $\omega_{[c,d]}$  is constant. If such an interval exists, define  $\pi : \mathbb{R} \rightarrow \mathbb{R}$  to be such that

$$\pi(t) = \begin{cases} t & \text{if } t \leq c, \\ c & \text{if } c < t \leq d, \\ t - (d - c) & \text{if } t > d. \end{cases}$$

Observe that  $\pi([a, b]) = [a, b - (d - c)]$  is an interval and that  $\pi$  is continuous and non-decreasing on this interval. Then reparametrize  $\omega$  into  $\omega' : [a, b - (d - c)] \rightarrow S$  such that  $\omega = \omega' \circ \pi$  holds, which is possible since  $\omega$  is constant on  $[c, d]$ .

By repeating this operation, we may assume that  $\omega$  is never locally constant and, in particular, that the map  $t \mapsto \varphi(t) := L(\omega_{[a,t]})/L(\omega)$  is strictly increasing on  $[a, b]$  and continuous by Lemma 7.1 and therefore invertible. In particular,  $\varphi^{-1}$  is also continuous, strictly increasing, and defined over  $[0, 1]$ . We define  $\bar{\omega} : [0, 1] \rightarrow S$  by  $\bar{\omega} = \omega \circ \varphi^{-1}$  which is a constant-speed reparametrization of  $\omega$ .  $\square$

Given  $x, y \in S$ , a path  $\omega : [a, b] \rightarrow S$  is said to *connect* (or *join*)  $x$  to  $y$  if  $\omega(a) = x$  and  $\omega(b) = y$ . By construction of the length function  $L$ ,  $d(x, y) \leq L(\omega)$  for any path  $\omega$  connecting  $x$  to  $y$ . The space  $S$  is called a *length space* if for all  $x, y \in S$ ,

$$d(x, y) = \inf_{\omega} L(\omega), \quad (7.5)$$

where the infimum is taken over all paths  $\omega$  connecting  $x$  to  $y$ . A length space is said to be a *geodesic space* if for all  $x, y \in S$ , the infimum on the right hand side of (7.5) is attained.

**Definition 7.3.** Let  $(S, d)$  be a length space. A geodesic between  $x$  and  $y$  is any path  $\omega : [0, 1] \rightarrow S$  attaining the infimum in (7.5).

In other words, a geodesic is a shortest path between two points. It follows from the minimizing property of a geodesic  $\omega$  that

$$d(\omega(s), \omega(t)) = L(\omega_{[s,t]}),$$

for all  $0 \leq s \leq t \leq 1$ . Together with (7.4) it yields the following useful characterization of *constant-speed geodesics*.

**Proposition 7.4.** A path  $\omega : [0, 1] \rightarrow S$  is a constant-speed geodesic if and only if

$$d(\omega(s), \omega(t)) = (t - s) d(\omega(0), \omega(1)),$$

for all  $0 \leq s \leq t \leq 1$ .

### 7.1.2 Midpoints

We now obtain a characterization of geodesic spaces in terms of midpoints. For any two points  $x, y$  in a metric space, a *midpoint* of  $(x, y)$  is any  $z \in S$  such that

$$d(x, z) = d(y, z) = \frac{1}{2} d(x, y).$$

**Proposition 7.5.** Let  $(S, d)$  be a complete metric space. Then the following are equivalent:

- (i)  $(S, d)$  is a geodesic space.
- (ii) Any two points  $x, y \in M$  admit a midpoint.

*Proof.* We begin with the easy direction: (i)  $\Rightarrow$  (ii). Let  $\omega$  be a geodesic that connects  $x$  to  $y$ , then clearly  $\omega(1/2)$  is a midpoint.

To prove (ii)  $\Rightarrow$  (i), we construct a path  $\omega : [0, 1] \rightarrow S$  such that  $\omega(0) = x$ ,  $\omega(1) = y$  and  $L(\omega) = d(x, y)$ . To that end, we first define  $\omega$  on the set  $\mathcal{D}$  of dyadic rationals of  $[0, 1]$  defined by

$$\mathcal{D} = \{k/2^m : m \geq 1, k \geq 0\} \cap [0, 1].$$

We proceed in a recursive fashion. Let  $z$  be a midpoint of  $(x, y) = (\omega(0), \omega(1))$  and define  $\omega(1/2) = z$ . Given  $H_m := \{\omega(\frac{k}{2^m}), k \in [2^m]\}$ , define  $H_{m+1} = \{\omega(\frac{k}{2^{m+1}}), k \in [2^{m+1}]\}$  by setting  $\omega(\frac{k}{2^{m+1}}) = \omega(\frac{k/2}{2^m}) \in$

$H_m$  if  $k$  is even and letting  $\omega(\frac{k}{2^{m+1}})$  be the midpoint of  $(\omega_{\frac{(k-1)}{2^m}/2}, \omega_{\frac{(k+1)}{2^m}/2})$  when  $k$  is odd. The union of  $H_m$ ,  $m \geq 0$  defines  $\omega$  on  $\mathcal{D}$ .

From our construction, for  $t, t' \in \mathcal{D}$ , it holds

$$d(\omega(t), \omega(t')) = |t - t'| d(x, y) \quad (7.6)$$

so that  $\omega$  is  $d(x, y)$ -Lipschitz on  $\mathcal{D}$ . We now show that  $\omega$  can be extended to a continuous function on  $[0, 1]$  that connects  $x$  to  $y$ . To that end, fix  $t \in [0, 1]$  and let  $(t_n)_{n \geq 0} \subseteq \mathcal{D}$  be a sequence of dyadic integers that converges to  $t$ . Observe that  $(\omega(t_n))_{n \geq 0}$  forms a Cauchy sequence in  $(S, d)$  since by (7.6) it holds

$$d(\omega(t_n), \omega(t_m)) \leq |t_n - t_m| d(x, y) \rightarrow 0, \quad n, m \rightarrow \infty.$$

Therefore since  $S$  is complete,  $(\omega(t_n))_{n \geq 0}$  converges and we set  $\omega(t)$  to be its limit. To see that such an  $\omega$  is continuous, note that for any  $t, u \in [0, 1]$ , there exists sequences  $(t_n)_{n \geq 0}, (u_n)_{n \geq 0} \subseteq \mathcal{D}$  such that  $t_n \rightarrow t$ ,  $u_n \rightarrow u$  and

$$\begin{aligned} d(\omega(t), \omega(u)) &= \lim_{n \rightarrow \infty} d(\omega(t_n), \omega(u_n)) \leq \lim_{n \rightarrow \infty} |t_n - u_n| d(x, y) \\ &= |t - u| d(x, y) \end{aligned}$$

where we used (7.6) in the equality. Therefore, we have constructed a path that connects  $x$  to  $y$ .

To conclude the proof, it suffices to observe that (7.1) and (7.6) imply that  $L(\omega) = d(x, y)$  as desired.  $\square$

### 7.1.3 Geodesics in Wasserstein space

We are now in a position to place the Wasserstein space  $\mathcal{W}_2$  within the framework of metric geometry. Compare the following theorem with Theorem 5.7.

**Theorem 7.6.** *The Wasserstein space  $\mathcal{W}_2$  is a geodesic space. Moreover, let  $\pi_t(x, y) := (1-t)x + t y$ ,  $t \in [0, 1]$ , and for any  $\mu, \nu \in \mathcal{W}_2$  let  $\gamma \in \Gamma_{\mu, \nu}$  be an optimal transport plan in the sense that*

$$\int \|x - y\|^2 \gamma(dx, dy) = W_2^2(\mu, \nu).$$

*Then the path  $\omega$  given by  $\omega(t) = (\pi_t)_{\#} \gamma$  is a constant-speed geodesic in  $\mathcal{W}_2$  connecting  $\omega(0) = \mu$  to  $\omega(1) = \nu$ .*

*Proof.* For any  $0 \leq s \leq t \leq 1$ , define the coupling  $\gamma_{s,t} := (\pi_s, \pi_t)_\# \gamma \in \Gamma_{\omega(s), \omega(t)}$ . Then

$$\begin{aligned} W_2^2(\omega(s), \omega(t)) &\leq \int \|x - y\|^2 \gamma_{s,t}(dx, dy) \\ &= \int \|\pi_s(x, y) - \pi_t(x, y)\|^2 \gamma(dx, dy) \\ &= \int \|(1-s)x + sy - ((1-t)x + ty)\|^2 \gamma(dx, dy) \\ &= (t-s)^2 \int \|x - y\|^2 \gamma(dx, dy) \\ &= (t-s)^2 W_2^2(\omega(0), \omega(1)). \end{aligned}$$

We have proved that

$$W_2(\omega(s), \omega(t)) \leq |t-s| W_2(\omega(0), \omega(1)).$$

To show that this inequality is in fact an equality, note that together with the triangle inequality, it yields

$$\begin{aligned} W_2(\omega(0), \omega(1)) &\leq W_2(\omega(0), \omega(s)) + W_2(\omega(s), \omega(t)) + W_2(\omega(t), \omega(1)) \\ &\leq (s + |t-s| + |1-t|) W_2(\omega(0), \omega(1)) \\ &= W_2(\omega(0), \omega(1)). \end{aligned}$$

Therefore, the above inequalities are equalities and in particular,

$$\begin{aligned} W_2(\omega(0), \omega(s)) + W_2(\omega(s), \omega(t)) + W_2(\omega(t), \omega(1)) \\ = s W_2(\omega(0), \omega(1)) + |t-s| W_2(\omega(0), \omega(1)) + |1-t| W_2(\omega(0), \omega(1)). \end{aligned}$$

Since each term on the left-hand side is smaller than its corresponding part in the right-hand side, we have that

$$W_2(\omega(s), \omega(t)) = |t-s| W_2(\omega(0), \omega(1)),$$

and the conclusion follows from Proposition 7.4. This explicit construction of geodesics joining any pair  $\mu, \nu \in \mathcal{W}_2$  readily implies that  $\mathcal{W}_2$  is indeed a geodesic space.  $\square$

For any constant-speed geodesic  $\omega$  connecting two measures  $\mu, \nu \in \mathcal{W}_2$  and any  $t \in [0, 1]$ , the measure  $\omega(t)$  is often called *displacement interpolation* after [McC97]. Crucially, if  $\mu$  and  $\nu$  have densities  $f_\mu$  and

$f_\nu$ , this interpolation differs from the usual interpolation given by the mixture with density  $(1-t)f_\mu + t f_\nu$ . This is a manifestation of the geometry of  $\mathcal{W}_2$ .

Note that the proof of Theorem 7.6 above implies the following interesting corollary.

**Corollary 7.7.** *Let  $\omega$  be any constant-speed geodesic in  $\mathcal{W}_2$  and let  $\gamma$  be an optimal coupling between  $\omega(0)$  and  $\omega(1)$ . Then for any  $0 \leq s \leq t \leq 1$ , the coupling  $\gamma_{s,t} := (\pi_s, \pi_t)_\# \gamma \in \Gamma_{\omega(s), \omega(t)}$ , where  $\pi_t(x, y) := (1-t)x + t y$ ,  $t \in [0, 1]$ , is optimal in the sense that*

$$\int \|x - y\|^2 \gamma_{s,t}(dx, dy) = W_2^2(\omega(s), \omega(t)).$$

Finally, in the case where the geodesic emanates from a distribution that admits a density, we get from Brenier's Theorem 1.16 the following useful corollary, which justifies Definition 5.8.

**Corollary 7.8.** *Let  $\mu, \nu \in \mathcal{W}_2$  be two probability measures such that  $\mu$  has a density and let  $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$  be the (unique) Brenier map such that  $T_\# \mu = \nu$ . Then, the constant-speed geodesic  $\omega : [0, 1] \rightarrow \mathcal{W}_2$  such that  $\omega(0) = \mu$  and  $\omega(1) = \nu$  is unique and given by*

$$\omega(t) = ((1-t)\text{id} + tT)_\# \mu, \quad \forall t \in [0, 1].$$

where  $\text{id} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  denotes the identity map.

In other words, if  $X \sim \mu$ , then  $(1-t)X + tT(X) \sim \omega(t)$ .

## 7.2 Curvature

### 7.2.1 Alexandrov curvature

Given a real number  $\kappa \in \mathbb{R}$ , a geodesic space of special interest is the (complete and simply connected) 2-dimensional Riemannian manifold with constant sectional curvature  $\kappa$ . For given  $\kappa \in \mathbb{R}$ , this metric space  $(M_\kappa, d_\kappa)$  is unique up to an isometry, and called a *model space*. For each  $\kappa \in \mathbb{R}$ , we use the following representative of the equivalence class generated by the group of isometries.

- If  $\kappa < 0$ ,  $(M_\kappa, d_\kappa)$  is the hyperbolic plane of constant curvature  $\kappa < 0$ .

- If  $\kappa = 0$ ,  $(M_0, d_0)$  is the Euclidean plane  $\mathbb{R}^2$  equipped with its Euclidean metric.
- If  $\kappa > 0$ ,  $(M_\kappa, d_\kappa)$  is the 2-dimensional Euclidean sphere of radius  $1/\sqrt{\kappa}$  equipped with the angular metric.

These model spaces play a central role in metric geometry. As described below, curvature bounds in general metric spaces are formulated by comparison arguments involving these model spaces as benchmarks.

The fundamental device allowing for this comparison is that of comparison triangles. Given a metric space  $(S, d)$ , we define a triangle as any set of three distinct points  $\{p, x, y\} \subset S$ . For  $\kappa \in \mathbb{R}$ , a *comparison triangle* for  $\{p, x, y\}$  in  $M_\kappa$  is an isometric embedding of  $\{p, x, y\}$  in  $M_\kappa$ , i.e., a set  $\{\bar{p}, \bar{x}, \bar{y}\} \subset M_\kappa$  such that

$$d_\kappa(\bar{p}, \bar{x}) = d(p, x), \quad d_\kappa(\bar{p}, \bar{y}) = d(p, y), \quad \text{and} \quad d_\kappa(\bar{x}, \bar{y}) = d(x, y).$$

When  $\kappa \leq 0$ , such a comparison triangle always exists (and is unique up to an isometry). When  $\kappa > 0$ , such a triangle exists (and is unique up to an isometry) provided it fits on the sphere of radius  $\kappa^{-1/2}$ . This condition may be specified in terms of its perimeter:

$$\text{peri}\{p, x, y\} := d(p, x) + d(p, y) + d(x, y) < \frac{2\pi}{\sqrt{\kappa}}. \quad (7.7)$$

For  $\kappa > 0$  say that a triangle  $\{p, x, y\}$  that satisfies (7.7) is *admissible*. When  $\kappa \leq 0$ , all triangles are admissible.

We are now in a position to define curvature bounds for general geodesic spaces.

**Definition 7.9.** Let  $\kappa \in \mathbb{R}$  and  $(S, d)$  be a geodesic space.

- We say that  $\text{curv}(S) \geq \kappa$  if for any admissible triangle  $\{p, x, y\} \subset S$  and any comparison triangle  $\{\bar{p}, \bar{x}, \bar{y}\} \subset M_\kappa$ , the following holds. For any constant-speed geodesics  $\omega : [0, 1] \rightarrow S$  and  $\bar{\omega} : [0, 1] \rightarrow M_\kappa$  joining  $x$  to  $y$  and  $\bar{x}$  to  $\bar{y}$  respectively, it holds

$$d(p, \omega(t)) \geq d_\kappa(\bar{p}, \bar{\omega}(t)), \quad \forall t \in [0, 1]. \quad (7.8)$$

- We say that  $\text{curv}(S) \leq \kappa$  if for any admissible triangle  $\{p, x, y\} \subset S$  and any comparison triangle  $\{\bar{p}, \bar{x}, \bar{y}\} \subset M_\kappa$ , the following holds. For any constant-speed geodesics  $\omega : [0, 1] \rightarrow S$  and  $\bar{\omega} : [0, 1] \rightarrow M_\kappa$  joining  $x$  to  $y$  and  $\bar{x}$  to  $\bar{y}$  respectively, it holds

$$d(p, \omega(t)) \leq d_\kappa(\bar{p}, \bar{\omega}(t)), \quad \forall t \in [0, 1]. \quad (7.9)$$

The previous definition admits a natural geometric interpretation: if  $\text{curv}(S) \geq \kappa$  (resp.  $\text{curv}(S) \leq \kappa$ ), a triangle  $\{p, x, y\}$  looks thicker (resp. thinner) than a corresponding comparison triangle  $\{\bar{p}, \bar{x}, \bar{y}\}$  in the model space  $M_\kappa$ .

The case  $\kappa = 0$  is of special interest since the model space of reference is flat. In that case, one compares our geometry to a familiar Euclidean one. We say that  $S$  is a space of non-positive curvature (NPC) when  $\text{curv}(S) \leq 0$ , and a space of non-negative curvature (NNC) when  $\text{curv}(S) \geq 0$ .

In the flat case the following lemma holds.

**Lemma 7.10.** *Let  $\mathbf{H}$  be a Hilbert space equipped with inner product  $\langle \cdot, \cdot \rangle$  and norm  $\|\cdot\|$ . Then, for any  $p, x, y \in \mathbf{H}$ , the constant-speed geodesic joining  $x$  to  $y$  is unique and given by  $\omega(t) = (1-t)x + t y$  and for any  $p \in \mathbf{H}$ ,*

$$\|p - \omega(t)\|^2 = (1-t) \|p - x\|^2 + t \|p - y\|^2 - t(1-t) \|x - y\|^2, \quad \forall t \in [0, 1].$$

In particular, this holds for the model space  $M_0 = \mathbb{R}^2$ .

*Proof.* It can be easily checked that  $\omega$  is indeed a constant-speed geodesic joining  $x$  to  $y$ . Fix  $t \in [0, 1]$ . To check the equality, observe that on the one hand

$$\begin{aligned} \|p - \omega(t)\|^2 &= \|p - (1-t)x - t y\|^2 \\ &= \|(1-t)(p - x) + t(p - y)\|^2 \\ &= (1-t)^2 \|p - x\|^2 + t^2 \|p - y\|^2 + 2t(1-t) \langle p - x, p - y \rangle. \end{aligned}$$

On the other hand,

$$\|x - y\|^2 = \|x - p + p - y\|^2 = \|p - x\|^2 + \|p - y\|^2 - 2 \langle p - x, p - y \rangle.$$

Putting the above two displays together yields

$$\begin{aligned} \|p - \omega(t)\|^2 &= (1-t)^2 \|p - x\|^2 + t^2 \|p - y\|^2 \\ &\quad + t(1-t) [\|p - x\|^2 + \|p - y\|^2 - \|x - y\|^2] \\ &= (1-t) \|p - x\|^2 + t \|p - y\|^2 - t(1-t) \|x - y\|^2. \end{aligned}$$

It remains to show that  $\omega$  is unique. To that end, let  $\omega'$  by any constant-speed geodesic joining  $x$  to  $y$  and fix  $t \in [0, 1]$ . Apply the above identity to  $p = \omega'(t)$  to get

$$\|\omega'(t) - \omega(t)\|^2 = (1-t) \|\omega'(t) - x\|^2 + t \|\omega'(t) - y\|^2 - t(1-t) \|x - y\|^2.$$

Since  $\omega'$  is a constant-speed geodesic joining  $x$  to  $y$ , we have by Proposition 7.4 that  $\|\omega'(t) - x\| = t \|x - y\|$  and  $\|\omega'(t) - y\| = (1-t) \|x - y\|$ . Therefore

$$\|\omega'(t) - \omega(t)\|^2 = ((1-t)t^2 + t(1-t)^2 - t(1-t)) \|x - y\|^2 = 0,$$

so that  $\omega' = \omega$ .  $\square$

Lemma 7.10 involves only squared distances and can be directly stated in geodesic spaces. It turns out that this generalization gives a useful characterization of NNC or NPC spaces. Note that this characterization does not extend to the cases where the reference space is not flat (i.e., curvature bounded by a non-zero quantity).

**Proposition 7.11.** *Let  $(S, d)$  be a geodesic space. Then  $\text{curv}(S) \geq 0$  if and only if for triangle  $\{p, x, y\} \in S$  and any constant-speed geodesic  $\omega$  joining  $x$  to  $y$ , we have*

$$d^2(p, \omega(t)) \geq (1-t) d^2(p, x) + t d^2(p, y) - t(1-t) d^2(x, y) \quad \forall t \in [0, 1]. \quad (7.10)$$

We have  $\text{curv}(S) \leq 0$  if and only if the same statement holds with the opposite inequality.

*Proof.* Consider a triangle  $\{p, x, y\}$  together with a comparison triangle  $\{\bar{p}, \bar{x}, \bar{y}\} \in \mathbb{R}^2$ .

Assume that  $\text{curv}(S) \geq 0$  in the sense of Definition 7.9. Then for any constant-speed geodesic  $\omega$  that connects  $x$  to  $y$  and  $\bar{\omega}$  the unique constant-speed geodesic that connects  $\bar{x}$  to  $\bar{y}$ , we have by Definition 7.9 and Lemma 7.10 respectively that

$$\begin{aligned} d^2(p, \omega(t)) &\geq \|\bar{p} - \bar{\omega}(t)\|^2 \\ &= (1-t) \|\bar{p} - \bar{x}\|^2 + t \|\bar{p} - \bar{y}\|^2 - t(1-t) \|\bar{x} - \bar{y}\|^2 \\ &= (1-t) d^2(p, x) + t d^2(p, y) - t(1-t) d^2(x, y). \end{aligned}$$

To prove the converse, note that (7.10) yields

$$\begin{aligned} d^2(p, \omega(t)) &\geq (1-t) d^2(p, x) + t d^2(p, y) - t(1-t) d^2(x, y) \\ &= (1-t) \|\bar{p} - \bar{x}\|^2 + t \|\bar{p} - \bar{y}\|^2 - t(1-t) \|\bar{x} - \bar{y}\|^2 \\ &= \|\bar{p} - \bar{\omega}(t)\|^2, \end{aligned}$$

by Lemma 7.10 so that Definition 7.9 holds.  $\square$

*Remark 7.12.* Comparing with the definition of  $\alpha$ -convexity in Appendix A, we see that  $\frac{1}{2}\|p - \cdot\|^2$  is 1-strongly convex in any Hilbert space. Similarly,  $(S, d)$  is an NPC space if and only if  $\frac{1}{2}d^2(p, \cdot)$  is 1-strongly convex *along the geodesics* of  $(S, d)$ . The notion of geodesic convexity was also used in Section 5.2, but here we work in the more general setting of geodesic spaces.

A geodesic space  $(S, d)$  with any curvature bound is called an *Alexandrov space*. If  $\text{curv}(S) \leq \kappa$  for some  $\kappa \in \mathbb{R}$ , then  $(S, d)$  is sometimes called a  $\text{CAT}(\kappa)$  space in reference to E. Cartan, A. D. Alexandrov, and V. A. Toponogov. As noted before, a  $\text{CAT}(0)$  space is also referred to as an NPC (non-positively curved) or sometimes *Hadamard* space. If  $\text{curv}(S) \geq 0$  we call the space *non-negatively curved* or NNC. It is worth noting that the previous definitions are of global nature as they require comparison inequalities to be valid for all triangles (that admit a comparison triangle in the relevant model space). Some definitions of curvature require the previous comparison inequalities to hold only locally. The local validity of these comparison inequalities is known, under suitable conditions depending on the value of  $\kappa$ , to imply their global validity (globalization theorems).

We conclude this subsection by giving a third equivalent definition of non-negative curvature.

**Proposition 7.13.** *Let  $(S, d)$  be a geodesic space. Then  $\text{curv}(S) \geq 0$  if and only if for any triangle  $\{p, x, y\} \subset S$ , comparison triangle  $\{\bar{p}, \bar{x}, \bar{y}\} \subset M_0$ , and any constant-speed geodesics  $\omega, \omega', \bar{\omega}, \bar{\omega}'$  joining  $p$  to  $x$ ,  $p$  to  $y$ ,  $\bar{p}$  to  $\bar{x}$ , and  $\bar{p}$  to  $\bar{y}$  respectively, we have*

$$d^2(\omega(s), \omega'(t)) \geq \|\bar{\omega}(s) - \bar{\omega}'(t)\|^2, \quad \forall s, t \in [0, 1]. \quad (7.11)$$

We have  $\text{curv}(S) \leq 0$  if and only if the same statement holds with the opposite inequality.

*Proof.* Assume first that  $\text{curv}(S) \geq 0$  and observe that by Proposition 7.11 and Definition 7.9 respectively, it holds

$$\begin{aligned} d^2(\omega(s), \omega'(t)) &\geq (1-s)d^2(p, \omega'(t)) + s d^2(x, \omega'(t)) - s(1-s)d^2(p, x) \\ &\geq (1-s)\|\bar{p} - \bar{\omega}'(t)\|^2 + s\|\bar{x} - \bar{\omega}'(t)\|^2 - s(1-s)\|\bar{p} - \bar{x}\|^2. \end{aligned}$$

The right-hand side of the above inequality is precisely  $\|\bar{\omega}(s) - \bar{\omega}'(t)\|^2$  by Lemma 7.10. We have proved (7.11).

Conversely, let  $\omega_x, \omega'_x$  be constant-speed geodesics joining  $x$  to  $y$  and  $x$  to  $p$ , respectively, and let  $\bar{\omega}_{\bar{x}}, \bar{\omega}'_{\bar{x}}$  be constant-speed geodesics joining  $\bar{x}$  to  $\bar{y}$  and  $\bar{x}$  to  $\bar{p}$  respectively. Then, taking  $s = 1$  in (7.11), we get for any  $t \in [0, 1]$ ,

$$\begin{aligned} d^2(p, \omega_x(t)) &= d^2(\omega'_x(1), \omega_x(t)) \\ &\geq \|\bar{p} - \bar{\omega}_{\bar{x}}(t)\|^2 \\ &= (1-t) \|\bar{p} - \bar{x}\|^2 + t \|\bar{p} - \bar{y}\|^2 - t(1-t) \|\bar{x} - \bar{y}\|^2 \\ &= (1-t) d^2(p, x) + t d^2(p, y) - t(1-t) d^2(x, y), \end{aligned}$$

which is the characterization of  $\text{curv}(S) \geq 0$  from Proposition 7.11.

The proof for  $\text{curv}(S) \leq 0$  follows using the same argument.  $\square$

### 7.2.2 Curvature of the Wasserstein space

Note that if  $d = 1$ , the space  $\mathcal{P}_2(\mathbb{R})$  equipped with the Wasserstein distance is actually *flat*.

**Proposition 7.14.** *The space  $\mathcal{W}_2(\mathbb{R})$  is flat in the sense that*

$$\text{curv}(\mathcal{W}_2(\mathbb{R})) \leq 0 \quad \text{and} \quad \text{curv}(\mathcal{W}_2(\mathbb{R})) \geq 0$$

and it can be isometrically embedded into a Hilbert space.

*Proof.* Recall from Proposition 1.18 that for any  $\mu, \nu \in \mathcal{W}_{2,\text{ac}}$ , it holds

$$W_2^2(\mu, \nu) = \int_0^1 |F_\mu^\dagger(u) - F_\nu^\dagger(u)|^2 du = \|F_\mu^\dagger - F_\nu^\dagger\|^2,$$

where  $\|\cdot\| := \|\cdot\|_{L^2([0,1])}$ . In particular, the map  $\mu \mapsto F_\mu^\dagger$  is an isometry from  $\mathcal{W}_{2,\text{ac}}$  to  $L^2([0, 1])$ .

Let now  $\omega$  be a constant-speed geodesic that connects  $\mu$  to  $\nu$  and recall from Proposition 1.18 and Theorem 7.6 that  $\omega$  is uniquely characterized by the fact that if  $V = (1-t)F_\mu^\dagger(U) + tF_\nu^\dagger(U)$ , where  $U \sim \text{Unif}([0, 1])$ , then  $W \sim \omega(t)$ . It yields that for any  $v \in \mathbb{R}$ ,

$$\mathbb{P}(V \leq v) = \mathbb{P}((1-t)F_\mu^\dagger(U) + tF_\nu^\dagger(U) \leq v) = ((1-t)F_\mu^\dagger + tF_\nu^\dagger)^\dagger(v).$$

Hence,

$$F_{\omega(t)}^\dagger = (1-t)F_\mu^\dagger + tF_\nu^\dagger.$$

Next, let  $\rho \in \mathcal{W}_2$  and  $t \in [0, 1]$ . Since  $L^2(\mathbb{R})$  is a Hilbert space, we get from Lemma 7.10 that

$$\begin{aligned} W_2^2(\rho, \omega(t)) &= \|F_\rho^\dagger - F_{\omega(t)}^\dagger\|^2 = \|F_\rho^\dagger - (1-t)F_\mu^\dagger - tF_\nu^\dagger\|^2 \\ &= (1-t)\|F_\rho^\dagger - F_\mu^\dagger\|^2 + t\|F_\rho^\dagger - F_\nu^\dagger\|^2 - t(1-t)\|F_\mu^\dagger - F_\nu^\dagger\|^2 \\ &= (1-t)W_2^2(\rho, \mu) + tW_2^2(\rho, \nu) - t(1-t)W_2^2(\mu, \nu). \end{aligned}$$

This completes the proof that  $\text{curv}(\mathcal{W}_{2,\text{ac}}(\mathbb{R})) = 0$ . In turn, one can show that this implies  $\text{curv}(\mathcal{W}_2(\mathbb{R})) = 0$  as well.  $\square$

More generally, for any  $d \geq 1$ ,  $\mathcal{W}_2(\mathbb{R}^d)$  is positively curved as indicated by the theorem below.

**Theorem 7.15.** *The 2-Wasserstein space  $\mathcal{W}_2$  is non-negatively curved,*

$$\text{curv}(\mathcal{W}_2) \geq 0,$$

i.e., for any  $\mu, \nu, \rho \in \mathcal{W}_2$  and any constant-speed geodesic  $\omega$  that connects  $\mu$  to  $\nu$ , it holds

$$W_2^2(\rho, \omega(t)) \geq (1-t)W_2^2(\rho, \mu) + tW_2^2(\rho, \nu) - t(1-t)W_2^2(\mu, \nu).$$

*Proof.* Let  $\gamma \in \Gamma_{\mu, \nu}$  be an optimal coupling and recall from Theorem 7.6 that for any  $t \in [0, 1]$ ,  $\omega(t) = (\pi_t)_\# \gamma$  where  $\pi_t(x, y) = (1-t)x + t y$ . In particular, if  $(X, Y) \sim \gamma$ , then  $V_t := (1-t)X + tY \sim \omega(t)$ . Next, let  $\gamma_t \in \Gamma_{\omega(t), \rho}$  be an optimal coupling between  $\omega(t)$  and  $\rho$ . In particular, it induces a conditional distribution on  $Z \sim \rho$  given  $V_t$ . We have described a joint distribution  $\Upsilon$  for  $(X, Y, Z)$  that has marginals  $\mu$ ,  $\nu$ , and  $\rho$  respectively (the reader will have recognized a variant of the gluing lemma, Lemma B.5).

With this notation, we have

$$\begin{aligned} W_2^2(\rho, \omega(t)) &= \int \|z - v\|^2 \gamma_t(dx, dv) \\ &= \int \|z - (1-t)x - t y\|^2 \Upsilon(dx, dy, dz). \end{aligned}$$

Next, observe that by Lemma 7.10 applied to  $\mathbf{H} = \mathbb{R}^d$ , we have

$$\|z - (1-t)x - t y\|^2 = (1-t)\|z - x\|^2 + t\|z - y\|^2 - t(1-t)\|x - y\|^2,$$

so that

$$\begin{aligned}
W_2^2(\rho, \omega(t)) &= \int [(1-t) \|z - x\|^2 + t \|z - y\|^2 \\
&\quad - t(1-t) \|x - y\|^2] \Upsilon(dx, dy, dz) \\
&\geq (1-t) W_2^2(\rho, \mu) + t W_2^2(\rho, \nu) \\
&\quad - t(1-t) \int \|x - y\|^2 \Upsilon(dx, dy, dz) \\
&= (1-t) W_2^2(\rho, \mu) + t W_2^2(\rho, \nu) - t(1-t) W_2^2(\mu, \nu),
\end{aligned}$$

where in the inequality, we used the suboptimality of the first two couplings and in the last equality, we used the optimality of the coupling between  $\mu$  and  $\nu$  induced by  $\Upsilon$ .  $\square$

### 7.3 Tangent cones

A geodesic space has a priori no differentiable structure but a surrogate for it may be built. It starts from the notion of *angle* which can be defined on any metric space by analogy to the Hilbert case, akin to our definition of curvature bounds. When applied to a geodesic space, angles allow us to define the notion of *direction*, which can be thought of as the initial velocity of a constant-speed geodesic. The collection of such directions forms the tangent cone.

#### 7.3.1 Angles

We first define angles on a metric space and show that they provide alternative characterizations of curvature bounds for geodesic spaces.

Recall that for any three points  $p, x, y \in \mathbb{R}^2$ , the cosine of the angle  $\angle_p(x, y)$ , formed by vectors  $\vec{px}$  and  $\vec{py}$  is given by

$$\cos \angle_p(x, y) = \frac{\langle x - p, y - p \rangle}{\|x - p\| \|y - p\|}.$$

Note that

$$\begin{aligned}
-2 \langle x - p, y - p \rangle &= \|(x - p) - (y - p)\|^2 - \|x - p\|^2 - \|y - p\|^2 \\
&= \|x - y\|^2 - \|x - p\|^2 - \|y - p\|^2.
\end{aligned}$$

Therefore, we can rewrite this definition only in terms of squared distances to obtain

$$\cos \angle_p(x, y) = \frac{\|x - p\|^2 + \|y - p\|^2 - \|x - y\|^2}{2 \|x - p\| \|y - p\|}.$$

This definition generalizes to any metric space.

**Definition 7.16.** Let  $(S, d)$  be a metric space and for any triangle  $\{p, x, y\}$  in  $S$ , define the angle  $\angle_p(x, y) \in [0, \pi]$  at  $p$  by

$$\cos \angle_p(x, y) := \frac{d^2(p, x) + d^2(p, y) - d^2(x, y)}{2 d(p, x) d(p, y)}.$$

Similar comparisons may be made with model spaces  $M_\kappa$  for  $\kappa \neq 0$  but are beyond the scope of these lectures.

The next result presents a characterization of positively curved spaces in terms of the *angle monotonicity*.

**Proposition 7.17 (Angle monotonicity).** Let  $(S, d)$  be a geodesic space. Then,  $\text{curv}(S) \geq 0$  in the sense of Definition 7.9, if and only if for any triangle  $\{p, x, y\}$  in  $S$  and any geodesics  $\omega$  and  $\omega'$  joining  $p$  to  $x$  and  $p$  to  $y$  respectively, the function

$$(s, t) \in [0, 1]^2 \mapsto \angle_p(\omega(s), \omega'(t))$$

is non-increasing in each variable when the other is fixed.

*Proof.* Assume first that  $\text{curv}(S) \geq 0$  and consider a triangle  $\{p, x, y\}$  in  $S$  and constant-speed geodesics  $\omega$  and  $\omega'$  joining  $p$  to  $x$  and  $p$  to  $y$  respectively. It is enough to prove that, for all  $(s, t) \in [0, 1]^2$ ,

$$\cos \angle_p(\omega(s), \omega'(t)) \leq \cos \angle_p(x, y).$$

Let  $\{\bar{p}, \bar{x}, \bar{y}\}$  be a comparison triangle for  $\{p, x, y\}$  in  $M_0 = \mathbb{R}^2$  and let  $\bar{\omega}$  and  $\bar{\omega}'$  be constant-speed geodesics in  $M_0$  connecting  $\bar{p}$  to  $\bar{x}$  and  $\bar{p}$  to  $\bar{y}$  respectively. It holds

$$\begin{aligned} \cos \angle_p(\omega(s), \omega'(t)) &= \frac{d^2(p, \omega(s)) + d^2(p, \omega'(t)) - d^2(\omega(s), \omega'(t))}{2 d(p, \omega(s)) d(p, \omega'(t))} \\ &= \frac{s^2 d^2(p, x) + t^2 d^2(p, y) - d^2(\omega(s), \omega'(t))}{2 s t d(p, x) d(p, y)} \\ &= \frac{s^2 \|\bar{p} - \bar{x}\|^2 + t^2 \|\bar{p} - \bar{y}\|^2 - d^2(\omega(s), \omega'(t))}{2 s t \|\bar{p} - \bar{x}\| \|\bar{p} - \bar{y}\|} \\ &\leq \frac{s^2 \|\bar{p} - \bar{x}\|^2 + t^2 \|\bar{p} - \bar{y}\|^2 - \|\bar{\omega}(s) - \bar{\omega}'(t)\|^2}{2 s t \|\bar{p} - \bar{x}\| \|\bar{p} - \bar{y}\|} \end{aligned}$$

$$\begin{aligned}
&= \frac{\|\bar{p} - \bar{\omega}(s)\|^2 + \|\bar{p} - \bar{\omega}'(t)\|^2 - \|\bar{\omega}(s) - \bar{\omega}'(t)\|^2}{2 \|\bar{p} - \bar{\omega}(s)\| \|\bar{p} - \bar{\omega}'(t)\|} \\
&= \cos \angle_{\bar{p}}(\bar{\omega}(s), \bar{\omega}'(t)) \\
&= \cos \angle_{\bar{p}}(\bar{x}, \bar{y}) \\
&= \cos \angle_p(x, y),
\end{aligned}$$

where in the inequality we used the fact that  $\text{curv}(S) \geq 0$  and Proposition 7.13. This completes the proof that the curvature lower bound implies angle monotonicity.

Conversely, assume that for any triangle  $\{p, x, y\}$  in  $S$ , any constant-speed geodesics  $\omega$  and  $\omega'$  connecting  $p$  to  $x$  and  $p$  to  $y$  respectively, and all  $(s, t) \in [0, 1]^2$ , we have

$$\cos \angle_p(\omega(s), \omega'(t)) \leq \cos \angle_p(x, y).$$

Then, the first part of the proof implies that

$$d^2(\omega(s), \omega'(t)) \geq \|\bar{\omega}(s) - \bar{\omega}'(t)\|^2, \quad \forall (s, t) \in [0, 1]^2$$

with the same notation as above, which is the characterization of  $\text{curv}(S) \geq 0$  from Proposition 7.13.  $\square$

### 7.3.2 Directions

From the notion of angles between points, we can readily define an angle between constant-speed geodesics.

Let  $(S, d)$  be a geodesic space such that  $\text{curv}(S) \geq 0$ ,  $p \in S$ , and  $\omega, \omega'$  two constant-speed geodesics connecting  $p$  to  $x$  and  $y$  respectively. We define the angle between  $\omega$  and  $\omega'$  as

$$\angle(\omega, \omega') := \lim_{s, t \searrow 0} \angle_p(\omega(s), \omega'(t)).$$

It follows from Proposition 7.17 that this limit exists under the assumption  $\text{curv}(S) \geq 0$ . In fact, under the same assumption,

$$\angle(\omega, \omega') = \lim_{t \searrow 0} \angle_p(\omega(t), \omega'(t)).$$

Given a third constant-speed geodesic  $\omega'' : [0, 1] \rightarrow S$  such that  $\omega''(0) = p$  and  $\omega''(1) = z$ , it can be shown (see Exercise 4) that we have the triangular inequality

$$\angle(\omega, \omega') \leq \angle(\omega, \omega'') + \angle(\omega'', \omega'), \quad (7.12)$$

so that  $\angle$  is a pseudo-metric on the set  $\mathcal{G}(p)$  of all constant-speed geodesics emanating from  $p$ . Next, we define the equivalence relation  $\sim$  on  $\mathcal{G}(p)$  by

$$\omega \sim \omega' \Leftrightarrow \angle(\omega, \omega') = 0.$$

We can turn  $\angle$  into a proper metric (still denoted  $\angle$ ) on the quotient  $\mathcal{G}(p)/\sim$ .

**Definition 7.18.** *The space of directions emanating from  $p$  is the completion  $(\Sigma_p, \angle)$  of  $(\mathcal{G}(p)/\sim, \angle)$ . An element of  $\Sigma_p$  is called a direction.*

### 7.3.3 Tangent cone

An analog of a tangent space for geodesic spaces is provided by the notion of a tangent cone.

**Definition 7.19 (Tangent cone).** *Let  $(S, d)$  be a geodesic space with positive curvature and fix  $p \in S$ . The tangent cone  $T_p S$  at  $p$  is the Euclidean cone over the space of directions  $(\Sigma_p, \angle)$ . In other words,  $T_p S$  is the metric space:*

- whose underlying set consists in equivalence classes in  $\Sigma_p \times [0, +\infty)$  for the equivalence relation  $\sim$  defined by

$$(\omega, s) \sim (\omega', t) \Leftrightarrow \begin{cases} s = t = 0 \\ \text{or } \omega = \omega' \text{ and } s = t \end{cases}$$

- and whose metric  $d_p$  is defined

$$d_p((\omega, s), (\omega', t)) := \sqrt{s^2 + t^2 - 2st \cos \angle(\omega, \omega')}.$$

For  $u = (\omega, s)$  and  $v = (\omega', t) \in T_p S$ , we write  $\|u - v\|_p := d_p(u, v)$ ,  $\|u\|_p := d_p(o_p, u)$ , where  $o_p = (\omega, 0) \in T_p S$  is the tip of the cone and

$$\langle u, v \rangle_p := \|u\|_p \|v\|_p \cos \angle(\omega, \omega') = \frac{1}{2} (\|u\|_p^2 + \|v\|_p^2 - \|u - v\|_p^2).$$

The terminology *cone* and the notation  $\|\cdot\|_p$  and  $\langle \cdot, \cdot \rangle_p$  introduced above is justified by the fact that the cone  $T_p S$  possesses a Hilbert-like structure described below. As often the case in metric geometry, the definition of  $d_p$  comes from rewriting a Euclidean notion using only notions that exist on a geodesic space, namely distances and angles in

this case. Indeed, if  $\omega, \omega'$  are points on the sphere and  $(\omega, s) := s \cdot \omega$ ,  $(\omega', t) := t \cdot \omega'$  then it follows from the law of cosines that their squared distance is given by

$$\|s \cdot \omega - t \cdot \omega'\|^2 = s^2 + t^2 - 2st \cos \angle(\omega, \omega'). \quad (7.13)$$

For a point  $u = (\omega, t)$  and  $\lambda \geq 0$ , we define  $\lambda \cdot u := (\omega, \lambda t)$ . Moreover, it may be checked using the previous definitions that, for any  $u, v \in T_p S$  and any  $\lambda \geq 0$ , we get

$$\|\lambda \cdot u\|_p = \lambda \|u\|_p \quad \text{and} \quad \langle \lambda \cdot u, v \rangle_p = \langle u, \lambda \cdot v \rangle_p = \lambda \langle u, v \rangle_p.$$

Note that the tangent cone may not be geodesic but in cases when it is, the sum of points  $u, v \in T_p S$  is defined as the midpoint of  $2 \cdot u$  and  $2 \cdot v$  as defined in Definition 7.5. In this case,  $T_p S$  is indeed a cone.

An example to keep in mind is when  $S$  is a filled-in square in the plane  $\mathbb{R}^2$ . Then, the tangent cone at one of the corners of  $S$  is “missing” some directions (the ones that would lead out of  $S$ ) and is therefore not a vector space, hence why we do not call it the tangent “space”.

The logarithmic map plays an important role in the sequel. For all  $x \in S$ , we denote  $\uparrow_p^x \subset \Sigma_p$  the set of all equivalence classes of constant-speed geodesics connecting  $p$  to  $x$  in  $S$ . Then for every  $x \in S$ , we arbitrarily choose one direction  $\uparrow_p^x \in \uparrow_p^x$ .

**Definition 7.20 (Logarithmic map).** *Let  $(S, d)$  be a positively curved geodesic space. Then, having chosen  $\uparrow_p^x \in \uparrow_p^x$  for every  $x \in S$ , the associated logarithmic map is defined by*

$$\log_p : x \in S \mapsto (\uparrow_p^x, d(p, x)) \in T_p S.$$

At this level of generality, the definition of  $\log_p(x)$  depends on the choice of directions  $\{\uparrow_p^x, x \in S\}$ . This ambiguity may be removed by restricting the  $\log_p$  map to an appropriate subset, namely the set of points  $x \in S$  for which there is only one equivalence class of directions of constant-speed geodesics connecting  $p$  to  $x$ . This set might be specified even more accurately by observing that, in an NNC space  $S$ , if constant-speed geodesics  $\omega$  and  $\omega'$  from  $p$  to  $x$  satisfy  $\angle(\omega, \omega') = 0$ , then  $\omega = \omega'$ . In other words, the set of points  $x \in S$  for which there is more than one equivalence class of constant-speed geodesics connecting  $p$  to  $x$  is exactly the set of points  $x$  connected to  $p$  by at least two distinct constant-speed geodesics. This set of points is denoted  $C(p)$  and called

the *cut-locus* of  $p$ . Then, for any  $x \in S \setminus C(p)$ ,  $\log_p(x)$  is defined without ambiguity as

$$\log_p(x) = (\omega_x, d(p, x)),$$

where  $\omega_x$  denotes the unique constant-speed geodesic from  $p$  to  $x$  therefore identified to its direction. With this notation, we get in particular, for all  $t \in [0, 1]$  and all  $x \in S \setminus C(p)$ , that

$$\log_p \omega_x(t) = t \log_p x.$$

More generally, if  $\omega : [0, 1] \rightarrow S$  is a constant-speed geodesic in  $S$  and  $p$  is such that  $p = \omega(t)$  for some  $t \in [0, 1]$ , i.e.,  $p$  is on the geodesic, then

$$\log_p \omega(s) = (1 - s) \log_p \omega(0) + s \log_p \omega(1). \quad (7.14)$$

In other words, the logarithmic maps turns geodesics into straight lines.

The following result shows that the  $\log_p$  map is expanding in a space of positive curvature.

**Proposition 7.21.** *Let  $(S, d)$  be a geodesic space and  $p \in S$  be fixed. If  $\text{curv}(S) \geq 0$ , then the logarithmic map is expansive in the sense that then for all  $x, y \in S$ ,*

$$d(x, y) \leq \|\log_p(x) - \log_p(y)\|_p,$$

with equality if  $x = p$  or  $y = p$ .

*Proof.* Let  $\omega, \omega'$  be two constant-speed geodesics connecting  $p$  to  $x$  and  $y$  respectively. Then if  $p \neq x$  and  $p \neq y$ , we have by definition of  $\|\cdot\|_p$  and angle monotonicity that for all  $s, t \in [0, 1]$ , it holds

$$\begin{aligned} \|\log_p(x) - \log_p(y)\|_p^2 &= d_p^2(\log_p(x), \log_p(y)) \\ &= d^2(p, x) + d^2(p, y) - 2d(p, x)d(p, y)\cos\angle(\omega, \omega') \\ &\geq d^2(p, x) + d^2(p, y) - 2d(p, x)d(p, y)\cos\angle_p(\omega(s), \omega'(t)). \end{aligned}$$

Applying Definition 7.16, we get

$$\begin{aligned} \|\log_p(x) - \log_p(y)\|_p^2 &\geq d^2(p, x) + d^2(p, y) - \frac{s^2d^2(p, x) + t^2d^2(p, y) - d^2(\omega(s), \omega'(t))}{st}. \end{aligned}$$

Letting  $s = t = 1$  yields

$$\|\log_p(x) - \log_p(y)\|_p^2 \geq d^2(x, y).$$

It is easy to check the equality cases from the definition of  $d_p$ .  $\square$

### 7.3.4 Tangent cone of the Wasserstein space

Going to the very definition of the tangent cone, we can show that it takes a very simple form in the Wasserstein case. To that end, let  $\mu, \nu, \rho \in \mathcal{W}_2$  be three probability distributions. Moreover, let  $\omega_\nu$  and  $\omega_\rho$  be two geodesics in the 2-Wasserstein space  $\mathcal{W}_2$  joining  $\mu$  to  $\nu$  and  $\mu$  to  $\rho$  respectively and recall that the tangent cone at  $\mu$  is the metric space of directions at  $\mu$  equipped with distance  $d_\mu$  such that

$$\begin{aligned} d_\mu^2((\omega_\nu, d(\mu, \nu)), (\omega_\rho, d(\mu, \rho))) \\ = W_2^2(\mu, \nu) + W_2^2(\mu, \rho) - 2 W_2(\mu, \nu) W_2(\mu, \rho) \cos \angle(\omega_\nu, \omega_\rho). \end{aligned}$$

What is the angle  $\angle(\omega_\nu, \omega_\rho)$  between these two Wasserstein geodesics? We can carry out a calculation assuming that  $\mu$  has a density so that Brenier's theorem ensures the existence of two optimal transport maps  $T_{\mu \rightarrow \nu}$  and  $T_{\mu \rightarrow \rho}$  so that

$$\begin{aligned} \cos \angle(\omega_\nu, \omega_\rho) &= \lim_{t \searrow 0} \frac{W_2^2(\mu, \omega_\nu(t)) + W_2^2(\mu, \omega_\rho(t)) - W_2^2(\omega_\nu(t), \omega_\rho(t))}{2 W_2(\mu, \omega_\nu(t)) W_2(\mu, \omega_\rho(t))} \\ &= \lim_{t \searrow 0} \frac{t^2 W_2^2(\mu, \nu) + t^2 W_2^2(\mu, \rho) - W_2^2(\omega_\nu(t), \omega_\rho(t))}{2t^2 W_2(\mu, \nu) W_2(\mu, \rho)} \\ &= \frac{W_2^2(\mu, \nu) + W_2^2(\mu, \rho) - \lim_{t \searrow 0} \frac{W_2^2(\omega_\nu(t), \omega_\rho(t))}{t^2}}{2 W_2(\mu, \rho) W_2(\mu, \nu)}. \end{aligned}$$

**Lemma 7.22.** *Let  $\mu \in \mathcal{P}_{2, \text{ac}}(\mathbb{R}^d)$  and denote by  $T_{\mu \rightarrow \nu}$  and  $T_{\mu \rightarrow \rho}$  the Brenier maps from  $\mu$  to  $\nu$  and  $\mu$  to  $\rho$  respectively. Then*

$$\lim_{t \searrow 0} \frac{W_2^2(\omega_\nu(t), \omega_\rho(t))}{t^2} = \|T_{\mu \rightarrow \nu} - T_{\mu \rightarrow \rho}\|_{L^2(\mu)}^2.$$

*Proof.* It is easy to show one of the required inequalities. Indeed, let  $X \sim \mu$  and observe that

$$\begin{aligned} X_t^\nu &:= (1-t)X + tT_{\mu \rightarrow \nu}(X) \sim \omega_\nu(t), \\ X_t^\rho &:= (1-t)X + tT_{\mu \rightarrow \rho}(X) \sim \omega_\rho(t). \end{aligned}$$

Therefore,

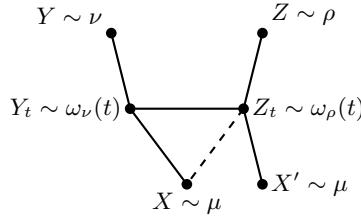
$$\begin{aligned} W_2^2(\omega_\nu(t), \omega_\rho(t)) &\leq \mathbb{E}\|X_t^\nu - X_t^\rho\|^2 \\ &= t^2 \mathbb{E}\|T_{\mu \rightarrow \nu}(X) - T_{\mu \rightarrow \rho}(X)\|^2 \end{aligned}$$

$$= t^2 \|T_{\mu \rightarrow \nu} - T_{\mu \rightarrow \rho}\|_{L^2(\mu)}^2. \quad (7.15)$$

To prove the converse, for any  $t \in [0, 1]$ , let  $\Upsilon_t$  be the following coupling between five random variables:  $(X, Y, Z, Y_t, Z_t) \sim \Upsilon_t$  if

1.  $X \sim \mu$  and  $Y = T_{\mu \rightarrow \nu}(X) \sim \nu$  are optimally coupled,
2.  $Y_t = (1-t)X + tY \sim \omega_{\nu}(t)$ ,
3.  $Z_t \sim \omega_{\rho}(t)$  and  $Y_t \sim \omega_{\nu}(t)$  are optimally coupled,
4.  $Z_t \sim \omega_{\rho}(t)$  and  $Z \sim \rho$  are optimally coupled,
5.  $Z_t \sim \omega_{\rho}(t)$  and  $X' \sim \mu$  are optimally coupled.

Figure 7.1 indicates that this joint coupling can be realized using the gluing lemma since the resulting graph is acyclic. Note in particular that  $Z_t = (1-t)X' + tZ$ .



**Fig. 7.1.** The coupling  $\Upsilon_t$  between  $(\mu, \nu, \rho, \omega_{\nu}(t), \omega_{\rho}(t))$ . Solid lines indicate optimal couplings. Dashed lines and missing lines indicate potentially suboptimal ones.

We have

$$\begin{aligned} W_2^2(\omega_{\nu}(t), \omega_{\rho}(t)) &= \mathbb{E}\|Y_t - Z_t\|^2 = \mathbb{E}\|(1-t)X + tY - Z_t\|^2 \\ &= (1-t)\mathbb{E}\|X - Z_t\|^2 + t\mathbb{E}\|Y - Z_t\|^2 - t(1-t)\mathbb{E}\|X - Y\|^2, \end{aligned} \quad (7.16)$$

where we used Lemma 7.10 for  $\mathbf{H} = \mathbb{R}^d$ . Now note that  $X$  and  $Z_t$  are potentially coupled in a suboptimal way so that

$$(1-t)\mathbb{E}\|X - Z_t\|^2 \geq (1-t)W_2^2(\mu, \omega_{\rho}(t)) = t^2(1-t)W_2^2(\mu, \rho).$$

Moreover, using Lemma 7.10 again, we get

$$\begin{aligned} t\mathbb{E}\|Y - Z_t\|^2 &= t\mathbb{E}\|Y - (1-t)X' + tZ\|^2 \\ &= t(1-t)\mathbb{E}\|Y - X'\|^2 + t^2\mathbb{E}\|Y - Z\|^2 - t^2(1-t)\mathbb{E}\|X' - Z\|^2 \\ &\geq t(1-t)\mathbb{E}\|Y - X\|^2 + t^2\mathbb{E}\|Y - Z\|^2 - t^2(1-t)\mathbb{E}\|X' - Z\|^2, \end{aligned}$$

where in the above inequality, we used the fact that  $X$  and  $Y$  are optimally coupled.

Plugging the above two displays in (7.16), we see that the terms in  $\mathbb{E}\|Y - X\|^2$  can cancel out. We get

$$\begin{aligned} W_2^2(\omega_\nu(t), \omega_\rho(t)) \\ \geq t^2(1-t)W_2^2(\mu, \rho) + t^2\mathbb{E}\|Y - Z\|^2 - t^2(1-t)\mathbb{E}\|X' - Z\|^2 \\ = t^2\mathbb{E}\|Y - Z\|^2. \end{aligned}$$

Recall from (7.15) that  $W_2(\omega_\nu(t), \omega_\rho(t)) = O(t)$ , so that  $\mathbb{E}\|Y_t - Z_t\|^2 = O(t^2)$ . Since

$$X = (Y_t - tY)/(1-t) \quad \text{and} \quad X' = (Z_t - tZ)/(1-t),$$

it implies  $\mathbb{E}\|X - X'\|^2 = O(t^2)$  as well. Assuming (without justification) that  $T_{\mu \rightarrow \rho}$  is Lipschitz<sup>1</sup>

$$\begin{aligned} \mathbb{E}\|Y - Z\|^2 &= \mathbb{E}\|T_{\mu \rightarrow \nu}(X) - T_{\mu \rightarrow \rho}(X')\|^2 \\ &= \mathbb{E}\|T_{\mu \rightarrow \nu}(X) - T_{\mu \rightarrow \rho}(X)\|^2 - O(t). \end{aligned}$$

This readily yields

$$\lim_{t \searrow 0} \frac{W_2^2(\omega_\nu(t), \omega_\rho(t))}{t^2} \geq \|T_{\mu \rightarrow \nu} - T_{\mu \rightarrow \rho}\|_{L^2(\mu)}^2,$$

which concludes the proof of our Lemma.  $\square$

It follows from Lemma 7.22 that

$$\begin{aligned} \cos \angle(\omega_\nu, \omega_\rho) &= \frac{W_2^2(\mu, \nu) + W_2^2(\mu, \rho) - \|T_{\mu \rightarrow \nu} - T_{\mu \rightarrow \rho}\|_{L^2(\mu)}^2}{2W_2(\mu, \nu)W_2(\mu, \rho)} \\ &= \frac{\|T_{\mu \rightarrow \nu} - \text{id}\|_{L^2(\mu)}^2 + \|T_{\mu \rightarrow \rho} - \text{id}\|_{L^2(\mu)}^2 - \|T_{\mu \rightarrow \nu} - T_{\mu \rightarrow \rho}\|_{L^2(\mu)}^2}{2\|T_{\mu \rightarrow \nu} - \text{id}\|_{L^2(\mu)}\|T_{\mu \rightarrow \rho} - \text{id}\|_{L^2(\mu)}} \\ &= \cos \angle(T_{\mu \rightarrow \nu} - \text{id}, T_{\mu \rightarrow \rho} - \text{id}), \end{aligned} \tag{7.17}$$

where the last cos is understood in the Hilbert space  $L^2(\mu)$ .

In turn, the law of cosines (7.13) implies that the metric on the tangent cone at  $\mu$  is given by

---

<sup>1</sup> This assumption be lifted via approximation arguments, at the cost of additional technicalities.

$$\begin{aligned}
d_\mu^2((\omega_\nu, s), (\omega_\rho, t)) &= s^2 + t^2 - 2st \cos \angle(\omega_\nu, \omega_\rho) \\
&= s^2 + t^2 - 2st \cos \angle(T_{\mu \rightarrow \nu} - \text{id}, T_{\mu \rightarrow \rho} - \text{id}) \\
&= \|s(T_{\mu \rightarrow \nu} - \text{id}) - t(T_{\mu \rightarrow \rho} - \text{id})\|_{L^2(\mu)}^2.
\end{aligned}$$

We have shown that the tangent cone equipped with the metric  $d_\mu$  is isometric to a Hilbert space. We have proved the following theorem which we had identified using the formalism of Otto calculus in Section 5.4.

**Theorem 7.23.** *Let  $\mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ . Then the tangent cone  $T_\mu \mathcal{W}_2(\mathbb{R}^d)$  at  $\mu$  is a convex subset of  $L^2(\mu)$ . Moreover, for any  $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ , we have*

$$\log_\mu(\nu) = T_{\mu \rightarrow \nu} - \text{id} \in L^2(\mu),$$

where  $T_{\mu \rightarrow \nu}$  is the Brenier map from  $\mu$  to  $\nu$ .

In cases such as the one above, where the tangent cone  $T_\mu \mathcal{W}_2$  equipped with  $\langle \cdot, \cdot \rangle$  from Definition 7.18 is, in fact, a convex subset of a Hilbert space, we call it, by abuse of notation, the *tangent space* at  $\mu$ . It follows readily from the definition of the logarithmic map that the inner product  $\langle \cdot, \cdot \rangle_\mu$  is given for any  $\nu, \rho \in \mathcal{P}_2(\mathbb{R}^d)$  by

$$\begin{aligned}
\langle \log_\mu(\nu), \log_\mu(\rho) \rangle_\mu &= \langle T_{\mu \rightarrow \nu} - \text{id}, T_{\mu \rightarrow \rho} - \text{id} \rangle_{L^2(\mu)} \\
&= \int \langle T_{\mu \rightarrow \nu}(x) - x, T_{\mu \rightarrow \rho}(x) - x \rangle \mu(dx).
\end{aligned}$$

## 7.4 Discussion

**§7.1.** Wasserstein geodesics are discussed in detail in [Vil03, Chapter 5] and [AGS08, Chapter 7]. More generally, the Wasserstein space over any length space is also a length space.

**§7.2.** The non-negative curvature of the Wasserstein space is proven in [AGS08, Section 7.3]. More generally, the Wasserstein space over a non-negatively curved Alexandrov space is also non-negatively curved.

The curve in Exercise 2 is called a *generalized geodesic* and it plays an important role in the theory of Wasserstein gradient flows, as well as occasionally in other applications of optimal transport.

As mentioned in the discussion notes for Section 6.1, there is a notion of synthetic Ricci curvature lower bounds which makes sense on geodesic spaces. It is a natural to ask whether this notion recovers the non-negative Alexandrov curvature of the Wasserstein space when

equipped with an appropriate measure. Unfortunately, [Cho12] shows that does not yield any finite lower bound even for even for the (flat) Wasserstein space on the real line.

**§7.3.** The tangent cone of the Wasserstein space and its relationship to the tangent space is discussed in [AGS08, Section 12.4].

## 7.5 Exercises

1. Show that the space of probability measures endowed with MMD (Definition 2.20) defines a flat geometry.
2. Let  $\mu, \nu, \rho \in \mathcal{P}_2(\mathbb{R}^d)$ . Prove that there *exists* a curve  $\omega : [0, 1] \rightarrow \mathcal{P}_2(\mathbb{R}^d)$  with  $\omega(0) = \mu, \omega(1) = \nu$  such that the opposite inequality to Theorem 7.15 holds, i.e.,

$$W_2^2(\rho, \omega(t)) \leq (1-t) W_2^2(\rho, \mu) + t W_2^2(\rho, \nu) - t(1-t) W_2^2(\mu, \nu).$$

*Hint:* for  $X_\mu \sim \mu, X_\nu \sim \nu, X_\rho \sim \rho$ , optimally couple  $(X_\mu, X_\rho)$  and  $(X_\nu, X_\rho)$ . Define  $\omega(t)$  to be the law of a suitable interpolation of  $X_\mu$  and  $X_\nu$ . Compare with Exercise 9 from Chapter 5.

3. Generalize the proof of Theorem 7.6 to show that for any  $p \geq 1$ , the space  $\mathcal{P}_p(\mathbb{R}^d)$  of probability measures with finite  $p$ -th moment, equipped with the  $p$ -Wasserstein metric  $W_p$ , is a geodesic space. What are the geodesics?
4. Generalizing Definition 7.19, use the following outline to show that if  $(S, d)$  is a metric space with diameter at most  $\pi$ , and  $\text{cone}(S)$  is the set  $X \times [0, \infty)$  with all points of the form  $(x, 0)$  identified, then

$$d((x, s), (y, t)) := \sqrt{s^2 + t^2 - 2st \cos d(x, y)}$$

defines a metric on  $\text{cone}(S)$ . To do so, let  $(x_1, r_1), (x_2, r_2), (x_3, r_3)$  be three points in the cone and construct three points  $y_1, y_2, y_3 \in \mathbb{R}^2$  so that their distances from the origin equal  $r_1, r_2, r_3$  respectively, and so that the angles between  $y_1$  and  $y_2$ , and between  $y_2$  and  $y_3$ , equal  $d(x_1, x_2)$  and  $d(x_2, x_3)$  respectively. Show that  $\|y_1 - y_2\| = d((x_1, r_1), (x_2, r_2))$  and  $\|y_2 - y_3\| = d((x_2, r_2), (x_3, r_3))$ . (Caution:  $\|y_1 - y_3\|$  does not necessarily equal  $d((x_1, r_1), (x_3, r_3))$ .) Now establish the triangle inequality for the cone metric, splitting into two cases according to whether or not  $d(x_1, x_2) + d(x_2, x_3) \leq \pi$ .

## Wasserstein barycenters

*Averaging* data is among the most fundamental of the statistician’s tools, but its implementation on non-Euclidean spaces, capturing data modalities that differ from the typical vector-valued covariates common in traditional statistical literature, often requires care. Suppose, for instance, that our dataset consists of *images* and we wish to define a suitable notion of an average image which captures representative aspects of the whole. This model problem arises in situations such as the aggregation of information from repeated MRI scans.

We can represent a  $p$ -pixel image via its values in the R, G, B channels for each pixel, thereby considering it as a vector taking values in  $\{0, 1, \dots, 255\}^{3p}$ . A naïve approach to the averaging problem would be to simply compute the usual average of these vector representations of the image. Attempting this method on a few images, however, should readily convince the reader that this notion of average is unsatisfactory.

A closer inspection of the naïve approach reveals the nature of the problem: when we average the vector representations of the image, we tacitly endow the space  $\mathbb{R}^{3p}$  with the Euclidean geometry, and there is no reason to expect that this geometry should be compatible with our embedding of images into  $\mathbb{R}^{3p}$ . Indeed, the representation of an image as a vector in  $\mathbb{R}^{3p}$  is an engineering choice, not an intrinsic quality of the image. A perhaps more principled approach would be to regard the images as living in an abstract space  $S$  endowed with a metric  $\mathsf{d}$  which captures closeness with respect to the attributes we regard as important for the data under consideration. Our task can then be formulated as follows: given points  $x_1, \dots, x_n$  inside a metric space  $(S, \mathsf{d})$ , what is a suitable notion for the *average* of  $x_1, \dots, x_n$ ?

Fortunately there is a general and useful answer to this question, which is motivated as follows. It is not hard to see that for  $x_1, \dots, x_n$  belonging to a Hilbert space  $\mathbf{H}$ , the average  $\frac{1}{n} \sum_{i=1}^n x_i$  is characterized as the unique minimizer of the functional

$$x \mapsto \frac{1}{n} \sum_{i=1}^n \|x_i - x\|^2.$$

This formulation only involves squared distances and is amenable to generalization to metric spaces.

**Definition 8.1 (Barycenter).** *Given any probability measure  $P$  over a metric space  $(S, \mathbf{d})$ , we say that  $b$  is a barycenter of  $P$  if it is a minimizer of the functional*

$$b \mapsto \int \mathbf{d}^2(b, x) P(dx).$$

In particular, if we take  $P$  to be an empirical measure  $\frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ , then a barycenter of  $P$  is an average of the points  $x_1, \dots, x_n$ . Note that at this level of generality, a barycenter may not exist, and even if one exists, it may not be unique.

The case when  $(S, \mathbf{d})$  is the Wasserstein space is already of interest and provides motivation for the theory we develop in this chapter. For example, Wasserstein barycenters provide a geometrically meaningful solution to the image averaging problem with which we opened, as well as to many other problems such as curve registration; see [RPDB12, CD14, GPC15, SdGP<sup>+</sup>15, BPC16, PZ16, SLD18, PC19b, LGLR20] and the references therein. However, since the framework we develop fits naturally within metric geometry, as developed in Chapter 7, we work in this setting and specialize later.

Here, we develop statistical theory to justify the use of geometric averaging methods in practice. Namely, assume that we have i.i.d. data  $X_1, \dots, X_n$  drawn from a distribution  $P$  over  $(S, \mathbf{d})$ , and let  $b^*$  denote the barycenter of  $P$ . This *population barycenter* is our quantity of interest and is unknown. In order for the statistical problem to be well-posed, we always work under assumptions which guarantee that  $b^*$  exists and is unique.

There is a natural plug-in estimator for this problem: the barycenter  $b_n$  of the empirical measure or the *empirical barycenter*, defined as the minimizer of the functional

$$b \mapsto \frac{1}{n} \sum_{i=1}^n d^2(b, X_i).$$

Our goal is to quantify the error  $d(b_n, b^*)$  based on natural geometric features of the space  $(S, d)$ .

## 8.1 The Hilbert case

In the case where  $(S, d)$  is a Hilbert space, the empirical barycenter converges at the so-called parametric rate. This is easy to see. To that end, let  $\mathbf{H}$  be a Hilbert space and recall that in this case

$$b_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad b^* = \mathbb{E}X = \int x P(dx),$$

where we used a Pettis integral to define  $b^*$ . We have

$$\begin{aligned} \mathbb{E}\|b_n - b^*\|^2 &= \frac{1}{n^2} \sum_{i,j=1}^n \mathbb{E}\langle X_i - \mathbb{E}X, X_j - \mathbb{E}X \rangle \\ &= \frac{1}{n} \text{var}(X), \quad \text{where} \quad \text{var}(X) = \mathbb{E}\|X - \mathbb{E}X\|^2, \end{aligned}$$

where we used the independence of the  $X_i$ 's and bilinearity of the inner product. Unfortunately, this proof, while concise and leading to an equality (!) is not very instructive since it makes crucial use of the closed form for the barycenter, as well as the inner product structure, which do not extend beyond Hilbert spaces.

Remarkably, the same parametric rate of estimation for the barycenter continue to hold in NPC spaces, see Exercise 2. Unfortunately, as we saw in Theorem 7.15, our main space of interest, namely the Wasserstein space, is an *NNC space*. Therefore, our goal is to develop statistical theory for the more difficult setting of  $\text{curv} \geq 0$ .

Returning to the Hilbert case for inspiration, we propose a second proof which still leads to qualitatively the same result but is off by a factor 4. By definition of  $b_n$ , we have

$$P_n\|b_n - \bullet\|^2 \leq P_n\|b^* - \bullet\|^2.$$

Here and in the sequel, we use the shorthand operator notation: for any integrable function  $f$ ,  $Pf(\bullet) = \int f(x) P(dx)$  and in particular

$$P_n f(\bullet) = \frac{1}{n} \sum_{i=1}^n f(X_i), \quad \text{where} \quad P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

denotes the empirical distribution of the  $X_i$ 's.

Next, note that

$$\|b_n - \bullet\|^2 = \|b_n - b^*\|^2 + \|b^* - \bullet\|^2 + 2 \langle b_n - b^*, b^* - \bullet \rangle.$$

Therefore, applying operator  $P_n$ , we get

$$\|b_n - b^*\|^2 + P_n \|b^* - \bullet\|^2 + 2P_n \langle b_n - b^*, b^* - \bullet \rangle \leq P_n \|b^* - \bullet\|^2,$$

so that

$$\|b_n - b^*\|^2 \leq 2P_n \langle b_n - b^*, \bullet - b^* \rangle.$$

Now the above inequality simply says that  $\|b_n - b^*\|^2 \leq 2 \|b_n - b^*\|^2$ , which is not very useful but we are going to keep going with it for the sake of argument.

Note first that by linearity of the inner product

$$P \langle b_n - b^*, \bullet - b^* \rangle = 0.$$

Therefore, we have

$$\|b_n - b^*\|^2 \leq 2 (P_n - P) \langle b_n - b^*, \bullet - b^* \rangle = 2 \langle b_n - b^*, (P_n - P)(\bullet - b^*) \rangle.$$

Dividing on both sides by  $\|b_n - b^*\|$  and applying the Cauchy–Schwarz inequality, we get

$$\mathbb{E} \|b_n - b^*\|^2 \leq 4 \mathbb{E} \|(P_n - P)(\bullet - b^*)\|^2 = \frac{4}{n} \text{var}(X). \quad (8.1)$$

What did we learn in this proof? First we have only an inequality and lost a factor 4. Our major gain was that we never used the closed form for  $b_n$ . Instead, we only used the fact that

$$\mathbb{E} \|(P_n - P)(\bullet - b^*)\|^2 \leq \text{var}(X)/n,$$

which applies more broadly. On the downside, we used the linearity of the inner product and more generally the Hilbert structure quite extensively. It turns out that this is quite necessary to obtain our results. Therefore, we force the Hilbert structure in through the tangent space of  $\mathcal{W}_2$  and keep track of how much we lose.

## 8.2 Barycenters on positively curved spaces

Let  $P$  be a probability measure on a positively curved geodesic space  $(S, d)$ . Let  $b^*$  be any barycenter of  $P$  and let  $b_n$  be an empirical barycenter built from  $n$  independent copies  $X_1, \dots, X_n$  of  $X \sim P$ :

$$b^* \in \arg \min_{b \in S} \int d^2(b, x) P(dx), \quad b_n \in \arg \min_{b \in S} \sum_{i=1}^n d^2(b, X_i).$$

Before we proceed, we discuss our overall approach. The argument in the Hilbert case rests on the inequality  $P_n d^2(b_n, \cdot) \leq P_n d^2(b^*, \cdot)$ , which holds true by definition of  $b_n$  and highlights that the empirical barycenter is an instance of the empirical risk minimization (ERM) framework within the statistical estimation literature. Using ERM techniques, we could hope to control the estimation error via measures of the “complexity” of the space  $(S, d)$ , and this approach has been pursued in the literature (see [ACLGP20]). However, for our application of interest in which  $(S, d)$  is taken to be the Wasserstein space, the complexity is prohibitively large and it leads to non-parametric rates of estimation; in particular, they suffer from the curse of dimensionality, similarly to what we saw in Chapter 2.

However, we have just seen that the rates of estimation in a Hilbert space escape the curse, despite the fact that Hilbert spaces can even be *infinite*-dimensional. Our intuition therefore leads us to believe that, even if we are working over a curved space  $(S, d)$ , as long we can restrict ourselves to sufficiently “flat” parts of  $S$ , then perhaps we could recover the Hilbertian rates. In the sequel, we seek natural geometric conditions—morally, they encode curvature bounds—which enable fast, *parametric* rates of estimation.

### 8.2.1 Master theorem

We begin with a general result that mimics the proof of Section 8.1 in the Hilbert case.

Before stating it, we introduce a quantity that measures how much the tangent space at the barycenter “hugs” the original space at the barycenter  $b^*$ .

**Definition 8.2 (Hugging).** *Let  $(S, d)$  be a geodesic space such that  $\text{curv}(S) \geq 0$ . For any  $b^*, b \in S$ , let  $h_{b^*}^b$  be the hugging function of  $S$  at  $b^*$  in direction  $b$  defined by*

$$h_{b^*}^b(x) = 1 - \frac{\|\log_{b^*}(x) - \log_{b^*}(b)\|_{b^*}^2 - d^2(x, b)}{d^2(b, b^*)}, \quad x \in S. \quad (8.2)$$

Note that it follows from Proposition 7.21 that  $h_{b^*}^b(x) \leq 1$  for all  $x \in S$ . Moreover, if  $S$  is a Hilbert space, then  $\|\log_{b^*}(x) - \log_{b^*}(b)\|_{b^*}^2 = d^2(x, b)$  and  $h_{b^*}^b \equiv 1$ . In general,  $h_{b^*}^b(x)$  may be negative when there is a lot of curvature around  $b^*$  but it remains non-negative in average when computed at barycenter  $b^*$ . This result follows from the following simple but important observation.

**Theorem 8.3 (Variance equality).** *Let  $(S, d)$  be a geodesic space with  $\text{curv}(S) \geq 0$ . Let  $Q \in \mathcal{P}_2(S)$  be a probability distribution on  $S$  with barycenter  $b^*$ . Assume further that the tangent cone of  $S$  at  $b^*$  equipped with  $\langle \cdot, \cdot \rangle_{b^*}$  is a convex subset of a Hilbert space. Then, for all  $b \in S$ ,*

$$d^2(b, b^*) \int h_{b^*}^b(x) Q(dx) = \int (d^2(x, b) - d^2(x, b^*)) Q(dx), \quad (8.3)$$

where  $h_{b^*}^b$  is the hugging function defined in (8.2).

*Proof.* By definition of  $h_{b^*}^b$ , we have

$$\begin{aligned} d^2(b, b^*) h_{b^*}^b(\bullet) &= d^2(b, b^*) + d^2(\bullet, b) - \|\log_{b^*} b - \log_{b^*} \bullet\|_{b^*}^2 \\ &= d^2(b, b^*) + d^2(\bullet, b) \\ &\quad - \|\log_{b^*} b\|_{b^*}^2 - \|\log_{b^*} \bullet\|_{b^*}^2 + 2 \langle \log_{b^*} \bullet, \log_{b^*} b \rangle_{b^*} \\ &= d^2(b, b^*) + d^2(\bullet, b) \\ &\quad - d^2(b, b^*) - d^2(b^*, \bullet) + 2 \langle \log_{b^*} \bullet, \log_{b^*} b \rangle_{b^*} \\ &= d^2(\bullet, b) - d^2(\bullet, b^*) + 2 \langle \log_{b^*} \bullet, \log_{b^*} b \rangle_{b^*}. \end{aligned}$$

Therefore applying the linear operator  $Q$ , we get

$$d^2(b, b^*) Q h_{b^*}^b(\bullet) = Q d^2(\bullet, b) - Q d^2(\bullet, b^*) + 2 \langle \log_{b^*} b^*, \log_{b^*} b \rangle_{b^*},$$

where we use the fact that  $Q \log_{b^*} \bullet = \log_{b^*} b^*$  or, in other words, that  $\log_{b^*} b^*$  is the barycenter of  $(\log_{b^*})_# Q$ . Indeed, we have by Proposition 7.21 that for all  $b \in S$ ,

$$Q \|\log_{b^*} \bullet - \log_{b^*} b^*\|_{b^*}^2 = Q d^2(\bullet, b^*) \leq Q d^2(\bullet, b) \leq Q \|\log_{b^*} \bullet - \log_{b^*} b\|_{b^*}^2$$

with equality if  $b = b^*$  so that  $\log_{b^*} b^*$  is a barycenter for  $(\log_{b^*})_# Q$  and therefore  $Q \log_{b^*} \bullet = \log_{b^*} b^*$ .

Finally since  $\log_{b^*} b^* = o_{b^*}$  is the tip of the tangent cone, we have  $\|\log_{b^*} b^*\|_{b^*} = 0$ , which, in turn, yields  $\langle \log_{b^*} b^*, \log_{b^*} b \rangle_{b^*} = 0$ . This completes the proof.  $\square$

A direct consequence of the variance equality is that if  $\int h_{b^*}^b dQ > 0$ , then  $b^*$  is the unique barycenter of  $Q$ . Moreover, since the right-hand side of the variance equality is non-negative by definition of a barycenter  $b^*$ , we readily get the following corollary.

**Corollary 8.4.** *Under the same assumptions as Theorem 8.3, we have for any  $b \in S$ ,*

$$0 \leq \int h_{b^*}^b(x) Q(dx) \leq 1.$$

Moreover, for any  $b, x \in S$ , we have  $h_{b^*}^b(x) \leq 1$ .

It turns out the hugging function at  $b^*$  plays a key role in obtaining parametric rates of convergence for empirical barycenters.

**Theorem 8.5 (Master theorem).** *Let  $P$  be a probability measure on a NNC geodesic space  $(S, d)$  and denote by  $b^*$  and  $b_n$  a barycenter of  $P$  and an empirical barycenter respectively. Assume further that the tangent cone of  $S$  at  $b^*$  equipped with  $\langle \cdot, \cdot \rangle_{b^*}$  is a convex subset of a Hilbert space. Then, the following holds: if for any  $b \in S$ ,*

$$h_{b^*}^b(\bullet) \geq h_{\min} > 0, \quad (8.4)$$

then  $b_n$  and  $b^*$  are both unique and

$$\mathbb{E}d^2(b_n, b^*) \leq \frac{4\sigma^2}{h_{\min}^2 n},$$

where  $\sigma^2$  denotes the variance of  $P$  defined by

$$\sigma^2 = \int d^2(b^*, x) P(dx). \quad (8.5)$$

*Proof.* Note first that uniqueness follows directly from the variance equality and (8.4).

Next, we start as in the Hilbert case by observing that

$$P_n d^2(b_n, \bullet) \leq P_n d^2(b^*, \bullet).$$

It yields

$$\begin{aligned} P_n d^2(b_n, \bullet) - P_n \|\log_{b^*} b_n - \log_{b^*} \bullet\|_{b^*}^2 \\ + P_n \|\log_{b^*} b_n - \log_{b^*} \bullet\|_{b^*}^2 - P_n d^2(b^*, \bullet) \leq 0. \end{aligned} \quad (8.6)$$

We now make use of the fact that the tangent cone has a Hilbert structure so that

$$\begin{aligned} \|\log_{b^*} b_n - \log_{b^*} \bullet\|_{b^*}^2 \\ = \|\log_{b^*} b_n\|_{b^*}^2 + \|\log_{b^*} \bullet\|_{b^*}^2 - 2 \langle \log_{b^*} b_n, \log_{b^*} \bullet \rangle_{b^*} \\ = d^2(b^*, b_n) + d^2(b^*, \bullet) - 2 \langle \log_{b^*} b_n, \log_{b^*} \bullet \rangle_{b^*} \end{aligned}$$

where in the second identity, we used twice the equality case in Proposition 7.21. Plugging this into (8.6) yields

$$d^2(b^*, b_n) \leq P_n [\|\log_{b^*} b_n - \log_{b^*} \bullet\|_{b^*}^2 - d^2(b_n, \bullet)] + 2P_n \langle \log_{b^*} b_n, \log_{b^*} \bullet \rangle_{b^*}.$$

Next, by definition of the hugging function, we get

$$\|\log_{b^*} b_n - \log_{b^*} \bullet\|_{b^*}^2 - d^2(b_n, \bullet) = (1 - h_{b^*}^{b_n}(\bullet)) d^2(b_n, b^*).$$

It yields

$$h_{\min} d^2(b_n, b^*) \leq 2P_n \langle \log_{b^*} b_n, \log_{b^*} \bullet \rangle_{b^*}.$$

The right-hand side is simply an average in a Hilbert space so, dividing by  $\|\log_{b^*} b_n\|_{b^*} = d(b_n, b^*)$  on both sides and applying Cauchy–Schwarz, we get

$$h_{\min}^2 \mathbb{E} d^2(b_n, b^*) \leq \frac{4\sigma^2}{n},$$

where

$$\sigma^2 = \int \|\log_{b^*} x\|_{b^*}^2 P(dx) = \int d^2(b^*, x) P(dx)$$

as desired.  $\square$

It follows from inspecting the proof of the master theorem that in order to obtain parametric rates of estimation for  $b^*$ , it suffices to have the weaker condition  $P_n h_{b^*}^{b_n}(\bullet) \geq h_{\min} > 0$ . Since  $P_n$  is a random measure, we prefer not to impose conditions on it and focus instead on the stronger condition (8.4). We are going to obtain such results using the notion of *extendable geodesics*.

### 8.2.2 Extendable geodesics

We now present a compelling synthetic geometric condition that implies this lower bound in the context of NNC spaces: the extendability, by a given factor, of all geodesics emanating from and arriving at the barycenter  $b^*$ .

**Definition 8.6 (Extendable geodesic).** Consider a constant-speed geodesic  $\omega : [0, 1] \rightarrow S$ . For  $(\lambda_{\text{in}}, \lambda_{\text{out}}) \in [0, \infty]^2$ , we say that  $\omega$  is  $(\lambda_{\text{in}}, \lambda_{\text{out}})$ -extendable if there exists a path  $\omega^+ : [-\lambda_{\text{in}}, 1 + \lambda_{\text{out}}] \rightarrow S$  which agrees with  $\omega$  on  $[0, 1]$ , called an extension of  $\omega$ , which remains a geodesic between its endpoints  $\omega^+(-\lambda_{\text{in}})$  and  $\omega^+(1 + \lambda_{\text{out}})$ .

Before we state the main result of this subsection, we need the following fact.

**Theorem 8.7.** Suppose that  $\text{curv}(S) \geq 0$ . Let  $Q \in \mathcal{P}_2(S)$  be a probability measure on  $S$  with a barycenter  $b^*$ . Suppose that, for each  $x \in \text{supp}(Q)$ , there exists a constant-speed geodesic  $\omega_x : [0, 1] \rightarrow S$  connecting  $b^*$  to  $x$  which is  $(0, \lambda)$ -extendable for  $\lambda > 0$ . Suppose in addition that  $b^*$  remains a barycenter of the distribution  $Q_\lambda = (e_\lambda)_\# Q$  where  $e_\lambda(x) = \omega_x^+(1 + \lambda)$ . Then for all  $b \in S$ ,

$$Qh_{b^*}^b(\bullet) \geq \frac{\lambda}{1 + \lambda}. \quad (8.7)$$

In particular, it implies that  $b^*$  is the unique barycenter of  $Q$ .

*Proof.* Fix  $y \in \text{supp}(Q)$  and define  $y_\lambda = e_\lambda(y)$ . Let  $\omega : [0, 1] \rightarrow S$  be a constant-speed geodesic connecting  $b^*$  to  $y_\lambda$ . By definition,  $\omega(\tau) = y$  for  $\tau = 1/(1 + \lambda)$ . Since  $\text{curv}(S) \geq 0$ , we have for any  $b \in S$ ,

$$\begin{aligned} d^2(b, y) &\geq (1 - \tau) d^2(b, b^*) + \tau d^2(b, y_\lambda) - \tau(1 - \tau) d^2(b^*, y_\lambda) \\ &= \frac{\lambda}{1 + \lambda} d^2(b, b^*) + \frac{1}{1 + \lambda} d^2(b, y_\lambda) - \frac{\lambda}{(1 + \lambda)^2} d^2(b^*, y_\lambda). \end{aligned}$$

Next, observe that

$$d^2(b^*, y_\lambda) = (1 + \lambda)^2 d^2(b^*, y)$$

so that

$$\begin{aligned} \frac{\lambda}{1 + \lambda} d^2(b, b^*) &\leq d^2(b, y) + \lambda d^2(b^*, y) - \frac{1}{1 + \lambda} d^2(b, y_\lambda) \\ &= (d^2(b, y) - d^2(b^*, y)) + (1 + \lambda) d^2(b^*, y) - \frac{1}{1 + \lambda} d^2(b, y_\lambda). \end{aligned} \quad (8.8)$$

Moreover,

$$(1 + \lambda) d^2(b^*, y) - \frac{1}{1 + \lambda} d^2(b, y_\lambda)$$

$$\begin{aligned}
&= \frac{1}{1+\lambda} \left( (1+\lambda)^2 d^2(b^*, y) - d^2(b, y_\lambda) \right) \\
&= \frac{1}{1+\lambda} \left( d^2(b^*, y_\lambda) - d^2(b, y_\lambda) \right).
\end{aligned}$$

Thus, writing  $Q_\lambda := (e_\lambda)_\# Q$ , we get

$$\begin{aligned}
&\int \left( (1+\lambda) d^2(b^*, y) - \frac{1}{1+\lambda} d^2(b, y_\lambda) \right) Q(dy) \\
&= \frac{1}{1+\lambda} \int (d^2(b^*, y) - d^2(b, y)) Q_\lambda(dy) \leq 0,
\end{aligned}$$

where in the last inequality, we used the fact that  $b^*$  remains a barycenter of  $Q_\lambda$ . Together with (8.8) integrated with respect to  $Q$ , we get

$$\frac{\lambda}{1+\lambda} d^2(b, b^*) \leq \int (d^2(b, y) - d^2(b^*, y)) Q(dy). \quad (8.9)$$

Combined with the variance equality (Theorem 8.3), this completes the proof.  $\square$

The above notion of extendable geodesics gives a lower bound on  $Ph_{b^*}^b(\bullet)$  uniformly in  $b$ . While this is already an attractive feature that implies uniqueness of the barycenter, it suffers from two deficiencies. First, we need to control  $P_n h_{b^*}^{b_n}(\bullet)$  and  $b_n$  is data-dependent, and it is unclear how to control the deviation  $|P_n h_{b^*}^{b_n} - Ph_{b^*}^b|$  in a suitable fashion. Second, the condition that  $P_\lambda = (e_\lambda)_\# P$  keeps the same barycenter is difficult to check and appears to be restrictive.

To overcome both limitations, we allow for geodesics emanating from  $b^*$  to be extendable in both directions.

**Theorem 8.8.** *Suppose that  $\text{curv}(S) \geq 0$  and let  $x, b, b^* \in S$ . Suppose that there exist  $\lambda_{\text{in}}, \lambda_{\text{out}} > 0$  and a geodesic connecting  $b^*$  to  $x$  which is  $(\lambda_{\text{in}}, \lambda_{\text{out}})$ -extendable. Then*

$$h_{b^*}^b(x) \geq h_{\min} = \frac{\lambda_{\text{out}}}{1 + \lambda_{\text{out}}} - \frac{1}{\lambda_{\text{in}}}.$$

*Proof.* Let  $\omega_x : [0, 1] \rightarrow S$  be a  $(\lambda_{\text{in}}, \lambda_{\text{out}})$ -extendable geodesic connecting  $b^*$  to  $x$  and denote by  $\omega_x^+ : [-\lambda_{\text{in}}, 1 + \lambda_{\text{out}}] \rightarrow S$  its extension. Let  $z = \omega_x^+(-\xi)$  where  $\xi = \lambda_{\text{in}}/(1 + \lambda_{\text{out}})$  and consider the measure  $Q$  defined by

$$Q := \frac{\xi}{1 + \xi} \delta_x + \frac{1}{1 + \xi} \delta_z.$$

Since  $Q$  is supported on  $\omega^+$  we can easily compute its barycenter. Indeed, note that  $x = \omega_x^+(1)$  so the barycenter of  $Q$  is given by

$$\omega_x^+\left(1 \cdot \frac{\xi}{1+\xi} - \xi \cdot \frac{1}{1+\xi}\right) = \omega^+(0) = b^*.$$

Now, we wish to apply Theorem 8.7 to  $Q$ . To that end, note that the constant-speed geodesics  $\omega_x$  connecting  $b^*$  to  $x$  and  $\sigma$  connecting  $b^*$  to  $z$  and defined by  $\sigma(t) = \omega_x^+(-t\xi)$  are both  $(0, 1 + \lambda_{\text{out}})$ -extendable by assumption and by construction respectively.

Finally, we check that  $b^*$  remains a barycenter of the probability measure  $Q_{\lambda_{\text{out}}} = (e_{\lambda_{\text{out}}})_# Q$  where  $e_{\lambda_{\text{out}}}(x) = \omega_x^+(1 + \lambda_{\text{out}})$ . Indeed, by construction,  $Q_{\lambda_{\text{out}}}$  is the two-point probability measure given by

$$Q_{\lambda_{\text{out}}} = \frac{\xi}{1+\xi} \delta_{\omega^+(1+\lambda_{\text{out}})} + \frac{1}{1+\xi} \delta_{\omega^+(-\lambda_{\text{in}})}.$$

Therefore, the barycenter is given by

$$\begin{aligned} & \omega^+\left((1 + \lambda_{\text{out}}) \cdot \frac{\xi}{1+\xi} - \lambda_{\text{in}} \cdot \frac{1}{1+\xi}\right) \\ &= \omega^+\left(\frac{(1 + \lambda_{\text{out}}) \lambda_{\text{in}}}{1 + \lambda_{\text{in}} + \lambda_{\text{out}}} - \frac{\lambda_{\text{in}} (1 + \lambda_{\text{out}})}{1 + \lambda_{\text{in}} + \lambda_{\text{out}}}\right) = \omega^+(0) = b^*. \end{aligned}$$

As a result, Theorem 8.7 implies that

$$\begin{aligned} \frac{\lambda_{\text{out}}}{1 + \lambda_{\text{out}}} &\leq Q h_{b^*}^b(\bullet) = \frac{\xi}{1+\xi} h_{b^*}^b(x) + \frac{1}{1+\xi} h_{b^*}^b(z) \\ &\leq \frac{\xi}{1+\xi} h_{b^*}^b(x) + \frac{1}{1+\xi}, \end{aligned}$$

where we used Corollary 8.4 to bound  $h_{b^*}^b(z) \leq 1$  for all  $b, z \in S$ .

Hence, we obtain

$$\begin{aligned} h_{b^*}^b(x) &\geq \frac{1+\xi}{\xi} \left( \frac{\lambda_{\text{out}}}{1 + \lambda_{\text{out}}} - \frac{1}{1+\xi} \right) \\ &= \frac{1+\xi}{\xi} \left( \frac{\lambda_{\text{out}}}{\lambda_{\text{in}}} \xi - \frac{1}{1+\xi} \right) \\ &= \frac{\lambda_{\text{out}}}{\lambda_{\text{in}}} (1 + \xi) - \frac{1}{\xi} \\ &= \frac{\lambda_{\text{out}}}{\lambda_{\text{in}}} + \frac{\lambda_{\text{out}}}{\lambda_{\text{in}}} \cdot \frac{\lambda_{\text{in}}}{1 + \lambda_{\text{out}}} - \frac{1 + \lambda_{\text{out}}}{\lambda_{\text{in}}} \\ &= \frac{\lambda_{\text{out}}}{1 + \lambda_{\text{out}}} - \frac{1}{\lambda_{\text{in}}}, \end{aligned}$$

which completes the proof.  $\square$

Note that Theorem 8.8 gives a lower bound on  $h_{b^*}^b(x)$  that is uniform in both  $b$  and  $x$ . It is of course possible to make this result depend on  $x$  only and get a result of the form

$$h_{b^*}^b(x) \geq \frac{\lambda_{\text{out}}(x)}{1 + \lambda_{\text{out}}(x)} - \frac{1}{\lambda_{\text{in}}(x)}.$$

If we assume that

$$P\left(\frac{\lambda_{\text{out}}(\bullet)}{1 + \lambda_{\text{out}}(\bullet)} - \frac{1}{\lambda_{\text{in}}(\bullet)}\right) > 0,$$

then standard concentration tools ensure that  $P_n h_{b^*}^{b_n}(\bullet) > 0$  for  $n$  large enough as desired.

Instead of going into these details, let us inspect the uniform bound more closely. From the master theorem, and Theorem 8.8, we get the following corollary.

**Corollary 8.9.** *Let  $P$  be a probability measure on an NNC geodesic space  $(S, d)$  and denote by  $b^*$  and  $b_n$  a barycenter of  $P$  and an empirical barycenter respectively. Assume that the tangent cone of at  $b^*$  equipped with  $\langle \cdot, \cdot \rangle_{b^*}$  is a convex subset of a Hilbert space. Moreover, let  $\lambda_{\text{in}}, \lambda_{\text{out}} \in [0, \infty]$  be such that*

$$h := \frac{\lambda_{\text{out}}}{1 + \lambda_{\text{out}}} - \frac{1}{\lambda_{\text{in}}} > 0$$

*and assume further that for any  $x \in \text{supp}(P)$ , there exists a geodesic connecting  $b^*$  to  $x$  that is  $(\lambda_{\text{in}}, \lambda_{\text{out}})$ -extendable. Then  $b^*$  is unique and the empirical barycenter satisfies*

$$\mathbb{E}[d^2(b_n, b^*)] \leq \frac{4\sigma^2}{hn}$$

*where  $\sigma^2$  denotes the variance of  $P$  that is defined in (8.5).*

As a result, we get parametric rates when geodesics may be sufficiently extended. In particular, if  $S$  is a Hilbert space, then all geodesics are infinitely extendable. Therefore  $h = 1$  and we recover (8.1).

### 8.3 Parametric rates for Wasserstein barycenters

To conclude these notes, we study Wasserstein barycenters as an example. Note that our result readily applies to this case. One may ask the

question: how does the condition of extendable geodesics translate in terms of optimal transport? It turns out that it can be characterized in terms of regularity conditions on the Brenier maps.

**Theorem 8.10.** *Let  $\mu, \nu \in \mathcal{W}_2$  be two probability measures such that  $\mu$  has a density and let  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$  be the convex function defined by  $\varphi(x) = (\|x\|^2 - f(x))/2$ , where  $f$  is the Kantorovich potential given in Definition 1.15. In particular,  $\nabla \varphi$  is defined  $\mu$ -almost surely and is the Brenier map. Recall that the unique constant-speed geodesic  $\omega$  connecting  $\mu$  to  $\nu$  is given by  $\omega(t) = ((1-t)\text{id} + t\nabla \varphi)_{\#}\mu$ . Then, for any  $\lambda > 0$ ,  $\omega$  is  $(0, \lambda)$ -extendable if and only if  $\varphi$  is  $\frac{\lambda}{1+\lambda}$ -strongly convex.*

*Proof.* Assume first that  $\omega$  is  $(0, 1 + \lambda)$ -extendable and let

$$\omega^+ : [0, 1 + \lambda] \rightarrow \mathcal{W}_2$$

denote its extension. Let  $Y_\lambda \sim \omega^+(1 + \lambda)$  and observe that there exists a convex function  $\varphi_\lambda$  defined  $\mu$ -almost everywhere such that  $Y_\lambda = \nabla \varphi_\lambda(X)$ , where  $X \sim \mu$ . Moreover, since  $\omega^+$  is a geodesic and  $\nabla \varphi(X) \sim \omega^+(1)$ , we also have

$$\nabla \varphi(X) = \frac{\lambda}{1 + \lambda} X + \frac{1}{1 + \lambda} Y_\lambda,$$

so that  $Y_\lambda = \nabla \varphi_\lambda(X) = (1 + \lambda) \nabla \varphi(X) - \lambda X$ . In particular, it means that we can choose

$$\varphi(x) = \frac{1}{1 + \lambda} \varphi_\lambda(x) + \frac{\lambda}{2(1 + \lambda)} \|x\|^2.$$

Since  $\varphi_\lambda$  is convex, so is  $\varphi_\lambda/(1 + \lambda)$  and the above display implies that  $\varphi$  is  $\frac{\lambda}{1 + \lambda}$ -strongly convex.

Conversely, assume that  $\varphi$  is  $\frac{\lambda}{1 + \lambda}$ -strongly convex and define  $Y_\lambda = (1 + \lambda) \nabla \varphi(X) - \lambda X$  where  $X \sim \mu$ . We are going to show that  $Y_\lambda$  and  $X$  are optimally coupled. To that end, note that  $Y_\lambda = \nabla \varphi_\lambda(X)$  where

$$\varphi_\lambda(x) = (1 + \lambda) \varphi(x) - \frac{\lambda}{2} \|x\|^2.$$

Since  $\varphi$  is  $\frac{\lambda}{1 + \lambda}$ -strongly convex,  $\varphi_\lambda$  is convex and thus  $\nabla \varphi_\lambda$  is the Brenier map. It follows that  $Y_\lambda$  and  $X$  are optimally coupled so that  $\omega^+ : [0, 1 + \lambda] \rightarrow \mathcal{W}_2$  defined by

$$\omega^+(t) = \left( \text{id} + \frac{t}{1+\lambda} (\nabla \varphi_\lambda - \text{id}) \right)_\# \mu$$

is a geodesic connecting  $\mu$  and the distribution of  $Y_\lambda$  such that  $\omega^+(t) = \omega(t)$  for  $t \in [0, 1]$ . Therefore  $\omega$  is  $(0, 1 + \lambda)$ -extendable.  $\square$

Recall that if  $\mu$  and  $\nu$  both have a density such that the Brenier map from  $\mu$  to  $\nu$  is given by  $\nabla \varphi$ , then the Brenier map from  $\nu$  to  $\mu$  is given by  $\nabla \varphi^*$ . Therefore, if  $\varphi$  is  $\beta$ -smooth in the sense that for any  $x, y \in \mathbb{R}^d$ ,

$$\varphi(x) - \varphi(y) \leq \langle \nabla \varphi(y), x - y \rangle + \frac{\beta}{2} \|x - y\|^2,$$

then  $\varphi^*$  is  $1/\beta$ -strongly convex (see Lemma A.9), which, in turn implies that the geodesic connecting  $\nu$  to  $\mu$  is  $(0, \frac{1}{\beta+1})$ -extendable.

These facts yield the following theorem but we provide an alternate, more direct, proof.

**Theorem 8.11.** *Let  $P$  be a probability measure on  $\mathcal{W}_2$  with a barycenter  $b^*$  that admits a density. Assume further that for any  $\mu \in \text{supp}(P)$  the Brenier map from  $b^*$  to  $\mu$  is  $\alpha$ -strongly convex and  $\beta$ -smooth with  $\beta - \alpha \in [0, 1)$ . Then  $b^*$  is unique and the empirical Wasserstein barycenter  $b_n$  satisfies*

$$\mathbb{E}[W_2^2(b_n, b^*)] \leq \frac{4\sigma^2}{(1 - (\beta - \alpha))^2 n}.$$

*Proof.* For any  $\mu \in \text{supp}(P)$ , let  $\varphi$  be such that  $\nabla \varphi$  is the Brenier map from  $b^*$  to  $\mu$ . For any  $b, \mu \in \mathcal{W}_2$ , let  $X, X' \sim \mu$ ,  $Y, Y' \sim b$  and  $Z, Z' \sim b^*$ . In view of the gluing lemma, we may assume that  $(X, Z)$  and  $(Y, Z)$  are optimally coupled whereas we assume that  $(X', Y')$  and  $(X', Z')$  are optimally coupled.

Note that rearranging terms in the definition of the hugging function, our goal is to prove that

$$\mathbb{E}\|X - Y\|^2 \leq \mathbb{E}\|X' - Y'\|^2 + (\beta - \alpha) \mathbb{E}\|Y - Z\|^2. \quad (8.10)$$

By assumption, for  $b^*$ -almost all  $z \in \mathbb{R}^d$  and any  $y \in \mathbb{R}^d$ , we have

$$\frac{\alpha}{2} \|y - z\|^2 \leq \varphi(y) - \varphi(z) - \langle \nabla \varphi(z), y - z \rangle \leq \frac{\beta}{2} \|y - z\|^2. \quad (8.11)$$

It holds

$$\mathbb{E}\|X - Y\|^2 = \mathbb{E}\|X - Z\|^2 + \mathbb{E}\|Y - Z\|^2 + 2\mathbb{E}\langle X - Z, Z - Y \rangle$$

$$= \mathbb{E}\|X - Z\|^2 + \mathbb{E}\|Y - Z\|^2 - 2\mathbb{E}\langle Z, Z - Y \rangle + 2\mathbb{E}\langle \nabla \varphi(Z), Z - Y \rangle. \quad (8.12)$$

Next, note that on the one hand, it follows from (8.11) that

$$\begin{aligned} 2\mathbb{E}\langle \nabla \varphi(Z), Z - Y \rangle &\leq 2\mathbb{E}\varphi(Z) - 2\mathbb{E}\varphi(Y) + \beta\mathbb{E}\|Y - Z\|^2 \\ &= 2\mathbb{E}\varphi(Z') - 2\mathbb{E}\varphi(Y') + \beta\mathbb{E}\|Y - Z\|^2 \\ &\leq 2\mathbb{E}\langle \nabla \varphi(Z'), Z' - Y' \rangle - \alpha\mathbb{E}\|Y' - Z'\|^2 + \beta\mathbb{E}\|Y - Z\|^2 \\ &= 2\mathbb{E}\langle X', Z' - Y' \rangle - \alpha\mathbb{E}\|Y' - Z'\|^2 + \beta\mathbb{E}\|Y - Z\|^2. \end{aligned}$$

Since  $\mathbb{E}\|Y' - Z'\|^2 \geq \mathbb{E}\|Y - Z\|^2$ , we get,

$$2\mathbb{E}\langle \nabla \varphi(Z), Z - Y \rangle \leq 2\mathbb{E}\langle X', Z' - Y' \rangle + (\beta - \alpha)\mathbb{E}\|Y - Z\|^2.$$

On the other hand,

$$\mathbb{E}\|Y - Z\|^2 - 2\mathbb{E}\langle Z, Z - Y \rangle = \mathbb{E}\|Y\|^2 - \mathbb{E}\|Z\|^2 = \mathbb{E}\|Y'\|^2 - \mathbb{E}\|Z'\|^2.$$

Together, with (8.12), the above two displays yield

$$\begin{aligned} \mathbb{E}\|X - Y\|^2 &\leq \mathbb{E}\|X' - Z'\|^2 + \mathbb{E}\|Y'\|^2 - \mathbb{E}\|Z'\|^2 \\ &\quad + 2\mathbb{E}\langle X', Z' - Y' \rangle + (\beta - \alpha)\mathbb{E}\|Y - Z\|^2 \\ &= \mathbb{E}\|X' - Y'\|^2 + (\beta - \alpha)\mathbb{E}\|Y - Z\|^2, \end{aligned}$$

which completes the proof of (8.10).

We have proved that  $h_{b^*}^b(\mu) \geq 1 - (\beta - \alpha)$  which, together with the master theorem, completes the proof.  $\square$

Barycenters are the equivalent of averages on curved spaces. As such they are the building block of many statistical tools including regression [CLM23], analysis of variance [DM19], change-point detection [DM20], discriminant analysis [FCCR18], and principal component analysis [BGKL17, CSB<sup>+</sup>18]. Despite initial work, many questions about the statistical properties of these statistical objects remain to be understood.

## 8.4 Discussion

**§8.1.** Beyond the setting of Hilbert spaces, quantitative laws of large numbers are obtained over Banach spaces in relation to the theory of

type and cotype, see [LT91]. Also, see the excellent exposition [Stu03] for barycenters over NPC spaces, from which Exercise 2 is taken.

**§8.2.** The material in this section is taken from [ACLGP20, LGPRS22].

**§8.3.** The basic theory of Wasserstein barycenters (existence, duality, etc.) was developed in [AC11]; see Exercise 3. Statistical consistency for Wasserstein barycenters was established in [LGL17]. The variance inequality in Exercise 4 is from [CMRS20].

Substantial attention has also been devoted to the computation of barycenters. For discrete distributions, the work of [ABA21, ABA22] established polynomial-time tractability of Wasserstein barycenters in fixed dimension, and **NP**-hardness in general dimension; see the references therein for other approaches, such as parametrization via neural networks and application of continuous optimization methods.

Another line of work, more closely related to Chapter 5, develops algorithms for computing the barycenter via gradient methods in the Wasserstein space [AEdBCAM16, ZP19, CMRS20, ACGS21, BVFRT22, KDLY22, BRT24]. The descent lemma in Exercise 5 is from [ZP19], which interpreted the fixed-point approach of [AEdBCAM16] as Wasserstein gradient descent.

Barycenters for Gaussians were studied earlier than the general case, dating back to [KS94, RU02]. Statistical estimation was studied in [KSS21], and non-asymptotic computational guarantees for Wasserstein gradient descent were given in [CMRS20, ACGS21].

Similarly to Chapter 4, one can add entropic regularization to the Wasserstein barycenter, at the level of the Wasserstein distance or the barycenter objective or both; see [Kro18, BCP19b, LGYS20, CEK21, Chi23, VC23].

## 8.5 Exercises

1. Let  $P$  be a distribution over  $\mathcal{P}_2(\mathbb{R})$ . Give a closed-form expression for the  $W_2$  barycenter of  $P$  in terms of the CDFs of the measures in  $\text{supp } P$ .
2. Suppose that  $P$  is a probability measure over an *NPC space*  $(S, d)$  with barycenter  $b^*$ . It turns out that statistical estimation of barycenters over NPC spaces is far easier, as we demonstrate in this exercise.
  - a) Show that for any  $b \in S$ ,

$$P[d^2(b, \cdot) - d^2(b^*, \cdot)] \geq d^2(b, b^*). \quad (8.13)$$

b) Suppose that  $(X_i)_{i=1}^n$  is an i.i.d. sequence drawn from  $P$  and form the following estimator  $b_n$  inductively: set  $b_1 = X_1$ , and for  $n \geq 2$  let  $b_n := \omega_{b_{n-1}, X_n}(1/n)$  where  $\omega_{x,y} : [0, 1] \rightarrow S$  is the constant-speed geodesic joining  $x$  to  $y$ . Prove by induction that for all  $n \geq 1$ ,

$$\mathbb{E} d^2(b_n, b^*) \leq \frac{\sigma^2}{n}, \quad \text{where } \sigma^2 = P d^2(b^*, \bullet).$$

*Hint:* Apply the NPC inequality from Proposition 7.11 together with the inequality (8.13).

3. Let  $\mu_1, \dots, \mu_n \in \mathcal{P}_2(\mathbb{R}^d)$  and let  $\Gamma(\mu_1, \dots, \mu_n)$  denote the set of couplings of  $\mu_1, \dots, \mu_n$ . Consider the *multi-marginal* optimal transport problem

$$\min_{\gamma \in \Gamma(\mu_1, \dots, \mu_n)} \int \sum_{i=1}^n \|x_i - \frac{1}{n} \sum_{j=1}^n x_j\|^2 \gamma(dx_1, \dots, dx_n).$$

Let  $\gamma^*$  denote an optimal solution. Prove that if  $(X_1, \dots, X_n) \sim \gamma^*$ , then the law of  $\frac{1}{n} \sum_{i=1}^n X_i$  is the Wasserstein barycenter of  $\mu_1, \dots, \mu_n$ .

4. Due to Theorems 8.7 and 8.10, in the case of the Wasserstein space we know that as long as the transport maps from the barycenter  $b^*$  to elements in the support of  $P$  are obtained from  $\alpha$ -strongly convex potentials, *and the barycenter of the extended distribution is still  $b^*$* , then  $Ph_{b^*}^b(\bullet) \geq \alpha$ . It turns out that due to the structure of the Wasserstein space, the second condition is unnecessary.

To prove this, use the following dual characterization of the Wasserstein barycenter: for each  $\mu \in \text{supp}(P)$ ,  $\varphi_\mu$  is such that  $(\nabla \varphi_\mu)_\# b^* = \mu$ , and  $\int (\frac{\|\cdot\|^2}{2} - \varphi_\mu) P(d\mu) = 0$ . Assume that each  $\varphi_\mu$  is  $\alpha(\mu)$ -strongly convex. Use this to show that

$$\varphi_\mu^*(x) + \varphi_\mu(y) \geq \langle x, y \rangle + \frac{\alpha(\mu)}{2} \|y - \nabla \varphi_\mu^*(x)\|^2.$$

By integrating this inequality, prove that (8.9) holds with  $\frac{\lambda}{1+\lambda}$  replaced by  $\int \alpha(\mu) P(d\mu)$ .

5. Let  $P$  be a probability measure over  $\mathcal{W}_2$  and let  $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$  denote the barycenter functional  $\mathcal{F}(b) := \frac{1}{2} PW_2^2(b, \bullet)$ . Using (5.30), the Wasserstein gradient of  $\mathcal{F}$  is given by  $\nabla \mathcal{F}(b) = \text{id} - PT_{b \rightarrow \bullet}$ , and a Wasserstein *gradient descent* step with step size  $h > 0$  is given by the iteration  $b^+ := (\text{id} - h \nabla \mathcal{F}(b))\# b$ . Prove the *descent lemma*

$$\mathcal{F}(b^+) - \mathcal{F}(b) \leq -h \left(1 - \frac{h}{2}\right) \|\mathbb{W}\mathcal{F}(b)\|_b^2,$$

which quantifies the progress made in one step of GD on  $\mathcal{F}$ . Deduce that  $h = 1$  is a reasonable choice of step size and write out the form of the GD updates in this case.

6. Specialize the GD updates (with step size  $h = 1$ ) from the previous exercise to the case when  $P$  is supported on centered, non-degenerate Gaussians. In particular, when initialized at a centered Gaussian, show that all of the iterates are centered Gaussians, and write down the update equations for the covariance matrix.

# A

---

## Convex analysis

In this appendix, we provide a quick review of convex analysis. We refer to the book [Roc97] for a comprehensive treatment.

### A.1 Convex functions, subdifferentials, and duality

**Definition A.1.** A function  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is convex if for all  $x, y \in \mathbb{R}^d$  and all  $t \in [0, 1]$ ,

$$f((1-t)x + t y) \leq (1-t)f(x) + t f(y).$$

Also, a set  $C \subseteq \mathbb{R}^d$  is convex if for all  $x, y \in \mathbb{R}^d$  and all  $t \in [0, 1]$ ,

$$(1-t)x + t y \in C.$$

The *domain* of  $f$ ,  $\text{dom}(f)$ , is the set  $\{f < \infty\}$  of points where  $f$  takes finite values. If  $f$  is convex, then  $\text{dom}(f)$  is a convex set.

We say that  $f$  is *proper* if it does not take the value  $-\infty$  (note that this is already assumed in the definition of convexity given above) and it is not identically  $+\infty$ . We assume without further mention that the convex functions we work with are proper. We say that  $f$  is *closed* or *lower semicontinuous* if for any sequence  $x_k \rightarrow x$  in  $\mathbb{R}^d$ , it holds that  $\liminf_{k \rightarrow \infty} f(x_k) \geq f(x)$ ; equivalently, all of the sublevel sets  $\{f \leq t\}$  for  $t \in \mathbb{R}$  are closed.

Suppose that  $f$  takes values in  $\mathbb{R}$ . Then, convexity of  $f$  implies that  $f$  is automatically continuous, and in fact locally Lipschitz, hence differentiable almost everywhere by Rademacher's theorem. If  $f$  is continuously differentiable, then convexity of  $f$  is equivalent to  $f$  always lying above its tangent line:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathbb{R}^d. \quad (\text{A.1})$$

If  $f$  is twice continuously differentiable, then convexity of  $f$  is equivalent to a condition on its Hessian:

$$\nabla^2 f(x) \succeq 0, \quad \forall x \in \mathbb{R}^d.$$

In general, a convex function may not be differentiable. One reason why differentiability can fail is simply because  $f$  takes on infinite values ( $\text{dom}(f) \neq \mathbb{R}^d$ ). However, as discussed above,  $f$  is always differentiable *almost* everywhere on the interior of its domain. Moreover, we can find a useful substitute for differentiability through the notion of a subgradient, which is based on the “above tangent line” property encapsulated in (A.1).

**Definition A.2 (Subdifferential).** Let  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  be convex and let  $x \in \mathbb{R}^d$ . We say that  $g$  is a subgradient of  $f$  at  $x$  if

$$f(y) \geq f(x) + \langle g, y - x \rangle, \quad \forall y \in \mathbb{R}^d.$$

The set of all subgradients of  $f$  at  $x$  is called the subdifferential of  $f$  at  $x$ , denoted  $\partial f(x)$ . Also,  $\partial f := \{(x, g) : x \in \mathbb{R}^d, g \in \partial f(x)\}$  is called the subdifferential of  $f$ .

Importantly, the following lemma holds:

**Lemma A.3.** If  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is convex and  $x$  lies in the interior of  $\text{dom}(f)$ , then  $\partial f(x)$  is non-empty. Also, if  $f$  is differentiable at  $x$ , then the subdifferential at  $x$  is single-valued and satisfies  $\partial f(x) = \{\nabla f(x)\}$ .

We next turn towards the crucial concept of duality.

**Definition A.4 (Convex conjugate).** For any function  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$ , we define its convex conjugate<sup>1</sup>  $f^*$  via

$$f^*(y) := \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - f(x)\}, \quad y \in \mathbb{R}^d.$$

*Example A.5.* Let  $A \succ 0$  be a positive definite matrix. Then, the convex conjugate of  $x \mapsto \frac{1}{2} \langle x, Ax \rangle$  is the function  $y \mapsto \frac{1}{2} \langle y, A^{-1}y \rangle$ . See Lemma A.13 for the proof. The reader is invited to write down other examples of convex functions and to compute their conjugates.

---

<sup>1</sup> The convex conjugate is also known as the *Fenchel–Legendre transform*, the *Fenchel dual* or variations of these terms.

As a supremum of affine functions, the convex conjugate of  $f$  is always a closed convex function, even if  $f$  is not. Conversely, if  $f$  is closed and convex, then  $f = f^{**}$ .

The inequality  $f(x) + f^*(y) \geq \langle x, y \rangle$  is trivial from the definition of the convex conjugate. However, it is important enough to deserve a name, and we need the equality case for later use.

**Theorem A.6 (Fenchel–Young inequality).** *For a convex function  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  and any  $x, y \in \mathbb{R}^d$ ,*

$$f(x) + f^*(y) \geq \langle x, y \rangle.$$

*Equality holds if and only if  $y \in \partial f(x)$ .*

Note that by symmetry, equality holds if and only if  $x \in \partial f^*(y)$ . In particular, when  $f$  and  $f^*$  are differentiable, then the subdifferentials are single-valued, so that the equality condition reads  $y = \nabla f(x)$  and  $x = \nabla f^*(y)$ . This says that the gradient mappings are inverse to each other:  $\nabla f^* = (\nabla f)^{-1}$ .

We conclude this section by proving Rockafellar's theorem (Theorem 1.10), which characterizes subdifferentials of closed convex functions as maximally monotone subsets of  $\mathbb{R}^d \times \mathbb{R}^d$ . In Section 1.4.1, we show that if  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex, then its subdifferential  $\partial \varphi$  is cyclically monotone. Here, we prove the converse.

*Proof of Theorem 1.10.* Let  $A$  be cyclically monotone and fix  $(x_0, y_0) \in A$ . Define for any  $x \in \mathbb{R}^d$  the function

$$\varphi(x) = \sup_{k \geq 0} \sup_{\substack{(x_i, y_i) \in A \\ i=1, \dots, k}} \{ \langle x_1 - x_0, y_0 \rangle + \langle x_2 - x_1, y_1 \rangle + \dots + \langle x - x_k, y_k \rangle \}.$$

Clearly  $\varphi$  is closed and convex as a supremum of affine functions. Moreover,  $\varphi(x_0) \leq 0$  by cyclical monotonicity<sup>2</sup> and  $\varphi(x_0) \geq 0$  (take  $k = 1$  and  $(x_1, y_1) = (x_0, y_0)$ ) so that  $\varphi(x_0) = 0$  and  $\varphi$  is a *proper* convex function. Finally note that for any  $(x, y) = (x_{k+1}, y_{k+1}) \in A$  and any  $z \in \mathbb{R}^d$ , it holds

$$\begin{aligned} \varphi(z) &\geq \sup_{k \geq 0} \sup_{\substack{(x_i, y_i) \in A, i=1, \dots, k}} \{ \langle x_1 - x_0, y_0 \rangle + \langle x_2 - x_1, y_1 \rangle + \dots \\ &\quad + \langle x - x_k, y_k \rangle + \langle z - x, y \rangle \} \\ &= \varphi(x) + \langle z - x, y \rangle. \end{aligned}$$

Therefore,  $y \in \partial \varphi(x)$ . □

<sup>2</sup> This is in fact the only place we use cyclical monotonicity!

## A.2 Strong convexity and smoothness

**Definition A.7.** A function  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is called  $\alpha$ -convex (for  $\alpha \in \mathbb{R}$ ) if for all  $x, y \in \mathbb{R}^d$  and all  $t \in [0, 1]$ ,

$$f((1-t)x + t y) \leq (1-t)f(x) + t f(y) - \frac{\alpha t(1-t)}{2} \|y - x\|^2.$$

The case when  $\alpha = 0$  corresponds to convexity, as in Definition A.1. When  $\alpha > 0$ , then  $f$  is called *strongly convex*, and the above inequality strengthens the usual convexity inequality. When  $\alpha < 0$ , then  $f$  is called *semi-convex*.

When  $f$  is continuously differentiable,  $\alpha$ -convexity is equivalent to the first statement below. When  $f$  is twice continuously differentiable,  $\alpha$ -convexity is equivalent to both statements below.

1.  $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|x - y\|^2$ , for all  $x, y \in \mathbb{R}^d$ .
2.  $\nabla^2 f(x) \succeq \alpha I$  for all  $x \in \mathbb{R}^d$ .

We also formulate the dual property of an upper bound on the second derivative.

**Definition A.8.** A continuously differentiable function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is called  $\beta$ -smooth ( $\beta \geq 0$ ) if for all  $x, y \in \mathbb{R}^d$ ,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2.$$

When  $f$  is twice continuously differentiable, then this is equivalent to  $\nabla^2 f(x) \preceq \beta I$  for all  $x \in \mathbb{R}^d$ . If, in addition,  $f$  is convex, then  $\nabla^2 f(x) \succeq 0$ , so in particular the operator norm of  $\nabla^2 f(x)$  is at most  $\beta$ . In turn, this is equivalent to the  $\beta$ -Lipschitzness of the mapping  $\nabla f : \mathbb{R}^d \rightarrow \mathbb{R}^d$  (thus, convex and smooth functions are often referred to as *gradient Lipschitz*).

The properties of strong convexity and smoothness are dual. For simplicity, we state the following lemma assuming that  $f$  is continuously differentiable, but the assumptions can be somewhat relaxed.

**Lemma A.9.** Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be continuously differentiable, convex, and  $\|\nabla f(x)\| \rightarrow \infty$  as  $\|x\| \rightarrow \infty$ . Let  $\alpha > 0$ . Then,  $f$  is  $\alpha$ -strongly convex if and only if its convex conjugate  $f^*$  is  $\frac{1}{\alpha}$ -smooth.

*Proof.* From classical results in convex analysis (see [Roc97, Theorem 25.5, Theorem 26.6, and Lemma 26.7]), under our assumptions,  $\nabla f : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is a diffeomorphism with inverse  $\nabla f^*$ .

( $\Rightarrow$ ) By taking the first-order condition for strong convexity and adding it to the inequality with  $x$  and  $y$  interchanged, we obtain

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|^2.$$

Let  $x = \nabla f^*(x')$  and  $y = \nabla f^*(y')$  and recall also that  $\nabla f \circ \nabla f^* = \text{id}$ . The above inequality yields, for all  $x', y' \in \mathbb{R}^d$ ,

$$\langle x' - y', \nabla f^*(x') - \nabla f^*(y') \rangle \geq \alpha \|\nabla f^*(x') - \nabla f^*(y')\|^2.$$

Applying Cauchy–Schwarz to the left-hand side and rearranging, it follows that  $\nabla f^*$  is  $\frac{1}{\alpha}$ -Lipschitz, which is equivalent to  $f^*$  being  $\frac{1}{\alpha}$ -smooth, as discussed above.

( $\Leftarrow$ ) By smoothness of  $f^*$ ,

$$\begin{aligned} f(y) &= \sup_{y' \in \mathbb{R}^d} \{ \langle y, y' \rangle - f^*(y') \} \\ &\geq \sup_{y' \in \mathbb{R}^d} \left\{ \langle y, y' \rangle - f^*(x') - \langle \nabla f^*(x'), y' - x' \rangle - \frac{1}{2\alpha} \|y' - x'\|^2 \right\} \\ &= -f^*(x') + \langle y, x' \rangle + \frac{\alpha}{2} \|y - \nabla f^*(x')\|^2. \end{aligned}$$

Choose  $x' = \nabla f(x')$  so that  $\nabla f^*(x') = x$  and recall that  $f(x) + f^*(x') = \langle x, x' \rangle$ . It yields

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq \frac{\alpha}{2} \|y - x\|^2,$$

completing the proof.  $\square$

Note that if  $f, f^*$  are twice continuously differentiable, then  $f$  is  $\alpha$ -strongly convex iff  $\nabla^2 f \succeq \alpha I$ , and  $f^*$  is  $\frac{1}{\alpha}$ -smooth iff  $\nabla^2 f^* = (\nabla^2 f)^{-1} \circ \nabla f^* \preceq \alpha^{-1} I$ , which provides a more transparent proof.

Strong convexity also implies the following property, which can be viewed as a strong quantitative form of the principle that locally optimal points are globally optimal under convexity.

**Definition A.10 (Polyak–Lojasiewicz inequality).** *We say that a continuously differentiable function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  satisfies a Polyak–Lojasiewicz (PL) inequality with constant  $\alpha > 0$  if for all  $x \in \mathbb{R}^d$ ,*

$$\|\nabla f(x)\|^2 \geq 2\alpha (f(x) - \inf f).$$

**Lemma A.11.** *If  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is continuously differentiable and  $\alpha$ -convex for  $\alpha > 0$ , then it satisfies a PL inequality with constant  $\alpha$ .*

*Proof.* Let  $x_*$  denote the minimizer of  $f$ . Then,

$$\inf f = f(x_*) \geq f(x) + \langle \nabla f(x), x_* - x \rangle + \frac{\alpha}{2} \|x_* - x\|^2.$$

By Cauchy–Schwarz and Young’s inequality,

$$\begin{aligned} \langle \nabla f(x), x_* - x \rangle &\geq -\|\nabla f(x)\| \|x_* - x\| \\ &\geq -\frac{1}{2\alpha} \|\nabla f(x)\|^2 - \frac{\alpha}{2} \|x_* - x\|^2. \end{aligned}$$

Substituting and rearranging finishes the proof.  $\square$

We also note that strong convexity implies quadratic growth around the minimizer.

**Lemma A.12.** *If  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is  $\alpha$ -convex, and if  $x_*$  denotes the minimizer of  $f$ , then for all  $x \in \mathbb{R}^d$ ,*

$$f(x) - f(x_*) \geq \frac{\alpha}{2} \|x - x_*\|^2. \quad (\text{A.2})$$

*Proof.* The strong convexity inequality gives

$$f((1-t)x_* + tx) \leq (1-t)f(x_*) + t f(x) - \frac{\alpha t (1-t)}{2} \|x - x_*\|^2,$$

or

$$\begin{aligned} 0 &\leq f((1-t)x_* + tx) - f(x_*) \\ &\leq t \left[ f(x) - f(x_*) - \frac{\alpha (1-t)}{2} \|x - x_*\|^2 \right]. \end{aligned}$$

Rearranging, dividing by  $t$ , and letting  $t \searrow 0$  proves the result.  $\square$

Actually, in Exercise 5 in Chapter 5, we refine this statement to show that the PL inequality itself implies the growth inequality (A.2). We refer the reader to [KNS16, Appendix A] for a concise exposition of the interplay between these inequalities.

### A.3 Convex conjugate of a quadratic function

In this section we establish the following useful lemma which states that the convex conjugate of a quadratic function is an explicit quadratic function.

**Lemma A.13.** *If  $f(x) = \frac{1}{2} x^\top A x + b^\top x$  for  $A \succ 0$ , then*

$$f^*(y) = \frac{1}{2} (y - b)^\top A^{-1} (y - b). \quad (\text{A.3})$$

*If  $A$  is not invertible, the same expression holds if we interpret  $A^{-1}(y - b)$  as the solution to  $y = Ax + b$  if it exists, and  $f^*(y) = +\infty$  otherwise.*

*Proof.* The definition of  $f^*(y)$  implies

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \left\{ y^\top x - \frac{1}{2} x^\top A x - b^\top x \right\}.$$

Differentiating the objective yields that if a maximizer  $x^*$  exists, then it satisfies

$$y = Ax^* + b,$$

which yields  $x^* = A^{-1}(y - b)$  if  $y - b \in \text{span}(A)$ . In this case, we obtain

$$f^*(y) = y^\top x^* - \frac{1}{2} (x^*)^\top A x^* - b^\top x^* = \frac{1}{2} (y - b)^\top A^{-1} (y - b).$$

On the other hand, if  $y - b \notin \text{span}(A)$ , then we can find a vector  $z \in \ker(A)$  such that  $z^\top (y - b) \neq 0$ . Considering  $x = \lambda z$  for  $\lambda \in \mathbb{R}$ , we obtain

$$\begin{aligned} f^*(y) &\geq \sup_{\lambda \in \mathbb{R}} \left\{ y^\top (\lambda z) - \frac{1}{2} (\lambda z)^\top A (\lambda z) - b^\top (\lambda z) \right\} \\ &= \sup_{\lambda \in \mathbb{R}} \lambda z^\top (y - b) = +\infty, \end{aligned}$$

as claimed.  $\square$



## B

---

### Probability

In this appendix, we gather together some background material on probability theory. See, e.g., the textbook [Bil99] for further discussion.

We begin with the notion of convergence of probability measures.

**Definition B.1.** *A sequence of probability measures  $(\mu_n)_n$  on  $\mathbb{R}^d$  is said to converge (weakly) to a probability measure  $\mu$  if for all bounded continuous functions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , it holds that*

$$\int f \, d\mu_n \rightarrow \int f \, d\mu.$$

For the topology of  $\mathbb{R}^d$ , we know exactly which subsets are compact: namely, a set  $A$  is compact if and only if it is closed and bounded (Heine–Borel theorem). This provides a useful criterion for when a sequence  $(x_n)_n$  in  $A$  converges, upon passing to a subsequence, to a point in  $A$ . The following definition and theorem characterize compact sets of probability measures in the topology of weak convergence.

**Definition B.2.** *A set  $\mathcal{A}$  of probability measures on  $\mathbb{R}^d$  is tight if for all  $\varepsilon > 0$ , there is a compact set  $K$  such that  $\mu(K^c) \leq \varepsilon$  for all  $\mu \in \mathcal{A}$ .*

**Theorem B.3 (Prokhorov's theorem).** *Any weakly convergent sequence of probability measures is tight. Conversely, any tight sequence of probability measures has a subsequential weak limit.*

*Equivalently, a set  $\mathcal{A}$  of probability measures on  $\mathbb{R}^d$  is compact if and only if it is closed and tight.*

We omit the proof, but the intuition can be gleaned via a simple example: on  $\mathbb{R}$ , let  $\mu_n = \delta_{x_n}$  for all  $n$ , where  $x_n \rightarrow \infty$ ; then,  $(\mu_n)_n$

clearly has no weakly convergent subsequence. The condition of tightness ensures that the mass does not run off to infinity, and once this is ensured then a weakly convergent subsequence is guaranteed.

The following theorem provides useful reformulations of weak convergence of measures.

**Theorem B.4 (Portmanteau theorem).** *Let  $(\mu_n)_n$  be a sequence in  $\mathcal{P}(\mathbb{R}^d)$  and let  $\mu \in \mathcal{P}(\mathbb{R}^d)$ . The following are equivalent.*

1.  $\mu_n \rightarrow \mu$  weakly.
2.  $\int f d\mu_n \rightarrow \int f d\mu$  for all bounded Lipschitz continuous  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ .
3.  $\int f d\mu \leq \liminf_{n \rightarrow \infty} \int f d\mu_n$  for all lower semicontinuous functions  $f : \mathbb{R}^d \rightarrow [0, \infty]$ .
4.  $\int f d\mu \geq \limsup_{n \rightarrow \infty} \int f d\mu_n$  for all upper semicontinuous functions  $f : \mathbb{R}^d \rightarrow [-\infty, 0]$ .
5.  $\mu(G) \leq \liminf_{n \rightarrow \infty} \mu_n(G)$  for all open  $G \subseteq \mathbb{R}^d$ .
6.  $\mu(F) \geq \limsup_{n \rightarrow \infty} \mu_n(F)$  for all closed  $F \subseteq \mathbb{R}^d$ .
7.  $\lim_{n \rightarrow \infty} \mu_n(A) = \mu(A)$  for all Borel  $A \subseteq \mathbb{R}^d$  such that  $\mu(\partial A) = 0$ .

*Proof.* (1)  $\Rightarrow$  (2) is trivial. Also, it is easy to see that (3) is equivalent to (4) by replacing  $f$  by  $-f$ , and that (5) is equivalent to (6) by taking complements.

(2)  $\Rightarrow$  (3): It is known that one can approximate  $f$  from below by a sequence  $(f_k)_k$  of Lipschitz continuous functions  $\mathbb{R}^d \rightarrow [0, \infty)$ . For any  $k, n \in \mathbb{N}$ , it holds that  $\int f_k d\mu_n \leq \int f d\mu_n$ . Taking the limit  $n \rightarrow \infty$ , we get  $\int f_k d\mu \leq \liminf_{n \rightarrow \infty} \int f d\mu_n$ . Then, take  $k \rightarrow \infty$  using the monotone convergence theorem.

(3)  $\Rightarrow$  (5): The indicator function  $\mathbb{1}_G$  is lower semicontinuous. Similarly, we obtain (4)  $\Rightarrow$  (6) since the indicator function  $\mathbb{1}_F$  is upper semicontinuous.

(5) and (6)  $\Rightarrow$  (7): The condition  $\mu(\partial A) = 0$  means that  $\mu(\text{int } A) = \mu(A) = \mu(\overline{A})$ . Applying (5) to the open set  $\text{int } A$  and (6) to the closed set  $\overline{A}$  proves (7).

(7)  $\Rightarrow$  (1): Let  $f$  be bounded and continuous; we may as well assume  $f$  is non-negative. Observe that for  $t \in \mathbb{R}$ , it holds that  $\partial f^{-1}([t, \infty)) \subseteq f^{-1}(\{t\})$ , and this set can have positive  $\mu$ -measure for at most countably many values of  $t$ . Applying (7), we obtain

$$\int f d\mu_n = \int_0^\infty \mu_n\{f \geq t\} dt \rightarrow \int_0^\infty \mu\{f \geq t\} dt = \int f d\mu,$$

where to justify the convergence we can use, e.g., bounded convergence (since the integral can actually be taken over a finite interval).  $\square$

The following lemma, as the name suggests, allows us to “glue” together couplings which share a marginal and is useful for some constructions in optimal transport (e.g., the proof of the triangle inequality for Wasserstein distances in Proposition 1.3).

**Lemma B.5 (Gluing lemma).** *Let  $\gamma$  and  $\gamma'$  be two measures on  $\mathbb{R}^d \times \mathbb{R}^d$  such that for any Borel set  $A \subset \mathbb{R}^d$ , it holds  $\gamma(\mathbb{R}^d \times A) = \gamma'(A \times \mathbb{R}^d)$ , i.e., the second marginal of  $\gamma$  coincides with the first marginal of  $\gamma'$ . Then there exists three random variables  $X, Y, Z \in \mathbb{R}^d$  such that  $(X, Z) \sim \gamma$  and  $(Z, Y) \sim \gamma'$ .*

*Proof.* We are going to explicitly construct such a triplet  $(X, Y, Z)$ . To that end, let  $Z$  be distributed according the second marginal of  $\gamma$  (which corresponds to the first marginal of  $\gamma'$  by assumption). Then  $\gamma$  (resp.  $\gamma'$ ) determines the conditional distribution of  $X$  (resp.  $Y$ ) given  $Z$ . For example, we may draw  $X$  and  $Y$  to be conditionally independent given  $Z$ . This gives a valid triplet  $(X, Y, Z)$ .  $\square$

$$X \xrightarrow{\text{opt}} Z \xrightarrow{\text{opt}} Y$$

**Fig. B.1.** The diagram above is a convenient way to represent couplings between multiple random variables. An edge represent the constraint that the coupling needs to be optimal. In general, any coupling described as a graph with no cycle can be realized using the gluing lemma.



---

## References

ABAM22. E. ABBE, E. BOIX-ADSERÀ, and T. MISIAKIEWICZ, The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks, in *Proceedings of Thirty Fifth Conference on Learning Theory* (P.-L. LOH and M. RAGINSKY, eds.), *Proceedings of Machine Learning Research* **178**, PMLR, 7 2022, pp. 4782–4887.

ABAM23. E. ABBE, E. BOIX-ADSERÀ, and T. MISIAKIEWICZ, SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics, in *Proceedings of Thirty Sixth Conference on Learning Theory* (G. NEU and L. ROSASCO, eds.), *Proceedings of Machine Learning Research* **195**, PMLR, 7 2023, pp. 2552–2623.

AC11. M. AGUEH and G. CARLIER, Barycenters in the Wasserstein space, *SIAM Journal on Mathematical Analysis* **43** (2011), 904–924.

ACLGP20. A. AHIDAR-COUTRIX, T. LE GOUIC, and Q. PARIS, Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics, *Probab. Theory Related Fields* **177** (2020), 323–368.

AKT84. M. AJTAI, J. KOMLÓS, and G. TUSNÁDY, On optimal matchings, *Combinatorica* **4** (1984), 259–264.

ACG<sup>+</sup>23. O. D. AKYILDIZ, F. R. CRUCINIO, M. GIROLAMI, T. JOHNSTON, and S. SABANIS, Interacting particle Langevin algorithm for maximum marginal likelihood estimation, *arXiv preprint 2303.13429* (2023), 1–38.

AKP22. S. ALEXANDER, V. KAPOVITCH, and A. PETRUNIN, Alexandrov geometry: foundations, *arXiv preprint 1903.08539* (2022), 1–301.

AR20. P. ALQUIER and J. RIDGWAY, Concentration of tempered posteriors and of their variational approximations, *Ann. Statist.* **48** (2020), 1475–1497.

ACGS21. J. ALTSCHULER, S. CHEWI, P. GERBER, and A. J. STROMME, Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent, in *Advances in Neural Information Processing Systems* (M. RANZATO, A. BEYGELZIMER, K. NGUYEN, P. S. LIANG, J. W. VAUGHAN, and Y. DAUPHIN, eds.), **34**, Curran Associates, Inc., 2021, pp. 22132–22145.

ABA21. J. M. ALTSCHULER and E. BOIX-ADSERÀ, Wasserstein barycenters can be computed in polynomial time in fixed dimension, *Journal of Machine Learning Research* **22** (2021), 1–19.

ABA22. J. M. ALTSCHULER and E. BOIX-ADSERÀ, Wasserstein barycenters are NP-hard to compute, *SIAM J. Math. Data Sci.* **4** (2022), 179–203.

AC24. J. M. ALTSCHULER and S. CHEWI, Faster high-accuracy log-concave sampling via algorithmic warm starts, *J. ACM* **71** (2024), 1–55.

ANWR17. J. M. ALTSCHULER, J. NILES-WEED, and P. RIGOLLET, Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration, in *Advances in Neural Information Processing Systems 30*, 2017, pp. 1961–1971.

Ama98. S.-I. AMARI, Natural gradient works efficiently in learning, *Neural Computation* **10** (1998), 251–276.

AN00. S.-I. AMARI and H. NAGAOKA, *Methods of information geometry*, *Translations of Mathematical Monographs* **191**, American Mathematical Society, Providence, RI, 2000, Translated from the 1993 Japanese original by Daishi Harada.

ABS21. L. AMBROSIO, E. BRUÉ, and D. SEMOLA, *Lectures on optimal transport*, *Unitext* **130**, Springer, 2021.

AG13. L. AMBROSIO and N. GIGLI, A user’s guide to optimal transport, in *Modelling and optimisation of flows on networks*, *Lecture Notes in Math.* **2062**, Springer, Heidelberg, 2013, pp. 1–155.

AGS08. L. AMBROSIO, N. GIGLI, and G. SAVARÉ, *Gradient flows in metric spaces and in the space of probability measures*, second ed., *Lectures in Mathematics ETH Zürich*, Birkhäuser Verlag, Basel, 2008.

AST19. L. AMBROSIO, F. STRA, and D. TREVISAN, A PDE approach to a 2-dimensional matching problem, *Probab. Theory Related Fields* **173** (2019), 433–477.

AKSG19. M. ARBEL, A. KORBA, A. SALIM, and A. GRETTON, Maximum mean discrepancy gradient flow, in *Advances in Neural Information Processing Systems* (H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D’ALCHÉ-BUC, E. FOX, and R. GARNETT, eds.), **32**, Curran Associates, Inc., 2019.

ACB17. M. ARJOVSKY, S. CHINTALA, and L. BOTTOU, Wasserstein generative adversarial networks, in *International Conference on Machine Learning*, 2017, pp. 214–223.

AL24. M. ARNESE and D. LACKER, Convergence of coordinate ascent variational inference for log-concave measures via optimal transport, *arXiv preprint 2404.08792* (2024), 1–28.

AFKL22. P.-C. AUBIN-FRANKOWSKI, A. KORBA, and F. LÉGER, Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM, in *Advances in Neural Information Processing Systems* (S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO, and A. OH, eds.), **35**, Curran Associates, Inc., 2022, pp. 17263–17275.

AR15. P. AWASTHI and A. RISTESKI, On some provably correct cases of variational inference for topic models, in *Advances in Neural Information Processing Systems* (C. CORTES, N. LAWRENCE, D. LEE, M. SUGIYAMA, and R. GARNETT, eds.), **28**, Curran Associates, Inc., 2015, pp. 1–9.

AJLS17. N. AY, J. JOST, H. V. LÊ, and L. SCHWACHHÖFER, *Information geometry, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]* **64**, Springer, Cham, 2017.

BNNR11. K. D. BA, H. L. NGUYEN, H. N. NGUYEN, and R. RUBINFELD, Sublinear time algorithms for earth mover's distance, *Theory Comput. Syst.* **48** (2011), 428–442.

BVFRT22. J. BACKHOFF-VERAGUAS, J. FONTBONA, G. RIOS, and F. TOBAR, Bayesian learning with Wasserstein barycenters, *ESAIM Probab. Stat.* **26** (2022), 436–472.

BDI<sup>+</sup>20. A. BACKURS, Y. DONG, P. INDYK, I. RAZENSHTEYN, and T. WAGNER, Scalable nearest neighbor search for optimal transport, in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. SINGH, eds.), *Proceedings of Machine Learning Research* **119**, PMLR, 7 2020, pp. 497–506.

BGL14. D. BAKRY, I. GENTIL, and M. LEDOUX, *Analysis and geometry of Markov diffusion operators, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **348**, Springer, Cham, 2014.

BCE<sup>+</sup>22. K. BALASUBRAMANIAN, S. CHEWI, M. A. ERDOGDU, A. SALIM, and M. S. ZHANG, Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo, in *Proceedings of Thirty Fifth Conference on Learning Theory* (P.-L. LOH and M. RAGINSKY, eds.), *Proceedings of Machine Learning Research* **178**, PMLR, 7 2022, pp. 2896–2923.

BB23. M. BALLU and Q. BERTHET, Mirror Sinkhorn: fast online optimization on transport polytopes, in *Proceedings of the 40th International Conference on Machine Learning* (A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO, and J. SCARLETT, eds.), *Proceedings of Machine Learning Research* **202**, PMLR, 7 2023, pp. 1595–1613.

dBCAMRR99. E. DEL BARRIO, J. A. CUESTA-ALBERTOS, C. MATRÁN, and J. M. RODRÍGUEZ-RODRÍGUEZ, Tests of goodness of fit based on the  $L_2$ -Wasserstein distance, *Ann. Statist.* **27** (1999), 1230–1239.

dBSLNW23. E. DEL BARRIO, A. G. SANZ, J.-M. LOUBES, and J. NILES-WEED, An improved central limit theorem and fast convergence rates for entropic transportation costs, *SIAM J. Math. Data Sci.* **5** (2023), 639–669.

BKR14. S. BASU, S. KOLOURI, and G. K. ROHDE, Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry, *Proceedings of the National Academy of Sciences* **111** (2014), 3448–3453.

BJL19. R. BHATIA, T. JAIN, and Y. LIM, On the Bures–Wasserstein distance between positive definite matrices, *Expo. Math.* **37** (2019), 165–191.

BCP19a. J. BIGOT, E. CAZELLES, and N. PAPADAKIS, Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications, *Electron. J. Stat.* **13** (2019), 5120–5150.

BCP19b. J. BIGOT, E. CAZELLES, and N. PAPADAKIS, Penalization of barycenters in the Wasserstein space, *SIAM J. Math. Anal.* **51** (2019), 2261–2285.

BGKL17. J. BIGOT, R. GOUET, T. KLEIN, and A. LÓPEZ, Geodesic PCA in the Wasserstein space by convex PCA, *Ann. Inst. Henri Poincaré Probab. Stat.* **53** (2017), 1–26.

Bil99. P. BILLINGSLEY, *Convergence of probability measures*, second ed., Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1999, A Wiley-Interscience Publication.

BB18. A. BLANCHET and J. BOLTE, A family of functional inequalities: Lojasiewicz inequalities and displacement convex functions, *J. Funct. Anal.* **275** (2018), 1650–1673.

BKM17. D. M. BLEI, A. KUCUKELBIR, and J. D. McAULIFFE, Variational inference: a review for statisticians, *Journal of the American Statistical Association* **112** (2017), 859–877.

BL19. S. BOBKOV and M. LEDOUX, One-dimensional empirical measures, order statistics, and Kantorovich transport distances, *Mem. Amer. Math. Soc.* **261** (2019), v+126.

BGL01. S. G. BOBKOV, I. GENTIL, and M. LEDOUX, Hypercontractivity of Hamilton–Jacobi equations, *J. Math. Pures Appl. (9)* **80** (2001), 669–696.

BL21. S. G. BOBKOV and M. LEDOUX, A simple Fourier analytic proof of the AKT optimal matching theorem, *Ann. Appl. Probab.* **31** (2021), 2567–2584.

Boe24. M. T. BOEDIHARDJO, Sharp bounds for the max-sliced Wasserstein distance, *arXiv preprint arXiv:2403.00666* (2024).

BLG14. E. BOISSARD and T. LE GOUIC, On the mean speed of convergence of empirical and occupation measures in Wasserstein distance, *Ann. Inst. Henri Poincaré Probab. Stat.* **50** (2014), 539–563.

BV05. F. BOLLEY and C. VILLANI, Weighted Csiszár–Kullback–Pinsker inequalities and applications to transportation inequalities, *Ann. Fac. Sci. Toulouse Math. (6)* **14** (2005), 331–352.

BLB24. S. BONNABEL, M. LAMBERT, and F. BACH, Low-rank plus diagonal approximations for Riccati-like matrix differential equations, *arXiv preprint 2407.03373* (2024), 1–21.

BPC16. N. BONNEEL, G. PEYRÉ, and M. CUTURI, Wasserstein barycentric coordinates: histogram regression using optimal transport, *ACM Trans. Graph.* **35** (2016), 10.

Bon13. N. BONNOTTE, *Unidimensional and evolution methods for optimal transportation*, Theses, Université Paris Sud – Paris XI; Scuola normale superiore (Pise, Italie), 12 2013.

BGR<sup>+</sup>06. K. M. BORGWARDT, A. GRETTON, M. J. RASCH, H.-P. KRIEGEL, B. SCHÖLKOPF, and A. J. SMOLA, Integrating structured biological data by kernel maximum mean discrepancy, *Bioinformatics* **22** (2006), e49–e57.

BLM13. S. BOUCHERON, G. LUGOSI, and P. MASSART, *Concentration inequalities*, Oxford University Press, Oxford, 2013, A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

BV04. S. BOYD and L. VANDENBERGHE, *Convex optimization*, Cambridge University Press, Cambridge, 2004.

BRT24. S. BRAHMACHARI, R. RUBBOLI, and M. TOMAMICHEL, A fixed-point algorithm for matrix projections with applications in quantum information, *arXiv preprint 2312.14615* (2024), 1–17.

BL76. H. J. BRASCAMP and E. H. LIEB, On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation, *J. Functional Analysis* **22** (1976), 366–389.

Bre87. Y. BRENIER, Décomposition polaire et rearrangement monotone des champs de vecteurs, *C. R. Acad. Sci. Paris Ser. I Math.* **305** (1987), 805–808.

Bub15. S. BUBECK, *Convex optimization: algorithms and complexity*, Now Publishers Inc., 2015.

BMPKC22. C. BUNNE, L. MENG-PAPAXANTHOS, A. KRAUSE, and M. CUTURI, Proximal optimal transport modeling of population dynamics, in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

BSG<sup>+</sup>23. C. BUNNE, S. G. STARK, G. GUT, J. S. DEL CASTILLO, M. LEVESQUE, K.-V. LEHMANN, L. PELKMANS, A. KRAUSE, and G. RÄTSCH, Learning single-cell perturbation responses using neural optimal transport, *Nature Methods* **20** (2023), 1759–1768.

BBI01. D. BURAGO, Y. BURAGO, and S. IVANOV, *A course in metric geometry*, *Graduate Studies in Mathematics* **33**, American Mathematical Society, Providence, RI, 2001.

BM03. S. BURER and R. D. C. MONTEIRO, A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization, *Math. Program.* **95** (2003), 329–357, Computational semidefinite and second order cone programming: the state of the art.

BM05. S. BURER and R. D. C. MONTEIRO, Local minima and convergence in low-rank semidefinite programming, *Math. Program.* **103** (2005), 427–444.

Bur69. D. BURES, An extension of Kakutani’s theorem on infinite product measures to the tensor product of semifinite  $w^*$ -algebras, *Trans. Amer. Math. Soc.* **135** (1969), 199–212.

CCCC20. T. CAI, J. CHENG, N. CRAIG, and K. CRAIG, Linearized optimal transport for collider events, *Phys. Rev. D* **102** (2020), 116019.

CKPJ24. R. CAPRIO, J. KUNTZ, S. POWER, and A. M. JOHANSEN, Error bounds for particle gradient descent, and extensions of the log-Sobolev and Talagrand inequalities, *arXiv preprint 2403.02004* (2024), 1–33.

CEK21. G. CARLIER, K. EICHINGER, and A. KROSHNIN, Entropic–Wasserstein barycenters: PDE characterization, regularity, and CLT, *SIAM J. Math. Anal.* **53** (2021), 5880–5914.

dC92. M. P. A. DO CARMO, *Riemannian geometry*, *Mathematics: Theory & Applications*, Birkhäuser Boston, Inc., Boston, MA, 1992, Translated from the second Portuguese edition by Francis Flaherty.

CSB<sup>+</sup>18. E. CAZELLES, V. SEGUY, J. BIGOT, M. CUTURI, and N. PAPADAKIS, Geodesic PCA versus log-PCA of histograms in the Wasserstein space, *SIAM J. Sci. Comput.* **40** (2018), B429–B456.

CRW23. F. CHEN, Z. REN, and S. WANG, Uniform-in-time propagation of chaos for mean field Langevin dynamics, *arXiv preprint 2212.03050* (2023), 1–66.

CRBD18. R. T. Q. CHEN, Y. RUBANOVA, J. BETTENCOURT, and D. K. DUVENAUD, Neural ordinary differential equations, *Advances in Neural Information Processing Systems* **31** (2018), 13.

CLTW24. S. CHEN, Q. LI, O. TSE, and S. J. WRIGHT, Accelerating optimization over the space of probability measures, *arXiv preprint 2310.04006* (2024), 1–40.

CLM23. Y. CHEN, Z. LIN, and H.-G. MÜLLER, Wasserstein regression, *Journal of the American Statistical Association* **118** (2023), 869–882.

CCSW22. Y. CHEN, S. CHEWI, A. SALIM, and A. WIBISONO, Improved analysis for a proximal algorithm for sampling, in *Proceedings of Thirty Fifth Conference on Learning Theory* (P.-L. LOH and M. RAGINSKY, eds.), *Proceedings of Machine Learning Research* **178**, PMLR, 7 2022, pp. 2984–3014.

CGP21. Y. CHEN, T. T. GEORGIOU, and M. PAVON, Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge, *SIAM Review* **63** (2021), 249–313.

CGT19. Y. CHEN, T. T. GEORGIOU, and A. TANNENBAUM, Optimal transport for Gaussian mixture models, *IEEE Access* **7** (2019), 6269–6278.

CB18. X. CHENG and P. BARTLETT, Convergence of Langevin MCMC in KL-divergence, in *Proceedings of Algorithmic Learning Theory* (F. JANOOS, M. MOHRI, and K. SRIDHARAN, eds.), *Proceedings of Machine Learning Research* **83**, PMLR, 4 2018, pp. 186–211.

CGHH17. V. CHERNOZHUKOV, A. GALICHON, M. HALLIN, and M. HENRY, Monge–Kantorovich depth, quantiles, ranks and signs, *Ann. Statist.* **45** (2017), 223–256.

Che23. S. CHEWI, *An optimization perspective on log-concave sampling and beyond*, Ph.D. thesis, MIT, 2023.

Che24. S. CHEWI, *Log-concave sampling*, Forthcoming, 2024, Available online at <https://chewisinho.github.io/>.

CLGL<sup>+</sup>20a. S. CHEWI, T. LE GOUIC, C. LU, T. MAUNU, and P. RIGOLLET, SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence, in *Advances in Neural Information Processing Systems* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. F. BALCAN, and H. LIN, eds.), **33**, Curran Associates, Inc., 2020, pp. 2098–2109.

CLGL<sup>+</sup>20b. S. CHEWI, T. LE GOUIC, C. LU, T. MAUNU, P. RIGOLLET, and A. J. STROMME, Exponential ergodicity of mirror-Langevin diffusions, in *Advances in Neural Information Processing Systems* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN, and H. LIN, eds.), **33**, Curran Associates, Inc., 2020, pp. 19573–19585.

CMRS20. S. CHEWI, T. MAUNU, P. RIGOLLET, and A. STROMME, Gradient descent algorithms for Bures–Wasserstein barycenters, in *Proceedings of Thirty Third Conference on Learning Theory* (J. ABERNETHY and S. AGARWAL, eds.), *Proceedings of Machine Learning Research* **125**, PMLR, 7 2020, pp. 1276–1304.

CP23. S. CHEWI and A.-A. POOLADIAN, An entropic generalization of Caffarelli’s contraction theorem via covariance inequalities, *Reports. Mathematical* **361** (2023), 1471–1482.

Chi23. L. CHIZAT, Doubly regularized entropic Wasserstein barycenters, *arXiv preprint 2303.11844* (2023), 1–27.

CB18. L. CHIZAT and F. BACH, On the global convergence of gradient descent for over-parameterized models using optimal transport, in *Advances in Neural Information Processing Systems* (S. BENGIO, H. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI, and R. GARNETT, eds.), **31**, Curran Associates, Inc., 2018.

CDV24. L. CHIZAT, A. DELALANDE, and T. VAŠKEVIČIUS, Sharper exponential convergence rates for Sinkhorn’s algorithm in continuous settings, *arXiv preprint 2407.01202* (2024), 1–36.

CPSV18. L. CHIZAT, G. PEYRÉ, B. SCHMITZER, and F.-X. VIALARD, An interpolating distance between optimal transport and Fisher–Rao metrics, *Found. Comput. Math.* **18** (2018), 1–44.

CRL<sup>+</sup>20. L. CHIZAT, P. ROUSSILLON, F. LÉGER, F. VIALARD, and G. PEYRÉ, Faster Wasserstein distance estimation with the Sinkhorn divergence, in *Advances in Neural Information Processing Systems 33* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN, and H. LIN, eds.), 2020.

Cho12. O. CHODOSH, Optimal transport and Ricci curvature: Wasserstein space over the interval, 2012, Cambridge Part III essay. Available at arXiv:1105.2883.

CLZ20. S.-N. CHOW, W. LI, and H. ZHOU, Wasserstein Hamiltonian flows, *J. Differential Equations* **268** (2020), 1205–1219.

CHD24. P. CLAVIER, T. HUIX, and A. DURMUS, VITS: variational inference Thompson sampling for contextual bandits, *arXiv preprint 2307.10167* (2024), 1–43.

CS91. E. G. COFFMAN, JR. and P. W. SHOR, A simple proof of the  $O(\sqrt{n} \log^{3/4} n)$  upright matching bound, *SIAM Journal on Discrete Mathematics* **4** (1991), 48–57.

CK07. H. COHN and A. KUMAR, Universally optimal distribution of points on spheres, *Journal of the American Mathematical Society* **20** (2007), 99–148.

CE02. D. CORDERO-ERAUSQUIN, Some applications of mass transport to Gaussian-type inequalities, *Arch. Ration. Mech. Anal.* **161** (2002), 257–269.

CFTC16. N. COURTY, R. FLAMARY, D. TUIA, and T. CORPETTI, Optimal transport for data fusion in remote sensing, in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pp. 3571–3574.

CFTR17. N. COURTY, R. FLAMARY, D. TUIA, and A. RAKOTOMAMONJY, Optimal transport for domain adaptation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39** (2017), 1853–1865.

CRMB24. C. CRISCITIELLO, Q. REBJOCK, A. D. MCRAE, and N. BOUMAL, Synchronization on circles and spheres with nonlinear interactions, 2024.

CTZ24. T. CUI, X. TONG, and O. ZAHM, Optimal Riemannian metric for Poincaré inequalities and how to ideally precondition Langevin dynamics, *arXiv preprint 2404.02554* (2024), 1–28.

Cut13. M. CUTURI, Sinkhorn distances: lightspeed computation of optimal transport, in *Advances in Neural Information Processing Systems 26* (C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI, and K. Q. WEINBERGER, eds.), Curran Associates, Inc., 2013, pp. 2292–2300.

CD14. M. CUTURI and A. DOUCET, Fast computation of Wasserstein barycenters, in *Proceedings of the 31st International Conference on Machine Learning* (E. P. XING and T. JEBARA, eds.), **32** Proceedings of Machine Learning Research no. 2, PMLR, Beijing, China, 6 2014, pp. 685–693.

Dal17. A. S. DALALYAN, Theoretical guarantees for approximate sampling from smooth and log-concave densities, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **79** (2017), 651–676.

DT12. A. S. DALALYAN and A. B. TSYBAKOV, Sparse regression learning by aggregation and Langevin Monte-Carlo, *J. Comput. System Sci.* **78** (2012), 1423–1443.

DN23. A. DAS and D. NAGARAJ, Provably fast finite particle variants of SVGD via virtual particle stochastic approximation, in *Advances in Neural Information Processing Systems* (A. OH, T. NAUMANN, A. GLOBERSON, K. SAENKO, M. HARDT, and S. LEVINE, eds.), **36**, Curran Associates, Inc., 2023, pp. 49748–49760.

DBTHD21. V. DE BORTOLI, J. THORNTON, J. HENG, and A. DOUCET, Diffusion Schrödinger bridge with applications to score-based generative modeling, in *Advances in Neural Information Processing Systems* (M. RANZATO, A. BEYGELZIMER, Y. DAUPHIN, P. LIANG, and J. W. VAUGHAN, eds.), **34**, Curran Associates, Inc., 2021, pp. 17695–17709.

DGS21. N. DEB, P. GHOSAL, and B. SEN, Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections, in *Advances in Neural Information Processing Systems 34* (M. RANZATO, A. BEYGELZIMER, Y. N. DAUPHIN, P. LIANG, and J. W. VAUGHAN, eds.), 2021, pp. 29736–29753.

DS23. N. DEB and B. SEN, Multivariate rank-based distribution-free non-parametric testing using measure transportation, *J. Amer. Statist. Assoc.* **118** (2023), 192–207.

DD20. J. DELON and A. DESOLNEUX, A Wasserstein-type distance in the space of Gaussian mixture models, *SIAM J. Imaging Sci.* **13** (2020), 936–970.

DSS13. S. DEREICH, M. SCHEUTZOW, and R. SCHOTTSTEDT, Constructive quantization: approximation by empirical measures, in *Annales de l'IHP Probabilités et Statistiques*, **49**, 2013, pp. 1183–1203.

DHS<sup>+</sup>19. I. DESHPANDE, Y. HU, R. SUN, A. PYRROS, N. SIDDIQUI, S. KOYEJO, Z. ZHAO, D. A. FORSYTH, and A. G. SCHWING, Max-sliced Wasserstein distance and its use for GANs, in *IEEE Conference on Computer Vision and Pattern Recognition*, Computer Vision Foundation / IEEE, 2019, pp. 10648–10656.

DBCS23. M. Z. DIAO, K. BALASUBRAMANIAN, S. CHEWI, and A. SALIM, Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein space, in *Proceedings of the 40th International Conference on Machine Learning* (A. KRAUSE, E. BRUNSKILL,

K. CHO, B. ENGELHARDT, S. SABATO, and J. SCARLETT, eds.), *Proceedings of Machine Learning Research* **202**, PMLR, 7 2023, pp. 7960–7991.

Div22. V. DIVOL, Measure estimation on manifolds: an optimal transport approach, *Probab. Theory Related Fields* **183** (2022), 581–647.

DNWP22. V. DIVOL, J. NILES-WEED, and A.-A. POOLADIAN, Optimal transport map estimation in general function spaces, *arXiv preprint 2212.03722* (2022), 1–68.

Dob70. R. L. DOBRUSHIN, Prescribing a system of random variables by conditional distributions, *Theory of Probability & Its Applications* **15** (1970), 458–486.

DEP23. C. DOMINGO-ENRICH and A.-A. POOLADIAN, An explicit expansion of the Kullback–Leibler divergence along its Fisher–Rao gradient flow, *Transactions on Machine Learning Research* (2023), 1–16.

Dom20. J. DOMKE, Provable smoothness guarantees for black-box variational inference, in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. SINGH, eds.), *Proceedings of Machine Learning Research* **119**, PMLR, 7 2020, pp. 2587–2596.

DGG23. J. DOMKE, R. GOWER, and G. GARRIGOS, Provable convergence guarantees for black-box variational inference, in *Advances in Neural Information Processing Systems* (A. OH, T. NEUMANN, A. GLOBERSON, K. SAENKO, M. HARDT, and S. LEVINE, eds.), **36**, Curran Associates, Inc., 2023, pp. 66289–66327.

DM19. P. DUBEY and H.-G. MÜLLER, Fréchet analysis of variance for random objects, *Biometrika* **106** (2019), 803–821.

DM20. P. DUBEY and H.-G. MÜLLER, Fréchet change-point detection, *The Annals of Statistics* **48** (2020), 3312–3335.

Dud67. R. M. DUDLEY, The sizes of compact subsets of Hilbert space and continuity of Gaussian processes, *J. Functional Analysis* **1** (1967), 290–330.

Dud69. R. M. DUDLEY, The speed of mean Glivenko–Cantelli convergence, *The Annals of Mathematical Statistics* **40** (1969), 40–50.

Dud99. R. M. DUDLEY, *Uniform central limit theorems*, Cambridge Studies in Advanced Mathematics **63**, Cambridge University Press, Cambridge, 1999.

Dud02. R. M. DUDLEY, *Real analysis and probability*, Cambridge Studies in Advanced Mathematics **74**, Cambridge University Press, Cambridge, 2002, Revised reprint of the 1989 original.

DNS23. A. DUNCAN, N. NUESKEN, and L. SZPRUCH, On the geometry of Stein variational gradient descent, *Journal of Machine Learning Research* **24** (2023), 1–39.

DE97. P. DUPUIS and R. S. ELLIS, *A weak convergence approach to the theory of large deviations*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1997, A Wiley-Interscience Publication.

DMM19. A. DURMUS, S. MAJEWSKI, and B. MIASOJEDOW, Analysis of Langevin Monte Carlo via convex optimization, *J. Mach. Learn. Res.* **20** (2019), Paper No. 73, 46.

DM17. A. DURMUS and E. MOULINES, Nonasymptotic convergence analysis for the unadjusted Langevin algorithm, *Ann. Appl. Probab.* **27** (2017), 1551–1587.

DGK18. P. DVURECHENSKY, A. GASNIKOV, and A. KROSHNIN, Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn’s algorithm, in *Proceedings of the 35th International Conference on Machine Learning* (J. DY and A. KRAUSE, eds.), *Proceedings of Machine Learning Research* **80**, PMLR, Stockholm Sweden, 7 2018, pp. 1367–1376.

EHL18. W. E, J. HAN, and Q. LI, A mean-field optimal control formulation of deep learning, *Research in the Mathematical Sciences* **6** (2018), 1–41.

AEdBCAM16. P. C. ÁLVAREZ ESTEBAN, E. DEL BARRIO, J. A. CUESTA-ALBERTOS, and C. MATRÁN, A fixed-point approach to barycenters in Wasserstein space, *J. Math. Anal. Appl.* **441** (2016), 744–762.

Fan91. J. FAN, On the optimal rates of convergence for nonparametric deconvolution problems, *The Annals of Statistics* **19** (1991), 1257–1272.

FYC23. J. FAN, B. YUAN, and Y. CHEN, Improved dimension dependence of a proximal algorithm for sampling, in *Proceedings of Thirty Sixth Conference on Learning Theory* (G. NEU and L. ROSASCO, eds.), *Proceedings of Machine Learning Research* **195**, PMLR, 7 2023, pp. 1473–1521.

FGP20. M. FATHI, N. GOZLAN, and M. PROD’HOMME, A proof of the Caffarelli contraction theorem via entropic regularization, *Calculus of Variations and Partial Differential Equations* **59** (2020), 96.

FSV<sup>+</sup>19. J. FEYDY, T. SÉJOURNÉ, F.-X. VIALARD, S.-I. AMARI, A. TROUVE, and G. PEYRÉ, Interpolating between optimal transport and MMD using Sinkhorn divergences, in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (K. CHAUDHURI and M. SUGIYAMA, eds.), *Proceedings of Machine Learning Research* **89**, PMLR, 4 2019, pp. 2681–2690.

FG23. A. FIGALLI and F. GLAUDO, *An invitation to optimal transport, Wasserstein distances, and gradient flows*, second ed., *EMS Textbooks in Mathematics*, EMS Press, Berlin, 2023.

FCCR18. R. FLAMARY, M. CUTURI, N. COURTY, and A. RAKOTOMAMONJY, Wasserstein discriminant analysis, *Machine Learning* **107** (2018), 1923–1945.

FLF20. R. FLAMARY, K. LOUNICI, and A. FERRARI, Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation, 2020.

FG15. N. FOURNIER and A. GUILLIN, On the rate of convergence in Wasserstein distance of the empirical measure, *Probab. Theory Related Fields* **162** (2015), 707–738.

GFPO21. T. GALY-FAJOU, V. PERRONE, and M. OPPER, Flexible and efficient inference with particles for the variational Gaussian approximation, *Entropy* **23** (2021), Paper No. 990, 34.

GM96. W. GANGBO and R. J. McCANN, The geometry of optimal transportation, *Acta Math.* **177** (1996), 113–161.

vdG87. S. VAN DE GEER, A new approach to least-squares estimation, with applications, *Ann. Statist.* **15** (1987), 587–602.

vdG02. S. VAN DE GEER,  $M$ -estimation using penalties or sieves, *J. Statist. Plann. Inference* **108** (2002), 55–69, C. R. Rao 80th birthday felicitation volume, Part II.

GCB<sup>+</sup>19. A. GENEVAY, L. CHIZAT, F. BACH, M. CUTURI, and G. PEYRÉ, Sample complexity of Sinkhorn divergences, in *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2019, pp. 1574–1583.

GPC18. A. GENEVAY, G. PEYRÉ, and M. CUTURI, Learning generative models with Sinkhorn divergences, in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1608–1617.

GLRT20. I. GENTIL, C. LÉONARD, L. RIPANI, and L. TAMANINI, An entropic interpolation proof of the HWI inequality, *Stoch. Process. Appl.* **130** (2020), 907–923.

Ger24. P. R. GERBER, *Likelihood-free hypothesis testing and applications of the energy distance*, Ph.D. thesis, MIT, 2024.

GLPR23. B. GESHKOVSKI, C. LETROUIT, Y. POLYANSKIY, and P. RIGOLLET, The emergence of clusters in self-attention dynamics, in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

GLPR24. B. GESHKOVSKI, C. LETROUIT, Y. POLYANSKIY, and P. RIGOLLET, A mathematical perspective on transformers, 2024.

GP22. L. GHODRATI and V. M. PANARETOS, Distribution-on-distribution regression via optimal transport maps, *Biometrika* **109** (2022), 957–974.

GN22. P. GHOSAL and M. NUTZ, On the convergence rate of Sinkhorn’s algorithm, *arXiv preprint 2212.06000* (2022), 1–28.

GS22. P. GHOSAL and B. SEN, Multivariate ranks and quantiles using optimal transport: consistency, rates and nonparametric testing, *Ann. Statist.* **50** (2022), 1012–1037.

GLNZ22. S. GHOSH, Y. LU, T. NOWICKI, and E. ZHANG, On representations of mean-field variational inference, *arXiv preprint 2210.11385* (2022), 1–19.

GN16. E. GINÉ and R. NICKL, *Mathematical foundations of infinite-dimensional statistical models*, Cambridge: Cambridge University Press, 2016 (English).

GGNWP20. Z. GOLDFELD, K. GREENEWALD, J. NILES-WEED, and Y. POLYANSKIY, Convergence of smoothed empirical measures with applications to entropy estimation, *IEEE Trans. Inform. Theory* **66** (2020), 4368–4391.

GKRS24. Z. GOLDFELD, K. KATO, G. RIOUX, and R. SADHU, Statistical inference with regularized optimal transport, *Inf. Inference* **13** (2024), Paper No. 13, 68.

GDGSCN23. J. GONZÁLEZ-DELGADO, A. GONZÁLEZ-SANZ, J. CORTÉS, and P. NEUVIAL, Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology, *Electron. J. Stat.* **17** (2023), 1547–1586.

GSLNW24. A. GONZALEZ-SANZ, J.-M. LOUBES, and J. NILES-WEED, Weak limits of entropy regularized optimal transport; potentials, plans and divergences, *arXiv preprint 2207.07427* (2024), 1–24.

GLL<sup>+</sup>23. S. GOPI, Y. T. LEE, D. LIU, R. SHEN, and K. TIAN, Algorithmic aspects of the log-Laplace transform and a non-Euclidean proximal sampler, in *Proceedings of Thirty Sixth Conference on Learning Theory* (G. NEU and L. ROSASCO, eds.), *Proceedings of Machine Learning Research* **195**, PMLR, 7 2023, pp. 2399–2439.

GPC15. A. GRAMFORT, G. PEYRÉ, and M. CUTURI, Fast optimal transport averaging of neuroimaging data, in *Information Processing in Medical Imaging* (S. OURSELIN, D. C. ALEXANDER, C.-F. WESTIN, and M. J. CARDOSO, eds.), Springer International Publishing, Cham, 2015, pp. 261–272.

GNCD23. G. GRECO, M. NOBLE, G. CONFORTI, and A. DURMUS, Non-asymptotic convergence bounds for Sinkhorn iterates and their gradients: a coupling approach., in *Proceedings of Thirty Sixth Conference on Learning Theory* (G. NEU and L. ROSASCO, eds.), *Proceedings of Machine Learning Research* **195**, PMLR, 7 2023, pp. 716–746.

GH23. M. GROPPE and S. HUNDRIESER, Lower complexity adaptation for empirical entropic optimal transport, *arXiv preprint 2306.13580* (2023), 1–51.

Hal17. M. HALLIN, *On distribution and quantile functions, ranks and signs in  $\mathbb{R}^d$* , Working Papers ECARES ECARES 2017-34, ULB – Universite Libre de Bruxelles, 9 2017.

Hal22. M. HALLIN, Measure transportation and statistical decision theory, *Annu. Rev. Stat. Appl.* **9** (2022), 401–424.

HdBCAM21. M. HALLIN, E. DEL BARRIO, J. CUESTA-ALBERTOS, and C. MATRÁN, Distribution and quantile functions, ranks and signs in dimension  $d$ : a measure transportation approach, *Ann. Statist.* **49** (2021), 1139–1165.

HMS21. M. HALLIN, G. MORDANT, and J. SEGERS, Multivariate goodness-of-fit tests based on Wasserstein distance, *Electron. J. Stat.* **15** (2021), 1328–1371.

HMJG21. A. HAN, B. MISHRA, P. K. JAWANPURIA, and J. GAO, On Riemannian optimization over positive definite matrices with the Bures–Wasserstein geometry, in *Advances in Neural Information Processing Systems* (M. RANZATO, A. BEYGEZIMER, Y. DAUPHIN, P. LIANG, and J. W. VAUGHAN, eds.), **34**, Curran Associates, Inc., 2021, pp. 8940–8953.

vH14. R. VAN HANDEL, Probability in high dimension, Lecture Notes (Princeton University), 2014.

HMHBE24. Y. HE, A. MOUSAVI-HOSSEINI, K. BALASUBRAMANIAN, and M. A. ERDOGDU, A separation in heavy-tailed sampling: Gaussian vs. stable oracles for proximal samplers, *arXiv preprint 2405.16736* (2024), 1–33.

Hoe48. W. HOEFFDING, A class of statistics with asymptotically normal distribution, *Ann. Math. Statist.* **19** (1948), 293–325.

HSM24. S. HUNDRIESER, T. STAUDT, and A. MUNK, Empirical optimal transport between different measures adapts to lower complexity, *Ann. Inst. Henri Poincaré Probab. Stat.* **60** (2024), 824–846.

HR21. J.-C. HÜTTER and P. RIGOLLET, Minimax estimation of smooth optimal transport maps, *Ann. Statist.* **49** (2021), 1166–1194.

JGH18. A. JACOT, F. GABRIEL, and C. HONGLER, Neural tangent kernel: convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems* (S. BENGIO, H. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI, and R. GARNETT, eds.), **31**, Curran Associates, Inc., 2018.

JMPC20. H. JANATI, B. MUZELLEC, G. PEYRÉ, and M. CUTURI, Entropic optimal transport between unbalanced Gaussian measures has a closed form, in *Advances in Neural Information Processing Systems* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN, and H. LIN, eds.), **33**, Curran Associates, Inc., 2020, pp. 10468–10479.

JCP24. Y. JIANG, S. CHEWI, and A.-A. POOLADIAN, Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space, in *Proceedings of Thirty Seventh Conference on Learning Theory* (S. AGRAWAL and A. ROTH, eds.), *Proceedings of Machine Learning Research* **247**, PMLR, 7 2024, pp. 2720–2721.

JGJS99. M. I. JORDAN, Z. GHAHRAMANI, T. S. JAAKKOLA, and L. K. SAUL, An introduction to variational methods for graphical models, *Mach. Learn.* **37** (1999), 183–233.

JKO98. R. JORDAN, D. KINDERLEHRER, and F. OTTO, The variational formulation of the Fokker–Planck equation, *SIAM J. Math. Anal.* **29** (1998), 1–17.

KLRS08. B. KALANTARI, I. LARI, F. RICCA, and B. SIMEONE, On the complexity of general matrix scaling and entropy minimization via the RAS algorithm, *Math. Program.* **112** (2008), 371–401.

Kan42. L. V. KANTOROVICH, On the translocation of masses, *Dokl. Akad. Nauk SSSR* **37** (1942), 227–229.

KNS16. H. KARIMI, J. NUTINI, and M. SCHMIDT, Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2016, pp. 795–811.

KR24. A. KATSEVICH and P. RIGOLLET, On the approximation accuracy of Gaussian variational inference, *Ann. Statist. (to appear)* (2024), 49.

KOW<sup>+</sup>23. K. KIM, J. OH, K. WU, Y. MA, and J. GARDNER, On the convergence of black-box variational inference, in *Advances in Neural Information Processing Systems* (A. OH, T. NEUMANN, A. GLOBERSON, K. SAENKO, M. HARDT, and S. LEVINE, eds.), **36**, Curran Associates, Inc., 2023, pp. 44615–44657.

KTM20. M. KLATT, C. TAMELING, and A. MUNK, Empirical regularized optimal transport: statistical theory and applications, *SIAM Journal on Mathematics of Data Science* **2** (2020), 419–443.

KS94. M. KNOTT and C. S. SMITH, On a generalization of cyclic monotonicity and distances among random vectors, *Linear Algebra Appl.* **199** (1994), 363–371.

Kol34. A. N. KOLMOGOROV, Zufällige Bewegungen (zur Theorie der Brownschen Bewegung), *Ann. of Math. (2)* **35** (1934), 116–117.

KNS<sup>+</sup>19. S. KOLOURI, K. NADJAH, U. SIMSEKLI, R. BADEAU, and G. K. ROHDE, Generalized sliced Wasserstein distances, in *Advances in Neural Information Processing Systems 32* (H. M. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D’ALCHÉ-BUC, E. B. FOX, and R. GARNETT, eds.), 2019, pp. 261–272.

KR15. S. KOLOURI and G. K. ROHDE, Transport-based single frame super resolution of very low resolution face images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4876–4884.

KTOR16. S. KOLOURI, A. B. TOSUN, J. A. OZOLEK, and G. K. ROHDE, A continuous linear optimal transport approach for pattern analysis in image datasets, *Pattern Recognition* **51** (2016), 453–462.

KMV16. S. KONDRATYEV, L. MONSAINGEON, and D. VOROTNIKOV, A new optimal transport distance on the space of finite Radon measures, *Advances in Differential Equations* **21** (2016), 1117 – 1164.

KVZ24. Y. KOOK, S. S. VEMPALA, and M. S. ZHANG, In-and-out: algorithmic diffusion for sampling convex bodies, *arXiv preprint 2405.01425* (2024), 1–32.

KZC<sup>+</sup>24. Y. KOOK, M. S. ZHANG, S. CHEWI, M. A. ERDOGDU, and M. B. LI, Sampling from the mean-field stationary distribution, in *Proceedings of Thirty Seventh Conference on Learning Theory* (S. AGRAWAL and A. ROTH, eds.), *Proceedings of Machine Learning Research* **247**, PMLR, 7 2024, pp. 3099–3136.

KSA<sup>+</sup>20. A. KORBA, A. SALIM, M. ARBEL, G. LUISE, and A. GRETTON, A non-asymptotic analysis for Stein variational gradient descent, in *Advances in Neural Information Processing Systems* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN, and H. LIN, eds.), **33**, Curran Associates, Inc., 2020, pp. 4672–4682.

Kro18. A. KROSHNIN, Fréchet barycenters in the Monge–Kantorovich spaces, *J. Convex Anal.* **25** (2018), 1371–1395.

KSS21. A. KROSHNIN, V. SPOKOINY, and A. SUVORIKOVA, Statistical inference for Bures–Wasserstein barycenters, *Ann. Appl. Probab.* **31** (2021), 1264–1298.

KDLY22. S. KUM, M. H. DUONG, Y. LIM, and S. YUN, A GPM-based algorithm for solving regularized Wasserstein barycenter problems in some spaces of probability measures, *Journal of Computational and Applied Mathematics* **416** (2022), 114588.

Lac23. D. LACKER, Independent projections of diffusions: gradient flows for variational inference and optimal mean field approximations, *arXiv preprint 2309.13332* (2023), 1–27.

Laf88. J. D. LAFFERTY, The density manifold and configuration space quantization, *Trans. Amer. Math. Soc.* **305** (1988), 699–741.

LBB24. M. LAMBERT, F. BACH, and S. BONNABEL, Variational dynamic programming for stochastic optimal control, *arXiv preprint 2404.14806* (2024), 1–17.

LBB23. M. LAMBERT, S. BONNABEL, and F. BACH, Variational Gaussian approximation of the Kushner optimal filter, in *Geometric science of information. Part I, Lecture Notes in Comput. Sci.* **14071**, Springer, Cham, [2023] ©2023, pp. 395–404.

LCB<sup>+</sup>22. M. LAMBERT, S. CHEWI, F. BACH, S. BONNABEL, and P. RIGOLLET, Variational inference via Wasserstein gradient flows, in *Advances in Neural Information Processing Systems* (A. H. OH, A. AGARWAL, D. BELGRAVE, and K. CHO, eds.), 2022.

LZ24. H. LAVENANT and G. ZANELLA, Convergence rate of random scan coordinate ascent variational inference under log-concavity, *arXiv preprint 2406.07292* (2024), 1–12.

Le 16. J.-F. LE GALL, *Brownian motion, martingales, and stochastic calculus*, French ed., *Graduate Texts in Mathematics* **274**, Springer, [Cham], 2016.

LGL17. T. LE GOUIC and J.-M. LOUBES, Existence and consistency of Wasserstein barycenters, *Probab. Theory Related Fields* **168** (2017), 901–917.

LGLR20. T. LE GOUIC, J.-M. LOUBES, and P. RIGOLLET, Projection to fairness in statistical learning, *arXiv preprint 2005.11720* (2020), 1–14.

LGPRS22. T. LE GOUIC, Q. PARIS, P. RIGOLLET, and A. J. STROMME, Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space, *J. Eur. Math. Soc.* (2022), 2229–2250.

Led17. M. LEDOUX, On optimal matching of Gaussian samples, *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **457** (2017), 226–264.

LT91. M. LEDOUX and M. TALAGRAND, *Probability in Banach spaces, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]* **23**, Springer-Verlag, Berlin, 1991, Isoperimetry and processes.

LZ21. M. LEDOUX and J.-X. ZHU, On optimal matching of Gaussian samples III, *Probab. Math. Statist.* **41** (2021), 237–265.

LST21. Y. T. LEE, R. SHEN, and K. TIAN, Structured logconcave sampling with a restricted Gaussian oracle, in *Proceedings of Thirty Fourth Conference on Learning Theory* (M. BELKIN and S. KPOTUFE, eds.), *Proceedings of Machine Learning Research* **134**, PMLR, 8 2021, pp. 2993–3050.

Lég21. F. LÉGER, A gradient descent perspective on Sinkhorn, *Appl. Math. Optim.* **84** (2021), 1843–1855.

LR05. E. L. LEHMANN and J. P. ROMANO, *Testing statistical hypotheses*, third ed., *Springer Texts in Statistics*, Springer, New York, 2005.

Léo14. C. LÉONARD, A survey of the Schrödinger problem and some of its connections with optimal transport, *Discrete Contin. Dyn. Syst.* **34** (2014), 1533–1574.

LGYS20. L. LI, A. GENEVAY, M. YUROCHKIN, and J. M. SOLOMON, Continuous regularized Wasserstein barycenters, in *Advances in Neural Information Processing Systems* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN, and H. LIN, eds.), **33**, Curran Associates, Inc., 2020, pp. 17755–17765.

Lia21. T. LIANG, How well generative adversarial networks learn distributions, *J. Mach. Learn. Res.* **22** (2021), Paper No. 228, 41.

LMS16. M. LIERO, A. MIELKE, and G. SAVARÉ, Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves, *SIAM J. Math. Anal.* **48** (2016), 2869–2911.

LMS18. M. LIERO, A. MIELKE, and G. SAVARÉ, Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures, *Invent. Math.* **211** (2018), 969–1117.

Lin83. B. G. LINDSAY, The geometry of mixture likelihoods: a general theory, *The Annals of Statistics* **11** (1983), 86–94.

Liu17. Q. LIU, Stein variational gradient descent as gradient flow, in *Advances in Neural Information Processing Systems* (I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN, and R. GARNETT, eds.), **30**, Curran Associates, Inc., 2017.

LW16. Q. LIU and D. WANG, Stein variational gradient descent: a general purpose Bayesian inference algorithm, in *Advances in Neural Information Processing Systems* (D. LEE, M. SUGIYAMA, U. LUXBURG, I. GUYON, and R. GARNETT, eds.), **29**, Curran Associates, Inc., 2016.

LV09. J. LOTT and C. VILLANI, Ricci curvature for metric-measure spaces via optimal transport, *Ann. of Math.* **169** (2009), 903–991.

LLN19a. J. LU, Y. LU, and J. NOLEN, Scaling limit of the Stein variational gradient descent: the mean field regime, *SIAM J. Math. Anal.* **51** (2019), 648–671.

LLN19b. Y. LU, J. LU, and J. NOLEN, Accelerating Langevin sampling with birth-death, arXiv preprint 1905.09863, 2019.

LSW23. Y. LU, D. SLEPČEV, and L. WANG, Birth-death dynamics for sampling: global convergence, approximations and their asymptotics, *Nonlinearity* **36** (2023), 5731–5772.

LGT22. Y. LUO and N. GARCÍA TRILLOS, Nonconvex matrix factorization is geodesically convex: global landscape analysis for fixed-rank matrix optimization from a Riemannian perspective, *arXiv preprint 2209.15130* (2022), 1–35.

MGM22. A. MALLASTO, A. GEROLIN, and H. Q. MINH, Entropy-regularized 2-Wasserstein distance between Gaussian measures, *Inf. Geom.* **5** (2022), 289–323.

MBNWW21. T. MANOLE, S. BALAKRISHNAN, J. NILES-WEED, and L. WASSERMAN, Plugin estimation of smooth optimal transport maps, 2021.

MBNWW23. T. MANOLE, S. BALAKRISHNAN, J. NILES-WEED, and L. WASSERMAN, Central limit theorems for smooth optimal transport maps, *arXiv preprint 2312.12407* (2023), 1–60.

MBW22. T. MANOLE, S. BALAKRISHNAN, and L. WASSERMAN, Minimax confidence intervals for the sliced Wasserstein distance, *Electron. J. Stat.* **16** (2022), 2252–2345.

MNW24. T. MANOLE and J. NILES-WEED, Sharp convergence rates for empirical optimal transport with smooth costs, *The Annals of Applied Probability* **34** (2024), 1108–1135.

MLGR23. T. MAUNU, T. LE GOUIC, and P. RIGOLLET, Bures–Wasserstein barycenters and low-rank matrix recovery, in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics* (F. RUIZ, J. DY, and J.-W. VAN DE MEENT, eds.), *Proceedings of Machine Learning Research* **206**, PMLR, 4 2023, pp. 8183–8210.

McC97. R. J. MCCANN, A convexity principle for interacting gases, *Advances in Mathematics* **128** (1997), 153–179.

McK66. H. P. MCKEAN, JR., A class of Markov processes associated with nonlinear parabolic equations, *Proc. Nat. Acad. Sci. U.S.A.* **56** (1966), 1907–1911.

MMN18. S. MEI, A. MONTANARI, and P.-M. NGUYEN, A mean field view of the landscape of two-layer neural networks, *Proc. Natl. Acad. Sci. USA* **115** (2018), E7665–E7671.

MNW19. G. MENA and J. NILES-WEED, Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem, in *Advances in Neural Information Processing Systems* (H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. d'ALCHÉ-BUC, E. FOX, and R. GARNETT, eds.), **32**, Curran Associates, Inc., 2019.

MM23. T. MISIAKIEWICZ and A. MONTANARI, Six lectures on linearized neural networks, 2023.

Mod17. K. MODIN, Geometry of matrix decompositions seen through optimal transport and information geometry, *J. Geom. Mech.* **9** (2017), 335–390.

Mon81. G. MONGE, Mémoire sur la théorie des déblais et des remblais, *Mém. de l'Ac. R. des Sc.* (1781), 666–704.

MFSS17. K. MUANDET, K. FUKUMIZU, B. SRIPERUMBUDUR, and B. SCHÖLKOPF, Kernel mean embedding of distributions: a review and beyond, *Foundations and Trends® in Machine Learning* **10** (2017), 1–141.

NDC<sup>+</sup>20. K. NADJAHI, A. DURMUS, L. CHIZAT, S. KOLOURI, S. SHAHROAMPUR, and U. SIMSEKLI, Statistical and topological properties of sliced probability divergences, in *Advances in Neural Information Processing Systems 33* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN, and H. LIN, eds.), 2020.

NY83. A. S. NEMIROVSKY and D. B. YUDIN, *Problem complexity and method efficiency in optimization, A Wiley-Interscience Publication*, John Wiley & Sons Inc., New York, 1983, Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Nes83. Y. E. NESTEROV, A method for solving the convex programming problem with convergence rate  $O(1/k^2)$ , *Dokl. Akad. Nauk SSSR* **269** (1983), 543–547.

Nes18. Y. NESTEROV, *Lectures on convex optimization, Springer Optimization and Its Applications* **137**, Springer, Cham, 2018.

NWKB23. M. NEYKOV, L. WASSERMAN, I. KIM, and S. BALAKRISHNAN, Nearly minimax optimal Wasserstein conditional independence testing, *arXiv preprint 2308.08672* (2023), 1–24.

NW18. J. NILES-WEED, Sharper rates for estimating differential entropy under Gaussian convolutions, *Massachusetts Institute of Technology (MIT), Tech. Rep* (2018), 1–2.

NWB19. J. NILES-WEED and F. BACH, Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance, *Bernoulli* **25** (2019), 2620–2648.

NWB22. J. NILES-WEED and Q. BERTHET, Minimax estimation of smooth densities in Wasserstein distance, *Ann. Statist.* **50** (2022), 1519–1540.

NWR22. J. NILES-WEED and P. RIGOLLET, Estimation of Wasserstein distances in the spiked transport model, *Bernoulli* **28** (2022), 2663–2688.

NW22. M. NUTZ and J. WIESEL, Entropic optimal transport: convergence of potentials, *Probab. Theory Related Fields* **184** (2022), 401–424.

OT11. S.-I. OHTA and A. TAKATSU, Displacement convexity of generalized relative entropies, *Adv. Math.* **228** (2011), 1742–1787.

OT13. S.-I. OHTA and A. TAKATSU, Displacement convexity of generalized relative entropies. II, *Comm. Anal. Geom.* **21** (2013), 687–785.

OI22. R. OKANO and M. IMAIZUMI, Inference for projection-based Wasserstein distances on finite spaces, *arXiv preprint 2202.05495* (2022), 1–29.

Oll10. Y. OLLIVIER, A survey of Ricci curvature for metric spaces and Markov chains, in *Probabilistic approach to geometry*, *Adv. Stud. Pure Math.* **57**, Math. Soc. Japan, Tokyo, 2010, pp. 343–381.

Oll13. Y. OLLIVIER, A visual introduction to Riemannian curvatures and some discrete generalizations, in *Analysis and geometry of metric measure spaces*, *CRM Proc. Lecture Notes* **56**, Amer. Math. Soc., Providence, RI, 2013, pp. 197–220.

OV12. Y. OLLIVIER and C. VILLANI, A curved Brunn–Minkowski inequality on the discrete hypercube, or: what is the Ricci curvature of the discrete hypercube?, *SIAM J. Discrete Math.* **26** (2012), 983–996.

vO22. J. VAN OOSTRUM, Bures–Wasserstein geometry for positive-definite Hermitian matrices and their trace-one subset, *Inf. Geom.* **5** (2022), 405–425.

Ott01. F. OTTO, The geometry of dissipative evolution equations: the porous medium equation., *Communications in Partial Differential Equations* **26** (2001), 101–174.

OV00. F. OTTO and C. VILLANI, Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality, *J. Funct. Anal.* **173** (2000), 361–400.

OV01. F. OTTO and C. VILLANI, 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** (2001), 697–700.

PZ16. V. M. PANARETOS and Y. ZEMEL, Amplitude and phase variation of point processes, *Ann. Statist.* **44** (2016), 771–812.

PZ20. V. M. PANARETOS and Y. ZEMEL, *An invitation to statistics in Wasserstein space*, Springer Nature, 2020.

PS23. S. PARK and D. SLEPČEV, Geometry and analytic properties of the sliced Wasserstein space, *arXiv preprint 2311.05134* (2023), 1–49.

PT18. S. PARK and M. THORPE, Representing and learning high dimensional data with the optimal transport map from a probabilistic viewpoint, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7864–7872.

PC19a. F. PATY and M. CUTURI, Subspace robust Wasserstein distances, in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019* (K. CHAUDHURI and R. SALAKHUTDINOV, eds.), *Proceedings of Machine Learning Research* **97**, PMLR, 2019, pp. 5072–5081.

PC19b. G. PEYRÉ and M. CUTURI, Computational optimal transport, *Foundations and Trends® in Machine Learning* **11** (2019), 355–607.

PKD07. F. PITIÉ, A. C. KOKARAM, and R. DAHYOT, Automated colour grading using colour distribution transfer, *Computer Vision and*

Image Understanding **107** (2007), 123–137, Special issue on color image processing.

Pol82. D. POLLARD, Quantization and the method of  $k$ -means, *IEEE Transactions on Information theory* **28** (1982), 199–205.

PW24. Y. POLYANSKIY and Y. WU, *Information theory: from coding to learning*, Cambridge University Press, Cambridge, 2024.

Pon23. D. PONNOPRAT, Universal consistency of Wasserstein  $k$ -NN classifier: a negative and some positive results, *Information and Inference: A Journal of the IMA* **12** (2023), 1997–2019.

PCNW22. A.-A. POOLADIAN, M. CUTURI, and J. NILES-WEED, Debiaseder beware: pitfalls of centering regularized transport maps, in *Proceedings of the 39th International Conference on Machine Learning* (K. CHAUDHURI, S. JEGELKA, L. SONG, C. SZEPESVARI, G. NIU, and S. SABATO, eds.), *Proceedings of Machine Learning Research* **162**, PMLR, 7 2022, pp. 17830–17847.

PDNW23. A.-A. POOLADIAN, V. DIVOL, and J. NILES-WEED, Minimax estimation of discontinuous optimal transport maps: the semi-discrete case, in *Proceedings of the 40th International Conference on Machine Learning* (A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO, and J. SCARLETT, eds.), *Proceedings of Machine Learning Research* **202**, PMLR, 7 2023, pp. 28128–28150.

PNW22. A.-A. POOLADIAN and J. NILES-WEED, Entropic estimation of optimal transport maps, *arXiv preprint 2109.12004* (2022), 1–37.

PBS24. V. PRISER, P. BIANCHI, and A. SALIM, Long-time asymptotics of noisy SVGD outside the population limit, *arXiv preprint 2406.11929* (2024), 1–28.

RPDB12. J. RABIN, G. PEYRÉ, J. DELON, and M. BERNOT, Wasserstein barycenter and its application to texture mixing, in *Scale Space and Variational Methods in Computer Vision* (A. M. BRUCKSTEIN, B. M. TER HAAR ROMENY, A. M. BRONSTEIN, and M. M. BRONSTEIN, eds.), Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 435–446.

RR98a. S. T. RACHEV and L. RÜSCHENDORF, *Mass transportation problems. Vol. I, Probability and its Applications* (New York), Springer-Verlag, New York, 1998, Theory.

RR98b. S. T. RACHEV and L. RÜSCHENDORF, *Mass transportation problems. Vol. II, Probability and its Applications* (New York), Springer-Verlag, New York, 1998, Applications.

RH17. P. RIGOLLET and J.-C. HÜTTER, High-dimensional statistics, Lecture notes, 2017.

RNW18. P. RIGOLLET and J. NILES-WEED, Entropic optimal transport is maximum-likelihood deconvolution, *Comptes Rendus Mathematique* **356** (2018), 1228–1235.

RNW19. P. RIGOLLET and J. NILES-WEED, Uncoupled isotonic regression via minimum Wasserstein deconvolution, *Inf. Inference* **8** (2019), 691–717.

RS22. P. RIGOLLET and A. J. STROMME, On the sample complexity of entropic optimal transport, *arXiv preprint 2206.13472* (2022), 1–28.

Roc66. R. T. ROCKAFELLAR, Characterization of the subdifferentials of convex functions, *Pacific J. Math.* **17** (1966), 497–510.

Roc97. R. T. ROCKAFELLAR, *Convex analysis*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1997, Reprint of the 1970 original, Princeton Paperbacks.

RVE22. G. ROTSKOFF and E. VANDEN-EIJNDEN, Trainability and accuracy of artificial neural networks: an interacting particle system approach, *Communications on Pure and Applied Mathematics* **75** (2022), 1889–1935.

RTG00. Y. RUBNER, C. TOMASI, and L. J. GUIBAS, The earth mover’s distance as a metric for image retrieval, *Int. J. Comput. Vision* **40** (2000), 99–121.

RU02. L. RÜSCENDORF and L. UCKELMANN, On the  $n$ -coupling problem, *J. Multivariate Anal.* **81** (2002), 242–258.

SKL20. A. SALIM, A. KORBA, and G. LUISE, The Wasserstein proximal gradient algorithm, in *Advances in Neural Information Processing Systems* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN, and H. LIN, eds.), **33**, Curran Associates, Inc., 2020, pp. 12356–12366.

SSR22. A. SALIM, L. SUN, and P. RICHTARIK, A convergence theory for SVGD in the population limit under Talagrand’s inequality T1, in *Proceedings of the 39th International Conference on Machine Learning* (K. CHAUDHURI, S. JEGELKA, L. SONG, C. SZEPESVARI, G. NIU, and S. SABATO, eds.), *Proceedings of Machine Learning Research* **162**, PMLR, 7 2022, pp. 19139–19152.

SABP22. M. E. SANDER, P. ABLIN, M. BLONDEL, and G. PEYRÉ, Sinkformers: transformers with doubly stochastic attention, in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics* (G. CAMPS-VALLS, F. J. R. RUIZ, and I. VALERA, eds.), *Proceedings of Machine Learning Research* **151**, PMLR, 3 2022, pp. 3515–3530.

San15. F. SANTAMBROGIO, *Optimal transport for applied mathematicians*, *Progress in Nonlinear Differential Equations and their Applications* **87**, Birkhäuser/Springer, Cham, 2015, Calculus of variations, PDEs, and modeling.

San17. F. SANTAMBROGIO, {Euclidean, metric, and Wasserstein} gradient flows: an overview, *Bull. Math. Sci.* **7** (2017), 87–154.

SST<sup>+</sup>19. G. SCHIEBINGER, J. SHU, M. TABAKA, B. CLEARY, V. SUBRAMANIAN, A. SOLOMON, J. GOULD, S. LIU, S. LIN, P. BERUBE, L. LEE, J. CHEN, J. BRUMBAUGH, P. RIGOLLET, K. HOCHEDLINGER, R. JAENISCH, A. REGEV, and E. S. LANDER, Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming, *Cell* **176** (2019), 928–943.

SC15. V. SEGUY and M. CUTURI, Principal geodesic analysis for probability measures under the optimal transport metric, in *Advances in Neural Information Processing Systems 28*, 2015, pp. 3312–3320.

SM23. J. SHI and L. MACKEY, A finite-particle convergence rate for Stein variational gradient descent, in *Advances in Neural Information Processing Systems* (A. OH, T. NAUMANN, A. GLOBERSON, K. SAENKO, M. HARDT, and S. LEVINE, eds.), **36**, Curran Associates, Inc., 2023, pp. 26831–26844.

SP18. S. SINGH and B. PÓCZOS, Minimax distribution estimation in Wasserstein distance, *arXiv preprint 1802.08855* (2018), 1–34.

SUL<sup>+</sup>18. S. SINGH, A. UPPAL, B. LI, C.-L. LI, M. ZAHEER, and B. POCZOS, Nonparametric density estimation under adversarial losses, in *Advances in Neural Information Processing Systems* (S. BENGIO, H. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI, and R. GARNETT, eds.), **31**, Curran Associates, Inc., 2018.

SS20. J. SIRIGNANO and K. SPILIOPOULOS, Mean field analysis of neural networks: a law of large numbers, *SIAM Journal on Applied Mathematics* **80** (2020), 725–752.

SdGP<sup>+</sup>15. J. SOLOMON, F. DE GOES, G. PEYRÉ, M. CUTURI, A. BUTSCHER, A. NGUYEN, T. DU, and L. GUIBAS, Convolutional Wasserstein distances: efficient optimal transportation on geometric domains, *ACM Trans. Graph.* **34** (2015), 66:1–66:11.

SLD18. S. SRIVASTAVA, C. LI, and D. B. DUNSON, Scalable Bayes via barycenter in Wasserstein space, *Journal of Machine Learning Research* **19** (2018), 1–35.

Ste01. J. M. STEELE, *Stochastic calculus and financial applications, Applications of Mathematics (New York)* **45**, Springer-Verlag, New York, 2001.

Str23. A. J. STROMME, Minimum intrinsic dimension scaling for entropic optimal transport, *arXiv preprint 2306.03398* (2023), 1–53.

Stu03. K.-T. STURM, Probability measures on metric spaces of nonpositive curvature, in *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*, *Contemp. Math.* **338**, Amer. Math. Soc., Providence, RI, 2003, pp. 357–390.

Stu06a. K.-T. STURM, On the geometry of metric measure spaces. I, *Acta Math.* **196** (2006), 65–131.

Stu06b. K.-T. STURM, On the geometry of metric measure spaces. II, *Acta Math.* **196** (2006), 133–177.

SNW23. T. SUZUKI, A. NITANDA, and D. WU, Uniform-in-time propagation of chaos for the mean-field gradient Langevin dynamics, in *The Eleventh International Conference on Learning Representations*, 2023.

Szn91. A.-S. SZNITMAN, Topics in propagation of chaos, in *École d’Été de Probabilités de Saint-Flour XIX—1989, Lecture Notes in Math.* **1464**, Springer, Berlin, 1991, pp. 165–251.

Tal96. M. TALAGRAND, Majorizing measures: the generic chaining, *Ann. Probab.* **24** (1996), 1049–1103.

Tal21. M. TALAGRAND, *Upper and lower bounds for stochastic processes—decomposition theorems*, second ed., *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]* **60**, Springer, Cham, [2021] ©2021.

Tan17. Y. S. TAN, Energy optimization for distributions on the sphere and improvement to the Welch bounds, *Electronic Communications in Probability* **22** (2017), 1–12.

Tan23. K. TANAKA, Accelerated gradient descent method for functionals of probability measures by new convexity and smoothness based on transport maps, *arXiv preprint 2305.05127* (2023), 1–31.

Tsy09. A. B. TSYBAKOV, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, 2009, Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

TR24. B. TZEN and M. RAGINSKY, Function approximation by neural nets in the mean-field regime: entropic regularization and controlled McKean–Vlasov dynamics, *arXiv preprint 2002.01987* (2024), 1–31.

USP19. A. UPPAL, S. SINGH, and B. PÓCZOS, Nonparametric density estimation & convergence rates for GANs under Besov IPM losses, in *Advances in Neural Information Processing Systems 32* (H. M. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. d’ALCHÉ-BUC, E. B. FOX, and R. GARNETT, eds.), 2019, pp. 9086–9097.

vdVW23. A. W. VAN DER VAART and J. A. WELLNER, *Weak convergence and empirical processes—with applications to statistics*, 2 ed., Springer Series in Statistics, Springer, Cham, 2023.

vdV98. A. W. VAN DER VAART, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics **3**, Cambridge University Press, Cambridge, 1998.

VMB<sup>+</sup>24. A. VACHER, B. MUZELLEC, F. BACH, F.-X. VIALARD, and A. RUDI, Optimal estimation of smooth transport maps with kernel SoS, *SIAM J. Math. Data Sci.* **6** (2024), 311–342.

VC23. T. VASKEVICIUS and L. CHIZAT, Computational guarantees for doubly entropic Wasserstein barycenters, in *Advances in Neural Information Processing Systems* (A. OH, T. NAUMANN, A. GLOBERSON, K. SAENKO, M. HARDT, and S. LEVINE, eds.), **36**, Curran Associates, Inc., 2023, pp. 12363–12388.

VSP<sup>+</sup>17. A. VASWANI, N. SHAZER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. U. KAISER, and I. POLOSKHIN, Attention is all you need, in *Advances in Neural Information Processing Systems* (I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN, and R. GARNETT, eds.), **30**, Curran Associates, Inc., 2017.

Ver18. R. VERSHYNIN, *High-dimensional probability*, Cambridge Series in Statistical and Probabilistic Mathematics **47**, Cambridge University Press, Cambridge, 2018, An introduction with applications in data science, With a foreword by Sara van de Geer.

Vil03. C. VILLANI, *Topics in optimal transportation*, Graduate Studies in Mathematics **58**, American Mathematical Society, Providence, RI, 2003.

Vil09a. C. VILLANI, Hypocoercivity, *Mem. Amer. Math. Soc.* **202** (2009), iv+141.

Vil09b. C. VILLANI, *Optimal transport*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] **338**, Springer-Verlag, Berlin, 2009, Old and new.

Wai19. M. J. WAINWRIGHT, *High-dimensional statistics: a non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.

WJ08. M. J. WAINWRIGHT and M. I. JORDAN, Graphical models, exponential families, and variational inference, *Foundations and Trends in Machine Learning* **1** (2008), 1–305.

WSB<sup>+</sup>13. W. WANG, D. SLEPČEV, S. BASU, J. A. OZOLEK, and G. K. ROHDE, A linear optimal transportation framework for quantifying and visualizing variations in sets of images, *Int. J. Comput. Vis.* **101** (2013), 254–269.

WL20. Y. WANG and W. LI, Information Newton’s flow: second-order optimization method in probability space, *arXiv preprint 2001.04341* (2020), 1–62.

WL22. Y. WANG and W. LI, Accelerated information gradient flow, *J. Sci. Comput.* **90** (2022), Paper No. 11, 47.

Wib18. A. WIBISONO, Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem, in *Conference on Learning Theory* (S. BUBECK, V. PERCHET, and P. RIGOLLET, eds.), *Proceedings of Machine Learning Research* **75**, PMLR, 2018, pp. 2093–3027.

XNW22. J. XI and J. NILES-WEED, Distributional convergence of the sliced Wasserstein process, in *Advances in Neural Information Processing Systems* (S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO, and A. OH, eds.), **35**, Curran Associates, Inc., 2022, pp. 13961–13973.

XH22. X. XU and Z. HUANG, Central limit theorem for the sliced 1-Wasserstein distance and the max-sliced 1-Wasserstein distance, *arXiv preprint 2205.14624* (2022), 1–37.

YWR24. Y. YAN, K. WANG, and P. RIGOLLET, Learning Gaussian mixtures using the Wasserstein–Fisher–Rao gradient flow, 2024.

YY23. R. YAO and Y. YANG, Mean-field variational inference via Wasserstein gradient flow, *arXiv preprint 2207.08074* (2023), 1–120.

ZP19. Y. ZEMEL and V. M. PANARETOS, Fréchet means and Procrustes analysis in Wasserstein space, *Bernoulli* **25** (2019), 932–976.

ZPFP20. K. S. ZHANG, G. PEYRÉ, J. FADILI, and M. PEREYRA, Wasserstein control of mirror Langevin Monte Carlo, in *Proceedings of Thirty Third Conference on Learning Theory* (J. ABERNETHY and S. AGARWAL, eds.), *Proceedings of Machine Learning Research* **125**, PMLR, 7 2020, pp. 3814–3841.



---

## Index

Alexandrov curvature, 211  
Alexandrov space, 215  
Atjai–Komlós–Tusnády theorem, 58  
attention, 198  
barycenter, 230  
Benamou–Brenier formula, 144  
Birkhoff polytope, 14  
birth-death sampling, 190  
Brenier’s theorem, 23  
improved, 36  
Burer–Monteiro, 166  
Bures–Wasserstein, 149  
gradient, 150  
CAT(0), *see* NPC  
chaining, 55  
change of variables, 169  
continuity equation, 136  
convex duality, 248  
cyclical monotonicity, 25  
dyadic partitioning, 49  
earth mover’s distance, 63  
empirical process, 55  
empirical risk minimization (ERM), 233  
entropic optimal transport, 109  
Fenchel–Young inequality, 249  
first variation, 145  
Fisher–Rao, 155  
Fokker–Planck equation, 180  
fundamental theorem of optimal transport, 34  
geodesic, 208  
extendable, 237  
generalized, 227  
Geodesic convexity, 142  
geodesic space, 205  
Gibbs variational principle, 112  
gluing lemma, 257  
goodness-of-fit testing, 64  
Grönwall’s inequality, 148  
Hadamard, *see* NPC  
Hellinger–Kantorovich, *see* Wasserstein–Fisher–Rao  
Hoeffding’s decomposition, 120  
hugging, 233

- integral probability metric (IPM), 75
- interaction energy, 147
- internal energy, 147
- Kantorovich potentials, 36
- Kantorovich problem, 11
- Kantorovich–Rubinstein distance, *see* Wasserstein distance
- Langevin diffusion, 178
  - mean-field, 189
  - mirror Langevin, 185
  - Newton–Langevin, 186
  - underdamped Langevin, 186
- length, 205
- logarithmic map, 222
- logistic function, 195
- maximum mean discrepancy (MMD), 76
- McKean–Vlasov, 188
- mean-field neural networks, 195
- metric derivative, 139
- midpoint, 208
- Monge problem, 10
- neural network, 195
- neural tangent kernel (NTK), 203
- NNC, 215
- non-parametric maximum likelihood (NPMLE), 191
- NPC, 215
- path, 205
- perceptron, 198
- plug-in estimator, 63
- Poincaré inequality, 101
- Polyak–Łojasiewicz (PL) inequality, 148, 251
- portmanteau theorem, 256
- potential energy, 146
- Prokhorov’s theorem, 255
- proximal sampler, 184
- rectifiable path, 206
- Rockafellar’s theorem, 25
- semi-discrete optimal transport, 83, 130
- semidual, 32
- stability, 91
- Sinkhorn divergence, 130
- Sinkhorn’s algorithm, 117
- Sobolev space, 71
- Stein variational gradient descent (SVGD), 175
- subdifferential, 248
- tangent cone, 218, 221
- transformers, 197
- unbalanced optimal transport, 155
- variance equality, 234
- variational inference (VI), 167
  - Gaussian, 171
  - mean-field, 172
- Wasserstein distance, 15, 16
  - sliced, 81
  - smoothed, 79
- Wasserstein gradient, 146
- Wasserstein gradient flow, 147
- Wasserstein–Fisher–Rao (WFR), 157, 190
- weak convergence, 255