

Interpolating Spline Curves of Measures

by

Julien Clancy

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mathematics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

© Massachusetts Institute of Technology 2021. All rights reserved.

Author
Department of Mathematics
April 27, 2021

Certified by.....
Philippe Rigollet
Professor, Mathematics
Thesis Supervisor

Accepted by
Jonathan Kelner
Graduate Chair, Applied Mathematics

Interpolating Spline Curves of Measures

by

Julien Clancy

Submitted to the Department of Mathematics
on April 27, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mathematics

Abstract

When dealing with classical, Euclidean data, the statistician’s toolkit is enviably deep: from linear and nonlinear regression, to dealing with sparse or structured data, to interpolation techniques, most any problem dealing with vector or matrix data is amenable to several different statistical methods. Yet modern data is often not Euclidean in nature. The semantic content of natural images does not have a vector structure; shifting an image one pixel to the right does not perceptibly change it, but its vector representation is very different. For model cross-validation or bootstrapping, each data point is a dataset in its own right, and one might want to consider an “average dataset”. In an ensemble method, experts may express their beliefs as prior distributions; how would we perform a statistical analysis of these?

Recently much attention has been paid to a framework which subsumes all of these problems: the Wasserstein space of measures with finite second moment. Works on point estimation [61], generalized means [19, 36, 5], and linear regression [11, 32] have appeared, as have some on smooth interpolation [7, 15], greatly expanding the statistical toolkit for modern data. In this vein, the present work is broadly a theoretical and computational exploration of curves of measures which in some sense minimize curvature while interpolating data, as splines do in Euclidean space. We answer several questions about the relationship between the intrinsic Wasserstein-Riemannian curvature of such curves and a particle flow-based, “fluid-dynamical” formulation, and provide fast and accurate smooth interpolation techniques. We also study a related probabilistic interpolation problem unique to the measure setting, which asks for particle trajectories that satisfy certain interpolation constraints and minimize a KL divergence, in analogy with the Schrödinger bridge problem [38, 56, 55]. We conclude with an extension of our methods to the case of unbalanced measures in the Wasserstein-Fisher-Rao space.

Thesis Supervisor: Philippe Rigollet

Title: Professor, Mathematics

Acknowledgments

I owe the most to my advisor, Philippe Rigollet. In my third year he graciously admitted me to his group, and this has been the single biggest influence on my skills as an effective researcher and a project leader. His group has truly felt like home and I cannot say how grateful I am to be a part of it. On that note I would also like to thank the other members of the group, especially those who formed the splines working group - Sinho, Austin, George, and Thibaut. The rollercoaster evolution of the project was probably the most fun and rewarding research project I've worked on, and it would have been impossible to do alone.

I would like to thank Laurent Demanet, first for convincing me to come to MIT, and second for providing a wide range of amazing problems to work on that broadened my conception of applied mathematics. The bulk of the evolution of my interests is due to him, and though none of this work is present in this dissertation, my time in his group grew my knowledge base tremendously, and on top of that the work I did there has been very helpful in my applied work. My fellow Laurent groupers — Matt, Felipe, and Sarah — brightened my experience at MIT and through their standards for progress helped me hone my tastes.

Angela has been with me from the start of the program, and has helped me more than anyone to complete the marathon of graduate school. My family, of course, in every way laid the groundwork for all of this to happen. The lifetime of encouragement and direction that they gave me was what allowed me to complete this work, or be at MIT at all.

Funding

This work was supported by the National Science Foundation, as well as a Dean of Science fellowship through MIT. A large part of this thesis comes from [18].

Contents

1	Introduction	11
2	Background on Optimal Transport	23
2.1	Duality	25
2.2	The Wasserstein Space W_2	26
2.2.1	Topology	26
2.2.2	Curves	27
2.2.3	Geodesics	28
2.2.4	Riemannian Metric	29
2.2.5	Covariant Derivative	30
3	Splines in W_2	35
3.1	Introduction	35
3.1.1	Prior Work	37
3.1.2	Contributions	38
3.2	Background on Splines	39
3.3	Results on P -splines and E -splines	41
3.3.1	Proof of Proposition 4	42
3.3.2	Proof of Proposition 5	44
3.4	Transport Splines	51
3.4.1	Relationship to E -Splines in One Dimension	53
3.5	The Gaussian Case	55
3.5.1	Failure of Equality Between Transport Splines and E -Splines	56

3.5.2	Approximation Guarantees	59
3.5.3	Proof of the Main Theorem	69
3.6	Thin-Plate Splines	72
4	Schrodinger’s Splines	77
4.1	Introduction and Problem Formulations	77
4.1.1	The Interpolation Problem	80
4.1.2	The Endpoint Problem	81
4.2	Technical Background	84
4.2.1	Γ -Convergence	84
4.2.2	Large Deviation Principles	86
4.3	Results	87
4.3.1	The Interpolation Problem	87
4.3.2	The Endpoint Problem	93
5	Splines in WFR	101
5.1	Unbalanced Optimal Transport	101
5.1.1	The Cone Space	103
5.1.2	The Covariant Derivative	106
5.2	Splines	109
5.3	Transport Splines	112
A	Details for Numerical Experiments	115
A.1	Figure 3-1	115
A.2	Figure 3-5	116
A.3	Figure 3-3	117
B	Details on Natural Cubic Splines	119
	Bibliography	123

List of Figures

3-1	Piecewise linear and cubic spline interpolation of four Gaussians. The interpolation knots are shown in red and the interpolated Gaussians are shown in orange. See Appendix A.1 for details.	36
3-2	A comparison of 50 trajectories sampled from P-splines and transport splines for the Gaussian interpolation problem in Proposition 4 (see Section 3.3.1 for a detailed discussion). The first figure shows trajectories drawn from the P-spline interpolation, while the second shows trajectories from our method.	38
3-3	Reconstruction of trajectories in a physical system. See Appendix A.3. . .	52
3-4	Transport splines interpolation for the four uniform distributions as in (3.8). The red line is the quantile of order $3/4$ for the interpolation and the orange dotted line represents the corresponding candidate $\bar{F}_t^\dagger(u)$ for $u = 3/4$ introduced in (3.9).	58
3-5	Thin-plate splines for California temperature data (in °F); in the left column are the quantiles, while in the right are the means of the interpolated measures for an increasing sample of observations. See Appendix A.2. . . .	76
4-1	Samples from the process Y_t , the integrated Brownian bridge conditioned on having endpoint data 0.	96
4-2	Samples from X_t as $\varepsilon \rightarrow 0$ for fixed $g(t)$	99
4-3	Samples from different noise levels of the Brownian bridge, illustrating the large deviations principle for this family.	99

Chapter 1

Introduction

For classical data, the statistician’s toolkit is enviably deep: from linear and nonlinear, parametric and nonparametric estimators, to regression of sparse or structured data, to interpolation techniques, most any problem dealing with numerical vector or matrix data is amenable to several different statistical methods. Moreover, all of these problems can be solved efficiently, often in time linear in the size of the data. However, an increasing amount of data is not naturally structured as a vector in \mathbb{R}^d , an essential property in the problems above. For instance, natural images are stored as vectors in, say, $\mathbb{R}^{1080 \times 1920}$, but their semantic content has nothing to do with this representation; shifting an image one pixel to the right does not change its meaning, but it drastically changes the vector representing it. In the same way, temperature data on the earth does not live in Euclidean space (we study this example in detail in Chapter 3). For model training and validation, say using cross-validation or bootstrapping, each data point might be a dataset in its own right; what would it mean to form an “average dataset”? Finally, we might imagine that in an ensemble method, each expert produces for us a prior representing their beliefs about a particular event; how would we perform a statistical analysis of these priors?

Much attention has been focused on non-Euclidean data via the so-called manifold hypothesis, which assumes that the data is drawn from a manifold. (This is believed to underlie some of the workings of neural networks, see for example [24].) While this framework is very general, this means statistical methods, theoretical and numerical,

are difficult to apply, owing to the wild variation within the class of manifolds. Even smooth interpolation with splines is highly nontrivial in this setting [49]. In contrast, in the statistical community much effort has been directed towards a different framework which is simultaneously general enough to encompass the problems mentioned above, and specific enough to admit theoretical analysis and fast numerical methods. This is the Wasserstein space \mathcal{P}_2 of probability measures over \mathbb{R}^d with finite second moment, defined roughly by

$$W_2^2(\mu, \nu) = \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}^d} \int |x - T(x)|^2 d\mu \quad \text{s.t.} \quad T_{\#}\mu = \nu$$

This seeks to find the least-distortion map from μ to ν . When greyscale images are viewed as measures on $[0, 1]^2$, with mass $d\mu(x, y)$ equal to the image intensity, this metric recovers semantic meaning in an image, allowing, for instance, denoising of unaligned brain scans [30, 21]. In bootstrapping and cross-validation, dataset folds can be viewed as empirical measures on a master dataset \mathcal{D} . Priors are already measures, so they fit naturally into this framework. This metric, by the name of optimal transport, has proved immensely fruitful in modern applications ranging from the understanding the evolution of cell types human development [54] to natural image superresolution [34]; see [33] for more. This is to say nothing of the enormous theoretical implications of optimal transport outside of statistics [58, 59].

A modern treatment of data science using the Wasserstein space requires analogues of all the techniques mentioned above. Works on point estimation [61], “means” (barycenters) [19, 36, 5], and linear regression [11, 32] have recently appeared, as have some on interpolation by splines [7, 15]. It is this latter problem with which we are concerned. This thesis can be seen broadly as an exploration of curves of measures, which are smooth and in some sense attempt to minimize the total curvature, in the same way a Euclidean spline does. We also cover a related probabilistic problem, of finding the most likely trajectories taken by a collection of particles subject to distributional constraints at certain times — an interpolation problem unique to the setting of measures. Broadly, this work can be seen as a theoretical study of measure

interpolation problems, coupled with fast methods and extensions to new problems.

A general formulation of the interpolation problem asks: given measures μ_i at times t_i , find a curve (μ_t) of probability measures that interpolates the data — that is, $\mu_{t_i} = \mu_i$ — and is smooth. In the Euclidean setting the solution to this problem is classical: the natural interpolant is a piecewise cubic polynomial.¹ To attack the problem of defining and computing splines of measures, we first proceed purely abstractly, considering the set \mathcal{P}_2 as a smooth manifold. Indeed it can be equipped with a Riemannian structure; we provide an overview in Chapter 2, and more details can be found in [28]. In any Riemannian manifold \mathcal{X} with covariant derivative $\nabla = \frac{\mathbf{D}}{dt}$, splines can be formulated via a variational problem, by minimizing the energy functional:

$$\min_{x_t} \int \|\nabla_{\dot{x}_t} \dot{x}_t\|_{x_t}^2 dt \text{ s.t. } x_{t_i} = x_i$$

where the norm $\|\cdot\|_{x_t}$ is that of the Riemannian metric. In Euclidean space, introducing a variation $x_t + \varepsilon\gamma_t$ and using integration by parts yields that the optimal x satisfies $\frac{d^4}{dt^4}x = 0$, so that x is a cubic. On a manifold, the same logic gives that (see [46])

$$\frac{\mathbf{D}^3}{dt^3} \dot{x}_t + R \left(\frac{\mathbf{D}}{dt} \dot{x}_t, \dot{x}_t \right) \dot{x}_t = 0$$

where R is the Riemann curvature tensor; in Euclidean space $R \equiv 0$, and this reduces to the equation characterizing cubics. This can be solved in certain cases with computationally expensive shooting schemes [13]. For merely approximating the minimizer, again in special cases, a different method known as the de Castel'jau algorithm can be used [49]. However, in a general manifold the spline problem essentially stops here, with no easy way to compute or approximate energy-minimizing splines.

The key to the Riemannian geometry on \mathcal{P}_2 , however, is that it is not a purely abstract notion; measures $\mu \in \mathcal{P}_2$ should be considered not only as points in a struc-

¹It is the unique piecewise cubic that interpolates the data, is \mathcal{C}^2 , and has zero curvature at the endpoints. Uniqueness can be checked by counting parameters.

tured space, but as distributions of indistinguishable particles in \mathbb{R}^d . The properties of objects in the Wasserstein space are then expressed in terms of the behavior of the underlying particles. As a starting point, given a curve of measures (μ_t) and a curve of velocity fields (v_t) with $v_t: \mathbb{R}^d \rightarrow \mathbb{R}^d$, we say these satisfy the *continuity equation* if

$$\partial_t \mu_t + \operatorname{div}(v_t \mu_t) = 0$$

Here $\partial_t \mu_t$ is the ‘‘Euclidean’’ time derivative; if μ_t has density ρ_t then $\partial_t \mu_t$ has density $\frac{d}{dt} \rho_t$. From the particle perspective, the fields v_t represent the instantaneous velocity of the particles, and the measures μ_t are their distributions. Specifically, define the *flow maps* φ_t by the ODE

$$\varphi'_t(x) = v_t(\varphi_t(x)), \quad \varphi_0(x) = x$$

Then particles travel along trajectories $t \mapsto \varphi_t(x)$, and $\mu_t = (\varphi_t)_\# \mu_0$. In this way we can move between a Lagrangian and Eulerian perspective on particle dynamics, the former emphasizing trajectories and the latter instantaneous velocity fields; given particle trajectories $\varphi_t(x)$ we can extract $v_t = \varphi'_t \circ \varphi_t^{-1}$, and then $\mu_t = (\varphi_t)_\# \mu_0$ will satisfy the continuity equation.²

The continuity equation provides the link between curves in \mathcal{P}_2 and their particle representations. The only issue is uniqueness; if w_t satisfies $\operatorname{div}(w_t \mu_t) = 0$ then $v_t + w_t$ satisfies the continuity equation just as well as v_t .³ This is solved by choosing the fields v_t that minimize the *kinetic energy*

$$v_t = \arg \min \int |v_t(x)|^2 d\mu_t(x)$$

while still solving the equation for the curve (μ_t) . This is exactly the kinetic energy in the physical sense.

²Incidentally, by using the divergence theorem the continuity equation can be seen as enforcing conservation of mass of the particle system.

³For example, if $\mu_t = \mu_0 = N(0, I)$ then we can choose $v_t = 0$ or $v_t(x) = Rx$, where R is any rotation; since the Gaussian is rotationally invariant, both will satisfy the continuity equation.

By considering the particle trajectories as the central object, one can obtain another relationship between curves of measures and curves of particles. A *path measure* is a measure P on the space Ω of continuous paths. A path measure P is a representation of a curve (μ_t) if

$$\mu_t = (X_t)_\# P$$

for each t , where $X_t(\omega) = \omega_t$ is the time- t evaluation functional on Ω . Strictly speaking this is more general than a continuity equation representation, since nothing here requires paths to be unique from their initial position; we can have $\omega_0^{(1)} = \omega_0^{(2)}$ for different paths in the support of P .

The Riemannian geometry on Wasserstein space uses the continuity equation to define the metric (see the Benamou-Brenier theorem below). A fundamental fact about Wasserstein space is that the intrinsic definition of geodesics, using the Riemannian formalism, coincides with the path measure perspective. Indeed, one might define a “path geodesic” between μ and ν by

$$\min_P \int \ell(\omega) dP \text{ s.t. } (X_0)_\# P = \mu, (X_1)_\# P = \nu$$

where ℓ is the length functional. The solutions to this problem are precisely the same as the intrinsic geodesics; see Chapter 2 below. The optimal P will be supported on Euclidean geodesics (straight lines), and the flow maps $\varphi_t(x)$ will form straight lines as well (the same ones). From this one can formulate the principle that to define a geometric object in Wasserstein space, one should look at the corresponding object in Euclidean space, and consider the Wasserstein object to be a measure on Euclidean objects. For example, a surface in Wasserstein space may be captured by a “surface measure” $P \in \mathcal{P}(\mathcal{Y})$, where \mathcal{Y} is the set of continuous surfaces in Euclidean space; for $p \in \mathbb{R}^2$ we would have $\mu_p = (X_p)_\# P$, where X_p is as before the evaluation functional. We study this exact problem in Chapter 3, under the name of thin-plate splines.

As mentioned in the last paragraph, for “first-order” objects — distances, lengths, geodesics — intrinsic Riemannian objects are perfectly described by lifting their Euclidean analogs, via both the continuity equation and a path measure. The first part

of this thesis is concerned with this question: for “second-order” objects — curvatures and splines — can we use the same principle? For example, a second-order path measure spline would be defined by

$$\min_P \int \int_0^1 |\ddot{\omega}_t|^2 dt dP(\omega) \text{ s.t. } (X_{t_i})_{\#} P = \mu_i$$

which was introduced simultaneously in [7, 15]. As we will see, intrinsic splines do differ from their path measure counterparts, even for Gaussian measures (Theorem 5), and imposing that the optimal path measure be a particle flow, with deterministic paths taken by each particle, creates yet another different solution (Theorem 4). Thus we resolve that for second-order objects the intrinsic Riemannian definition, the Lagrangian path-measure perspective, and the Eulerian particle-evolution perspective differ. However, all is not lost. We go on to use the Eulerian perspective to define a new spline of measures, which we term the *transport spline*. This turns out to be both much cheaper to compute, and a closer approximation of the intrinsic spline, even equaling it in certain situations (see Theorem 10). We then demonstrate, as mentioned above, that the same techniques can be applied to more general objects, such as interpolating surfaces of measures.

In Chapter 4 we study a stochastic formulation of measure splines. To motivate it, we recall a circle of ideas around entropic optimal transport and the Schrödinger bridge problem.

Though discovered early in the 20th century, entropic optimal transport was recently popularized in [22] as a means of speeding up the computation of ordinary optimal transport. The entropic OT problem is

$$\min_{\pi \in \Pi(\mu, \nu)} \int |x - y|^2 d\pi - \varepsilon E[\pi]$$

where $\varepsilon > 0$ is some parameter and E is the ordinary differential entropy (so in par-

ticular π must be absolutely continuous if the objective is finite). Just as optimal transport has the static formulation $\min \int |x - y|^2 d\pi$ and the dynamical formulation given by the continuity equation, one might hope that entropic OT also has a dynamical form. Indeed, it was found in the 1930s by Schrödinger [56, 55] that the entropic OT problem is precisely equivalent to the Schrödinger bridge problem

$$\min_P \text{KL}(P \parallel Q^\epsilon) \text{ s.t. } (X_0)_\#P = \mu, (X_1)_\#P = \nu$$

where Q^ϵ is the *reference process* of Brownian motion started at μ , that is, $Q_t^\epsilon = x_0 + \sqrt{\epsilon}B_t$ where x_0 is sampled from μ . In physical terms, the bridge problem asks: if I observe a gas in configuration μ at time $t = 0$ and configuration ν at $t = 1$, then what was the most likely distribution of paths for the particles of gas to have taken? Of course, as stated this problem is ill-formed since the distribution at $t = 1$ must be precisely $\mu * \mathcal{N}(0, \epsilon I)$, but by considering a finite discretization of μ and asking that at time 1 it be close in some sense to ν , this problem can be given meaning, and as the discretization grows finer this problem will converge to the Schrödinger bridge problem; see [38] for more details. With this in hand, we have a complete picture of optimal transport, describing entropic regularization and the static-dynamic equivalence (this is the table (4.7) from Chapter 4):

	Dynamical	Static
$\epsilon > 0$	Schrödinger bridge	entropically regularized OT
$\epsilon = 0$	dynamical formulation of OT	OT

Horizontally we have equivalence, while vertically we have convergence as $\epsilon \rightarrow 0$; this is proved in [38, 39]. Convergence is not merely in the value, but in the arg-optima as well. For instance, the lower-left problem is

$$\min_P \int \int_0^1 |\dot{\omega}_t|^2 dt dP(\omega) \text{ s.t. } (X_0)_\#P = \mu, (X_1)_\#P = \nu$$

If P^ϵ optimizes the Schrödinger bridge problem and P solves the dynamical OT

problem, then $P^\varepsilon \rightarrow P$ weakly (hence in the Wasserstein metric on path measures). recall from above that the optimal P will be supported on straight lines; this means that P^ε will be supported on Brownian bridges with variance going to zero, which in some sense converge to straight lines.

Whereas the above discussion focuses on geodesics, we develop a similar understanding of splines of measures. For the purposes of this discussion we focus on path measure splines, defined above by minimizing $\int \int_0^1 |\ddot{\omega}_t|^2 dt dP(\omega)$. Furthermore, the case that admits the most complete solution is not an interpolating spline, but a so-called endpoint spline. In Euclidean space, the following variational problem

$$\min \int_0^1 |\ddot{x}(t)|^2 dt \text{ s.t. } x(0) = x_0, x(1) = x_1, \dot{x}(0) = v_0, \dot{x}(1) = v_1$$

can be seen to admit as its solution the unique cubic polynomial satisfying the endpoint constraints.⁴ Likewise, we can define an endpoint measure spline by

$$\min_P \int \int_0^1 |\ddot{\omega}_t|^2 dt dP(\omega) \text{ s.t. } P \in \Pi(\mu, \nu, V, W)$$

where V and W are fixed vector fields and the constraint set $\Pi(\mu, \nu, V, W)$ is the set of path measures P with marginals μ at 0 and ν at 1, and for each $\omega \in \text{Supp}(P)$ we have $\dot{\omega}_0 = V(\omega_0)$ and $\dot{\omega}_1 = W(\omega_1)$. In other words, the initial and final positions dictate the initial and final velocities. This is a dynamical problem, and it can be shown to be equivalent to a static optimal transport problem $\min \int c(x, y) d\pi$, where the cost c is the optimal value of the Euclidean endpoint spline problem above (with velocities $V(x)$ and $W(y)$). In turn this can be entropically regularized in the usual way. The remaining piece of the puzzle is to find a stochastic dynamical problem that is equivalent to the regularized one.

We show in Chapter 4 that the right notion of stochastic dynamical spline is given

⁴This is a subproblem of the spline interpolation problem. By solving the endpoint problem on each interval for fixed velocities, and then optimizing over the velocities, we obtain exactly the cubic spline interpolant. Incidentally, the optimal value of the interpolation problem, considered as a function of (x_0, x_1, v_0, v_1) , is a positive definite quadratic form, so expressing the spline problem as an iterated optimization makes it the same as optimizing a quadratic form.

by altering the reference process in the Schrödinger bridge problem. Specifically, instead of Brownian motion, we consider the *integrated Brownian motion*

$$Q_t^\varepsilon = x_0 + tV(x_0) + \sqrt{\varepsilon} \int_0^t B_s ds$$

with noise level ε , where x_0 is an initial position sampled from μ . This yields a path measure problem which we call the *Schrödinger spline*

$$\min \text{KL}(P \parallel Q^\varepsilon) \text{ s.t. } P \in \Pi(\mu, \nu, V, W)$$

By standard methods this can be seen to be equivalent to the entropically regularized endpoint spline, and adapting arguments from [38, 39] we can show that as $\varepsilon \rightarrow 0$ this converges to the path measure problem above.

Furthermore, we can characterize fully the solution P^ε ; we show that it is the product of an OT coupling with a certain “spline bridge”. This bridge is the sum of an ordinary cubic spline and a special noise process (which is scaled by $\sqrt{\varepsilon}$). In this way we can completely understand the convergence of the Schrödinger spline to the path spline, through the noise process. See the figures in Chapter 4 for more details.

Finally, we expand our study of splines in Wasserstein space to splines of measure of different total masses. The natural setting for such measures is the so-called Wasserstein-Fisher-Rao space introduced in [41, 40, 20, 35]. Whereas the continuity equation above encodes conservation of mass, we are explicitly interested in non-conservation of mass, since the total mass of the measures must change along the curve. We begin with the *non-conservative continuity equation*

$$\partial_t \mu_t + \text{div}(v_t \mu_t) = 4\alpha_t \mu_t$$

We now have a triple (μ_t, v_t, α_t) representing the distribution of particles, the particles’ instantaneous velocity, and an instantaneous relative growth rate. Physically, this

describes (among other things) an evolving system of chemical compounds which are undergoing reactions, increasing or decreasing its mass locally by α_t as well as convecting by v_t . This equation does not have a unique solution; whereas before we disambiguated the solution by minimizing the kinetic energy, we now select for each μ_t the pair (v_t, α_t) minimizing the energy

$$\min_{v_t, \alpha_t} \int |v_t(x)|^2 + 4\alpha_t(x)^2 d\mu_t$$

(The factor of 4 is convenient later). In this way a Riemannian structure on this space can be defined; see [35]. The natural analog of the flow maps φ_t from Wasserstein space adds a mass term: with $\varphi'_t(x) = v_t(\varphi_t(x))$ as above, and defining

$$m'_t(x) = 4\alpha_t(\varphi_t(x))m_t(x), \quad m_0 = 1$$

we have under some smoothness conditions that

$$\mu_t = (\varphi_t)_\#(m_t \cdot \mu_0)$$

This provides a Lagrangian perspective on curves in WFR. Contrasted with the W_2 case, this gives to each particle a mass $m_t(x)$, and the particle evolves as $((\varphi_t(x), m_t(x)))$.

The space of (point, mass) pairs is called the *cone space* $\mathfrak{C} = \mathbb{R}^d \times \mathbb{R}_+ / \mathbb{R}^d \times \{0\}$, and all points of zero mass are identified. Curves in WFR space can then be naturally written as path measures on the set $\Omega_{\mathfrak{C}}$ of continuous paths in the cone space, in the same way that curves in Wasserstein space are written as measures on the space of euclidean paths. Indeed, geodesics in WFR are intimately related to path measures on \mathfrak{C} ; see Chapter 5 for more information.

We begin our study of WFR splines by deriving the covariant derivative associated to the Riemannian structure and using this to define an intrinsic spline. We then define a Lagrangian notion of splines as measures on $\Omega_{\mathfrak{C}}$, and show that it is a relaxation of intrinsic splines (which most likely is not tight, as we demonstrated in the Wasserstein case). Proceeding as in W_2 , we use the flow maps to define transport

splines in WFR, and describe a simple numerical algorithm for computing them. This requires computing the intrinsic spline on \mathfrak{C} ; it is not known how to do this exactly, even in the one-dimensional case, so we use a simple approximation. In addition to numerical examples of our method, for the benefit of the reader we also provide examples of WFR geodesics between some one-dimensional measures.

Chapter 2

Background on Optimal Transport

The material in this chapter is based on the excellent sources [2, 28, 53].

The starting point for optimal transport is the so-called *Monge problem* of finding the “best” mapping from an initial measure μ onto another measure ν :

$$\min_T \int |x - T(x)|^2 d\mu \text{ s.t. } T\#\mu = \nu \tag{2.1}$$

where $T\#\mu$ is the measure defined by $T\#\mu(A) = \mu(T^{-1}(A))$. Here, “best” is defined as minimizing total distance squared — one can think of moving a pile of sand into a different configuration, grain by grain, and minimizing effort spent doing so. The intuitive appeal of this problem is counterbalanced by the intractability of the constraint $T\#\mu = \nu$. Indeed, by the change-of-variables formula, assuming that $d\mu = f dx$ and $d\nu = g dx$ this can be phrased as

$$f(x) = g(T(x)) |\det \nabla T(x)|$$

∇ being the Jacobian. This is highly nonlinear and largely intractable both theoretically and practically. Furthermore, it is not even guaranteed that a feasible T for (2.1) exists; for instance, if $\mu = \delta_0$ and $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, there can be no mapping from μ to ν .

For this reason, though practically a solution to (2.1) is desired, it is better to

consider the *Kantorovich problem*

$$\min_{\gamma \in \Pi(\mu, \nu)} \int c(x, y) d\gamma \quad (2.2)$$

where the case $c(x, y) = |x - y|^2$ corresponds to the Monge problem. The set $\Pi(\mu, \nu)$ is defined as those probability measures $\gamma \in \mathcal{P}(X^2)$ such that $\pi_x \gamma = \mu$ and $\pi_y \gamma = \nu$, where $\pi_x \gamma$ is the first marginal of γ , so $\pi_x \gamma(A) = \gamma(A \times X)$ and likewise for π_y . Whereas (2.1) is concerned with mapping the mass of μ onto ν , (2.2) finds a *plan* of mass movement, which can possibly split the mass of μ as it moves it to ν . In contrast with (2.1) the constraints are each linear (so the constraint set is convex) and the objective is linear as well, considered as a function of γ , so it is simple to analyze and easily solved. Furthermore, it is a relaxation of (2.1), since if T is feasible for (2.1) then $\gamma = (\text{Id}, T)_\# \mu$ is feasible for (2.2). Theoretically, much is gained:

Theorem 1. *The problem (2.2) is always feasible. If $c(x, y): X \times X \rightarrow [0, \infty]$ is lower semicontinuous, (2.2) admits a solution.*

It also transpires that little is lost; the injection $T \mapsto (\text{Id}, T)_\# \mu$ from (2.1) to (2.2) is reversible at optimality, as often as can be expected.

Definition 1. *A cost function $c(x, y)$ satisfies the **twist condition** if c is differentiable in x at every point, and the map $y \mapsto \nabla_x c(x, y)$ is injective for every x . If the cost function is \mathcal{C}^2 , this is equivalent to $\det((\partial_{x_i} \partial_{y_j} c)_{ij}) \neq 0$.*

For the cost $c(x, y) = |x - y|^2$ we have $\nabla_x c(x, y) = 2(x - y)$, which is injective for fixed x .

Theorem 2. *If c satisfies the twist condition, and μ is absolutely continuous (or at least places zero mass on sets of Hausdorff dimension at most $d - 1$), then (2.2) has a unique solution γ^* , and it is induced by a map T , so that $\gamma^* = (\text{Id}, T)_\# \mu$.*

As (2.2) is a relaxation of (2.1), the T supplied from the theorem above must minimize (2.1). For the specific case of $c = |x - y|^2$, which will be our almost exclusive concern in this work, more can be said:

Theorem 3 (Brenier). *For the quadratic cost, the map T is the gradient of a convex function.*

For two measures μ and ν , we now know there is a unique optimal map T between them, and it must be the gradient of a convex function. Without making reference to μ and ν , however, we can still say that a map T is optimal. We have

Proposition 1. *If $T = \nabla\varphi$ where φ is convex, then T is optimal between μ and $T_{\#}\mu$ for any measure μ .*

This can be seen, for instance, from Brenier's polar factorization theorem, stated in a simple setting below.

Theorem 4 (Brenier Polar Factorization). *Let Ω be a compact domain and $T: \Omega \rightarrow \mathbb{R}^d$ some mapping. Let λ be the unit-scaled Lebesgue measure on Ω , and suppose that $T_{\#}\lambda$ is absolutely continuous. Then there is a convex function φ and a measure-preserving map S (meaning that $S_{\#}\lambda = \lambda$) such that $T = (\nabla\varphi) \circ S$, and furthermore S solves*

$$\max \int \langle T(x), S(x) \rangle d\lambda(x) \text{ s.t. } S_{\#}\lambda = \lambda$$

or equivalently

$$\min \int |T(x) - S(x)|^2 d\lambda(x) \text{ s.t. } S_{\#}\lambda = \lambda$$

This idea will recur later in the discussion of the dynamic formulation of optimal transport.

2.1 Duality

As remarked, both the objective and constraints in (2.2) are linear in γ , so that it is an infinite-dimensional linear program. Following the standard procedure, its dual program is seen to be

$$\max_{\varphi, \psi} \int \varphi d\mu + \int \psi d\nu \text{ s.t. } \varphi(x) + \psi(y) \leq c(x, y) \tag{2.3}$$

The duality is more than formal.

Theorem 5. *Assume that there are functions $c_X \in L^1(\mu)$ and $c_Y \in L^1(\nu)$ such that $c(x, y) \leq c_X(x) + c_Y(y)$. Then the maximum in (2.3) is attained, and we have the strong duality statement $\min(2.2) = \max(2.3)$. If the cost is merely lower semicontinuous and lower-bounded, then strong duality still holds, though dual optimality may not be attained.*

In the quadratic case, the optimal Monge map can be read from the dual solution. If φ is dual optimal, then $\frac{1}{2}|x|^2 - \varphi(x)$ is convex, and $T(x) = x - \nabla\varphi$.

2.2 The Wasserstein Space W_2

The optimization problem (2.2) defines a metric on the space \mathcal{P}_2 of probability measures with finite second moment.¹ Specifically,

$$W_2^2(\mu, \nu) = \min_{\gamma \in \Pi(\mu, \nu)} \int c d\gamma$$

where $c = \frac{1}{2}|x - y|^2$ as usual. As a metric space, (\mathcal{P}_2, W_2) has a rich structure.

2.2.1 Topology

To begin with, the topology of W_2 is easily described.

Definition 2. *The weak topology on $\mathcal{P}(X)$ is defined in duality with $\mathcal{C}_b(X)$, the set of bounded continuous functions. Specifically, μ_n converges weakly to μ , denoted $\mu_n \xrightarrow{w} \mu$, if $\int f d\mu_n \rightarrow \int f d\mu$ for all $f \in \mathcal{C}_b(X)$.*

(This is also known as the *narrow topology*.)

Theorem 6. *Take $X \subset \mathbb{R}^n$. Let $\mu_n, \mu \in \mathcal{P}_2$. Then $\mu_n \xrightarrow{W_2} \mu$ if and only if $\mu_n \xrightarrow{w} \mu$ and $\int |x|^2 d\mu_n \rightarrow \int |x|^2 d\mu$, where \xrightarrow{w} denotes weak convergence.*

An analogous statement holds for non-Euclidean X . In brief, W_2 metrizes the weak topology.

¹If X is not a subset of \mathbb{R}^d then this condition can be phrased as $\int d(x, x_0)^2 d\mu < \infty$ for some x_0 , or equivalently for any x_0 .

2.2.2 Curves

Generally, a curve (x_t) in a metric space (X, d) , abusing notation for the moment, is called *absolutely continuous* if there is $f \in L^1$ such that

$$d(x_s, x_t) \leq \int_s^t f(r) dr$$

If (x_t) is absolutely continuous, its *metric derivative* is defined by

$$|\dot{x}_t| = \lim_{h \rightarrow 0} \frac{d(x_{t+h}, x_t)}{h}$$

and as (x_t) is absolutely continuous this limit exists almost everywhere.

Since W_2 is a metric these definitions apply readily, but we are interested in the specific interpretation of its points as measures representing the density of particles, and curves of its points as movement of these particles. In this way the Wasserstein space has a natural interpretation in terms of fluid dynamics. Let $\mu_0 \in \mathcal{P}_2$, and fix velocity fields $v_t: X \rightarrow TX$. By solving the Cauchy problem

$$\varphi_t(x)' = v_t(x), \varphi_0(x) = x \tag{2.4}$$

we obtain the so-called *flow maps* $\varphi_t(x)$. From this we can define a curve of measures by

$$\mu_t = (\varphi_t)_\# \mu_0$$

Physically, the flow maps define the movement of infinitesimal masses $\mu_0(x)$ originating at x , and the measures μ_t describe the density of the resulting configuration at time t . It is a consequence of the divergence theorem that the measures μ_t solve the *continuity equation*

$$\partial_t \mu_t + \operatorname{div}(v_t \mu_t) = 0 \tag{2.5}$$

Indeed, this equation characterizes regular curves in Wasserstein space.

Theorem 7. *If (μ_t) is an absolutely continuous curve, then there is a family of vector fields v_t solving the continuity equation such that $\|v_t\|_{L^2(\mu_t)} \leq |\dot{\mu}_t|$.*

Conversely, if (μ_t, v_t) solves the continuity equation and $\int_0^1 \|v_t\|_{L^2(\mu_t)} dt < \infty$, then the curve (μ_t) is absolutely continuous and $|\dot{\mu}_t| \leq \|v_t\|_{L^2(\mu_t)}$.

The reason for the reversing inequalities is that even for fixed (μ_t) , the solution to (2.5) is not uniquely defined. We may add to a solution any family w_t with $\operatorname{div}(w_t \mu_t) = 0$ (which can be thought of as divergence-free fields, experiencing only curl), and $(\mu_t, v_t + w_t)$ will again solve (2.5).

2.2.3 Geodesics

Definition 3. *Let (X, d) be a metric space. A curve $(x_t)_{t \in [0,1]} \subset X$ is a **geodesic** if $d(x_s, x_t) = |s - t|d(x_0, x_1)$ for all $s, t \in [0, 1]$. A metric space (X, d) is a **geodesic space** if for all $x, y \in X$ there is a geodesic between them.*

The central result is that if X is a geodesic space, then (\mathcal{P}_2, W_2) is a geodesic space as well, and the geodesic is easily described. Let $\mu_0, \mu_1 \in \mathcal{P}_2$, and let γ be the optimal coupling between them. For x, y , let $\ell^{x,y}(t)$ be the geodesic between them in X . Using this, define $X_t(x, y) = \ell^{x,y}(t)$, and let

$$\mu_t = (X_t)_\# \gamma \tag{2.6}$$

Theorem 8. *The curve (μ_t) defined in (2.6) is a geodesic in W_2 . In $X = \mathbb{R}^d$, if γ is induced by a map T from μ_0 to μ_1 , this curve has the simpler description*

$$\mu_t = ((1 - t) \operatorname{Id} + tT)_\# \mu_0$$

One can also arrive at the description of geodesics by considering an optimization program over paths in X . Let Ω be the set of continuous paths in X , and let $\ell: \Omega \rightarrow [0, \infty]$ the length functional. Then consider the optimization program

$$\min_{P \in \mathcal{P}(\Omega)} \int \ell dP \text{ s.t. } (X_0)_\# P = \mu_0, (X_1)_\# P = \mu_1 \tag{2.7}$$

where $X_t(\gamma) = \gamma(t)$ is the evaluation function.

Proposition 2. *Let P be the solution to (2.7). Then P is supported on geodesics, and $(X_0, X_1)_\#P$ is the optimal coupling between μ_0 and μ_1 , so $\mu_t = (X_t)_\#P$ is the geodesic described above.*

2.2.4 Riemannian Metric

Theorem 7 strongly suggests that if \mathcal{P}_2 were to have a Riemannian structure, then its tangent vectors would be given by velocity fields v_t , and the norm of this vector would be $\|v_t\|_{L^2(\mu_t)}$. This intuition is strengthened by the following result.

Theorem 9 (Benamou-Brenier). *Let $\mu_0, \mu_1 \in \mathcal{P}_2$. Then*

$$W_2^2(\mu_0, \mu_1) = \inf_{\mu_t, v_t} \int_0^1 \|v_t\|_{L^2(\mu_t)}^2 dt$$

where the infimum is taken over families (μ_t, v_t) satisfying the continuity equation.

This is exactly analogous to the Riemannian definition of geodesics as length-minimizing curves under the hypothesis above that we should take $\|v_t\|_{\mu_t} = \|v_t\|_{L^2(\mu_t)}$. One problem remains, however: non-uniqueness of solutions to the continuity equation. Theorems 7 and 9 suggest that we should privilege vector fields v_t solving the continuity equation that have minimal $L^2(\mu_t)$ norm. Let v_t be some fixed solution to the continuity equation. Then the problem

$$\min \|\tilde{v}_t\|_{L^2(\mu_t)}^2 \text{ s.t. } \partial_t \mu_t + \operatorname{div}(\tilde{v}_t \mu_t) = 0$$

has a unique solution for each t , as the norm is strongly convex. From this, applying the continuity equation shows that the optimal field v_t satisfies

$$\langle v_t, w_t \rangle_{\mu_t} = 0 \text{ for all } w \text{ s.t. } \operatorname{div}(w_t \mu_t) = 0$$

this implies that v_t is a gradient, or rather in its closure in $L^2(\mu_t)$. Thus we are led

to define

$$T_\mu(\mathcal{P}_2) = \text{clos}_{L^2(\mu)} \{ \nabla \varphi \mid \varphi \in \mathcal{C}_c^\infty \} \quad (2.8)$$

With this in hand, the right Riemannian metric on Wasserstein space is given by

$$\langle v, w \rangle_\mu = \langle v, w \rangle_{L^2(\mu)} = \int \langle v(x), w(x) \rangle d\mu(x)$$

A more intuitive explanation of the tangent space is that it is composed, as mentioned above, of optimal maps, which are gradients of convex functions. Specifically, if (μ_t, v_t) satisfies the continuity equation and v_t is $L^2(\mu_t)$ -minimal — that is, it is in the tangent space — we have the following limiting result:

$$v_t = \lim_{h \rightarrow 0} \frac{T_{\mu_t \rightarrow \mu_{t+h}} - \text{Id}}{h} \quad (2.9)$$

where $T_{\mu \rightarrow \nu}$ is the optimal map from μ to ν , and the limit occurs in $L^2(\mu_t)$. Tangent vectors should point along geodesics, and as discussed above geodesics are given by moving mass along straight lines. Likewise, if $v \in T_\mu(\mathcal{P}_2)$ is small then $T = \text{Id} + v$ is the gradient of a convex function, and causing mass to travel from μ in the direction of v is the same as applying the map T . Thus, in a sense, the exponential map is given transporting mass along geodesics in the direction v , and the logarithmic map is given by solving the Monge problem to recover the transport map.

2.2.5 Covariant Derivative

As \mathcal{P}_2 is not a true Riemannian manifold — it is infinite-dimensional, and has singular points at singular measures — the covariant derivative and Levi-Civita connection are not immediate from the metric and must be verified to exist. We present a new construction here that takes pieces from [28], as it will prove useful later. The Levi-Civita connection is has two properties:

1. It must respect the metric. If (μ_t) is a curve and v_t, w_t are two vector fields

along it, then

$$\frac{d}{dt}\langle v_t, w_t \rangle_{\mu_t} = \left\langle \frac{\mathbf{D}}{dt} v_t, w_t \right\rangle_{\mu_t} + \left\langle v_t, \frac{\mathbf{D}}{dt} w_t \right\rangle_{\mu_t}$$

2. It must be torsion-free. If X and Y are vector fields, then

$$\nabla_X Y - \nabla_Y X = [X, Y]$$

Let (u_t) be the tangent field of the curve (μ_t) (as noted above, the minimal field is unique and is a gradient²), and let $(v_t), (w_t)$ be two other fields along (μ_t) . The product rule yields

$$\begin{aligned} \frac{d}{dt}\langle v_t, w_t \rangle_{\mu_t} &= \frac{d}{dt} \int \langle v_t, w_t \rangle d\mu_t \\ &= \int \langle \partial_t v_t, w_t \rangle + \langle v_t, \partial_t w_t \rangle d\mu_t + \int \langle v_t, w_t \rangle d(\partial_t \mu_t) \end{aligned}$$

Because (μ_t, u_t) solves the continuity equation, the dual definition of $\text{div}(u_t \mu_t)$ gives

$$\int \langle v_t, w_t \rangle d(\partial_t \mu_t) = \int \langle \nabla v_t \cdot u_t, w_t \rangle + \langle \nabla w_t \cdot u_t, v_t \rangle d\mu_t$$

If the covariant derivative respects the metric, then grouping terms yields

$$\left\langle \frac{\mathbf{D}}{dt} v_t, w_t \right\rangle + \left\langle v_t, \frac{\mathbf{D}}{dt} w_t \right\rangle = \langle \partial_t v_t + \nabla v_t \cdot u_t, w_t \rangle + \langle v_t, \partial_t w_t + \nabla w_t \cdot u_t \rangle$$

where all inner products are with respect to μ_t . From here it is almost required that

$$\frac{\mathbf{D}}{dt} v_t = \mathcal{P}_{\mu_t} \left(\frac{D}{dt} v_t \right) = \mathcal{P}_{\mu_t} (\partial_t v_t + \nabla v_t \cdot u_t) \quad (2.10)$$

where \mathcal{P}_{μ_t} is the orthogonal projection onto $T_{\mu_t}(\mathcal{P}_2)$ in $L^2(\mu_t)$. We call $\frac{D}{dt}$ the *total derivative*. Note that if $v_t = u_t = \nabla \varphi_t$ then $\partial_t v_t + \nabla v_t \cdot v_t = \nabla (\partial_t \varphi_t + \frac{1}{2} |\nabla \varphi_t|^2) \in T_{\mu_t}(\mathcal{P}_2)$, so no projection is necessary, and the total and covariant derivatives coincide;

²In general it is merely in the closure of the set of gradients, but if all measures involved are absolutely continuous then it is truly a gradient.

in other words,

$$\frac{\mathbf{D}^2}{dt^2}\mu_t = \partial_t v_t + \nabla v_t \cdot v_t \quad (2.11)$$

We now examine the torsion-free property, and we follow Gigli's argument [Gigli page 73], which we partially repeat for convenience of reference. It is quite technical to give meaning directly to a smooth vector field on all of \mathcal{P}_2 , and so to the Levi-Civita connection, but it can be indirectly defined via the covariant derivative, which is all that will be necessary in this work. Let $(\mu_t^{(1)})$ and $(\mu_t^{(2)})$ be two (absolutely continuous) curves such that $\mu_0^{(1)} = \mu_0^{(2)} = \mu$, and let their velocity fields be $(u_t^{(1)})$ and $(u_t^{(2)})$. Define two new tangent fields along these curves by

$$\begin{aligned} v_t^{(1)} &= u_0^{(2)} \\ v_t^{(2)} &= u_0^{(1)} \end{aligned}$$

Under reasonable assumptions (for instance, of all the measures involved are absolutely continuous) $u_0^{(i)}$ are gradients, so they are in $T_\mu(\mathcal{P}_2)$ for every μ , so these indeed define tangent fields.

With this definition, it is reasonable to interpret

$$\nabla_{v_0^{(1)}} v_t^{(2)} \Big|_{t=0} = \frac{\mathbf{D}}{dt} v_t^{(2)} \Big|_{t=0}$$

with the derivative of course being taken along $\mu_t^{(2)}$, and similarly for $\nabla_{v_0^{(2)}} v_t^{(1)}$. Now, fix φ and consider the functional $F: \mu \mapsto \int \varphi d\mu$. By the continuity equation, the derivative of F along $v_t^{(2)}$ is

$$\partial_t F[\mu_t^{(2)}] = \int \varphi d(\partial_t \mu_t^{(2)}) = \int \langle \nabla \varphi, v_t^{(2)} \rangle d\mu_t^{(2)}$$

Then since the covariant derivative above respects the metric,

$$\begin{aligned}
v_0^{(1)}(v^{(2)}(F))[\mu] &= \frac{d}{dt} \langle \nabla \varphi, v_t^{(2)} \rangle_{\mu_t^{(2)}} \Big|_{t=0} \\
&= \left\langle \frac{\mathbf{D}}{dt} \nabla \varphi, v_t^{(2)} \right\rangle_{\mu_t^{(2)}} + \left\langle \nabla \varphi, \frac{\mathbf{D}}{dt} v_t^{(2)} \right\rangle_{\mu_t^{(2)}} \Big|_{t=0} \\
&= \left\langle \nabla^2 \varphi \cdot v_0^{(1)}, v_0^{(2)} \right\rangle_{\mu} + \left\langle \nabla \varphi, \nabla_{v_0^{(1)}} v_t^{(2)} \right\rangle_{\mu}
\end{aligned}$$

Performing the same calculation for $v_0^{(2)}(v^{(1)}(F))[\mu]$ and subtracting, since $\nabla^2 \varphi$ is symmetric the first terms cancel and we get

$$v_0^{(1)}(v^{(2)}(F))[\mu] - v_0^{(2)}(v^{(1)}(F))[\mu] = \left\langle \nabla \varphi, \nabla_{v_0^{(1)}} v_t^{(2)} - \nabla_{v_0^{(2)}} v_t^{(1)} \right\rangle_{\mu}$$

Since gradients $\nabla \varphi$ are dense in $T_{\mu}(\mathcal{P}_2)$ this means that $\frac{\mathbf{D}}{dt}$ is indeed torsion-free. We will return to this calculation when characterizing the covariant derivative in Wasserstein-Fisher-Rao space.

Chapter 3

Splines in W_2

3.1 Introduction

Smooth interpolation is a fundamental tool in numerical analysis that plays a central role in data science.¹ While this task is traditionally studied on the flat Euclidean space \mathbb{R}^d , recent applications have called for interpolation of points living on curved spaces such as smooth manifolds [46] and, more recently, the Wasserstein space of probability measures. An important application arises in single-cell genomic data analysis where the measure μ_t represents a population of cells at time t of a biological process such as differentiation, and the cells of an organism specialize over the course of early development. In this context, two main questions arise: 1) to infer the profile of the population at unobserved times; and more importantly 2) to reconstruct the trajectories of individual cells in gene space, that is: given a cell at time t , determine its (likely) history and fate. [51] argue that cellular trajectory reconstruction is crucial to unlocking the promises of single-cell genomics. A breakthrough in this direction was recently achieved using optimal transport by [54], but their work does not produce *smooth* trajectories. To illustrate, we display in Figure 3-1 a comparison of their approach with the smooth interpolation methodology developed in the present work. Although we are mainly motivated by cell trajectory reconstruction, we are confident that the flexibility and efficiency of the method will allow it to find applications

¹This work appears in [17].

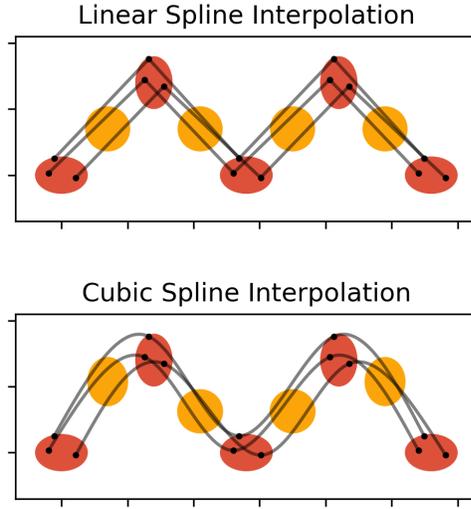


Figure 3-1: Piecewise linear and cubic spline interpolation of four Gaussians. The interpolation knots are shown in red and the interpolated Gaussians are shown in orange. See Appendix A.1 for details.

beyond this scope.

While the first question above is a natural extension of interpolation to the space of probability measures, the second question calls for a specific type of interpolation: one that also reconstructs the (smooth) trajectories of individual particles. These two concepts are linked by the continuity equation (2.5) and the flow maps (2.4). If (μ_t) is an interpolating curve of measures, then the tangent velocity field $v_t \in T_{\mu_t} \mathcal{P}_2$ yields canonical particle trajectories (X_t) via (2.4), but there are other choices that lead to alternative trajectories while maintaining the same measures. We are led to the following problem.

The problem. Let $(X_t^*)_{t \in [0,1]}$ be a stochastic process on \mathbb{R}^d with \mathcal{C}^2 sample paths and marginal laws $X_t^* \sim \mu_t^*$, $t \in [0, 1]$. Given $\mu_{t_0}^*, \mu_{t_1}^*, \dots, \mu_{t_N}^*$ at times $0 = t_0 < t_1 < \dots < t_N = 1$, the task is to construct a stochastic process $(X_t)_{t \in [0,1]}$ such that X_t has \mathcal{C}^2 sample paths and the distribution μ_t of X_t interpolates the given measures, so $\mu_{t_i} = \mu_{t_i}^*$ for $i = 0, 1, \dots, N$.

3.1.1 Prior Work

This work is at the intersection of interpolation and optimal transport. On the one hand, interpolation in \mathbb{R}^d is very well-developed, with fast and accurate methods ranging from interpolating polynomials and splines to more exotic non-parametric approaches [60], and with renewed interest due to recent theoretical results [6]. Our methodology can accommodate all of these options, but we focus on cubic spline interpolations due to their simplicity, theoretical guarantees, and their *curvature-minimizing* property (see Section 3.2). On the other hand, optimal transport has become a useful tool in the analysis of observations represented in the form of probability measures. Recent computational advances [23, 1, 48] have led to the development of many methods in statistical optimal transport, from barycenters to geodesic PCA. The present work extends this toolbox by developing a method for smooth interpolation over the Wasserstein space of probability measures.

Splines in Wasserstein space were considered concurrently and independently by [15] and [7]. Both papers converge to the same notion of splines, which we call P-splines. Though motivated by particle dynamics, P-splines solve an optimal transport problem that is not guaranteed to have a Monge solution (see Theorem 2). Instead, it outputs stochastic processes $(X_t)_{t \in [0,1]}$ for which X_t is not a deterministic function of X_0 . In other words, given an initial position, there is no unique particle trajectory emanating from this position but rather a superposition of such trajectories; see Figure 3-2 and the discussion in Section 3.4.1. We show that this is not an isolated phenomenon arising from pathological data but applies even to the canonical example of one-dimensional Gaussian distributions. This limitation, together with a relatively heavy computational cost, severely hinders the deployment of P-splines in applications, ours included, especially where interpretation is a priority. We discuss these prior works in the sequel.

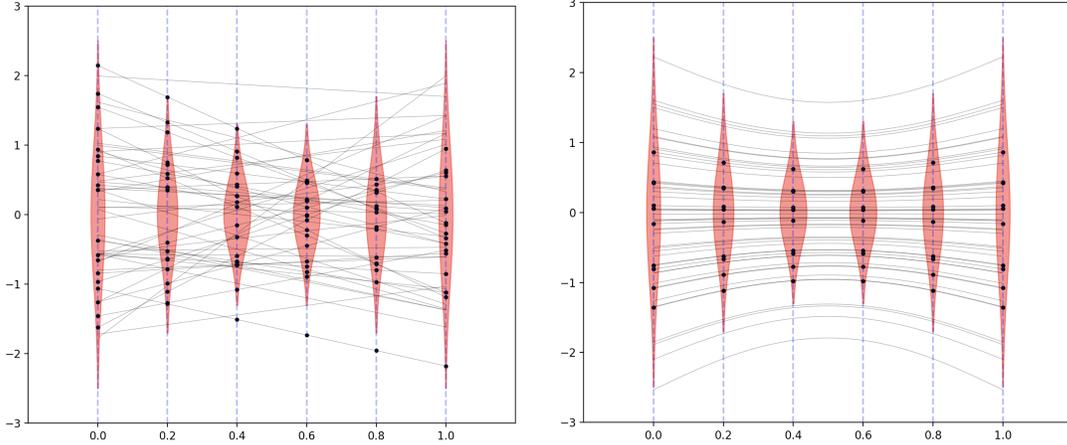


Figure 3-2: A comparison of 50 trajectories sampled from P-splines and transport splines for the Gaussian interpolation problem in Proposition 4 (see Section 3.3.1 for a detailed discussion). The first figure shows trajectories drawn from the P-spline interpolation, while the second shows trajectories from our method.

3.1.2 Contributions

To overcome the aforementioned issues, we propose in Section 3.4 a new method for constructing measure-valued splines. Our method outputs Monge solutions, and moreover enjoys significant computational advantages: it only requires N evaluations of Monge maps and standard Euclidean cubic spline fitting to output trajectories. In the case where all of the measures are Gaussian, our approach is more interpretable and scalable than the SDP-based approach of [15].

In particular, for Gaussian measures, our method only requires one $d \times d$ matrix inversion and $O(1)$ multiplications per sample point $\mu_{t_i}^*$. In comparison, the method of [15] solves an SDP with N coupled $4d \times 4d$ matrix variables. In the general case we still only need to perform N pairwise OT computations, which can be done efficiently [1], while the competing algorithms in [7] require time exponential in either N or d .

Our new method comes with a theoretical study of its approximation error. In the Gaussian setting, we introduce new techniques for studying quantitative approximation of transport maps and vector fields. In turn, it yields an approximation guarantee analogous to the classical setting (Theorem 11), but adapted to the geometry of the space. This paves the way for a principled theory of approximation on Wasserstein

space that mirrors classical Euclidean results. In a forthcoming work, we build upon these ideas to develop higher-order approximation schemes.

A key feature of our approach is its flexibility, which allows us to easily extend our method to fitting thin-plate splines for measures indexed by high-dimensional covariates. We study the case of two-dimensional spatial covariates in Section 3.6.

3.2 Background on Splines

We first recall the definition of *natural cubic splines* in Euclidean space. Given points (x_0, x_1, \dots, x_N) in \mathbb{R}^d to interpolate at a sequence of times $0 = t_0 < t_1 < \dots < t_N = 1$, consider the variational problem

$$\min_{(\gamma_t)} \int_0^1 |\ddot{\gamma}_t|^2 dt \text{ s.t. } \gamma_{t_i} = x_i \text{ for all } i. \quad (3.1)$$

It is classical that the solution to this minimization problem is a piece-wise cubic polynomial that is globally \mathcal{C}^2 and has zero acceleration at times $t_0 = 0$ and $t_N = 1$.

Based on this energy-minimizing property, there is a natural generalization of cubic splines to Riemannian manifolds: in (3.1) the acceleration $\ddot{\gamma}$ is replaced with its Riemannian analogue, the second covariant derivative $\frac{\mathbf{D}^2}{dt^2}\gamma = \nabla_{\dot{\gamma}}\dot{\gamma}$, and the norm is given by the Riemannian metric. However, unlike its Euclidean counterpart, there is no general algorithm to fit Riemannian cubic splines, leading to alternative proposals [29]. Regardless, in analogy with the Riemannian setting, it is natural to define *energy splines* (*E-splines* in short) via

$$\inf_{(\mu_t, v_t)} \int_0^1 \left\| \frac{\mathbf{D}}{dt} v_t \right\|_{\mu_t}^2 dt \text{ s.t. } \mu_{t_i} = \mu_i \text{ for all } i \quad (3.2)$$

where the minimization is taken over all curves (μ_t) and their tangent vectors (v_t) , and the measures μ_i are the fixed measures to interpolate. This definition appeared simultaneously in (Chen et al and Benamou et al).

This is an intuitively appealing definition of an interpolating curve, but there is, at present, no way of computing (3.2). Thus, (Chen et al and Benamou et al) also

introduced a relaxation which we call *path splines* (*P-splines* in short):

$$\inf_{P \in \mathcal{P}(\Omega)} \int \int_0^1 |\ddot{X}|^2 dt dP(X) \text{ s.t. } (X_{t_i})_{\#} P = \mu_i \quad (3.3)$$

where Ω is the set of twice-differentiable paths $X: [0, 1] \rightarrow \mathbb{R}^n$, and X_t is the evaluation map $X_t(\omega) = \omega_t$. We may equally well rewrite (3.2) as

$$\inf_{(X_t)} \mathbb{E} |\ddot{X}|^2 \text{ s.t. } X_{t_i} \sim \mu_{t_i}^*$$

which emphasizes the stochastic process rather than a path measures; this is the perspective we take in some of our proofs, and we refer to the two interchangeably. This is indeed a relaxation of (3.2), as proved in (Chen et al and Benamou et al). Specifically, if (μ_t, v_t) is feasible for (3.2), letting X_t be the flow maps defined by $X'_t = v_t(X_t)$ and $X_0 = \text{Id}$, and defining P to place mass $\mu_0(x)$ on the path $t \mapsto X_t(x)$, then by definition these flow maps govern this curve of measures and

$$\mathbf{E}[(\mu_t)] = \mathbf{P}[P]$$

where \mathbf{E} is the E-spline cost in (3.2) and \mathbf{P} is the cost in (3.3). Thus (3.3) \leq (3.2). As we will show, this inequality is strict, even for very well-behaved classes of measures.

It is instructive to compare P-splines to the discussion of geodesics in section 2.2.3. Proposition 2 reduces Wasserstein geodesics to distributions of geodesics in the underlying space; while the same procedure for curvature-minimizing paths does not lead to a tight characterization, as we mentioned above, it is nevertheless profitable to consider.

The program (3.3) can be reduced to a multimarginal optimization problem as follows. Let $c(x_0, \dots, x_N) = \int_0^1 |\ddot{S}_t[x_0, \dots, x_N]|^2 dt$, where $S_t[x_0, \dots, x_N]$ is the (natural) Euclidean spline interpolating $(t_0, x_0), \dots, (t_N, x_N)$, with the times kept implicit. Then Chen et al prove

Proposition 3. *The P-spline problem is equivalent to*

$$\min_{\pi} \int c d\pi \text{ s.t. } \pi_{x_i} = \mu_i \tag{3.4}$$

where π_{x_i} is the i th marginal. If π^* is a solution to (3.4) then the measure P placing mass $\pi^*(x_0, \dots, x_N)$ on the cubic spline $S_t[x_0, \dots, x_n]$ is optimal for (3.3), and any optimal solution of (3.3) induces thusly an optimal solution of (3.4).

This simply says that the optimal P-spline measure will be supported on Euclidean splines.

Unfortunately, though solvable in principle, (3.4) remains difficult to compute, scaling exponentially in N , and its solution is not necessarily induced by a deterministic map; that is, there is no guarantee of a deterministic function $\varphi_t: \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $X_t = \varphi_t(X_0)$. This point is particularly problematic for inference of trajectories as illustrated in Figure 3-2.

3.3 Results on P-splines and E-splines

Given the various definitions of splines, some natural questions arise, some raised in [7] and [15]. Specifically, these papers left open the question of whether E-splines coincide with P-splines, and whether the solution to the P-spline problem is necessarily induced by Monge maps. This section resolves these questions in the negative.

Proposition 4 (informal). *There exist non-degenerate Gaussian data $\mu_{t_0}^*, \dots, \mu_{t_N}^*$ such that there is a unique jointly Gaussian solution to the P-spline problem (3.3) and it is not induced by a deterministic map.*

Proposition 5 (informal). *There exist non-degenerate Gaussian data $\mu_{t_0}^*, \dots, \mu_{t_N}^*$ for which the E-spline (3.2) and P-spline (3.3) interpolations do not coincide.*

These propositions require a sequence of lemmas, some interesting in their own right, which we collect in separate sections for organizational purposes.

3.3.1 Proof of Proposition 4

We begin by remarking that in general, there is no reason to expect that solutions of the P-spline problem (3.3) are deterministic. Indeed, consider the following.

Proposition 6. *Let μ_0^* and μ_1^* be any probability measures. Then, any coupling (X_0, X_1) of the two measures induces an optimal P-spline solution (X_t) to (3.3) with data μ_0^* and μ_1^* .*

Proof. Indeed, simply set $X_t := (1 - t)X_0 + tX_1$. Since $t \mapsto X_t$ is a line traversed at constant speed, it incurs zero P-spline cost and is therefore optimal for (3.3). \square

As this example shows, the P-spline problem with two measures is quite degenerate; in particular, it does not recover the W_2 geodesic joining μ_0 to μ_1 , and X_1^* is not guaranteed to be a deterministic function of X_0^* . A slight modification of this simple example yields:

Proposition 7. *Let μ_0^* be any absolutely continuous measure. Then, there exist absolutely continuous data $(\mu_{i/N}^*)_{i=1}^N$ and an optimal solution (X_t) to the P-spline problem (3.3) for $(\mu_{i/N}^*)_{i=0}^N$ such that X_1 is not a deterministic function of X_0 .*

Proof. Indeed, let $T, \bar{T} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be two mappings which are μ_0^* -a.e. distinct, i.e., $T \neq \bar{T}$. Draw $X_0 \sim \mu_0^*$. Then, we either set $X_t = (1 - t)X_0 + tT(X_0)$ or else $X_t = (1 - t)X_0 + t\bar{T}(X_0)$ with probability 1/2 each (with the choice being made independently of the draw of X_0). Set $\mu_{i/N}^* := \text{law}(X_{i/N})$.

By construction, the marginals of the process (X_t) at times $0, 1/N, \dots, 1$ do indeed interpolate the data. Also, since $t \mapsto X_t$ is a straight line traversed at constant speed, then (X_t) incurs zero P-spline cost and is optimal for (3.3)

Since T and \bar{T} are distinct, X_1 is not a deterministic function of X_0 . Also, the mappings T and \bar{T} can easily be chosen to make the data all absolutely continuous (e.g., by taking them to be gradients of uniformly convex functions; c.f. the proof of [58, Proposition 5.9]). \square

(Compare this with Proposition 7 and the subsequent remark in [7].)

We next turn towards the Gaussian case. As noted, the P-spline problem is equivalent to the problem (3.4) and the optimal stochastic process X_t for the P-spline problem is supported on natural cubic splines. Furthermore, the function c in (3.4) is quadratic in x_0, \dots, x_N , so the cost $\int c d\pi$ depends only on the mean and covariance matrix of π . Thus, if the data $\mu_{t_i}^*$ is Gaussian, for any coupling π there is a jointly Gaussian coupling $\tilde{\pi}$ with the same mean and covariance, and thus the same cost $\int c d\tilde{\pi}$. This $\tilde{\pi}$ then induces a Gaussian process which is optimal for (3.3) — in specific, the law of X_t is Gaussian for all t . Thus it is natural to restrict ourselves to solutions of (3.3) that are Gaussian processes. We call this a *Gaussian solution* to (3.3). We now state the counterexample that proves Proposition 4.

Proposition 8. *Assume $N > 1$. For $i = 0, \dots, N$, let $\mu_{t_i}^* = \mathcal{N}(0, (1 - t_i)^2 + t_i^2)$. Then there is a unique Gaussian solution to the P-spline problem (3.3) and it is not induced by a deterministic map.*

Proof. The key observation is that the marginals $\mu_{t_i}^*$ arise from the curve of measures formed as the law of $X_t^* := (1 - t)X_0^* + tX_1^*$ for independent standard Gaussians X_0^* and X_1^* . If we consider the distribution on paths which is the law of (X_t^*) , then it is supported on straight lines traversed at constant speed and so it must be optimal for the P-spline problem (3.3), having zero objective value.

Consider some other stochastic process (X_t) such that the law of $(X_{t_i})_{i=0}^N$ is jointly Gaussian. For (X_t) to be an optimal solution to the P-spline problem (3.3), it must also have zero objective value and hence be supported on straight lines almost surely. Thus, we must have $X_t = (1 - t)X_0 + tX_1$. By the marginal constraints we have $\mathbb{E}[X_0^2] = \mathbb{E}[X_1^2] = 1$ and so long as $N > 1$, for $i = 1, \dots, N - 1$, it holds that $t_i \notin \{0, 1\}$ and

$$\begin{aligned} (1 - t_i)^2 + t_i^2 &= \mathbb{E}[\left((1 - t_i)X_0 + t_iX_1\right)^2] \\ &= (1 - t_i)^2 + t_i^2 + 2t_i(1 - t_i)\mathbb{E}[X_0X_1]. \end{aligned}$$

Therefore $\mathbb{E}[X_0X_1] = 0$ and (X_t) has the same distribution as (X_t^*) . Consequently, the unique jointly Gaussian solution to the P-spline problem is (X_t^*) . Clearly, the

path (X_t^*) is not a deterministic function of X_0^* . Indeed, X_1^* is *independent* of X_0^* . \square

Note that the uniqueness assertion is false when $N = 1$, highlighting the degeneracy of the case of two marginals.

3.3.2 Proof of Proposition 5

Understanding E-splines requires a few technical results, which we first collect before moving on to the proof. We remark that, prior to this work, little was known about E-splines. In particular, it was not known whether the E-spline interpolation of Gaussian measures consists only of Gaussian measures.

Throughout, it will be convenient to consider the E-spline problem over the closed convex set of curves taking values in a closed convex set K of a Hilbert space:

$$\min_{\gamma: [0,1] \rightarrow K} \int_0^1 \|\ddot{\gamma}(t)\|^2 dt \quad \text{s.t.} \quad \gamma(t_i) = x_i \quad \text{for all } i \quad (\text{E}_K)$$

Denote by $E[\gamma] = \int_0^1 \|\ddot{\gamma}(t)\|^2 dt$ the objective function in (E_K) . It follows from the triangle inequality and strict convexity of the function $x \mapsto x^2$ that E is strictly convex on the convex set of admissible curves, so the solution must be unique if it exists. We denote this unique solution by γ_K .

Proposition 9. *Let H be a Hilbert space, and let $L \subseteq H$ be a closed linear subspace. Take points $x_0, \dots, x_N \in L$. Then the solution γ_H of the E-spline problem (E_H) on H satisfies $\gamma_H(t) = \gamma_L(t) \in L$ for all t .*

Proof. Let P be the orthogonal projection onto L , and suppose γ interpolates the points $(x_i)_{i=0}^N$. Then for any admissible curve $\gamma(t) = P\gamma(t) + (I - P)\gamma(t)$, so $\ddot{\gamma}(t) = P\ddot{\gamma}(t) + (I - P)\ddot{\gamma}(t)$ as well. Since these two terms are orthogonal, we have

$$\|\ddot{\gamma}(t)\|^2 = \|P\ddot{\gamma}(t)\|^2 + \|(I - P)\ddot{\gamma}(t)\|^2.$$

Thus, on the one hand, if $\bar{\gamma}(t) = P\gamma_H(t)$ then $E[\bar{\gamma}] \leq E[\gamma_H]$, and $\bar{\gamma}$ is interpolating because $x_i \in L$. On the other hand, $E[\gamma_H] \leq E[\gamma_L] \leq E[\bar{\gamma}]$ and by uniqueness,

$$\gamma_H = \gamma_L. \quad \square$$

Proposition 10. *Let K be a convex subset of a Hilbert space H whose span is closed, and let $x_1, \dots, x_n \in K$. If $\gamma_K(t)$ lies in the relative interior of K for all times t , then $\gamma_K = \gamma_H$.*

Proof. Let L be the linear span of K , which is closed. In light of Proposition 9, it suffices to prove that $\gamma_K = \gamma_L$ so replacing H by L we may assume that K is of full dimension.

Let $f: [0, 1] \rightarrow H$ be a twice differentiable perturbation such that $f(t_i) = 0$ for all i . Hence, $\gamma_K + \epsilon f$ is admissible for (E_H) . Since γ_K lies in the interior of K and K is full-dimensional, a standard compactness argument shows that for any such f there exists an $\epsilon > 0$ with $\gamma_K(t) + \epsilon f(t) \in K$ for all t . By optimality of γ_K we then have $E[\gamma_K + \epsilon f] \geq E[\gamma_K]$. Thus γ_K is stationary for E considered on H , and because E is strictly convex it follows that γ_K is optimal for (E_H) and is therefore equal to γ_H by uniqueness. \square

Proposition 11. *Let $\mu^*t_0, \mu^*t_1, \dots, \mu^*t_N$ be Gaussian measures on \mathbb{R} . Consider the Gaussian version of the E -spline problem on \mathbb{R} :*

$$\min_{(\gamma_t)} \int_0^1 \|\nabla_{v_t} v_t\|_{L^2(\gamma_t)}^2 dt \quad \text{s.t.} \quad \gamma_{t_i} = \mu^*t_i, \quad i = 1, \dots, N$$

where the minimization is taken over curves (γ_t) of Gaussian measures with their corresponding tangent vectors $v_t \in T_{\gamma_t} \mathcal{P}_2(\mathbb{R})$. That is, it is the Wasserstein E -spline problem (3.2) in $\mathcal{P}_2(\mathbb{R})$ with the added constraint that the measures are Gaussian. If there is an optimal solution (γ_t^*) which is a non-degenerate Gaussian for all time, then it is also the solution to the E -spline problem (3.2).

Proof. It is known that $\mathcal{P}_2(\mathbb{R})$ is isometric to a closed convex subset S of the Hilbert space $H = L^2[0, 1]$ (see the discussion following Lemma 9.1.4 in [3]). This isometry is given by $\mu \mapsto F_\mu^\dagger$, where F_μ^\dagger denotes the quantile function of μ . Let K be the image of the mean-zero Gaussian measures under this isometry; it is immediate that K is convex, since the Gaussian measures form a geodesically convex set in $\mathcal{P}_2(\mathbb{R})$, and it

has closed span because it is finite-dimensional. In light of this isometry the E-spline problem (3.2) is equivalent to (E_S) while the Gaussian E-spline problem stated in the proposition is equivalent to (E_K) and $\gamma^* = \gamma_K$ (the preservation of E-splines under isometry is discussed in Section 3.3.2).

Applying Proposition 10 to $\gamma^* = \gamma_K$, we deduce that $\gamma^* = \gamma_H$. Moreover, $E[\gamma_H] \leq E[\gamma_S] \leq E[\gamma^*]$, whence by uniqueness we get that $\gamma^* = \gamma_S$ as well. \square

We also require a technical lemma regarding P-splines which remain Gaussian for all times, which follows from considerations of several-variable complex functions.

Lemma 1. *Let (μ_t) be a P-spline with initial and final data μ_0 and μ_1 which are Gaussian, and assume:*

1. μ_t is a Gaussian distribution for all times $t \in [0, 1]$,
2. (μ_t) has zero cost for the P-spline objective.

Then (μ_t) is induced by a jointly Gaussian coupling of μ_0 and μ_1 .

Proof. Since (μ_t) has zero cost it must be supported on straight lines, so if we let $X_t \sim \mu_t$ where these are coupled according to the (μ_t) coupling, then

$$X_t = (1 - t)X_0 + tX_1 \tag{3.5}$$

and by assumption this variable is Gaussian. Let Z be the Gaussian with the same covariance structure as X . Scaling (3.5) by a positive constant, we get, for all $a, b \geq 0$

$$\langle (a, b), X \rangle \stackrel{d}{=} \langle (a, b), Z \rangle$$

where we mean equality in distribution. This implies

$$\varphi_X(a, b) = \varphi_Z(a, b)$$

where φ_Y denotes the characteristic function of Y and is defined by $\varphi_Y(z) = \mathbb{E}[e^{i\langle z, Y \rangle}]$. Now, it is well-known that if $\mathbb{E}e^{m|Y|} < \infty$ for some $m > 0$ then φ_Y continues to a

holomorphic function in the strip $\{z \mid |\operatorname{Im} z_i| < m \ \forall i\}$ [37, Theorem 2.7.1]. In particular, if Y has sub-Gaussian tails, ϕ_Y is entire.

Functions of several complex variables admit an identity theorem, similar to the univariate complex case, which can be found in [50, Remark 1.20].² This is:

Theorem (identity theorem). *Let f and g be holomorphic functions of several complex variables in a domain $\Omega \subseteq \mathbb{C}^d$, and let $z \in \Omega$. A real cube of radius r about z is defined as*

$$\{(z_1 + x_1, \dots, z_d + x_d) \in \mathbb{C}^d \mid |\Re x_i| < r \text{ for } i = 1, \dots, d\}.$$

If f and g agree on a real cube of positive radius about z , then $f \equiv g$ on all of Ω .

Now, X has sub-Gaussian tails. Indeed,

$$M_X(t) = \mathbb{E}e^{\langle t, X \rangle} = \mathbb{E}e^{t_1 X_0 + t_2 X_1} \leq (\mathbb{E}e^{2t_1 X_0} \mathbb{E}e^{2t_2 X_1})^{1/2} = e^{t_1^2 \operatorname{var} X_0 + t_2^2 \operatorname{var} X_1}$$

where M_X denotes the moment generating function of X . Thus φ_X is entire, along with φ_Z , and it is clear from the above discussion that they agree on the real cube about $z = (1, 1)$ with radius $r = 1$. The identity theorem then implies that $\varphi_X \equiv \varphi_Z$, so $X \stackrel{d}{=} Z$. Thus X is jointly Gaussian. \square

Proposition 5 is implied by the following result.

Proposition 12. *For $i = 0, \dots, N$, let $\mu_{t_i}^* = \mathcal{N}(0, \sigma_{t_i}^2)$, where $\sigma_t^2 = (1-t)^2 + t^2$. Then for all $N \geq 2$, the E-spline (3.2) and P-spline (3.3) interpolations do not coincide.*

Before starting the proof, we dispense with a possible source of confusion. The solution to the P-spline problem (3.3) is a stochastic process (X_t) ; on the other hand, the E-spline solution yields a natural stochastic process, namely the (X_t^*) induced by the continuity equation with the intrinsic tangent velocity fields. In the proposition,

²The careful reader will note that the hypothesis of this theorem is much stronger than the single-variable requirement that f and g agree merely on a set with an accumulation point. For several complex variables this is not sufficient; indeed, several-variable holomorphic functions never have isolated zeros.

we are not asserting that the processes (X_t) and (X_t^*) are different (indeed this is an easier statement to prove since the P-spline solution is often not even deterministic; see Section 3.3.1). Instead, we are asserting that the *interpolated measures* associated with the E- and P-splines are different, which is strictly stronger statement.

Proof. First, the manifold of mean-zero Gaussian measures on \mathbb{R} equipped with the W_2 metric is isometric to the ray $[0, \infty)$ equipped with the standard Euclidean metric. Indeed, we have

$$W_2(\mathcal{N}(0, \sigma_0^2), \mathcal{N}(0, \sigma_1^2)) = |\sigma_0 - \sigma_1|.$$

Suppose we have data $\mu^* t_i = \mathcal{N}(0, \sigma_i^2)$ at times t_i and let $t \mapsto \gamma(t)$ be the Euclidean spline interpolation of $(t_i, \sigma_i)_{i=0}^N$ on \mathbb{R} . It is possible that $\gamma(t) \leq 0$ at some t , but if $\gamma(t) > 0$ for all t , then by Proposition 10 it must also be the spline considered on the ray $[0, \infty)$. Since covariant derivatives are preserved under isometry (see 3.3.2 for a formal verification in our setting), the function $E[\cdot]$ is also preserved under isometry, and so its minimizers — E-splines — are preserved as well. This means that the Gaussian-constrained E-spline is

$$\mu_t^E = \mathcal{N}(0, \gamma(t)^2), \quad t \in [0, 1],$$

and by Proposition 11 this must coincide with the Wasserstein E-spline (3.2). This is all under the hypothesis that $\gamma(t) > 0$.

Now substitute our example, with $\sigma_i^2 = (1 - t_i)^2 + t_i^2$. We need to check that $\gamma(t)$ remains strictly positive for all times. From [31, Theorem 5], we see that for all t

$$|\gamma(t) - \sqrt{t^2 + (1 - t)^2}| \leq \frac{5}{384} \cdot 24\sqrt{2} \cdot \frac{1}{N^4}.$$

For $N \geq 2$ this is less than 0.03. The smallest value of $\sqrt{t^2 + (1 - t)^2}$ is $\sqrt{1/2} \approx 0.7071$, so the spline is bounded below by 0.704 for all times.

Let (μ_t^P) be an interpolating P-spline. It is possible that this is not unique, but if μ_t^P is not Gaussian for some t then we are done, since μ_t^E is Gaussian by Proposition 11. Applying Lemma 1, we see that μ_t^P must be induced by a jointly Gaussian coupling

of μ_0^* and μ_1^* , so by Proposition 8 it must be that $\mu_t^P = \mathcal{N}(0, (1-t)^2 + t^2)$.

The standard deviation of μ_t^E is $\gamma(t)$ and this is locally a cubic polynomial in t . The standard deviation of the P-spline μ_t^P , however, is given by $\sqrt{(1-t)^2 + t^2}$, which cannot be locally represented by a polynomial, so they must differ. \square

From the final steps of our proof, we see that (in the Gaussian case) P-splines and E-splines will most likely differ generically, since their interpolated variances are polynomial splines of different orders.

Preservation of Splines Under Isometry

In this section, we give a formal³ verification of the assertion that the E-spline functional is preserved under the isometry between $\mathcal{P}_2(\mathbb{R})$ and its image in $H = L^2[0, 1]$. Formally, this assertion can be viewed as a manifestation of a classical fact from Riemannian geometry: the covariant derivative (associated with the Levi-Civita connection) depends only on the Riemannian metric, and is thus preserved under isometries.⁴

In the derivation below, we make all necessary regularity assumptions (e.g., we can assume that the measures are compactly supported) in order to convey the intuition. Suppose (μ_t) is a curve of measures in $\mathcal{P}_2(\mathbb{R})$ and let $v_t \in T_{\mu_t}\mathcal{P}_2(\mathbb{R})$ be the corresponding tangent vectors. If F_μ denotes the CDF of μ , then (2.5) implies

$$\partial_t F_{\mu_t}(x) = \partial_t \int_{-\infty}^x d\mu_t = - \int_{-\infty}^x (\mu_t v_t)' = -\mu_t(x)v_t(x).$$

³The word *formal* here, meaning that the argument proceeds by manipulating the *form* of the expressions, is not a synonym for “rigorous”.

⁴In fact, this is related to Gauss’s famous *Theorema Egregium*, see [27, §4.3] and [26, Remark 2.7].

Next, if we differentiate the relation $F_{\mu_t}^{-1}(F_{\mu_t}(x)) = x$, we obtain

$$\begin{aligned} 0 &= (\partial_t F_{\mu_t}^{-1})(F_{\mu_t}(x)) + (F_{\mu_t}^{-1})'(F_{\mu_t}(x)) \\ &= (\partial_t F_{\mu_t}^{-1})(F_{\mu_t}(x)) + \frac{1}{F_{\mu_t}'(x)} \\ &= (\partial_t F_{\mu_t}^{-1})(F_{\mu_t}(x)) + \frac{1}{\mu_t(x)}, \end{aligned}$$

where we have applied the inverse function theorem. Thus,

$$(\partial_t F_{\mu_t}^{-1})(\alpha) = v_t(F_{\mu_t}^{-1}(\alpha)) \tag{3.6}$$

Differentiating again,

$$\begin{aligned} (\partial_t^2 F_{\mu_t}^{-1})(\alpha) &= (\partial_t v_t)(F_{\mu_t}^{-1}(\alpha)) + v_t'(F_{\mu_t}^{-1}(\alpha))(\partial_t F_{\mu_t}^{-1})(\alpha) \\ &= (\partial_t v_t + v_t'v_t)(F_{\mu_t}^{-1}(\alpha)) \end{aligned}$$

However, we recognize $\partial_t v_t + v_t'v_t$ as the covariant derivative $\nabla_{v_t} v_t$ in $\mathcal{P}_2(\mathbb{R})$ (see for example the discussion in [15, §5.1]). In particular, it implies

$$\begin{aligned} \int_0^1 |\partial_t^2 F_{\mu_t}^{-1}|^2 &= \int_0^1 |(\partial_t v_t + v_t'v_t) \circ F_{\mu_t}^{-1}|^2 \\ &= \int |\partial_t v_t + v_t'v_t|^2 d\mu_t \\ &= \|\nabla_{v_t} v_t\|_{L^2(\mu_t)}^2 \end{aligned}$$

where we use the fact that the pushforward of the uniform distribution on $[0, 1]$ under $F_{\mu_t}^{-1}$ is μ_t . This equation shows that the norm (measured in H) of the acceleration of the curve $t \mapsto F_{\mu_t}^{-1}$ in H is the same as the norm (measured in $\mathcal{P}_2(\mathbb{R})$) of the acceleration of the curve $t \mapsto \mu_t$ in $\mathcal{P}_2(\mathbb{R})$, and thus the E-spline cost functional is preserved by the embedding $\mathcal{P}_2(\mathbb{R}) \hookrightarrow H$.

From the equation (3.6), we can also read off the isometry between the tangent space of H and the tangent space of $\mathcal{P}_2(\mathbb{R})$.

The reader who is uncomfortable with the formal derivation above can instead use

the isometric embedding $\mathcal{P}_2(\mathbb{R}) \hookrightarrow L^2[0, 1]$ as the definition of the geometry of $\mathcal{P}_2(\mathbb{R})$ (and thus, the definition of E-splines on $\mathcal{P}_2(\mathbb{R})$). Indeed, a rigorous development of second-order calculus on Wasserstein space faces significant technical hurdles [28], and such a definition is actually more convenient for the purposes of this paper.

3.4 Transport Splines

To address the difficulties discussed in the previous section, we propose a new method for measure interpolation, which we call *transport splines*. Our framework decouples the interpolation problem into two steps:

1. Couple the given measures, that is, construct a random vector $(X_{t_0}, X_{t_1}, \dots, X_{t_N})$ with the specified marginal laws $\mu_{t_0}^*, \mu_{t_1}^*, \dots, \mu_{t_N}^*$.
2. Apply a Euclidean interpolation algorithm to the points $X_{t_0}, X_{t_1}, \dots, X_{t_N}$.

A convenient choice for the second step is to use cubic splines, but our framework works equally well with other standard Euclidean methods and can be adapted to the application at hand. We illustrate this point in Section 3.6, where we construct surfaces interpolating one-dimensional measures using thin-plate splines.

A simple and practical choice for the first step, which we explore in the present paper, is to couple the random variables $X_{t_0}, X_{t_1}, \dots, X_{t_N}$ successively using the Monge maps between them. That is, we draw $X_{t_0} \sim \mu_{t_0}^*$, and for each $i = 1, \dots, N$ we set $X_{t_i} = T_i(X_{t_{i-1}})$, where T_i is the Monge map from $\mu_{t_{i-1}}^*$ to $\mu_{t_i}^*$. The second step then reduces to interpolating $X_{t_0}, T_1(X_{t_0}), \dots, T_N \circ \dots \circ T_1(X_{t_0})$ in Euclidean space. The interpolation property of transport splines follows readily from the definition of Monge maps since $T_i \circ \dots \circ T_1(X_{t_0}) \sim \mu_{t_i}^*$.

For the task of outputting sample trajectories from the transport spline, we summarize our method in Algorithm 1, and we display an application to the reconstruction of trajectories in a many-body physical system in Figure 3-3.

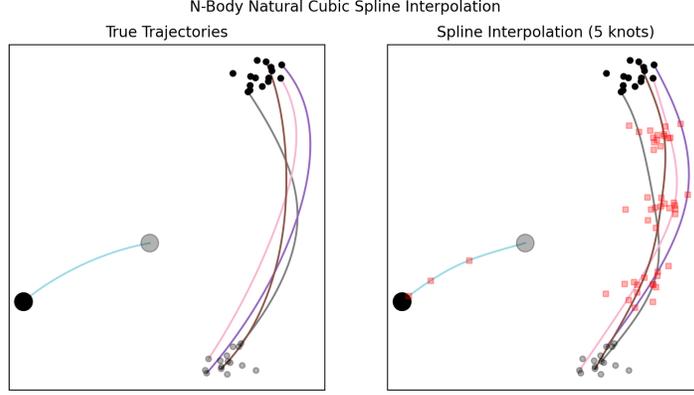


Figure 3-3: Reconstruction of trajectories in a physical system. See Appendix A.3.

Algorithm 1 Sample Transport Spline Trajectories

- 1: **procedure** INTERPOLATE($((t_i)_{i=0}^N, (\mu_{t_i}^*)_{i=0}^N)$)
 - 2: Draw $X_{t_0} \sim \mu_{t_0}^*$
 - 3: **for** $i = 1, \dots, N$ **do**
 - 4: Set $X_{t_i} = T_i(X_{t_{i-1}})$, where T_i is the Monge map from $\mu_{t_{i-1}}^*$ to $\mu_{t_i}^*$
 - 5: **end for**
 - 6: Interpolate the points $X_{t_0}, X_{t_1}, \dots, X_{t_N}$ to obtain a curve (X_t)
 - 7: **output** (X_t)
 - 8: **end procedure**
-

The choice of coupling in the first step of our method is motivated by the geometry of $\mathcal{P}_2(\mathbb{R}^d)$. If the observations $\mu_{t_0}^*, \dots, \mu_{t_N}^*$ sit along a curve of measures (μ_t^*) , then there are flow maps (X_t^*) satisfying $\dot{X}_t^* = v_t^*(X_t^*)$, where v_t are the tangent fields. Thus if $\delta = t_1 - t_0$, then $X_{t_1}^* = X_{t_0}^* + \delta v_{t_0}^*(X_{t_0}^*) + o(\delta)$. On the other hand, from (2.9) the Monge map T_1 gives a first-order approximation to $v_{t_0}^*$: $T_1 - \text{Id} = \delta v_{t_0}^* + o(\delta)$ (see [3, Proposition 8.4.6]). Combining these approximations we get $T_1(X_{t_0}^*) = X_{t_1}^* + o(\delta)$. From this heuristic discussion, one expects that as the mesh size $\max_{i \in [N]}(t_i - t_{i-1})$ tends to zero, the coupling $X_{t_0}, X_{t_1}, \dots, X_{t_N}$ obtained via successive Monge maps is a good approximation to the coupling along the flow maps $(X_{t_0}^*, X_{t_1}^*, \dots, X_{t_N}^*)$.

3.4.1 Relationship to E -Splines in One Dimension

Although E -splines are in general intractable, in the one-dimensional case it turns out that there are many situations of interest in which E -splines coincide with transport splines. Indeed, suppose that the measures $\mu_{t_0}^*, \mu_{t_1}^*, \dots, \mu_{t_N}^*$ are all one-dimensional, and for a measure μ let F_μ^\dagger denote its quantile function.⁵ Let (G_t) be the natural cubic spline in $L^2[0, 1]$ interpolating the quantile functions $F_{\mu_{t_0}^*}^\dagger, F_{\mu_{t_1}^*}^\dagger, \dots, F_{\mu_{t_N}^*}^\dagger$. Then:

Theorem 10. *Suppose that for all t , G_t is a valid⁶ quantile function. Then the transport spline and the E -spline both coincide with the curve (μ_t) , where μ_t has quantile function G_t . Furthermore, if (X_t) is the stochastic process associated with the transport spline and (X_t^*) is the flow map for the E -spline, then (X_t) and (X_t^*) have the same distribution as the law of $(G_t(U))$, where U is a uniform random variable on $[0, 1]$.*

Before moving on to the proof, we emphasize that, in light of the counterexamples above, the P -spline and E -spline are likely to differ generically, even in the Gaussian case. Therefore, it appears that the transport spline is more suitable as a relaxation of the E -spline when interpolating univariate distributions.

Proof. Let U be a uniform random variable on $[0, 1]$, and define the random variables

$$X_{t_i} := F_{\mu_{t_i}^*}^\dagger(U) \sim \mu_{t_i}^*, \quad i = 0, \dots, N.$$

These random variables are simultaneously optimally coupled, as can be seen in Section 3.6. In particular, each successive pair of these random variables is coupled via a Monge map. It follows from the definition of a transport spline that the stochastic process (X_t) associated with the transport spline can be realized as the (Euclidean) cubic spline interpolating the points $(X_{t_i})_{i=0}^N$.

Since each X_{t_i} is a function of U , so is the interpolation X_t , so we can write $X_t = \tilde{G}_t(U)$. It follows that (\tilde{G}_t) is the cubic spline in $H = L^2[0, 1]$ which interpolates

⁵Under our assumption that the measures are absolutely continuous, the quantile function F_μ^\dagger simply coincides with the inverse CDF F_μ^{-1} , but we use the quantile function notation here to reflect the general embedding $\mathcal{P}_2(\mathbb{R}) \hookrightarrow L^2[0, 1]$.

⁶A valid quantile function $G_t: [0, 1] \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is increasing and right-continuous.

the quantiles $(F_{\mu_{t_i}^*}^\dagger)_{i=0}^N$, that is, $(\tilde{G}_t) = (G_t)$. At this point, we have established one of the assertions of the theorem, namely, the explicit description of the process (X_t) associated with the transport spline.

Next, since $X_t = G_t(U)$, by hypothesis G_t is an increasing function that pushes forward the uniform distribution to the law μ_t of X_t . By the characterization of Monge maps in one dimension as increasing mappings (see again Section 3.6), it follows that $G_t = F_{\mu_t}^\dagger$.

Since (G_t) is a cubic spline, then it minimizes curvature, i.e., it solves the problem

$$\inf_{(G_t)} \int_0^1 \|\ddot{G}_t\|_{L^2[0,1]}^2 dt, \quad \text{s.t.} \quad G_{t_i} = F_{\mu_{t_i}^*}^\dagger \text{ for all } i.$$

From our characterization $G_t = F_{\mu_t}^\dagger$, it is clear that (μ_t) solves the problem

$$\inf_{(\mu_t)} \int_0^1 \|\partial_t^2 F_{\mu_t}^\dagger\|_{L^2[0,1]}^2 dt, \quad \text{s.t.} \quad \mu_{t_i} = \mu_{t_i}^* \text{ for all } i,$$

since the the first problem is a relaxation of the second (given a solution (μ_t) of the second problem, we can obtain a solution $(G_t) = (F_{\mu_t}^\dagger)$ for the first problem). Indeed, the second problem can be interpreted as the first problem with the additional constraint that the functions G_t must be quantile functions. Next, in light of the isometry described in 3.3.2, the latter problem is equivalent to

$$\inf_{(\mu_t, v_t)} \int_0^1 \|\nabla_{v_t} v_t\|_{L^2(\mu_t)}^2 dt, \quad \text{s.t.} \quad \mu_{t_i} = \mu_{t_i}^* \text{ for all } i,$$

where the infimum is taken over curves (μ_t) in $\mathcal{P}_2(\mathbb{R})$ and their corresponding tangent vectors (v_t) . This problem is seen to be the E-spline problem (3.2).

We have thus shown that (μ_t) is an E-spline. Actually, in light of Proposition 9 and the fact that (G_t) is the spline in H , then the E-spline is unique. Thus, the E-spline and transport spline coincide.

Finally, it remains to show that the flow map coupling (X_t^*) associated with the E-spline has the same law as (X_t) . For this, we can simply appeal to the embedding $\mathcal{P}_2(\mathbb{R}) \hookrightarrow H$ again. Indeed, since $\dot{X}_t = \partial_t F_{\mu_t}^\dagger(U)$, the calculation in 3.3.2 shows that

$\dot{X}_t = v_t(X_t)$ where (v_t) is the tangent vector to (μ_t) , so in fact (X_t) is the flow map coupling of (μ_t) . \square

3.5 The Gaussian Case

We now focus on the Gaussian case and we assume that we employ natural cubic splines in Step 2 of our algorithm. For simplicity, we can assume that the measures are centered.⁷ A centered non-degenerate Gaussian can be identified with its covariance matrix, and the Wasserstein distance induces a Riemannian metric on the space of positive definite matrices. The resulting manifold is called the Bures-Wasserstein space; see [8] for a comprehensive survey.

It is known that the Monge map from Gaussian $\mathcal{N}(0, \Sigma_1)$ to $\mathcal{N}(0, \Sigma_2)$ is the linear map T given by

$$T(X) = \Sigma_1^{-1/2} (\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \Sigma_1^{-1/2} X \quad (3.7)$$

Cubic splines have the property that the interpolation evaluated at time t is a linear function of the interpolated points $(x_{t_i})_{i=0}^N$. That is, the map S_t referenced above, taking x_0, \dots, x_N to the natural cubic spline interpolant, is linear in x_0, \dots, x_N for each t .⁸ This has important consequences for our algorithm:

1. It implies that our algorithm outputs a process (X_t) such that X_t is a linear function of $X_{t_0}, X_{t_1}, \dots, X_{t_N}$. On the other hand, each X_{t_i} is a linear function of X_{t_0} , which follows from the description of Step 1 of our algorithm and the fact that Monge maps between Gaussians are linear (3.7).

Since a linear function of a Gaussian is also Gaussian, we conclude that *the transport spline interpolating Gaussian measures only passes through Gaussian measures*.

2. From the previous point, it is clear that the covariance matrix of X_t can be computed in terms of S_t, Σ_{t_0} , and the Monge maps (which have the closed-form

⁷The discussion here extends easily to incorporate non-centered measures.

⁸Note that the matrix representing S_t is independent of $(x_{t_i})_{i=0}^N$, but depends on the time grid $(t_i)_{i=0}^N$.

expression (3.7)). We conclude that in this setting, not only can we output sample trajectories as in Algorithm 1, but we can also efficiently output the covariance matrices of the interpolated measures.

Furthermore, this discussion extends to any other interpolation method with this linearity property, such as higher-order splines, polynomial interpolation, and thin-plate splines. We also remark that in the case where the data consists of *one-dimensional* Gaussian distributions, then in many cases the transport spline and the E-spline coincide.

Proposition 13. *Suppose that $\mu_{t_0}^*, \mu_{t_1}^*, \dots, \mu_{t_N}^*$ are one-dimensional Gaussians. Then, if the transport spline (μ_t) interpolating these data is never degenerate, i.e., μ_t is a non-degenerate Gaussian for each $t \in [0, 1]$, then the conditions of Theorem 10 hold.*

Proof. Since the Gaussian measures form a 2 dimensional half-subspace of $L^2[0, 1]$ with the usual identification $\mathcal{P}_2(\mathbb{R}) \hookrightarrow L^2[0, 1]$, the E-spline interpolation between Gaussian measures is the transport spline if transport splines is not degenerate at any time (i.e., the transport lies in the relative interior of Gaussian measures within $\mathcal{P}_2(\mathbb{R})$). \square

3.5.1 Failure of Equality Between Transport Splines and E-Splines

In this section we give some examples that show that E-splines and transport splines differ when the spline (G_t) described in Theorem 10 does not stay within $\mathcal{P}_2(\mathbb{R}) \subset L^2[0, 1]$. First, we give a Gaussian counterexample.

Proposition 14. *Let $\delta > 0$ be sufficiently small and consider the measures*

$$\mu_0^* = \mu_1^* = \mathcal{N}(0, 1), \quad \mu_{1/3}^* = \mu_{2/3}^* = \mathcal{N}(0, \delta^2).$$

Then, the E-spline interpolation (μ_t^E) and transport spline interpolation (μ_t^T) do not coincide for this data.

Proof. Let (X_t) denote the stochastic process corresponding to the transport spline. It is easy to see that $(X_0, X_{1/3}, X_{2/3}, X_1) = (X_0, \delta X_0, \delta X_0, X_0)$ is the optimal coupling at the knots. Recalling that S_t is the linear mapping which produces the spline, it follows that

$$X_t = S_t(X_0, \delta X_0, \delta X_0, X_0) = S_t(1, \delta, \delta, 1)X_0,$$

so that $\mu_t^\top = \mathcal{N}(0, S_t(1, \delta, \delta, 1)^2)$.

If we identify the space of Gaussians with the half-ray $[0, \infty)$, then the transport spline corresponds to the curve of standard deviations $t \mapsto |S_t(1, \delta, \delta, 1)|$. However, because the spline curve $t \mapsto S_t(1, 0, 0, 1)$ becomes negative between $1/3$ and $2/3$, then so does the curve $t \mapsto S_t(1, \delta, \delta, 1)$ for small δ . It can be checked that at time $1/3$, the curve $t \mapsto |S_t(1, \delta, \delta, 1)|$ is not \mathcal{C}^2 differentiable and therefore cannot be an E-spline. \square

This counterexample, however, is somewhat degenerate because the transport spline passes through a degenerate measure, and thus it is not clear if the E-spline exists, and if so whether it remains non-degenerate. We now give another example where the transport spline does not coincide with the E-spline, but the transport spline remains non-degenerate; hence, we believe that the E-spline problem is well-posed for these data.

For this example, we take $\delta > 0$ and let

$$\begin{aligned} \mu_0^* &= \mu_1^* = \text{uniform on } [-(1 + \delta), -1] \cup [1, 1 + \delta] \\ \mu_{1/4}^* &= \mu_{3/4}^* = \text{uniform on } [-\delta, \delta] \end{aligned} \tag{3.8}$$

As in the proof of Proposition 11, $\mathcal{P}_2(\mathbb{R})$ is seen as a convex subset of $L^2[0, 1]$ where probability measures are identified as their quantile function. Thus our E-spline interpolation can be reformulated as the problem

$$\inf_{(\mu_t)} \int_0^1 \int_0^1 |\ddot{F}_t^\dagger(u)|^2 du dt \text{ s.t. } \mu_t = \mu_t^* \text{ for all } t \in \{0, 1/4, 3/4, 1\},$$

where F_t^\dagger denotes the quantile function of μ_t . In particular, the E-spline interpolation

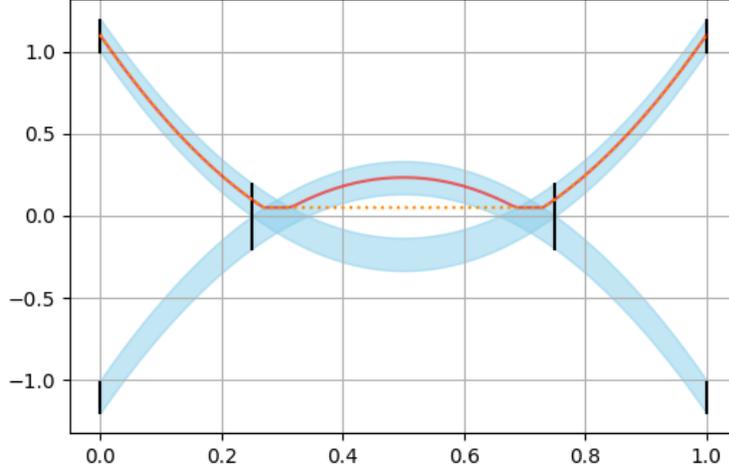


Figure 3-4: Transport splines interpolation for the four uniform distributions as in (3.8). The red line is the quantile of order 3/4 for the interpolation and the orange dotted line represents the corresponding candidate $\bar{F}_t^\dagger(u)$ for $u = 3/4$ introduced in (3.9).

problem can be seen as the transport spline interpolation with the extra constraint that the trajectories of the particles must stay ordered (see Theorem 10).

Denote by (X_t) the random process given by the transport spline problem. One can check that

$$X_t = \text{sign}(X_0) \left[\frac{16}{3}(t - 1/2)^2 - \frac{1}{3} + |X_0| - 1 \right].$$

Clearly, for δ small enough the quantiles $F_t^\dagger(u)$ of order $u > 1/2$ associated to the transport spline interpolation decrease before $t = 1/4$ and increase after $t = 3/4$. In particular, for each $u > 1/2$, there exists $1/4 < t_u^- < t_u^+ < 3/4$ such that $\partial_t F_t^\dagger(u)|_{t=t_u^-} = \partial_t F_t^\dagger(u)|_{t=t_u^+} = 0$ and $|\partial_t^2 F_t^\dagger(u)| > 0$ for $t \in (t_u^-, t_u^+)$. One can check then that the function $u \mapsto \bar{F}_t^\dagger$ at time $t \in [0, 1]$ defined by

$$\bar{F}_t^\dagger(u) = \begin{cases} F_{t_u^-}^\dagger(u), & u \in (t_u^-, t_u^+) \\ F_t^\dagger(u), & \text{otherwise} \end{cases} \quad (3.9)$$

is a quantile function. In particular, the measures with quantiles \bar{F}_t^\dagger interpolate the

measures (3.8) and

$$|\partial_t^2 \bar{F}_t^\dagger(u)| = \begin{cases} 0, & u \in (t_u^-, t_u^+) \\ |\partial_t^2 F_t^\dagger(u)|, & \text{otherwise,} \end{cases}$$

ensuring that \bar{F}_t^\dagger has a lower cost than the transport spline. Thus, the transport spline is not the E-spline.

Since the transport spline is non-degenerate for this example, we believe that the E-spline also exists and is non-degenerate. Therefore, we expect that the failure of transport splines to equal E-splines in general is not simply due to the fact that E-splines can be ill-posed.

To summarize: when the trajectories of the transport spline remain ordered throughout the interpolation, then it coincides with the E-spline. Otherwise, there is no reason to expect the two notions of spline to coincide.

3.5.2 Approximation Guarantees

Our method is the first to provide approximation guarantees on Wasserstein space. In order to obtain strong quantitative results, we focus on the Bures-Wasserstein setting detailed in the previous section, where all measures $\mu_{t_i}^*$ are centered non-degenerate Gaussian distributions.

The Bures-Wasserstein space has already been used in works such as [44, 19] as a prototypical setting in which to understand the behavior of algorithms set on the general Wasserstein space. Although the Bures-Wasserstein space is a Riemannian manifold and transport splines can in principle be studied using purely Riemannian techniques, we give proofs inspired by optimal transport so that the analysis may be more easily extended to other settings of interest.

We now state our main approximation result.

Theorem 11. *Let (μ_t^*) be a curve of measures in Bures-Wasserstein space, and let $(X_t^*) \sim (\mu_t^*)$ be the flow map coupling. Let:*

- $L := \sup_{t \in [0,1]} \|\dot{X}_t^*\|_{L^2(\mathbb{P})}$ be the Lipschitz constant of the curve, and
- $R := \sup_{t \in [0,1]} \|\ddot{X}_t^*\|_{L^2(\mathbb{P})}$ be an upper bound on its curvature, and
- λ_{\min} be a lower bound on the eigenvalues of the covariance matrices of $\mu_{t_0}^*, \mu_{t_1}^*, \dots, \mu_{t_N}^*$.

where $\|X\|_{L^2(\mathbb{P})} = \sqrt{\mathbb{E}[\|X\|^2]}$ Let (μ_t) be the cubic transport spline interpolating $\mu_{t_0}^*, \dots, \mu_{t_N}^*$ and assume

$$\alpha\delta \leq t_i - t_{i-1} \leq \delta, \quad \text{for } i = 1, \dots, N \quad (3.10)$$

where $\alpha, \delta > 0$. Then, provided that $\delta < \sqrt{\lambda_{\min}}/(2L)$, we have the following approximation guarantee:

$$\sup_{t \in [0,1]} W_2(\mu_t, \mu_t^*) \leq \frac{58}{\alpha^3} R\delta^2.$$

The proof comprises the remainder of this section. We begin by describing the general strategy. Consider the interval $[t_{i-1}, t_i]$, let (X_t^*) denote the flow map coupling for $(\mu_t^*)_t$, and let (X_t) be the stochastic process associated with the transport spline. Since $\mu_{t_{i-1}} = \mu_{t_{i-1}}^*$, we can couple the two processes together so that $X_{t_{i-1}} = X_{t_{i-1}}^*$. By the definition of the Wasserstein distance, we can bound $W_2(\mu_t, \mu_t^*) \leq \|X_t - X_t^*\|_{L^2(\mathbb{P})}$, so it suffices to show that the trajectories (X_t) and (X_t^*) are close on the interval $[t_{i-1}, t_i]$.

We will use a basic deterministic fact: if two curves x and y defined on $[0, \delta]$ are such that:

- $x(0) = y(0)$,
- $\dot{x}(0) = \dot{y}(0) + O(\delta)$, and
- the two curves satisfy the curvature bound

$$\sup_{t \in [0, \delta]} \{|\ddot{x}(t)| \vee |\ddot{y}(t)|\} \leq R,$$

then it follows that $\sup_{t \in [0, \delta]} |x(t) - y(t)| \leq CR\delta^2$, where C is a numerical constant.

1. the velocities of X_t and X_t^* at time $t = t_{i-1}$ are within $O(\delta)$ of each other (Proposition 16);
2. the trajectory (X_t) has curvature $O(R)$ (Proposition 17);
3. the trajectory (X_t^*) has curvature $O(R)$;

The last step is immediate from our assumptions; the point of the second step is to control the curvature of the interpolated process (X_t) in terms of the curvature of the true process (X_t^*) .

Putting these pieces together, we give the proof of Theorem 11.

Notation

Since we study the approximation guarantee in the Bures-Wasserstein setting, we can equivalently think in terms of the probability measure (a Gaussian), or in terms of the covariance matrix. It will be useful to employ the language of matrices, so we fix notational conventions here.

Associated with the curve (μ_t^*) , we have a corresponding curve of covariance matrices (Σ_t) such that $\mu_t^* = \mathcal{N}(0, \Sigma_t)$.

Given a matrix $A \in \mathbb{R}^{d \times d}$, we define the norm

$$\|A\|_{\Sigma} := \sqrt{\langle A, \Sigma A \rangle}$$

The norm is defined so that if $X^* \sim \mathcal{N}(0, \Sigma)$, then $\|AX^*\|_{L^2(\mathbb{P})} = \|A\|_{\Sigma}$. From our eigenvalue bound we have $\|A\|_{\Sigma} \geq \sqrt{\lambda_{\min}(\Sigma)} \|A\|_{\mathbb{F}}$.

The Monge map T between two Gaussians is the linear map $T(X)$ given in (3.7) and abusing notation slightly, we identify the map T with the corresponding matrix, and we write $T(x) = Tx$. In particular, linearity of the Monge maps implies that the velocity vector field (v_t^*) associated to the Lagrangian coupling of the curve, is also linear for each t : v_t^* is a symmetric linear mapping $\mathbb{R}^d \rightarrow \mathbb{R}^d$, that is, there exists a symmetric matrix $V_t^* \in \mathbb{R}^{d \times d}$ such that $v_t^*(x) = V_t^*x$.

Control of the Velocities

We write $\delta_i := t_{i+1} - t_i$ and $\delta := \max_{i \in [N]} \delta_i$. The first step is to prove a quantitative bound on how well the Monge map T_i approximates $\text{Id} + \delta_i v_{t_{i-1}}$. We prove a more general approximation result which may be of independent interest.

Theorem 12. *Let $t, t+h \in [0, 1]$, where $h \neq 0$. Write $\delta := |h|$ and assume $\delta \leq c\sqrt{\lambda_{\min}(\Sigma_t)}/L$, for some constant $0 < c < 1$. Let T denote the Monge map from μ_t^* to μ_{t+h}^* , and let $\bar{T} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be another linear mapping satisfying the following properties:*

1. \bar{T} can be identified with a symmetric matrix.
2. $\|\bar{T}X_t^* - X_{t+h}^*\|_{L^2(\mathbb{P})} \leq c\sqrt{\lambda_{\min}(\Sigma_t)}$.

Then,

$$\|TX_t^* - \bar{T}X_t^*\|_{L^2(\mathbb{P})} \leq \frac{1+2c}{1-c} \|\bar{T}X_t^* - X_{t+h}^*\|_{L^2(\mathbb{P})}$$

Proof. Let $e := X_{t+h}^* - \bar{T}X_t^*$.

Consider the quadratic function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by $\phi(x) := \langle x, Ax \rangle$, where $A := (T - \bar{T})/\|T - \bar{T}\|_{\Sigma_t}$. Note that A is symmetric (since T and \bar{T} are). Then,

$$\mathbb{E}\phi(TX_t^*) = \mathbb{E}\phi(X_{t+h}^*) = \mathbb{E}\phi(\bar{T}X_t^* + e)$$

Expanding this out,

$$\begin{aligned} 0 &= \mathbb{E}\langle (T + \bar{T})X_t^* + e, A\{(T - \bar{T})X_t^* - e\} \rangle \\ &= \mathbb{E}\langle (T + \bar{T})X_t^*, A(T - \bar{T})X_t^* \rangle + \text{error} \end{aligned}$$

We next bound the error term. First, note that by our assumption,

$$\begin{aligned} \|T - I_d\|_{\Sigma_t} &= W_2(\mu_t^*, \mu_{t+h}^*) \leq L\delta \leq c\sqrt{\lambda_{\min}}, \\ \|\bar{T} - I_d\|_{\Sigma_t} &\leq c\sqrt{\lambda_{\min}} \end{aligned}$$

where we write $\lambda_{\min} = \lambda_{\min}(\Sigma_t)$. The error term is split into two further terms. For the first term,

$$\begin{aligned}
|\mathbb{E}\langle e, A(T - \bar{T})X_t^* \rangle| &\leq \|e\|_{L^2(\mathbb{P})} \|A(T - \bar{T})\|_{\Sigma_t} \\
&\leq \|e\|_{L^2(\mathbb{P})} \|A\|_{\text{F}} \|T - \bar{T}\|_{\Sigma_t} \\
&\leq \|e\|_{L^2(\mathbb{P})} \frac{1}{\sqrt{\lambda_{\min}}} (\|T - I_d\|_{\Sigma_t} + \|\bar{T} - I_d\|_{\Sigma_t}) \\
&\leq 2c \|e\|_{L^2(\mathbb{P})}
\end{aligned}$$

where we used the fact that $\|A\|_{\Sigma_t} \leq 1$ implies that $\|A\|_{\text{F}} \leq 1/\sqrt{\lambda_{\min}}$. The second term is bounded by

$$\begin{aligned}
|\mathbb{E}\langle (T + \bar{T})X_t^* + e, Ae \rangle| &\leq |\mathbb{E}\langle TX_t^* + X_{t+h}^* - 2X_t^*, Ae \rangle| + 2|\mathbb{E}\langle X_t^*, Ae \rangle| \\
&\leq \{\|A\|_{\text{F}} (\|T - I_d\|_{\Sigma_t} + \|X_{t+h}^* - X_t^*\|_{L^2(\mathbb{P})}) + 2\|A\|_{\Sigma_t}\} \|e\|_{L^2(\mathbb{P})} \\
&\leq 2(1 + c) \|e\|_{L^2(\mathbb{P})}
\end{aligned}$$

where we used

$$\|X_{t+h}^* - X_t^*\|_{L^2(\mathbb{P})}^2 = \mathbb{E} \left[\left\| \int_t^{t+h} \dot{X}_s^* ds \right\|^2 \right] \leq \delta \left| \int_t^{t+h} \|\dot{X}_s^*\|_{L^2(\mathbb{P})}^2 ds \right| \leq L^2 \delta^2$$

Thus, we have

$$\begin{aligned}
2\|T - \bar{T}\|_{\Sigma_t} &= 2\mathbb{E}\langle X_t^*, A(T - \bar{T})X_t^* \rangle \\
&= -\mathbb{E}\langle (T + \bar{T} - 2I_d)X_t^*, A(T - \bar{T})X_t^* \rangle + \text{error} \\
&\leq (\|T - I_d\|_{\Sigma_t} + \|\bar{T} - I_d\|_{\Sigma_t}) \|A\|_{\text{F}} \|T - \bar{T}\|_{\Sigma_t} + \text{error} \\
&\leq 2c\|T - \bar{T}\|_{\Sigma_t} + (2 + 4c) \|e\|_{L^2(\mathbb{P})}
\end{aligned}$$

which finally yields

$$\|T - \bar{T}\|_{\Sigma_t} \leq \frac{1 + 2c}{1 - c} \|e\|_{L^2(\mathbb{P})}$$

as required. □

corollary 1. *Let $t, t + h \in [0, 1]$, where $h \neq 0$, and write $\delta := |h|$. Let $k \in \{0, 1, 2\}$, and suppose δ is small enough so that*

$$\sum_{i=1}^k \frac{R_i \delta^i}{i!} \leq c \sqrt{\lambda_{\min}(\Sigma_t)},$$

where we set $R_i := \sup_{t \in [0, 1]} \|\partial^i X^*\|_{L^2(\mathbb{P})}$. Then,

$$\left\| TX_t^* - \sum_{i=0}^k \frac{h^i}{i!} (\partial^i X^*)_t \right\|_{L^2(\mathbb{P})} \leq \frac{1 + 2c}{1 - c} \frac{R_{k+1} \delta^{k+1}}{(k+1)!}$$

Proof. We apply Theorem 12 with

$$\bar{T}X_t^* = \sum_{i=0}^k \frac{h^i}{i!} (\partial^i X^*)_t$$

Using $\dot{X}_t^* = V_t^* X_t^*$, where V_t^* is symmetric, we obtain:

$$\begin{aligned} \dot{X}_t^* &= V_t^* X_t^*, \\ \ddot{X}_t^* &= \dot{V}_t^* X_t^* + V_t^{*2} X_t^* = (\dot{V}_t^* + V_t^{*2}) X_t^*, \\ \dddot{X}_t^* &= (\dot{V}_t^* + 2\dot{V}_t^* V_t^* + V_t^* \dot{V}_t^* + V_t^{*3}) X_t^*, \\ &\vdots \end{aligned}$$

Observe that the i th derivative of $t \mapsto X_t^*$ at t is indeed a linear function of X_t^* , but for $i \geq 3$ it is no longer given by a symmetric matrix, so it no longer satisfies the first assumption of Theorem 11; this is why we restrict ourselves to $k = 0, 1, 2$.

For the third assumption of Theorem 11, note that

$$\|\bar{T}X_t^* - X_t^*\|_{L^2(\mathbb{P})} = \left\| \sum_{i=1}^k \frac{h^i}{i!} (\partial^i X^*)_t \right\|_{L^2(\mathbb{P})} \leq \sum_{i=1}^k \frac{\delta^i R_i}{i!} \leq c \sqrt{\lambda_{\min}(\Sigma_t)},$$

by our assumption on δ .

Finally, the error $e := X_{t+h}^* - \bar{T}X_t^*$ is controlled via Taylor's theorem:

$$\begin{aligned} \|e\|_{L^2(\mathbb{P})} &= \left\| X_{t+h}^* - \sum_{i=0}^k \frac{h^i}{i!} (\partial^i X^*)_t \right\|_{L^2(\mathbb{P})} \\ &= \left\| \int_t^{t+h} \frac{(\partial^{k+1} X^*)_s}{k!} (s-t)^k ds \right\|_{L^2(\mathbb{P})} \\ &\leq \frac{R_{k+1} \delta^{k+1}}{(k+1)!} \end{aligned}$$

□

One remark: if we let $\delta \searrow 0$, we can also take $c \searrow 0$, obtaining

$$\limsup_{\delta \searrow 0} \frac{1}{\delta^{k+1}} \left\| TX_t^* - \sum_{i=0}^k \frac{h^i}{i!} (\partial^i X^*)_t \right\|_{L^2(\mathbb{P})} \leq \frac{R_{k+1}}{(k+1)!}$$

Comparing this to a Euclidean Taylor expansion, this is apparently sharp.

Corollary 1 says that in order to prove our desired result $\dot{X}_{t_{i-1}} = \dot{X}_{t_{i-1}}^* + O(\delta)$, it suffices to show that $\dot{X}_{t_{i-1}} = (T_i X_{t_{i-1}} - X_{t_{i-1}})/\delta_i + O(\delta)$ (since the RHS of both expressions equals $V_{t_{i-1}}^* X_{t_{i-1}} = V_{t_{i-1}}^* X_{t_{i-1}}^*$ up to $O(\delta)$). Since the latter statement involves only the process (X_t) , it is easier to prove.

However, there is still a major difficulty to overcome: $\dot{X}_{t_{i-1}}$ is the velocity of an interpolating cubic spline, which depends on all of the interpolated points $X_{t_0}, X_{t_1}, \dots, X_{t_N}$. In Appendix B, we show that the derivative of the cubic spline interpolation can be understood in terms of the linear system of equations involving the quantities

$$\Delta_i := \frac{X_{t_{i+1}} - X_{t_i}}{\delta_{i+1}} - \frac{X_{t_i} - X_{t_{i-1}}}{\delta_i}, \quad i \in [N-1]$$

Therefore, we next control these quantities.

Proposition 15. *Assume $\delta \leq \sqrt{\lambda_{\min}}/(2L)$. For each $i \in [N-1]$, it holds that*

$$\left\| \frac{X_{t_{i+1}} - X_{t_i}}{\delta_{i+1}} - \frac{X_{t_i} - X_{t_{i-1}}}{\delta_i} \right\|_{L^2(\mathbb{P})} \leq \frac{25}{4} R \delta$$

Proof. From Corollary 1,

$$\left\| \frac{X_{t_i} - X_{t_{i-1}}}{\delta_i} - V_{t_{i-1}}^* X_{t_{i-1}} \right\|_{L^2(\mathbb{P})} = \left\| \frac{T_i - I_d}{\delta_i} - V_{t_{i-1}}^* \right\|_{\Sigma_{t_{i-1}}} \leq 2R\delta_i$$

where we use the fact that $X_{t_{i-1}} \sim \mu_{t_{i-1}}^*$ and that $X_{t_i} = T_i X_{t_{i-1}}$. Similarly,

$$\left\| \frac{X_{t_{i+1}} - X_{t_i}}{\delta_{i+1}} - V_{t_i}^* X_{t_i} \right\|_{L^2(\mathbb{P})} \leq 2R\delta_{i+1}$$

Therefore,

$$\|\Delta_i\|_{L^2(\mathbb{P})} \leq 4R\delta + \|V_{t_i}^* X_{t_i} - V_{t_{i-1}}^* X_{t_{i-1}}\|_{L^2(\mathbb{P})}$$

Since $X_{t_i} = T_i X_{t_{i-1}}$, we replace T_i by $I_d + \delta_i V_{t_{i-1}}^*$

$$\begin{aligned} & \|V_{t_i}^* X_{t_i} - V_{t_{i-1}}^* X_{t_{i-1}}\|_{L^2(\mathbb{P})} \\ & \leq \|V_{t_i}^* (T_i - I_d - \delta_i V_{t_{i-1}}^*) X_{t_{i-1}}\|_{L^2(\mathbb{P})} + \|V_{t_i}^* (I_d + \delta_i V_{t_{i-1}}^*) X_{t_{i-1}} - V_{t_{i-1}}^* X_{t_{i-1}}\|_{L^2(\mathbb{P})} \end{aligned}$$

We control the first term using Corollary 1:

$$\begin{aligned} \|V_{t_i}^* (T_i - I_d - \delta_i V_{t_{i-1}}^*) X_{t_{i-1}}\|_{L^2(\mathbb{P})} & \leq \|V_{t_i}^*\|_{\mathbb{F}} \|(T_i - I_d - \delta_i V_{t_{i-1}}^*) X_{t_{i-1}}\|_{L^2(\mathbb{P})} \\ & \leq \frac{L}{\sqrt{\lambda_{\min}}} \|T_i - I_d - \delta_i V_{t_{i-1}}^*\|_{\Sigma_{t_{i-1}}} \\ & \leq \frac{L}{\sqrt{\lambda_{\min}}} \cdot 2R\delta_i^2 \leq R\delta_i \end{aligned}$$

where we used $\|V_{t_i}^*\|_{\mathbb{F}} \leq \lambda_{\min}^{-1/2} \|V_{t_i}^*\|_{\Sigma_{t_i}} \leq L\lambda_{\min}^{-1/2}$ by our Lipschitz assumption. Now for the second term. Introduce the random trajectory (X_t^*) sampled from the true curve (μ_t^*) with the Lagrangian coupling, and couple the process (X_t) with (X_t^*) by setting $X_{t_{i-1}} = X_{t_{i-1}}^*$. Thus,

$$\begin{aligned} & \|V_{t_i}^* (I_d + \delta_i V_{t_{i-1}}^*) X_{t_{i-1}} - V_{t_{i-1}}^* X_{t_{i-1}}\|_{L^2(\mathbb{P})} \\ & \leq \|V_{t_i}^* X_{t_i}^* - V_{t_{i-1}}^* X_{t_{i-1}}^*\|_{L^2(\mathbb{P})} + \|V_{t_i}^* \{(I_d + \delta_i V_{t_{i-1}}^*) X_{t_{i-1}}^* - X_{t_i}^*\}\|_{L^2(\mathbb{P})} \end{aligned}$$

It is easy to control

$$\|V_{t_i}^* X_{t_i}^* - V_{t_{i-1}}^* X_{t_{i-1}}^*\|_{L^2(\mathbb{P})} = \left\| \int_{t_{i-1}}^{t_i} \ddot{X}_t^* dt \right\|_{L^2(\mathbb{P})} \leq R\delta_i$$

Lastly,

$$\begin{aligned} \|V_{t_i}^* \{(I_d + \delta_i V_{t_{i-1}}^*) X_{t_{i-1}}^* - X_{t_i}^*\}\|_{L^2(\mathbb{P})} &\leq \|V_{t_i}^*\|_{\mathbb{F}} \|X_{t_i}^* - X_{t_{i-1}}^* - \delta_i V_{t_{i-1}}^* X_{t_{i-1}}^*\|_{L^2(\mathbb{P})} \\ &\leq \frac{L}{\sqrt{\lambda_{\min}}} \left\| \int_{t_{i-1}}^{t_i} \int_{t_{i-1}}^t \ddot{X}_s^* ds dt \right\|_{L^2(\mathbb{P})} \\ &\leq \frac{L}{\sqrt{\lambda_{\min}}} \cdot \frac{R\delta_i^2}{2} \leq \frac{R\delta_i}{4} \end{aligned}$$

Putting it all together, we obtain

$$\|\Delta_i\|_{L^2(\mathbb{P})} \leq \frac{25}{4} R\delta$$

□

To match notation with Appendix B, we set

$$M_i := \ddot{X}_{t_{i-1}}, \quad i \in [N+1]$$

Lemma 2. *Assume $\delta \leq \sqrt{\lambda_{\min}}/(2L)$. It holds that*

$$\|M_i\|_{L^2(\mathbb{P})} \leq \frac{75(1+\alpha)^2}{4\alpha^3} R$$

Proof. As described in Appendix B, we know that $M = 6\mathbf{T}^{-1}\Delta$, where the entries of

\mathbf{T}^{-1} are bounded in Lemma 10. Thus,

$$\begin{aligned}
\|M_i\|_{L^2(\mathbb{P})} &= 6 \left\| \sum_{j=1}^{N-1} (\mathbf{T}^{-1})_{i,j} \Delta_j \right\|_{L^2(\mathbb{P})} \\
&\leq 6 \sum_{j=1}^{N-1} |(\mathbf{T}^{-1})_{i,j}| \|\Delta_j\|_{L^2(\mathbb{P})} \\
&\leq 6 \sum_{j=1}^{N-1} \frac{1}{4\alpha^2\delta} \frac{1}{(1+\alpha)^{|i-j|-1}} \frac{25}{4} R\delta \\
&\leq \frac{75R}{4\alpha^2} \sum_{k=0}^{\infty} \frac{1}{(1+\alpha)^{k-1}} = \frac{75(1+\alpha)^2}{4\alpha^3} R
\end{aligned}$$

where we use Proposition 15. □

Finally, we are ready to state our control on the velocity of the trajectory (X_t) .

Proposition 16. *Assume $\delta \leq \sqrt{\lambda_{\min}}/(2L)$. Then,*

$$\|\dot{X}_{t_{i-1}} - \dot{X}_{t_{i-1}}^*\|_{L^2(\mathbb{P})} \leq \frac{16\alpha^3 + 75(1+\alpha)^2}{8\alpha^3} R\delta$$

Proof. It holds that

$$\dot{X}_{t_{i-1}} = \frac{X_{t_i} - X_{t_{i-1}}}{\delta_i} - \frac{M_{i+1} + 2M_i}{6} \delta_i$$

(see Appendix B). Therefore,

$$\left\| \dot{X}_{t_{i-1}} - \frac{X_{t_i} - X_{t_{i-1}}}{\delta_i} \right\|_{L^2(\mathbb{P})} \leq \frac{\|M_{i+1}\|_{L^2(\mathbb{P})} + 2\|M_i\|_{L^2(\mathbb{P})}}{6} \delta \leq \frac{75(1+\alpha)^2}{8\alpha^3} R\delta$$

by Lemma 2. Next, we recall that $X_{t_i} = T_i X_{t_{i-1}}$, and that (X_t) and (X_t^*) are coupled so that $X_{t_{i-1}} = X_{t_{i-1}}^*$. Thus,

$$\begin{aligned}
\|\dot{X}_{t_{i-1}} - \dot{X}_{t_{i-1}}^*\|_{L^2(\mathbb{P})} &\leq \left\| \dot{X}_{t_{i-1}} - \frac{T_i X_{t_{i-1}} - X_{t_{i-1}}}{\delta_i} \right\|_{L^2(\mathbb{P})} + \left\| \dot{X}_{t_{i-1}}^* - \frac{T_i X_{t_{i-1}}^* - X_{t_{i-1}}^*}{\delta_i} \right\|_{L^2(\mathbb{P})} \\
&\leq \frac{75(1+\alpha)^2}{8\alpha^3} R\delta + 2R\delta
\end{aligned}$$

where we invoke Corollary 1 again. \square

Curvature of the Transport Spline

Next, we must bound the curvature of (X_t) , but this is an easy task given what we have established so far.

Proposition 17. *Assume $\delta \leq \sqrt{\lambda_{\min}}/(2L)$. Then,*

$$\sup_{t \in [0,1]} \|\ddot{X}_t\|_{L^2(\text{Pr})} \leq \frac{75(1+\alpha)^2}{4\alpha^3} R.$$

Proof. Indeed, $t \mapsto \ddot{X}_t$ is a piecewise linear function (see Appendix B), so it is maximized at the knots. For $t \in [t_{i-1}, t_i]$, it follows that

$$\begin{aligned} \|\ddot{X}_t\|_{L^2(\text{Pr})} &= \left\| \frac{t_i - t}{\delta_i} \ddot{X}_{t_{i-1}} + \frac{t - t_{i-1}}{\delta_i} \ddot{X}_{t_i} \right\|_{L^2(\text{Pr})} \\ &\leq \frac{t_i - t}{\delta_i} \|\ddot{X}_{t_{i-1}}\|_{L^2(\text{Pr})} + \frac{t - t_{i-1}}{\delta_i} \|\ddot{X}_{t_i}\|_{L^2(\text{Pr})} \\ &\leq \|\ddot{X}_{t_{i-1}}\|_{L^2(\text{Pr})} \vee \|\ddot{X}_{t_i}\|_{L^2(\text{Pr})} \\ &= \|M_i\|_{L^2(\text{Pr})} \vee \|M_{i+1}\|_{L^2(\text{Pr})} \\ &\leq \frac{75(1+\alpha)^2}{4\alpha^3} R, \end{aligned}$$

by Lemma 2. \square

3.5.3 Proof of the Main Theorem

Proof of Theorem 11. Let $t \in [t_{i-1}, t_i]$, and let the processes (X_t) and (X_t^*) be coupled with $X_{t_{i-1}} = X_{t_{i-1}}^*$. Then,

$$\begin{aligned} \|X_t - X_t^*\|_{L^2(\text{Pr})} &\leq \delta_i \|\dot{X}_{t_{i-1}} - \dot{X}_{t_{i-1}}^*\|_{L^2(\text{Pr})} + \left\| \int_{t_{i-1}}^{t_i} \int_{t_{i-1}}^t (\ddot{X}_s - \ddot{X}_s^*) ds dt \right\|_{L^2(\text{Pr})} \\ &\leq \frac{16\alpha^3 + 75(1+\alpha)^2}{8\alpha^3} R\delta^2 + \frac{\delta^2}{2} \sup_{t \in [0,1]} (\|\ddot{X}_t\|_{L^2(\text{Pr})} + \|\ddot{X}_t^*\|_{L^2(\text{Pr})}) \\ &\leq \frac{10\alpha^3 + 75(1+\alpha)^2}{4\alpha^3} R\delta^2 \leq \frac{115}{2\alpha^3} R\delta^2, \end{aligned}$$

where we have used Proposition 16 and Proposition 17. \square

Piecewise Geodesic Interpolation

In this section, we study the approximation error of piecewise geodesic interpolation.

Namely, we define a stochastic process, still denoted (X_t) , as follows.

1. Draw $X_{t_0} \sim \mu_{t_0}$.
2. For $i = 1, \dots, N$, set $X_{t_i} := T_i(X_{t_{i-1}})$.
3. We join the points $X_{t_0}, X_{t_1}, \dots, X_{t_N}$ via straight lines. Namely, for $t \in [t_{i-1}, t_i]$ we set

$$X_t = \frac{t_i - t}{t_i - t_{i-1}} X_{t_{i-1}} + \frac{t - t_{i-1}}{t_i - t_{i-1}} X_{t_i}.$$

Let μ_t denote the law of X_t .

Theorem 13. *Let the notation and assumptions of Theorem 11 hold (except for the definition of (μ_t)). Then,*

$$\sup_{t \in [0,1]} W_2(\mu_t, \mu_t^*) \leq \frac{5}{2} R \delta^2.$$

Proof. As in Appendix 3.5.3, we have

$$\begin{aligned} \|X_t - X_t^*\|_{L^2(\text{Pr})} &\leq \delta_i \|\dot{X}_{t_{i-1}} - \dot{X}_{t_{i-1}}^*\|_{L^2(\text{Pr})} + \left\| \int_{t_{i-1}}^{t_i} \int_{t_{i-1}}^t \ddot{X}_s^* ds dt \right\|_{L^2(\text{Pr})} \\ &\leq 2R\delta^2 + \frac{1}{2}R\delta^2. \end{aligned}$$

Here, we use several facts: (1) $\dot{X}_{t_{i-1}}^+$, the derivative of $(X_t)_t$ at t_{i-1} from the right, equals

$$(T_i - I_d)X_{t_{i-1}} = (T_i - I_d)X_{t_{i-1}}^*,$$

and so we can apply Corollary 1; (2) the curve (X_t) , consisting of piecewise straight lines, has no acceleration. This finishes the proof. \square

Formally, Theorem 13 is a slightly better approximation guarantee than Theorem 11. Theorem 13 can also be strengthened asymptotically to

$$\limsup_{\delta \searrow 0} \frac{1}{\delta^2} \sup_{t \in [0,1]} W_2(\mu_{\delta,t}, \mu_t^*) \leq R$$

as above. Of course, we do not advocate for using piecewise geodesic interpolation because it is unsuitable for trajectory estimation (see Figure 3-1).

Some remarks:

1. The definition of L in the theorem agrees with the Lipschitz constant of (μ_t^*) in the metric sense, as can be seen from [3, Theorem 8.3.1].
2. The quantity λ_{\min}^{-1} can be interpreted as a bound on the curvature of Bures-Wasserstein space at the interpolation points; see [43] for details.
3. The $O(\delta^2)$ rate of convergence is optimal given our assumptions: a bound R on the second covariant derivative of the curve (μ_t^*) . Indeed, this matches classical approximation results for cubic splines on Euclidean space [12]. We remark that under these assumptions, piecewise geodesic interpolation, where trajectories are piecewise linear and not differentiable, also achieves the $O(\delta^2)$ rate, and we give the proof of this in 3.5.3. Of course, despite achieving the optimal rate in this class of curves, such interpolation is unsuitable for many applications (especially ones in which interpretation and visualization are a priority; see Figure 3-1).
4. We did not attempt to optimize the constant factor in Theorem 11 and it appears that it can, in fact, be improved.
5. Cubic splines achieve higher-order approximation rates in the Euclidean setting, albeit over a restricted class of curves. For approximation of functions $f \in \mathcal{C}^k, k \leq 4$, cubic splines enjoy a $O(\delta^k)$ approximation rate with explicit dependence on $\|f^{(k)}\|_{\text{sup}}$. It is then natural to ask whether it is possible to obtain rates better than $O(\delta^2)$ through a variant of transport splines. This can

indeed be done by using more accurate approximations to the velocity vector fields (v_t); this study will be reported in a forthcoming work.

3.6 Thin-Plate Splines

To demonstrate the flexibility of our method, we use transport splines to define a class of smooth interpolating surfaces on Wasserstein space. We first recall classical thin-plate splines. For a more complete account see [60].

Thin-plate splines are the surface analog of cubic splines, and are useful in spatial problems where measurements are taken on a plane. Here, the times t_i are replaced with points $x_i \in \mathbb{R}^2$ at which we observe real values z_i . To account for this additional dimension the energy functional $\int_0^1 \|\ddot{\gamma}_t\|^2 dt$ that appears in the variational definition (3.1) of cubic splines is replaced by its bivariate counterpart. Thin-plate splines are defined as parametrized surfaces f that solve

$$\inf_f \int_{\mathbb{R}^2} \|\nabla^2 f\|_F^2 \quad \text{s.t.} \quad \begin{cases} f : \mathbb{R}^2 \rightarrow \mathbb{R} \\ f(x_i) = z_i, \quad i = 0, \dots, N \end{cases} \quad (3.11)$$

where $\nabla^2 f$ is the Hessian of f , $\|\cdot\|_F$ denotes the Frobenius norm, and the interpolation data $(x_i, z_i) \in \mathbb{R}^2 \times \mathbb{R}$ is given. (Just as before, f is constrained to be \mathcal{C}^2 .) It can be shown that (3.11) has a unique solution given by

$$f(x) = c_0 + c_1 x^{(1)} + c_2 x^{(2)} + \sum_{i=0}^N \alpha_i \varphi(\|x - x_i\|)$$

where we use $x^{(i)}$ to denote coordinates, and

$$\varphi(r) = r^2 \log r.$$

This leads to a closed form for the coefficients as follows. Let $K = (\varphi(\|x_i - x_j\|))_{i,j=0}^N$ be the “kernel matrix” of the data, and define $P \in \mathbb{R}^{(N+1) \times 3}$ to have i th row $(1, x_i^{(1)}, x_i^{(2)})$.⁹

⁹The function φ plays the role of a kernel for the reproducing kernel Hilbert space of twice-

Then let $L \in \mathbb{R}^{(N+4) \times (N+4)}$ be

$$L = \begin{bmatrix} K & P \\ P^\top & 0_{3 \times 3} \end{bmatrix}$$

Letting $b = (z_0, \dots, z_N, 0, 0, 0)$ be the padded data and $w = (\alpha_0, \dots, \alpha_N, c_0, c_1, c_2)$ the coefficients from (3.6), these solve $Lw = b$. This can be inverted explicitly using the Schur complement, and in particular the resulting coefficients are linear in the data $(z_i)_{i=0}^N$.

We now consider the measure-valued analog of the interpolation problem, namely, at each point x_i we observe a measure $\mu_{x_i}^*$ and our goal is to find a smooth interpolating surface $x \mapsto \mu_x$ of measures.

As in the definition of E-splines, 3.11 can be generalized to Wasserstein space, but it is intractable for the same reasons. In contrast, applying Algorithm 1 is straightforward. Step 2 simply requires the fitting of a Euclidean thin-plate spline. For Step 1 we need only produce couplings between the observed measures $\mu_{x_i}^*$.

One possibility is to mimic the sequential coupling technique described in Section 3.4, namely we fix the ordering x_0, x_1, \dots, x_N and use the system of Monge maps $T_{i-1,i}$ taking $\mu_{x_{i-1}}^*$ to $\mu_{x_i}^*$. As before, we can draw $X_{x_0} \sim \mu_{x_0}^*$ and then successively compute the random variables $X_{x_i} = T_{i-1,i}(X_{x_{i-1}}) \sim \mu_{x_i}^*$ for all i . Sequential coupling is unsuitable here, however, because it distorts the geometry of the plane. To circumvent this issue, we next turn towards the special case when the measures $\mu_{x_i}^*$ are defined over \mathbb{R} , which is already interesting enough to capture a breadth of applications.

As leveraged extensively in Section 3.4.1, $\mathcal{P}_2(\mathbb{R})$ is isometric to a convex subset of a Hilbert space, which simplifies analysis. Indeed, the special structure of $\mathcal{P}_2(\mathbb{R})$ has already been used fruitfully in many prior applications of optimal transport, such as curve registration [47], geodesic principal components [9], estimation of barycenters [10], and uncoupled isotonic regression [52].

For our purposes, we will use the following key property of $\mathcal{P}_2(\mathbb{R})$: there is

differentiable, finite-curvature surfaces, but it is *not* a kernel because it is not positive definite.

a unique coupling of all of the measures $\mu_{x_0}^*, \mu_{x_1}^*, \dots, \mu_{x_N}^*$ which is *simultaneously* optimal for every pair of measures. In other words, there exist random variables $X_{x_0}, X_{x_1}, \dots, X_{x_N}$ such that for any $i, j = 0, 1, \dots, N$, we have $X_{x_j} = T_{i,j}(X_{x_i})$, where $T_{i,j}$ is the Monge map from $\mu_{x_i}^*$ to $\mu_{x_j}^*$. Sampling from this coupling can be done using either of the of the following equivalent procedures:

1. Draw $X_{x_0} \sim \mu_{x_0}^*$, and for each $i \in [N]$ let $X_{x_i} = T_{0,i}(X_{x_0})$ (the choice of x_0 does not affect the coupling).
2. Let U be uniform random on $[0, 1]$, and for $i = 0, 1, \dots, N$ set $X_{x_i} = F_{\mu_{x_i}^*}^{-1}(U)$, where F_μ denotes the CDF of μ .

Indeed, the Monge map $T_{i,j}$ from $\mu_{x_i}^*$ to $\mu_{x_j}^*$ is characterized as the ($\mu_{x_i}^*$ -a.e.) unique mapping which both pushes $\mu_{x_i}^*$ forward to $\mu_{x_j}^*$ and is the gradient of a convex function (see Theorem 3). In one dimension, the latter condition simply means that $T_{i,j}$ is an increasing function. It is easily checked that $F_{\mu_{x_j}^*}^{-1} \circ F_{\mu_{x_i}^*}$ satisfies these properties, and thus¹⁰

$$T_{i,j} = F_{\mu_{x_j}^*}^{-1} \circ F_{\mu_{x_i}^*}.$$

Now, a composition of increasing maps is increasing, which implies that $T_{j,k} \circ T_{i,j}$ must be the Monge map $T_{i,k}$. This key fact directly implies the existence of the simultaneously optimal coupling of the measures. In higher dimensions, this breaks down because the composition of Monge maps is no longer necessarily a Monge map (that is, the composition of gradients of functions is not necessarily the gradient of a function).

In Figure 3-5 we display an application of thin-plate transport splines to temperature data. In the left-hand column we plot the quantiles of the interpolated measures. This is especially convenient when all of the measures are Gaussian, where we have a simple expression. Let Φ be the standard Gaussian CDF, so the quantile

¹⁰The inverse CDFs described here exist because of our assumption of absolute continuity of the measures.

of $\mathcal{N}(m, \sigma^2)$ is $m + \sigma\Phi^{-1}(\alpha)$. Now, suppose the measures $\mu_{x_i}^*$, $i = 0, 1, \dots, N$, are all one-dimensional Gaussians, and write $\mu_{x_i}^* = \mathcal{N}(m_{x_i}, \sigma_{x_i}^2)$. We have

Proposition 18. *Let $(m_x)_{x \in \mathbb{R}^2}$ be the Euclidean thin-plate spline interpolating the means m_{x_i} , and $(s_x)_{x \in \mathbb{R}^2}$ the Euclidean thin-plate spline interpolating the standard deviations σ_{x_i} . Then the quantile function of the thin-plate transport spline interpolating $(\mu_{x_i}^*)$ is $\alpha \mapsto m_x + |s_x|\Phi^{-1}(\alpha)$.*

Proof. It is standard that there is a linear mapping S_x such that the Euclidean thin-plate spline interpolating through $(x_i, z_i)_{i=0}^N$ is given by $S_x(z_0, z_1, \dots, z_N)$.

It follows from (3.7) that the Monge map from $\mu_{x_0}^*$ to $\mu_{x_i}^*$ is the increasing map $z \mapsto (\sigma_{x_i}/\sigma_{x_0})(z - m_{x_0}) + m_{x_i}$. Thus,

$$\begin{aligned} X_x &= S_x(X_{x_0}, X_{x_1}, \dots, X_{x_N}) \\ &= S_x\left(X_{x_0}, \frac{\sigma_{x_1}}{\sigma_{x_0}}(X_{x_0} - m_{x_0}) + m_{x_1}, \dots, \frac{\sigma_{x_N}}{\sigma_{x_0}}(X_{x_0} - m_{x_0}) + m_{x_N}\right) \\ &= S_x(m_{x_0}, m_{x_1}, \dots, m_{x_N}) + S_x\left(X_{x_0} - m_{x_0}, \frac{\sigma_{x_1}}{\sigma_{x_0}}(X_{x_0} - m_{x_0}), \dots, \frac{\sigma_{x_N}}{\sigma_{x_0}}(X_{x_0} - m_{x_0})\right) \\ &= m_x + \frac{X_{x_0} - m_{x_0}}{\sigma_{x_0}} S_x(\sigma_{x_0}, \sigma_{x_1}, \dots, \sigma_{x_N}) \\ &= m_x + s_x \frac{X_{x_0} - m_{x_0}}{\sigma_{x_0}} \sim \mathcal{N}(m_x, s_x^2) = \mu_x. \end{aligned}$$

This is the desired result. □

We conclude this section with a few remarks about the case of higher-dimensional measures, in which case there is no simultaneous optimal coupling of the measures. If we wish to use Monge map couplings as in Algorithm 1, one possibility is to first construct a tree graph whose vertices are the data $\mu_{x_i}^*$, and use Monge map couplings along the edges of the tree. Here, the tree should be chosen to adequately capture the two-dimensional geometry of the spatial covariates. This consideration becomes especially relevant when the spatial covariates are sampled from a manifold, and it is of interest to combine our methodology with existing results on approximation of manifolds via graphs [57].

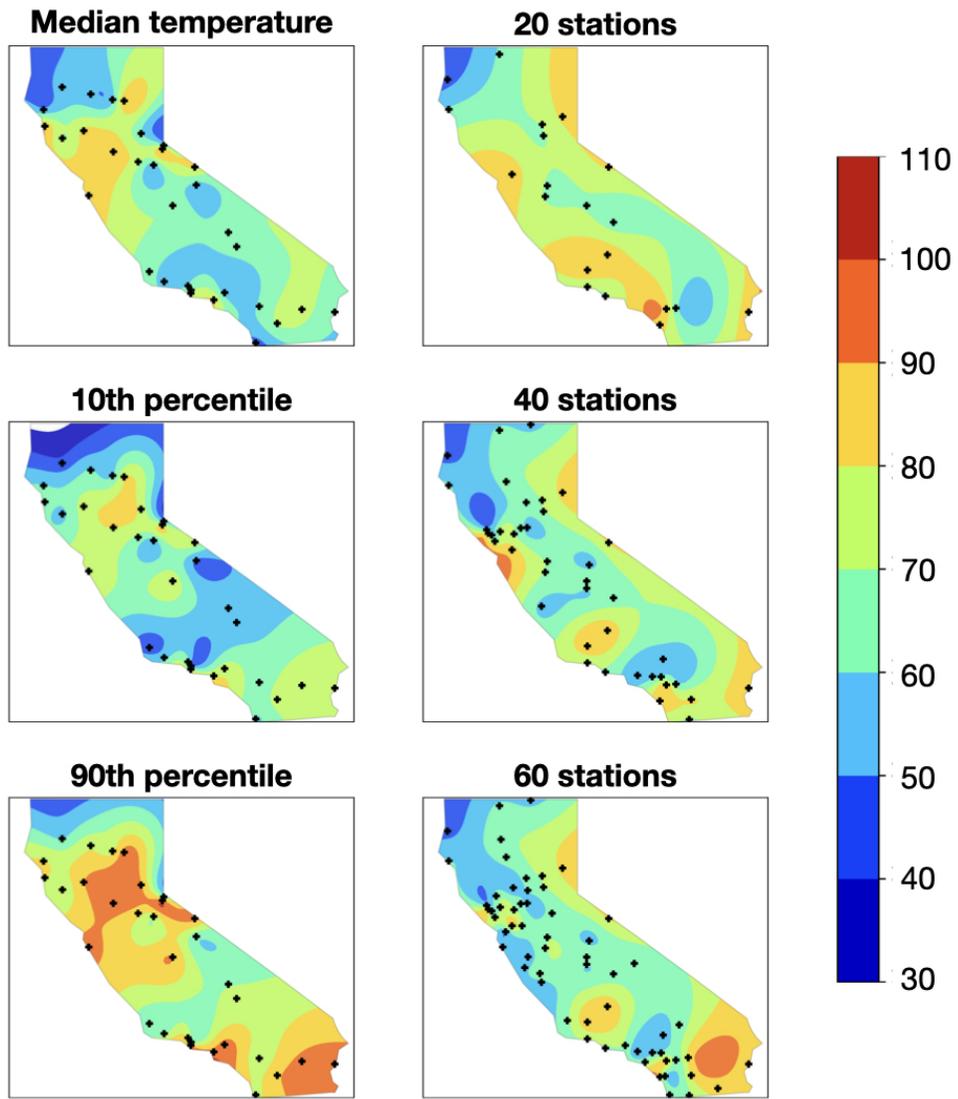


Figure 3-5: Thin-plate splines for California temperature data (in °F); in the left column are the quantiles, while in the right are the means of the interpolated measures for an increasing sample of observations. See Appendix A.2.

Chapter 4

Schrodinger's Splines

This chapter is concerned with an inertial version of the Schrodinger problem, and its relationship with the P -spline problem (3.3) considered in Chapter 2. First, we describe the Schrodinger problem and its solution, and then define two inertial problem formulations: one with several marginals corresponding to natural cubic spline interpolation, and one with two marginals corresponding to an endpoint-constrained spline problem. Then we show that as the noise level tends to zero, the former converges in a strong sense to the P -spline problem (3.3). We then focus on the second, two-marginal problem, proving analogous convergence results as the noise level $\varepsilon \rightarrow 0$, and we furthermore provide a clean characterization of its solution for fixed ε and show that it is equivalent to an entropically regularized OT problem. Our convergence results are of a similar flavor to those in [38, 39], and we use many of their techniques. While in preparation we became aware of the paper [16], which formulates a similar multimarginal problem, though they do not provide results about it and focus more on an optimal control formulation.

4.1 Introduction and Problem Formulations

The Schrodinger bridge problem, originally formulated in [56, 55], is an old problem which has received new interest in part due to its connection with optimal transport. It asks: if a collection of particles was observed to have density μ_0 at time 0 and μ_1

at time 1, and assuming that particles evolve under Brownian motion, what is the most likely trajectory that the particles took? More precisely, we are asked to solve

$$\min \text{KL}(P \parallel Q) \text{ s.t. } P \in \mathcal{P}(\Omega), (X_i)_{\#}P = \mu_i, i = 0, 1 \quad (4.1)$$

Here Ω is the set of continuous paths and the *reference process* Q is the Brownian motion started at μ_0

$$dQ(\omega) = \int dB^x(\omega) d\mu_0(x)$$

and of course B^x is Brownian motion started at x . With the following classical proposition, the Schrodinger problem can be completely solved.

Proposition 19. *Let $\theta: X \rightarrow Y$ be measurable and let $P, Q \in \mathcal{P}(X)$. Then we have the chain rule*

$$\text{KL}(P \parallel Q) = \text{KL}(\theta_{\#}P \parallel \theta_{\#}Q) + \int \text{KL}(P^y \parallel Q^y) d(\theta_{\#}P)(y) \quad (4.2)$$

where $P^y(\cdot) = P(\cdot \mid \theta = y)$, and likewise for Q .

Let $\theta: \omega \mapsto (\omega_0, \omega_1)$, let $P_{01} = \theta_{\#}P$, and define $P^{xy} = P(\cdot \mid \theta = (x, y))$ (and likewise for Q). Then the objective in (4.1) becomes

$$\text{KL}(P_{01} \parallel Q_{01}) + \int \text{KL}(P^{xy} \parallel Q^{xy}) dP_{01}(x, y)$$

It is clear that these two terms are independent, and setting $P^{xy} = Q^{xy}$ the second term vanishes; then P^{xy} is simply the Brownian bridge from x to y . Now, note

$$dQ_{01}(x, y) = c_d e^{-\frac{|x-y|^2}{2}} d\mu_0(x)$$

where $c_d = (2\pi)^{-\frac{d}{2}}$, so

$$\text{KL}(P_{01} \parallel Q_{01}) = -H(P_{01}) + \frac{1}{2} \int |x - y|^2 dP_{01}(x, y) + c$$

with $c = H(\mu_0) - \log c_d$ a constant independent of P_{01} , and H as usual is the entropy

functional $H[\mu] = -\int \mu d\mu$. The entire Schrodinger problem then reduces to

$$\min \frac{1}{2} \int |x - y|^2 dP_{01}(x, y) - H(P_{01}) \text{ s.t. } (P_{01})_i = \mu_i \quad (4.3)$$

This is precisely the entropically regularized optimal transport problem (see for example [48]) with regularization parameter $\lambda = 1$. In summary, to solve the Schrodinger problem we first solve the standard OT problem with entropic regularization to obtain the marginal law P_{01} , then interpolate using the Brownian bridge B^{xy} between each pair (x, y) . To sample a trajectory from P , we first sample endpoints $(x, y) \sim P_{01}$, then sample a Brownian bridge from x to y .

By considering $B^{x,\varepsilon} = x + \sqrt{\varepsilon}B$ and instead minimizing

$$\min \varepsilon \text{KL}(P \parallel Q^\varepsilon) \text{ s.t. } P \in \mathcal{P}(\Omega), P_i = \mu_i \quad (4.4)$$

with $Q^\varepsilon = \int B^{x,\varepsilon} d\mu_0(x)$, the same analysis reduces the problem to

$$\min \frac{1}{2} \int |x - y|^2 dP_{01}(x, y) - \varepsilon H(P_{01}) \quad (4.5)$$

One would hope from this that as $\varepsilon \rightarrow 0$ we have $(4.5) \rightarrow W_2^2(\mu_0, \mu_1)$, and this is proved in [38]. Specifically, they show

Theorem 14 ([38, Theorem 3.3]). *Let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)$. Then as $\varepsilon \rightarrow 0$, we have $(4.5) \rightarrow W_2^2(\mu_0, \mu_1)$. Furthermore, for small ε (4.5) admits a unique solution \hat{P}_{01}^ε , and any limit point of the sequence $(\hat{P}_{01}^\varepsilon)$ is an optimal coupling in $\min \frac{1}{2} \int |x - y|^2 d\pi$. In particular, if the optimal W_2 coupling $\hat{\pi}$ is unique, as it is for μ_0 and μ_1 absolutely continuous, then $\hat{P}_{01}^\varepsilon \rightarrow \hat{\pi}$ (in the W_2 topology).*

They also prove the stronger statement that (4.4) converges to

$$\min_{P \in \mathcal{P}(\Omega)} \int C dP \text{ s.t. } (X_i)_\# P = \mu_i \quad (4.6)$$

where $C(\omega) = \frac{1}{2} \int |\dot{\omega}_t|^2 dt$ is the (half) squared-length functional on absolutely continuous curves, taking the value infinity otherwise, and X_t is the evaluation functional

on paths; see [38, Theorem 3.6]. In this case too the arg-optima converge.

We wish to emphasize naturality of (4.5) and (4.4). Entropically regularized optimal transport was rediscovered in a completely different setting than here, focused on computational speed [22], and subsequently exploited broadly in computational transport (see again [48]). The results above say exactly that the limit as the regularization parameter tends to zero is the Wasserstein distance. The problem (4.4) in path space is nothing but a dynamical formulation of entropic OT, a la Benamou-Brenier. Thus the Schrodinger problem completes a set of formulations of optimal transport, summarized in the table below.

	Dynamical	Static	
$\epsilon > 0$	Schrödinger bridge	entropically regularized OT	(4.7)
$\epsilon = 0$	dynamical formulation of OT	OT	

4.1.1 The Interpolation Problem

Now we define our inertial analogs of the Schrodinger problem, starting with the multimarginal formulation. Our starting point is the integrated Brownian motion with initial position and velocity:

$$Q_t^{\epsilon, \mathbf{x}_0} = x_0 + tv_0 + \sqrt{\epsilon} \int_0^t B_s ds \tag{4.8}$$

where we write the shorthand $\mathbf{x}_0 = (x_0, v_0)$. Whereas displacements for Brownian motion are white noise, here the particle has momentum which changes according to white noise, and this in turn affects position through integration. Fix an initial distribution $\mu \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ on initial positions and velocities — for convenience of results we choose it to be standard normal, though for the same results it is only required that $\mu \ll dx$ and $dx \ll \mu$, where dx is the Lebesgue measure. We then define our reference process as

$$Q^\epsilon = \int Q^{\epsilon, \mathbf{x}_0} d\mu(\mathbf{x}_0)$$

We pause briefly to discuss the initial distribution μ . While for the Schrodinger problem (4.4) we have complete information about the starting distribution $(X_0)_\#Q^\varepsilon$, this is not the case in our setting. We can think of splines as formulated in phase space, with initial position and velocity, and only the initial position distribution μ_0 is known, not the joint initial position-velocity distribution. Indeed, in light of Proposition 4 in Chapter 3, v_0 will almost certainly not be a deterministic function of x_0 , and it cannot be independent either in general, so we cannot assume anything about it a priori.

Let $\mu_i \in \mathcal{P}(\mathbb{R}^d)$ be fixed measures and $t_i \in [0, 1]$ fixed times. The *multimarginal Schrodinger spline* problem is

$$\min \varepsilon \text{KL}(P \parallel Q^\varepsilon) \text{ s.t. } P \in \mathcal{P}(\Omega), (X_{t_i})_\#P = \mu_i \quad (4.9)$$

Now, let

$$C(\omega) = \frac{1}{2} \int |\ddot{\omega}_t|^2 dt$$

be the curvature cost functional on Ω , defined to be ∞ if ω does not have absolutely continuous first derivative. Recall that the P-spline problem (3.3) is (up to a factor of $\frac{1}{2}$)

$$\min_{P \in \mathcal{P}(\Omega)} \int C dP \text{ s.t. } (X_{t_i})_\#P = \mu_i \quad (4.10)$$

We will show a result analogous to Theorem 14 for this problem, that is, that this is the limit of (4.9) as $\varepsilon \rightarrow 0$.

4.1.2 The Endpoint Problem

Referring to the table (4.7), in the case of splines the lower-left and lower-right entries correspond to the formulations (3.3) and (3.4) of P -splines. The top-right problem should then be the entropically regularized version of (3.4)

$$\min \int c d\pi - \varepsilon H(\pi) \text{ s.t. } \pi_{x_i} = \mu_i \quad (4.11)$$

where π_{x_i} is the i -th marginal. It is natural to suggest that (4.9) become the top-left entry; however, (4.9) is *not* equivalent to (4.11) (among other things, the arbitrariness of the initial distribution of positions and velocities prevents this). One can work backwards and formulate a Gaussian process which when taken to be the reference process for a Schrodinger problem becomes equivalent to (4.11), but it is difficult to characterize this process explicitly. For this reason we formulate another inertial Schrodinger problem that admits a complete solution.

We first revisit the Euclidean spline problem. Recall that the cost c in (3.4) and (4.11) is the cost of the full spline problem (3.1), as a function of the data (x_i) , and it is quadratic (indeed it is positive semidefinite). The matrix representing it is complicated, but by restricting the problem the cost can be greatly simplified. Fixing now velocities v_i as well as positions x_i , define the *endpoint cubic spline* problem by

$$\min \int_0^1 |\ddot{x}_t|^2 dt \text{ s.t. } x_{t_i} = x_i, \dot{x}_{t_i} = v_i \quad (4.12)$$

This decouples across intervals $[t_i, t_{i+1}]$ into problems of the form

$$\min \int_0^1 |\ddot{x}_t|^2 dt \text{ s.t. } x_0 = x_0, x_1 = x_1, \dot{x}_0 = v_0, \dot{x}_1 = v_1 \quad (4.13)$$

Integration by parts yields that the optimal curve x_t is a cubic polynomial — indeed, the only cubic that satisfies the boundary conditions. Furthermore, as a function of (x_i, v_i) , the objective value of (4.13) is quite simply

$$c(x_0, x_1, v_0, v_1) = 12|x_1 - x_0 - v_0|^2 - 12\langle x_1 - x_0 - v_0, v_1 - v_0 \rangle + 4|v_1 - v_0|^2$$

This can be seen, for instance, in [15]. Now let v_0, v_1 be vector fields on \mathbb{R}^d . We generalize (4.13) to the *endpoint P-spline* problem

$$\min \int C dP \text{ s.t. } P \in \Pi(\mu_0, v_0, \mu_1, v_1) \quad (4.14)$$

where the constraint set $\Pi(\mu_0, v_0, \mu_1, v_1)$ is

$$\left\{ P \in \mathcal{P}(\Omega) \mid (X_0, \dot{X}_0)_\# P = (\text{Id}, v_0)_\# \mu_0, (X_1, \dot{X}_1)_\# P = (\text{Id}, v_1)_\# \mu_1 \right\}$$

In a particle interpretation, we prescribe the initial and final distribution of particles, as well as the initial and final velocities of each particle as a function of its position. The motivation for this is considering v_0, v_1 to be tangent vectors in Wasserstein space. Suppose a particle flow has trajectories (x_t) , and let μ_t be the induced curve of measures as densities of the particles. Letting $v_t = \dot{x}_t$, it must be that (μ_t, v_t) solves the continuity equation. The endpoint constraints in $\Pi(\mu_0, v_0, \mu_1, v_1)$ then exactly enforce that v_0 and v_1 are the endpoint derivatives of the curve μ_t from the continuity equation.

Proposition 20. *Fix deterministic vector fields v_0, v_1 . The problem (4.14) is equivalent to*

$$\min \int c d\pi \text{ s.t. } \pi \in \Pi(\mu_0, \mu_1) \tag{4.15}$$

where $\Pi(\mu_0, \mu_1)$ is the ordinary set of couplings, and $c = c(x_0, v_0(x_0), x_1, v_1(x_1))$ is the cost (4.1.2) considered as a function of the two variables x_0, x_1 . If π^* is optimal for (4.16) then the measure P^* defined by giving mass $\pi^*(x, y)$ to the cubic with endpoint data $(x, v_0(x), y, v_1(y))$ solves (4.14), and conversely every optimal P^* arises in this way.

Proof. If P^* is optimal, then for each endpoint pair (x, y) it must be that P^* places zero mass on any curve with this data except the cubic with $(x, v_0(x), y, v_1(y))$, since otherwise the total curvature cost C could be reduced by substituting this cubic. The remainder follows from the definition of c . \square

The problem (4.15) can be entropically regularized as

$$\min \int c d\pi - \varepsilon H(\pi) \text{ s.t. } \pi \in \Pi(\mu_0, \mu_1) \tag{4.16}$$

It turns out that this does correspond to a Schrodinger-type problem related to

(4.14). To define it, we begin again with the integrated Brownian motion, but now we know precisely the distribution of the initial state (x_0, v_0) . The reference process is

$$Q^\varepsilon = \int Q^{\varepsilon, (x_0, v_0(x_0))} d\mu_0(x_0)$$

which leads us to the *endpoint Schrodinger spline* problem

$$\min \varepsilon \text{KL}(P \parallel Q^\varepsilon) \text{ s.t. } P \in \Pi(\mu_0, v_0, \mu_1, v_1) \quad (4.17)$$

With this we can complete a table like (4.7):

	Dynamical	Static	
$\varepsilon > 0$	(4.17)	(4.16)	(4.18)
$\varepsilon = 0$	(4.14)	(4.15)	

Our results can be thought of as establishing equality in the horizontal axis, and convergence in the vertical axis (as $\varepsilon \rightarrow 0$).

4.2 Technical Background

We require some technical background before presenting the proofs of our results, which we collect here for convenience.

4.2.1 Γ -Convergence

As noted in [38], to prove convergence of optimization programs it does not suffice to prove pointwise convergence of the objective functions; even uniform convergence is not enough if convergence of the arg-optima is desired as well. A different notion of convergence is required, and this is exactly Γ -convergence. This material is taken from [45, 14]. We assume all spaces are complete and separable metric spaces.

Definition 4. A function f is **lower semicontinuous** (*lsc*) if for all x

$$f(x) = \sup_{U \ni x} \inf_{y \in U} f(y)$$

Equivalently, for all sequences $x_n \rightarrow x$

$$f(x) \leq \liminf f(x_n)$$

Consider a sequence of optimization problems $P_n = \min f_n(x)$ and $P = \min f(x)$. In order to take the limit $P_n \rightarrow P$ or $\arg P_n \rightarrow \arg P$, the right form of convergence is that f_n Γ -converges to f , which we write as $f_n \xrightarrow{\Gamma} f$.

Definition 5. We say $f_k \xrightarrow{\Gamma} f$ if

$$f(x) = \sup_{U \ni x} \lim_{k \rightarrow \infty} \inf_{y \in U} f_k(y)$$

Equivalently, for all sequences $x_n \rightarrow x$

$$f(x) \leq \liminf f_n(x_n)$$

and there is a sequence $x_n \rightarrow x$ (called the *recovery sequence*) such that equality holds.

Definition 6. A function f is **coercive** if for all α the set $\{f \leq \alpha\}$ is precompact; if f is *lsc*, this is equivalent to compactness.

A family of functions $\{f_n\}$ is **equicoercive** if for all α there is a compact set K such that $\{f_n \leq \alpha\} \subset K$ for all n .

Theorem 15. If f is *lsc* and coercive, $\inf f$ is achieved.

Theorem 16. If $\{f_n\}$ is equicoercive and $f_n \xrightarrow{\Gamma} f$, then

1. f is coercive
2. $\inf f_n(x) \rightarrow \inf f(x)$

3. Letting $x_n^* = \arg \inf f_n$, any subsequential limit point of (x_n^*) is a minimizer of f .

Incidentally, the topology of Γ -convergence makes the set of all functions compact; any sequence of functions has a subsequence that Γ -converges.

We will make use of the following theorems from Leonard.

Theorem 17 ([38, Theorem 7.1]). *Suppose $f_n \xrightarrow{\Gamma} f$ and $\{f_n\}$ is equicoercive, and let $\theta: \mathcal{X} \rightarrow \mathcal{Y}$ be a continuous function. Define*

$$\psi_n(y) = \inf_x f_n(x) \text{ s.t. } \theta(x) = y$$

and similarly for ψ from f . Then $\{\psi_n\}$ is equicoercive and $\psi_n \xrightarrow{\Gamma} \psi$.

Theorem 18 ([38, Corollary 6.4]). *Let \mathcal{Y} be a normed space and \mathcal{X} its topological dual. Let $g_n, g: \mathcal{Y} \rightarrow \mathbb{R}$ be a sequence of functions satisfying*

1. g_n is convex
2. $g_n \rightarrow g$ pointwise
3. There is a constant c such that $|g_n(y)| \leq c(1 + \|y\|)$

Define $f_n = g_n^$ and $g = g^*$. Then $f_n \xrightarrow{\Gamma} f$ and there is a compact set K such that $\text{Dom}(f_n), \text{Dom}(f) \subset K$, both with respect to the weak topology induced by \mathcal{Y} on \mathcal{X} .*

4.2.2 Large Deviation Principles

The other key ingredient in our results is so-called large deviation principles for families of stochastic processes, most notably Brownian motion. This material appears in [25].

The crux of the proof is the satisfaction of a so-called large deviation principle.

Definition 7. *A family of measures $\{\mu_\varepsilon\}$ and a function C satisfies the **large deviation principle (LDP)** if for all measurable E*

$$- \inf_{\omega \in E^\circ} C(\omega) \leq \liminf \varepsilon \log \mu_\varepsilon(E) \leq \limsup \varepsilon \log \mu_\varepsilon(E) \leq - \inf_{\omega \in \bar{E}} C(\omega)$$

where E° is the interior of E and \bar{E} is the closure of E . It is required that C is lsc, and it is referred to as the rate function.

The most important such family $\{\mu_\varepsilon\}$ is Brownian motion.

Theorem 19 (Schilder). *Let $B_t^\varepsilon = \sqrt{\varepsilon}B_t$ be rescaled Brownian motion. Then $\{B^\varepsilon\}$ satisfies an LDP with rate function $C(\omega) = \frac{1}{2} \int |\dot{\omega}_t|^2 dt$.*

We will use a few other results about LDP families.

Proposition 21 (Contraction). *Suppose $\{\mu_\varepsilon\}$ satisfies an LDP with rate function C , and let $\theta: \mathcal{X} \rightarrow \mathcal{Y}$ be continuous. Define $\nu_\varepsilon = \theta_\# \mu_\varepsilon$ and $D(y) = \inf_{\theta(x)=y} C(x)$. Then $\{\nu_\varepsilon\}$ satisfies an LDP with rate function D . Also, if C is coercive, then so is D .*

Theorem 20 (Laplace-Varadhan). *Suppose $\{\mu_\varepsilon\}$ satisfies an LDP with coercive rate function C , and let f be a bounded continuous function. Then*

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \int e^{f/\varepsilon} d\mu_\varepsilon = \sup_{\omega} f - C$$

4.3 Results

We now present and prove our results, first for the interpolation problem then for the endpoint problem.

4.3.1 The Interpolation Problem

Our main result in this section is the following.

Theorem 21. *Let (μ_i) be absolutely continuous probability measures. Then there are measures $\mu_k^\varepsilon \xrightarrow{W_2} \mu_k$ such that as $\varepsilon \rightarrow 0$, we have (4.9) \rightarrow (4.10). If (4.10) is finite, then for all small ε the problem (4.9) (with the data μ_k^ε) admits a unique solution \hat{P}^ε , and furthermore any limit point of (\hat{P}^ε) is a solution of (4.10). In particular, if (4.10) admits a unique solution \hat{P} , then $\hat{P}^\varepsilon \rightarrow \hat{P}$ in the W_2 topology.*

The proof leverages the properties of Γ -convergence mentioned above. Recall that

$$Q^\varepsilon = \int Q^{\varepsilon, \mathbf{x}_0} d\mu(\mathbf{x}_0)$$

where μ is standard Gaussian on $\mathbb{R}^d \times \mathbb{R}^d$. Consider the sequence of functionals

$$F_\varepsilon(P) = \varepsilon \text{KL}(P \parallel Q^\varepsilon)$$

which are just the objectives in (4.9) (without constraints). Though these are defined only on the space $\mathcal{P}(\Omega)$, we extend them to another space $\mathcal{X} \supset \mathcal{P}(\Omega)$ by $F_\varepsilon + \mathbb{1}_{\mathcal{P}(\Omega)}$. This space is defined as follows: Let \mathcal{Y} be the set of bounded Lipschitz functions on Ω , and \mathcal{X} its dual. By the Riesz representation theorem we have that $\mathcal{P}(\Omega) \subset \mathcal{X}$. (However, there is no explicit description of the entirety of \mathcal{X} .)

Let $F(P) = \int C dP$ be the total curvature cost, where $C(\omega) = \frac{1}{2} \int_0^1 |\ddot{\omega}_t|^2 dt$ is the squared curvature as before. The strategy of proof is the same as in [38]: we will show that $F_\varepsilon^* \rightarrow F^*$ pointwise, where G^* denotes the convex dual of G , and then use Theorem 18 to conclude that $F^* \varepsilon \xrightarrow{\Gamma} F$, and then reimpose the constraints using Theorem 17. From there the conclusion is simple.

To calculate F_ε^* we use the following classical result.

Lemma 3. *Let P and Q be probability distributions on a metric space with its Borel σ -algebra. Then*

$$\text{KL}(P \parallel Q) = \sup_f \int f dP - \log \int e^f dQ$$

where the infimum may be taken over continuous functions, bounded continuous functions, or bounded Lipschitz functions.

This implies that

$$F_\varepsilon^*(f) = \varepsilon \log \int e^{f/\varepsilon} dQ^\varepsilon$$

The following lemma calculating F^* appears in [38].

Lemma 4. *For any function C , the convex dual of the convex functional acting on*

probability measures by

$$P \mapsto \int C dP + \mathbb{1}_{\mathcal{P}(\Omega)}(P)$$

is the functional

$$f \mapsto \sup_{\omega} f(\omega) - C(\omega)$$

We now aim to prove the pointwise convergence of F_{ε}^* to F^* , which we will do via the Laplace-Varadhan Lemma (20).

Lemma 5. *For any $\mathbf{x}_0 \in \mathbb{R}^d \times \mathbb{R}^d$, the family $Q^{\varepsilon, \mathbf{x}_0}$ satisfies an LDP with coercive rate function $C(\omega) = \frac{1}{2} \int |\ddot{\omega}_t|^2 dt$ on the space $\Omega^{\mathbf{x}_0}$ of paths with initial condition \mathbf{x}_0 . The family Q^{ε} satisfies an LDP with the same rate function on the space Ω of all paths, but the rate function is not coercive on this space.*

Proof. Applying the contraction principle (21) with the function $\theta(\omega)_t = x_0 + tv_0 + \int_0^t \omega_s ds$, Schilder's theorem implies that the family $Q^{\varepsilon, \mathbf{x}_0}$ satisfies an LDP with the required rate function, so the first claim is proved.

For the second, first of all, if $\mu_{\varepsilon} = \mu$ for all ε then it is easy to check that the "family" μ_{ε} satisfies an LDP with the null rate function $C \equiv 0$. Now suppose that μ_{ε} satisfies an LDP with rate function C and ν_{ε} satisfies an LDP with rate function D . Then $\mu_{\varepsilon} \otimes \nu_{\varepsilon}$ satisfies an LDP with rate function $C \oplus D$. To see this, take measurable sets E and F ; then $(E \times F)^{\circ} = E^{\circ} \times F^{\circ}$ and $\overline{E \times F} = \overline{E} \times \overline{F}$. Simply adding the two LDPs for μ_{ε} and ν_{ε} yields the statement for sets of the form $E \times F$, and since these generate the product topology and σ -algebra this yields the claim. Now, as $Q^{\varepsilon} = Q^{\varepsilon, \mathbf{x}_0} \otimes \mu$, the lemma is proved. \square

By itself this is not sufficient, since the Laplace-Varadhan Lemma requires a coercive rate function. Still, the measures Q^{ε} satisfy the conclusions of the lemma.

Recall the classical Egorov theorem.

Theorem 22 (Egorov). *Let μ be a finite measure on a measure space and let f_n be a sequence of functions such that $f_n \rightarrow f$ μ -a.e. Then for all $\delta > 0$ there is a measurable set E such that $\mu(E^c) < \delta$ and $f_n \rightarrow f$ uniformly on E .*

Lemma 6. *Let μ be a finite measure and suppose that $f_\varepsilon \rightarrow f$ μ -a.e. and that there is a constant C such that f_ε and f are bounded by C . Suppose further that $\limsup_{\varepsilon \rightarrow 0} \sup f_\varepsilon \leq \sup f$. Then*

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \int e^{f_\varepsilon/\varepsilon} d\mu = \sup f \quad (4.19)$$

All suprema are taken with respect to μ , that is they are μ -essential suprema.

Proof. By rescaling we can assume that μ is a probability measure. Then since $|f_\varepsilon| \leq C$ we have $-C \leq \varepsilon \log \int e^{f_\varepsilon/\varepsilon} d\mu \leq C$, so the \limsup and \liminf in the left-hand side of (4.19) exist; we will show both are equal to $\sup f$.

Let us first assume that $f_n \rightarrow f$ uniformly. Take $\delta > 0$ and let ε be small enough that $|f_\varepsilon - f| < \delta$. Then

$$\varepsilon \log \int e^{(f-\delta)/\varepsilon} d\mu < \varepsilon \log \int e^{f_\varepsilon/\varepsilon} d\mu < \varepsilon \log \int e^{(f+\delta)/\varepsilon} d\mu$$

As $\varepsilon \rightarrow 0$ the Laplace method says that the left-hand side converges to $\sup f - \delta$ and the right-hand side to $\sup f + \delta$. Letting $\delta \rightarrow 0$ concludes.

In general, fixing again $\delta > 0$, let E be the set yielded by Egorov's theorem. Then

$$\varepsilon \log \int e^{f_\varepsilon/\varepsilon} d\mu \geq \varepsilon \log \int_E e^{f_\varepsilon/\varepsilon} d\mu \rightarrow \sup_E f$$

where we have used the first part of the lemma on the measure $\mu|_E$. Letting $\delta \rightarrow 0$ the right-hand side converges to $\sup f$ and so

$$\liminf_{\varepsilon \rightarrow 0} \varepsilon \log \int e^{f_\varepsilon/\varepsilon} d\mu \geq \sup f$$

For the other inequality, we have

$$\varepsilon \log \int e^{f_\varepsilon/\varepsilon} d\mu \leq \varepsilon \log \sup e^{f_\varepsilon/\varepsilon} = \sup f_\varepsilon$$

Taking the $\limsup_{\varepsilon \rightarrow 0}$ of both sides and using the second assumption of the lemma

concludes. □

A remark is in order for the curious reader. Though one might think a sort of dominated convergence result might allow to drop the second hypothesis — essentially, that $\sup f_\varepsilon \rightarrow \sup f$ — this cannot be so, and the hypothesis is essential. For example, letting $f_\varepsilon(x) = \mathbb{1}_{(0,\varepsilon)}(x)$, we have $f_n(x) \rightarrow 0$ pointwise, but it can be computed that $\varepsilon \log \int_0^1 e^{f_\varepsilon/\varepsilon} dx \rightarrow 1$. This also shows that a monotone convergence assumption does not suffice.

Proposition 22. *For any bounded Lipschitz function f on Ω , we have*

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \int e^{f/\varepsilon} dQ^\varepsilon = \sup_{\omega} f - C$$

where C is the squared curvature cost.

Proof. To leverage Varadhan's lemma we need to separate the integral for each \mathbf{x}_0 individually, in order to get a coercive rate function. Write the integral as

$$\varepsilon \log \int \exp \left[\frac{1}{\varepsilon} \left(\varepsilon \log \int e^{f(\omega)/\varepsilon} dQ^{\varepsilon, \mathbf{x}_0}(\omega) \right) \right] d\mu(\mathbf{x}_0)$$

where μ is standard normal, as mentioned above. Define

$$f_\varepsilon(\mathbf{x}_0) = \varepsilon \log \int e^{f(\mathbf{x}_0 + \omega)/\varepsilon} dQ^{\varepsilon, \mathbf{0}}(\omega)$$

where $\mathbf{x}_0 + \omega$ has the clear meaning of linearly offsetting the path ω by the curve $x_0 + tv_0$ to have initial data \mathbf{x}_0 . Then the integral becomes

$$\varepsilon \log \int e^{f_\varepsilon(\mathbf{x}_0)/\varepsilon} d\mu(\mathbf{x}_0) \tag{4.20}$$

The Laplace-Varadhan Lemma says that for each \mathbf{x}_0

$$f_\varepsilon(\mathbf{x}_0) \rightarrow f_0(\mathbf{x}_0) := \sup_{\omega \in \Omega_0} f(\mathbf{x}_0 + \omega) - C(\omega)$$

and furthermore we have

$$\begin{aligned}
\sup_{\mathbf{x}_0} f_\varepsilon(\mathbf{x}_0) &= \sup_{\mathbf{x}_0} \varepsilon \log \int e^{f(\mathbf{x}_0+\varepsilon)/\varepsilon} dQ^{\varepsilon, \mathbf{0}}(\omega) \\
&\leq \varepsilon \log \int e^{\sup_{\mathbf{x}_0} f(\mathbf{x}_0+\omega)/\varepsilon} dQ^{\varepsilon, \mathbf{0}}(\omega) \\
&\rightarrow \sup_{\omega \in \Omega^{\mathbf{0}}} \sup_{\mathbf{x}_0} f(\mathbf{x}_0 + \omega) - C(\omega) \\
&= \sup_{\omega} f(\omega) - C(\omega) \\
&= \sup_{\mathbf{x}_0} f_0(\mathbf{x}_0)
\end{aligned}$$

where in the third line we have again applied the Laplace-Varadhan Lemma to the function $\omega \rightarrow \sup_{\mathbf{x}_0} f(\mathbf{x}_0 + \omega)$. Since f is bounded and Lipschitz, this function is also bounded and Lipschitz with the same constants, as the supremum of Lipschitz functions with uniform constant is again Lipschitz. We have also used the fact that C is invariant under translation by \mathbf{x}_0 . Applying Lemma 6 to (4.20) then gives the result. \square

Proposition 23. *The functionals F_ε are equicoercive and $F_\varepsilon \xrightarrow{\Gamma} F$. Furthermore, let*

$$\begin{aligned}
L_\varepsilon[\mu_i] &= \inf \varepsilon \text{KL}(P \parallel Q^\varepsilon) \text{ s.t. } (X_{t_i})_\# P = \mu_i \\
L[\mu_i] &= \inf \int C dP \text{ s.t. } (X_{t_i})_\# P = \mu_i
\end{aligned}$$

Then $L_\varepsilon \xrightarrow{\Gamma} L$. In particular, there are measures $\mu_i^\varepsilon \xrightarrow{W_2} \mu_i$ such that $L_\varepsilon[\mu_i^\varepsilon] \rightarrow L[\mu_i]$.

Proof. That $F_\varepsilon^* \rightarrow F^*$ pointwise is implied by Proposition 22. The first statement then follows from Theorem 18. For the second, we apply Theorem 17 to the function $\theta(P) = (X_{t_i})_\# P$. The final statement follows from the definition of Γ -convergence. \square

Proof of Theorem 21. The bulk of the statements follow from the definition of Γ -convergence and the previous proposition. Uniqueness of \hat{P}^ε follows from the strict convexity of the KL divergence. \square

4.3.2 The Endpoint Problem

Throughout this section we write $\mathbf{x} = (x_0, v_0(x_0), x_1, v_1(x_1))$, as opposed to above where \mathbf{x}_0 corresponded only to initial data. We also write $Q^{\varepsilon|\mathbf{x}}$ for the process $Q^{\varepsilon, (x_0, v_0(x_0))}$ conditioned on $(X_1, \dot{X}_1) = (x_1, v_1(x_1))$, and $Q_{\mathbf{x}}^{\varepsilon}$ for the distribution of $(X_0, \dot{X}_0, X_1, \dot{X}_1)$ under Q^{ε} . These notations are extended similarly for P . Since μ_0, μ_1, v_0, v_1 are fixed, we interchangeably refer to the cost c , measures π , etc. as functions of two or four variables, with the understanding that the velocity variables v are deterministic functions of the position variables x .

We first show that the Schrodinger endpoint spline problem is equivalent to an entropically regularized OT problem.

Proposition 24. *The problem (4.17) is equivalent to (4.16) (after reparametrizing $\varepsilon \rightarrow 2\varepsilon$). Specifically, if $\hat{\pi}$ is the optimal coupling for (4.16), then the optimal path measure \hat{P} for (4.17) is the semidirect product*

$$\hat{P} = \int Q^{\varepsilon|\mathbf{x}} d\pi(\mathbf{x})$$

Furthermore, the optimal values are the same up to a factor of 2 and an additive constant which is $O(\varepsilon)$.

Proof. First apply the KL divergence chain rule (4.2) to the function $\theta: \omega \mapsto (\omega_0, \dot{\omega}_0, \omega_1, \dot{\omega}_1)$, we get

$$\text{KL}(P \parallel Q^{\varepsilon}) = \text{KL}(P_{\mathbf{x}} \parallel Q_{\mathbf{x}}^{\varepsilon}) + \int \text{KL}(P^{\mathbf{x}} \parallel Q^{\varepsilon|\mathbf{x}}) dP_{\mathbf{x}}(\mathbf{x})$$

These two terms decouple, and it is clear that it is optimal to set $P^{\mathbf{x}} = Q^{\varepsilon|\mathbf{x}}$, so the optimization reduces to just the first term. We now compute the density of $Q_{\mathbf{x}}^{\varepsilon}$. Requiring that Q^{ε} have endpoint values \mathbf{x} means

$$\begin{aligned} x_1 &= x_0 + v_0(x_0) + \sqrt{\varepsilon} \int_0^1 B_s ds \\ v_1(x_1) &= v_0(x_0) + \sqrt{\varepsilon} B_1 \end{aligned}$$

Now, $\mathbb{E} B_1 \int_0^1 B_s ds = \frac{1}{2}$ and $\mathbb{E} \left(\int_0^1 B_s ds \right)^2 = \frac{1}{3}$ so $\left(\int_0^1 B_s ds, B_1 \right) \sim \mathcal{N}(0, A)$ where

$$A = \varepsilon \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

From this we can read off

$$dQ_{\mathbf{x}}^\varepsilon = c_A \cdot d\mu_0(x_0) \cdot \exp \left(\frac{1}{2\varepsilon} c(x_0, v_0(x_0), x_1, v_1(x_1)) \right)$$

where c_A is the Gaussian normalization constant. Explicitly writing the KL divergence yields

$$\varepsilon \text{KL}(P \parallel Q^\varepsilon) = \frac{1}{2} \int c dP_{\mathbf{x}} - \varepsilon H(P_{\mathbf{x}}) + \varepsilon(\log c_A + H(\mu_0))$$

This is the result. □

We now characterize the process $Q^{\varepsilon|\mathbf{x}}$. Consider first the case of $\mathbf{x} = 0$, so our goal is to condition the process $I_t = \int_0^t B_s ds$ on $I_1 = \dot{I}_1 = 0$. As this is a Gaussian process, we can use the Gaussian conditioning lemma.

Lemma 7 (Gaussian conditioning). *If*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left(0, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \right)$$

then

$$(X|Y = y) \sim \mathcal{N}(BC^{-1}y, A - BC^{-1}B^\top)$$

We furthermore have the following simple consequence of Fubini's theorem.

Lemma 8. *If X_t is a Gaussian process with covariance function $a_X(s, t)$, then $Z_t = \int_0^t X_s ds$ has covariance function*

$$a_Z(s, t) = \int_0^s \int_0^t a_X(s', t') ds' dt'$$

Conditioning I_t first on $\dot{I}_1 = 0$ we obtain the process $Z_t = \int_0^t (BB)_s ds$, where BB

is the standard Brownian bridge. To compute a_Z we need to integrate a_{BB} . As the Brownian bridge has covariance $a_{BB}(s, t) = s(1 - t)$ for $s \leq t$, calculus reveals that for $0 \leq s \leq t \leq 1$,

$$a_Z(s, t) = \frac{1}{2}s^2 \left(t - \frac{1}{3}s \right) - \frac{1}{4}s^2t^2$$

Now we must condition Z on $Z_1 = 0$. It is easy to compute

$$((Z_t)_t, Z_1) \sim \begin{bmatrix} a_Z(s, t) & \frac{1}{12}s^2(3 - 2s) \\ \frac{1}{12}t^2(3 - 2t) & \frac{1}{12} \end{bmatrix}$$

We use matrix notation for clarity; the “last” row and column corresponds to $s = 1$ and $t = 1$, respectively. Letting $Y = (Z \mid Z_1 = 0) = (I \mid I_1 = 0, \dot{I}_1 = 0)$, the conditioning lemma 7 yields

$$a_Y(s, t) = a_Z(s, t) - \frac{1}{12}s^2(3 - 2s)t^2(3 - 2t)$$

Though this expression is opaque, we can use it to test an ansatz. From the similar expression for the Brownian bridge as $BB_t = B_t - tB_1$, we hypothesize that Y_t is equal to

$$Y_t^f = Z_t - f(t)Z_1$$

for some function f . This has covariance function

$$a_{Y^f}(s, t) = a_Z(s, t) - f(t)a_Z(s, 1) - f(s)a_Z(t, 1) + \frac{1}{12}f(s)f(t)$$

For this to equal $a_Y(s, t)$, and thus for Y^f to equal Y , we must have

$$\frac{1}{12}s^2(3s - 2t)t^2(3 - 2t) = f(t)a_Z(s, 1) + f(s)a_Z(t, 1) - \frac{1}{12}f(s)f(t)$$

As $a_Z(s, 1) = \frac{1}{12}s^2(3 - 2s)$, this is satisfied with

$$f(t) = 3t^2 - 2t^3$$

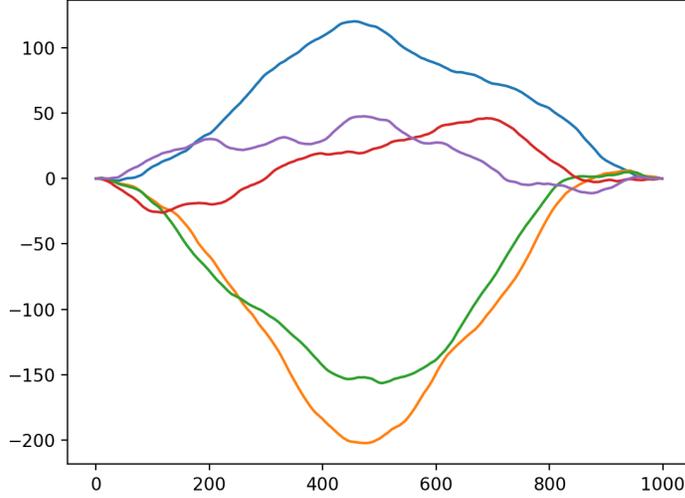


Figure 4-1: Samples from the process Y_t , the integrated Brownian bridge conditioned on having endpoint data 0.

By the uniqueness of Gaussian processes given the covariance function we have identified the process. Notice that $f(t)$ is exactly the cubic with endpoint constraints $f(0) = f'(0) = f'(1) = 0$, $f(1) = 1$. Thus we have shown (inserting the factor of $\sqrt{\varepsilon}$ where it belongs)

Lemma 9. *If $\mathbf{x} = 0$, then the process $Q^{\varepsilon|\mathbf{x}}$ is*

$$X_t^\varepsilon = \sqrt{\varepsilon} Y_t$$

where

$$Y_t = \int_0^t (BB)_s ds - (3t^2 - 2t^3) \int_0^1 (BB)_s ds \quad (4.21)$$

with BB_s being the Brownian bridge.

In figure 4-1 we show a few samples from this process with fixed ε ; its smoothness is to be contrasted with Brownian motion, and notice its endpoint data.

With the same argument we can prove the following.

Proposition 25. *For any \mathbf{x} , the process $Q^{\varepsilon|\mathbf{x}}$ is*

$$X_t^\varepsilon = g(t) + \sqrt{\varepsilon}Y_t$$

where g_t is the unique cubic with endpoint data \mathbf{x} and Y_t is (4.21).

Proof. By subtracting an affine function we can assume that $x_0 = v_0 = 0$, and so we are at the starting point of the argument above with the process $I_t = \int_0^t B_s ds$. Conditioning on $\dot{I}_1 = v_1$ yields the process $\int_0^t (B_s - sB_t + sv_1) ds = \frac{1}{2}t^2v_1 + \int_0^t (BB)_s ds$, so we may subtract this quadratic term as well (since it will become part of the cubic). We are thus reduced to the case where $x_0 = v_0 = v_1 = 0$, $x_1 = 1$. This is easily shown to be true by calculating the mean term in the application of the Gaussian conditioning lemma 7 above, and this term is precisely $t^2(3-2t)$, a cubic. Uniqueness of the admissible cubic concludes. \square

Now, to show a convergence result like Theorem 21, we must show a large deviation principle for X_t^ε , and thus for Y_t^ε .

Proposition 26. *For each \mathbf{x} , the family $Q^{\varepsilon|\mathbf{x}}$ satisfies a large deviation principle on $\Omega^\mathbf{x}$, the set of paths with the requisite endpoint constraints, with coercive rate function*

$$C(\omega) = \frac{1}{2} \int_0^1 |\ddot{\omega}_t - \ddot{g}(t)|^2 dt$$

where $g(t)$ is the unique cubic satisfying the endpoint constraints.

Proof. To begin with, the contraction principle (21) applied to Brownian motion with the function $\theta(\omega)_t = \gamma_t = \omega_t - t\omega_1$ implies that the Brownian bridge family $\sqrt{\varepsilon}(BB)_t$ satisfies an LDP with rate function

$$\begin{aligned} \inf_{\omega_t - t\omega_1 = \gamma_t} \frac{1}{2} \int_0^1 |\dot{\omega}_t|^2 dt &= \inf_{\omega_1} \frac{1}{2} \int_0^1 |\dot{\gamma}_t + \omega_1|^2 dt \\ &= \inf_{\omega_1} \frac{1}{2} \int_0^1 |\dot{\gamma}_t|^2 dt + \omega_1 \int_0^1 \dot{\gamma}_t dt + \frac{1}{2}\omega_1^2 \\ &= \frac{1}{2} \int_0^1 |\dot{\gamma}_t|^2 dt \end{aligned}$$

(Since $\gamma_0 = \gamma_1 = 0$, the middle term on the second line vanishes.) Thus the Brownian bridge satisfies the same LDP as Brownian motion, on the smaller space where both endpoints are zero.

Now, we show that the family $\sqrt{\varepsilon}Y_t$ satisfies a large deviation principle with rate function

$$C(\omega) = \frac{1}{2} \int_0^1 |\ddot{\omega}_t|^2 dt \quad (4.22)$$

on Ω^0 . This is enough to conclude, since we may apply the contraction principle to the function $\theta: \omega \mapsto \omega + g(t)$ to obtain the result.

Letting $Z_t = \int_0^1 (BB)_s ds$, by the exact same logic as for the integrated Brownian motion, $\sqrt{\varepsilon}Z_t$ satisfies an LDP with rate C in (4.22). Next, considering $\theta(\omega)_t = \gamma_t = \omega_t - f(t)\omega_1$ where $f(t) = (3t^2 - 2t^3)$, by the argument in the earlier part of this proof we see $\{\sqrt{\varepsilon}Y_t\}$ has rate function

$$\begin{aligned} \inf_{\omega_t - f(t)\omega_1 = \gamma_t} \frac{1}{2} \int_0^1 |\ddot{\omega}_t|^2 dt &= \inf_{\omega_1} \frac{1}{2} \int_0^1 |\ddot{\gamma}_t + 6\omega_1(1-t)|^2 dt \\ &= \inf_{\omega_1} C(\gamma) + 6\omega_1 \int_0^1 \ddot{\gamma}_t \cdot (1-t) dt + 6\omega_1^2 \\ &= C(\gamma) \end{aligned}$$

where integrating by parts we see the middle term in the second line vanishes. This concludes. \square

In figure 4-2 we illustrate the large deviation principle for our process with a specific cubic (the endpoint data is initial and final position 0, and initial and final velocity 1). Compare this to figure 4-3, which shows convergence for the Brownian bridge.

By the same strategy as before, we can prove an analog of Theorem 21.

Theorem 23. *Let $L_\varepsilon[\mu, v]$ be (4.17) and let $L[\mu, v]$ be (4.14). Then there is a sequence $\mu_1^k \rightarrow \mu_1$ converging in the weak topology such that*

1. $L_\varepsilon[\mu_0, v_0, \mu_1^k, v_1] \rightarrow \frac{1}{2}L[\mu_0, v_0, \mu_1, v_1]$

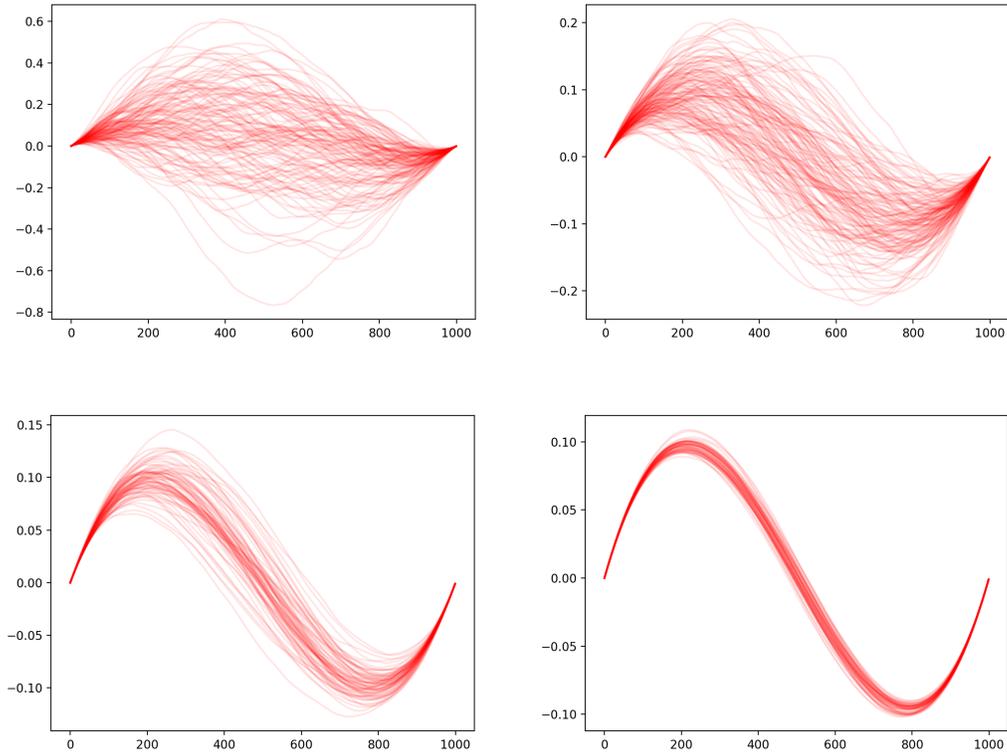


Figure 4-2: Samples from X_t as $\epsilon \rightarrow 0$ for fixed $g(t)$.

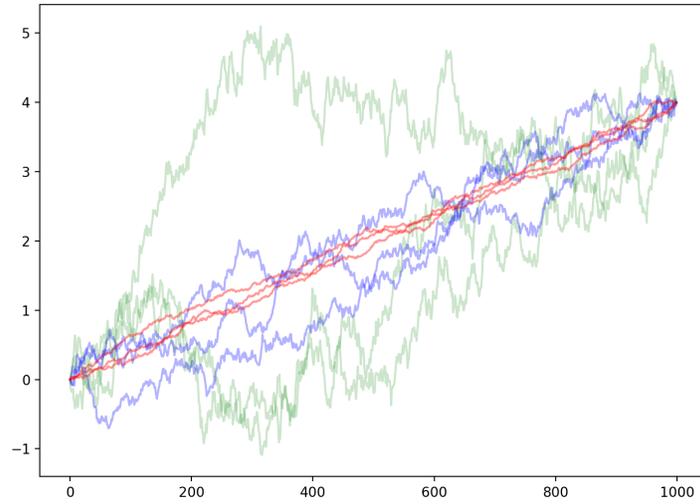


Figure 4-3: Samples from different noise levels of the Brownian bridge, illustrating the large deviations principle for this family.

2. If P_ε^* optimizes $L[\mu_0, v_0, \mu_1^k, v_1]$, then any limit point of the sequence (P_ε^*) optimizes $L[\mu_0, v_0, \mu_1, v_1]$. In particular, if L admits a unique minimizer, then $P_\varepsilon^* \rightarrow P^*$.

Furthermore L admits a unique minimizer when v_0 and v_1 are sufficiently small and μ_0 and μ_1 are absolutely continuous.

Proof. We first prove the final statement; by the equivalence of (4.14) and (4.15), it suffices to show the latter has a unique solution, and we will do so by showing that $c(x, v(x), y, w(y))$, as a function of (x, y) , satisfies the twist condition in Definition 1. We have

$$\nabla_x c(x, y) = \nabla_x c(x, v(x), y, w(y)) + \nabla v(x) \cdot \nabla_v c(x, v(x), y, w(y))$$

and

$$\begin{aligned} \nabla_x c &= -24(y - x - v(x)) + 12(w(y) - v(x)) \\ \nabla_v c &= \nabla_x c - 8(w(y) - v(x)) \end{aligned}$$

so

$$\nabla_{c_x}(x, y) = (\text{Id} + \nabla v(x)) \cdot \nabla_x c(x, v(x), y, w(y)) - 8 \nabla v(x) \cdot (w(y) - v(x))$$

Assuming v and w are sufficiently small, it is sufficient to show that

$$y \mapsto \nabla_x c(x, v(x), y, w(y))$$

is injective. But using that w is small again this is clear from the expression for $\nabla_x c$ above, as it is nearly linear.

In light of the large deviation principle for $Q^{\varepsilon|\mathbf{x}}$, the remainder of the proof is identical to the proof of Theorem 21, so we leave it to the interested reader. \square

Chapter 5

Splines in WFR

5.1 Unbalanced Optimal Transport

By the nature of optimal transport, it is only possible to transport one measure μ to another one ν if $\mu(X) = \nu(X)$; otherwise there can be no feasible coupling between them. Nevertheless, in applications it is sometimes necessary to consider this situation, for instance if the fluid-dynamical perspective above is modified so that the fluid is reactive, and it may increase or decrease in total mass as it evolves. It is tempting to simply rescale and transport $\frac{1}{\mu(X)}\mu$ to $\frac{1}{\nu(X)}\nu$, but this obscures the effects of creation and destruction of mass unique to the problem. For this purpose the Wasserstein-Fisher-Rao metric on non-negative measures, also known as the Hellinger-Kantorovich metric, was introduced simultaneously in [41, 40, 20, 35].

The starting point of the theory is the so-called *non-conservative continuity equation*

$$\partial_t \mu_t + \operatorname{div}(v_t \mu_t) = 4\alpha_t \mu_t \quad (5.1)$$

This modifies the continuity equation with a relative growth term α_t , taking inspiration from the reactive fluid system described above. The Benamou-Brenier theorem indicates that we should define

$$\operatorname{WFR}(\mu_0, \mu_1) = \inf_{(\mu_t, v_t, \alpha_t)} \int_0^1 \int \|v_t(x)\|^2 + 4|\alpha_t(x)|^2 d\mu_t dt \quad (5.2)$$

the infimum being taken over solutions to the non-conservative continuity equation with the prescribed endpoints. (The factor of 4 allows for much simpler notation later.) It was shown simultaneously in [41, 20] that (5.2) defines a metric on $\mathcal{M}_+(\mathbb{R}^d)$, the space of non-negative measures, and that it turns this into a geodesic space. Furthermore, [35] it has the structure of a pseudo-Riemannian manifold analogous to the Riemannian structure on W_2 (see [28]). As in the W_2 case, to define the tangent space we must consider the functions v_t, α_t that minimize the $L^2(\mu_t)$ norm.

Proposition 27. *Let (μ_t, v_t, α_t) solve (5.1), and consider the problem*

$$\min \|\tilde{v}_t\|_{L^2(\mu_t)}^2 + 4\|\tilde{g}_t\|_{L^2(\mu_t)}^2 \text{ s.t. } \operatorname{div}(v_t\mu_t) - 4\alpha_t\mu_t = \operatorname{div}(\tilde{v}_t\mu_t) - 4\tilde{g}_t\mu_t$$

This has a unique solution \bar{v}_t, \bar{g}_t , and

$$(\bar{v}_t, \bar{g}_t) \in \operatorname{clos}_{L^2(\mu_t)} \{(\nabla\varphi, \varphi) \mid \varphi \in \mathcal{C}_c^\infty\}$$

Thus we define the WFR tangent space to be

$$T_\mu(\mathcal{P}_2) = \operatorname{clos}_{L^2(\mu_t)} \{(\nabla\varphi, \varphi) \mid \varphi \in \mathcal{C}_c^\infty\}$$

and equip it with the (pseudo-) Riemannian metric

$$\langle (v, g), (w, h) \rangle_\mu = \langle v, w \rangle_{L^2(\mu)} + 4\langle g, h \rangle_{L^2(\mu)}$$

Notice that this is the same tangent space as W_2 (up to isomorphism) but with the full $H^1(\mu)$ norm of φ , whereas the W_2 metric contains only the first derivatives $\nabla\varphi$. For the benefit of the reader, a few observations are in order:

1. The formula (5.2) defines geodesics between any two non-negative measures of arbitrary mass. If $\mu_1 = \alpha\mu_0$ — the final measure is a simple rescaling of the initial measure — then it can be checked that the WFR geodesic between them

is (assuming without loss of generality that $\alpha < 1$)

$$\mu_t = (1 - t)^2 \mu_0$$

parametrized as $t \in [0, 1 - \sqrt{\alpha}]$. This is exactly the geodesic in the closely related Fisher-Rao metric, which we will not discuss.

2. From this one might expect that the geodesic from μ_0 to $\mu_1(x) = \mu_0(x - x_0)$ is a simple translation, with no creation or destruction of mass. This is not the case. The (non-conservative) continuity equation together with the characterization of $T_\mu(\mathcal{P}_2)$ states that

$$\partial_t \mu_t + \operatorname{div}(\nabla \alpha_t \mu_t) = 4 \alpha_t \mu_t$$

If $\nabla \alpha_t$ is not locally zero, meaning that there is translation, it must be that α_t is not locally constant, and in specific it is not 1, thus mass must be modified.

The WFR metric has several equivalent characterizations that we will exploit in the sequel.

5.1.1 The Cone Space

The *cone space* \mathfrak{C} is the manifold $\mathbb{R}^d \times \mathbb{R}_+$, with $\mathbb{R}^d \times \{0\}$ identified as a single point.¹ Points $(x, r) \in \mathfrak{C}$ are thought of as tuples of position and mass. The metric (as a metric space) is given by

$$d_{\mathfrak{C}}((x_0, r_0), (x_1, r_1))^2 = r_0^2 + r_1^2 - 2r_0 r_1 \cos(|x_0 - x_1| \wedge \pi)$$

Call the vertex of the cone (the point $\mathbb{R}^d \times \{0\}$) as \mathfrak{o} . Notice that if $|x_0 - x_1| \geq \pi$ then $d_{\mathfrak{C}}((x_0, r_0), (x_1, r_1)) = r_0 + r_1$. Since for any (x, r) we have $d_{\mathfrak{C}}((x, r), \mathfrak{o}) = r$, this reflects that fact that if x_0 and x_1 are far away then the shortest path is to the vertex and back out.

¹Fraktur font characters will be reserved for objects relating to the cone space, consistent with [41, 40].

Writing $(v, p) = \frac{d}{dt}(x, r)$, the Riemannian metric on the cone space is given by

$$\langle (v_1, p_1), (v_2, p_2) \rangle_{(x,r)} = \langle v_1, v_2 \rangle r^2 + p_1 p_2 \quad (5.3)$$

Given a measure $\lambda \in \mathcal{M}(\mathfrak{C})$, we can project it to a measure $\mathfrak{P}\lambda \in \mathcal{M}(\mathbb{R}^d)$ via

$$\int f(x) d\mathfrak{P}\lambda(x) = \int r^2 f(x) d\gamma(x, r)$$

The measure λ is then called a lift of $\mathfrak{P}\lambda$. The presence of the term r^2 , as opposed to r , is to simplify the parametrization of \mathfrak{C} . Observe that there are many possible liftings of a measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ to $\lambda \in \mathcal{M}(\mathfrak{C})$, the most obvious of which is $d\lambda(x, r) = \delta_1(r) \cdot d\mu(x)$. As \mathfrak{C} is (save for the point \mathfrak{o}) a Riemannian manifold we can define its W_2 metric as usual

$$W_{\mathfrak{C},2}^2(\lambda, \eta) = \inf_{\gamma \in \Pi(\lambda, \eta)} \int d_{\mathfrak{C}}^2 d\gamma$$

The following theorem is proved in [41].

Theorem 24. *For any measures $\mu_0, \mu_1 \in \mathcal{M}_+(\mathbb{R}^d)$, we have*

$$\text{WFR}(\mu_0, \mu_1) = \inf \{W_{\mathfrak{C},2}(\lambda_0, \lambda_1) \mid \lambda_i \in \mathcal{P}(\mathfrak{C}), \mathfrak{P}\lambda_0 = \mu_0, \mathfrak{P}\lambda_1 = \mu_1\}$$

Furthermore, there are optimal lifts λ_0, λ_1 , and an optimal coupling γ between them.

This allows to characterize WFR geodesics in the same way as in Euclidean space.

Theorem 25. *Let $\mu_0, \mu_1 \in \mathcal{M}_+(\mathbb{R}^d)$, and let $\gamma \in \mathcal{M}(\mathfrak{C}^2)$ be the optimal coupling of the optimal lifts in Theorem 24. For each pair of points $z_0, z_1 \in \mathfrak{C}$, let $g_t^{z_0, z_1}$ be the geodesic between them. Then the measures defined by*

$$\mu_t = \mathfrak{P} [(g_t^{z_1, z_2})_{\#} \gamma]$$

compose a geodesic in WFR between μ_0 and μ_1 .

There is yet another characterization of the WFR metric that will be useful in the

sequel. The Wasserstein metric can be phrased as

$$\inf_{\eta} \int d(x, y)^2 d\eta(x, y) + \chi_{\pi_0\eta=\mu_0}(\eta) + \chi_{\pi_1\eta=\mu_1}(\eta)$$

where χ is the convex analysis indicator (taking values 0 and ∞). We might relax the tight constraints as $\text{KL}(\pi_0\eta \parallel \mu_0)$ — extending the KL divergence to arbitrary nonnegative measures in the obvious way — in order to accommodate measures of non-unit mass. This leads to

Theorem 26. *For any measures μ_0, μ_1 , we have*

$$\text{WFR}^2(\mu_0, \mu_1) = \inf_{\eta \in \mathcal{M}_+} \int c(x, y) d\eta + \text{KL}(\pi_0\eta \parallel \mu_0) + \text{KL}(\pi_1\eta \parallel \mu_1) \quad (5.4)$$

where

$$c(x, y) = -2 \log \cos(|x - y| \wedge \frac{\pi}{2})$$

Optimality is achieved, and if μ_0 and μ_1 are absolutely continuous the optimizing η^ is unique and is induced by a map. Furthermore, if $\alpha_i = d\eta_i^*/d\mu_i$, then for (x, y) on the support of η^* we have $|x - y| < \frac{\pi}{2}$ and*

$$\alpha_0(x)\alpha_1(y) = \cos^2(|x - y|) > 0$$

Also, for μ_0 -a.e. x and μ_1 -a.e. y , $\alpha_0(x)\alpha_1(y) \leq \cos^2(|x - y|)$.

The cost c differs from the term in the cone metric $d_{\mathfrak{C}}$ by taking the minimum against $\frac{\pi}{2}$, not π . This is the difference between transport of points in \mathfrak{C} and measures on \mathfrak{C} ; to transport $z_0 = (x_0, r_0)$ to $z_1 = (x_1, r_1)$ in \mathfrak{C} we may reduce the mass to 0, then increase it again to r_1 , whereas to transport δ_{z_0} to δ_{z_1} these can be done simultaneously, so that the WFR geodesic is at each time a combination of two deltas.

The optimal η in Theorem 26 and the optimal coupling γ in Theorem 24 are intimately related.

Theorem 27. *Suppose η minimizes the objectives in Theorem 26, and let $\eta_i = \pi_i \eta$ be its marginals. Write*

$$\mu_i = \sigma_i \eta_i + \mu_i^\perp$$

where $\sigma_i = d\eta_i/d\mu_i$ and μ_i is mutually singular with μ_i^\perp . Define the plan γ_η by

$$\begin{aligned} d\gamma_\eta(z_0, z_1) &= \delta_{\sqrt{\sigma_0(x_0)}}(r_0) \cdot \delta_{\sqrt{\sigma_1(x_1)}}(r_1) \cdot d\eta(x_0, x_1) \\ &\quad + \delta_1(r_0) \cdot d\mu_0^\perp(x_0) \cdot \delta_0(z_1) + \delta_1(r_1) \cdot d\mu_1^\perp(x_1) \cdot \delta_0(z_0) \end{aligned}$$

Then γ_η is optimal for the objective in Theorem 24.

Let us remark on the somewhat surprising nature of this result. In principle, the optimal coupling of the optimal lifts of μ_0 and μ_1 has no reason to be induced by a map *in the x variable*. Even supposing that the OT problem in Theorem 24 has a Monge solution, it is only Monge in \mathfrak{C} ; there are many points (x, r) that project down to each $x \in \mathbb{R}^d$. Ignoring the singular component, Theorem 27 says that each x_0 maps to a unique x_1 , and also a unique r_0 and r_1 . The singular component governs the mass that is to be created or destroyed completely, and if both measures have the same support or are sufficiently close, then this is zero.

5.1.2 The Covariant Derivative

In this section we find the covariant derivative in WFR; recall from Section 2.2.5 that we must show that it respects the metric and is torsion-free. Repeating the argument from that section, let $\mathbf{u}_t^i = (u_t^i, \beta_t^i)$ be two tangent fields along a curve μ_t , which has derivative (v_t, α_t) . Metric compatibility then reads

$$\begin{aligned} \frac{d}{dt} \langle \mathbf{u}_t^1, \mathbf{u}_t^2 \rangle_{\mu_t} &= \int \langle \partial_t u_t^1, u_t^2 \rangle + \langle u_t^1, \partial_t u_t^2 \rangle + 4\partial_t \beta_t^1 \beta_t^2 + 4\beta_t^1 \partial_t \beta_t^2 d\mu_t \\ &\quad + \int \langle u_t^1, u_t^2 \rangle + 4\beta_t^1 \beta_t^2 d(\partial_t \mu_t) \\ &= T_1 + T_2 \end{aligned}$$

By the continuity equation and the dual definition of $\text{div } v_t \mu_t$, the second integral becomes

$$T_2 = \int \langle \nabla u_t^1 \cdot v_t, u_t^2 \rangle + \langle u_t^2, \nabla u_t^2 \cdot v_t \rangle + 4\alpha_t \langle u_t^1, u_t^2 \rangle \\ + \int 4\beta_t^2 \langle \nabla \beta_t^1, v_t \rangle + 4\beta_t^1 \langle \nabla \beta_t^2, v_t \rangle + 16\alpha_t \beta_t^1 \beta_t^2 d\mu_t$$

Now, we need this to be equal to $\langle \mathbf{u}_t^1, \frac{\mathbf{D}}{dt} \mathbf{u}_t^2 \rangle + \langle \frac{\mathbf{D}}{dt} \mathbf{u}_t^1, \mathbf{u}_t^2 \rangle$. Some terms are uniquely attributable to one or the other, such as $\int \langle \partial_t u_t^1, u_t^2 \rangle$, but some are not, such as $4\beta_t^2 \langle \nabla \beta_t^1, v_t \rangle$. This can arise from either of the two terms: as $\nabla \beta_t^i = u_t^i$,

$$\langle (0, \langle \nabla \beta_t^1, v_t \rangle), \mathbf{u}_t^2 \rangle_{\mu_t} = \langle \mathbf{u}_t^1, (4\beta_t^2 v, 0) \rangle_{\mu_t}$$

Indeed, by splitting these terms and gathering the others, the possible covariant derivatives that satisfy metric compatibility are of the form

$$\frac{\mathbf{D}}{dt} \mathbf{u}_t = \mathcal{P}_{\mu_t} \begin{pmatrix} \partial_t u_t + \nabla u_t \cdot v_t + 2\alpha_t u_t + 4p\beta_t v_t \\ \partial_t \beta_t + (1-p)\langle \nabla \beta_t, v_t \rangle + 2\alpha_t \beta_t \end{pmatrix}$$

for real p . Let us check the torsion-free identity with $p = 1/2$. In this case, if $\mathbf{u}_t = (\nabla \varphi, \varphi)$ is constant in time, then

$$\frac{\mathbf{D}}{dt} \begin{pmatrix} \nabla \varphi \\ \varphi \end{pmatrix} = \mathcal{P}_{\mu_t} \begin{pmatrix} \nabla^2 \varphi \cdot v_t + 2\alpha_t \nabla \varphi + 2\varphi v_t \\ \frac{1}{2} \langle \nabla \varphi, v_t \rangle + 2\alpha_t \varphi \end{pmatrix}$$

Now, with the setup as in the W_2 case, defining $F: \varphi \mapsto \int \varphi d\mu$ and writing $\varphi = (\nabla \varphi, \varphi)$, we have from the continuity equation

$$\partial_t F[\mu_t^i] = \int \langle \varphi, \mathbf{v}_t^i \rangle d\mu_t^i$$

where \mathbf{v}_t^i is the derivative of μ_t^i . As above, we have

$$\begin{aligned}\mathbf{u}_0^1(\mathbf{u}^2(F))[\mu] &= \frac{d}{dt} \langle \varphi, \mathbf{u}_t^2 \rangle_{\mu_t^2} \Big|_{t=0} \\ &= \left\langle \frac{\mathbf{D}}{dt} \varphi, \mathbf{u}_t^2 \right\rangle_{\mu_t^2} + \left\langle \varphi, \nabla_{\mathbf{u}_0^1} \mathbf{u}_t^2 \right\rangle_{\mu_t^2} \Big|_{t=0}\end{aligned}$$

Recalling that at $t = 0$ we have $\mathbf{u}_0^2 = \mathbf{v}_0^1 = (v_0^1, \alpha_0^1)$, the first term becomes (we may ignore the projection, since \mathbf{u}_t^2 is already tangent)

$$Q_1 = \left\langle \left(\begin{array}{c} \nabla^2 \varphi \cdot v_0^2 + 2\alpha_0^2 \nabla \varphi + 2\varphi v_0^2 \\ \frac{1}{2} \langle \nabla \varphi, v_0^2 \rangle + 2\alpha_0^2 \varphi \end{array} \right), \begin{pmatrix} v_0^1 \\ \alpha_0^1 \end{pmatrix} \right\rangle_{\mu}$$

while the corresponding term from $\mathbf{u}_0^2(\mathbf{u}^2(F))[\mu]$ is

$$Q_2 = \left\langle \left(\begin{array}{c} \nabla^2 \varphi \cdot v_0^1 + 2\alpha_0^1 \nabla \varphi + 2\varphi v_0^1 \\ \frac{1}{2} \langle \nabla \varphi, v_0^1 \rangle + 2\alpha_0^1 \varphi \end{array} \right), \begin{pmatrix} v_0^2 \\ \alpha_0^2 \end{pmatrix} \right\rangle_{\mu}$$

and we must check that these agree. The “top-left”, “top-right”, and “bottom-right” terms are identical for both. The top-middle of the first is equal to the bottom-left for the second, and vice-versa. Thus we have shown that the covariant derivative is given by

$$\frac{\mathbf{D}}{dt} \mathbf{u}_t = \mathcal{P}_{\mu_t} \begin{pmatrix} \partial_t u_t + \nabla u_t \cdot v_t + 2\alpha_t u_t + 2\beta_t v_t \\ \partial_t \beta_t + \frac{1}{2} \langle \nabla \beta_t, v_t \rangle + 2\alpha_t \beta_t \end{pmatrix}$$

and in specific,

$$\frac{\mathbf{D}^2}{dt^2} \mu_t = \begin{pmatrix} \partial_t v_t + \nabla v_t \cdot v_t + 4\alpha_t v_t \\ \partial_t \alpha_t + \frac{1}{2} |\nabla \alpha_t|^2 + 2\alpha_t^2 \end{pmatrix} \quad (5.5)$$

This quantity is tangent, so no projection is necessary.

This generalizes the results in [41, 20] that characterize geodesics in WFR space by the equation

$$\partial_t \alpha_t + \frac{1}{2} |\nabla \alpha_t|^2 + 2\alpha_t^2 = 0$$

5.2 Splines

As in the W_2 case, we can define an intrinsic notion of curvature-minimizing interpolators, which we also term *E-splines*, by

$$\inf_{(\mu_t, \mathbf{v}_t)} \int_0^1 \left\| \frac{\mathbf{D}}{dt} \mathbf{v}_t \right\|_{\mu_t}^2 dt \text{ s.t. } \mu_{t_i} = \mu_i \quad (5.6)$$

Though the characterization (5.5) yields an explicit objective function, there is no practical way to optimize (5.6). Thus we first wish to define an analog of *P-splines*. From the discussion of Section 5.1.1, we define *P-splines* in WFR by

$$\inf_{P \in \mathcal{P}(\Omega_{\mathfrak{C}})} \int \int_0^1 |\ddot{z}_t|^2 dt dP(z) \text{ s.t. } \mathfrak{P}[(e_{t_i})_{\#} P] = \mu_i \quad (5.7)$$

We prove that this is indeed a relaxation of the *E-spline* problem (5.6).

Proposition 28. *Let μ_t be a sufficiently smooth curve in WFR. Then there is a measure $P \in \mathcal{P}(\Omega_{\mathfrak{C}})$ such that $\mathfrak{P}[(e_t)_{\#} P] = \mu_t$ for all t , and the *E-cost* of μ is equal to the *P-cost* of P . The measure P is induced by the flow maps associated to the curve μ_t .*

To present the proof (and complete the description of the proposition) we must define the flow maps on WFR. These have the same particle trajectories as in 2.4, but with varying mass. Similarly to W_2 , there is an identification of (sufficiently smooth) paths in WFR with their flow maps, due to [42].

Proposition 29. *Let $v \in L^1(W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d), [0, 1])$ be a vector field and $\alpha \in \mathcal{C}(\mathbb{R}^d \times [0, 1])$ a bounded locally Lipschitz scalar function. For $\mu_0 \in \mathcal{M}_+(\mathbb{R}^d)$, there is a unique weak solution to the nonconservative continuity equation (5.1) with initial measure μ_0 . Furthermore, this satisfies*

$$\mu_t = (X_t)_{\#}(R_t^2 \cdot \mu_0), \quad (5.8)$$

for the flow map (X_t) and scalar field (R_t) which solve the ODE system

$$\begin{cases} \dot{X}_t = v_t(X_t), & X_0 = \text{Id} \\ \dot{R}_t = 2\alpha_t(X_t) R_t, & R_0 = 1 \end{cases}$$

We are now in a position to present the proof of Proposition 28.

Proof of proposition 28. Making the definition explicit, we wish of our measure P that the cost of (5.7) is equal to

$$\int_0^1 \int |\partial_t v_t + \nabla v_t \cdot v_t + 4\alpha_t v_t|^2 + 4 \left(\partial_t \alpha_t + \frac{1}{2} |\nabla \alpha_t|^2 + 2\alpha_t^2 \right)^2 d\mu_t dt \quad (5.9)$$

Let $\lambda_0 \in \mathcal{P}(\mathfrak{C})$ be a lift of μ_0 . By Proposition 29, the measure $\lambda_t = (X_t, R_t)_\# \lambda_0$ is a lift of μ_t .

In order to compute the second covariant derivative of a curve on the cone, we compute the Christoffel symbols. These are given by the formulas

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k, \quad \Gamma_{ij}^k = \frac{1}{2} g^{kl} \left(\frac{\partial}{\partial_j} g_{il} + \frac{\partial}{\partial_i} g_{jl} - \frac{\partial}{\partial_l} g_{ij} \right).$$

Let $\partial_1, \dots, \partial_n, \partial_r$ be the coordinate basis of the tangent space on the cone, hence

$$(g_{ij})_{ij} = \begin{pmatrix} r^2 I_n & 0 \\ 0 & 1 \end{pmatrix}, \quad \frac{\partial}{\partial_k} g_{ij} = \begin{cases} 2r & i = j \neq r, k = r, \\ 0, & \text{else.} \end{cases}$$

Since (g^{ij}) and $\partial_r g_{ij}$ are only defined along the diagonal and $\partial_l g_{ij} = 0$ for $l \neq r$, the only terms that do not vanish are

$$\Gamma_{rj}^k = r^{-1} \delta_{jk}, \quad \Gamma_{ij}^r = -r \delta_{ij}$$

thus the Levi-Civita connection on the cone is given by

$$\nabla_{\partial_r} \partial_r = 0, \quad \nabla_{\partial_r} X = r^{-1} X, \quad \nabla_{X_1} X_2 = -r \langle X_1, X_2 \rangle \partial_r.$$

From this, the second covariant derivative of a curve $z_t = (x_t, r_t)$ on the cone is given by

$$\ddot{z}_t = \nabla_{\dot{z}_t} \dot{z}_t = \left(\ddot{x}_t + 2\frac{\dot{r}_t}{r_t} \dot{x}_t, \ddot{r}_t - r_t |\dot{x}_t|^2 \right)$$

thus from the Riemannian metric (5.3)

$$|\ddot{z}_t|^2 = |r_t \ddot{x}_t + 2\dot{r}_t \dot{x}_t|^2 + |\ddot{r}_t - r_t |\dot{x}_t|^2|^2$$

Let λ_0 be any lift of μ_0 and define P to place mass $\lambda_0(z_0)$ on the path $z_t = (X_t(x_0), R_t(r_0))$, so that P is supported on the flow map curves in $\Omega_{\mathfrak{C}}$. By applying the total derivative to the defining equations of the flow maps, these curves satisfy

$$\begin{aligned} \dot{x}_t &= v_t(x_t) \\ \ddot{x}_t &= \partial_t v_t + \nabla v_t \cdot v_t \\ \dot{r}_t &= 2r_t \alpha_t(x_t) \\ \ddot{r}_t &= 2r_t (\partial_t \alpha_t + \nabla \alpha_t \cdot v_t + 2\alpha_t^2) \end{aligned}$$

Now, we have

$$\int_{\Omega_{\mathfrak{C}}} \int_0^1 |\ddot{z}_t|^2 dt dP(z) = \int_{\mathfrak{C}} \int_0^1 |\ddot{z}_t|^2 dt d\lambda_0(z_0)$$

Expanding out, this is

$$\int_{\mathfrak{C}} \int_0^1 |r_t \ddot{x}_t + 2\dot{r}_t \dot{x}_t|^2 + |\ddot{r}_t - r_t |\dot{x}_t|^2|^2 dt d\lambda_0(z_0)$$

Let us deal with each term separately so the expressions do not become unwieldy.

For the first

$$\begin{aligned}
& \int_{\mathfrak{E}} \int_0^1 |r_t \ddot{x}_t + 2\dot{r}_t \dot{x}_t|^2 dt d\lambda_0(z_0) \\
&= \int_{\mathfrak{E}} \int_0^1 r_t^2 |\partial_t v_t + \nabla v_t \cdot v_t + 4\alpha_t v_t|^2(x_t) dt d\lambda_0(z_0) \\
&= \int_0^1 \int_{\mathfrak{E}} |\partial_t v_t + \nabla v_t \cdot v_t + 4\alpha_t v_t|^2(x_t) d(r_t^2 \cdot \lambda_0)(z_0) dt \\
&= \int_0^1 \int |\partial_t v_t + \nabla v_t \cdot v_t + 4\alpha_t v_t|^2(x) d\mu_t(x) dt
\end{aligned}$$

We have used that $(X_t, R_t)_\# \lambda_0 = \lambda_t$ and $\mathfrak{P}\lambda_t = \mu_t$. The second term is dealt with in exactly the same way. \square

5.3 Transport Splines

To alleviate problems of computational tractability and interpretability in applications, above we introduced transport splines on W_2 , and in this section we will define a similar family of curves.

To begin with we need a map between measures on W_2 associated to a geodesic. The characterization in Theorem 24 is not directly useful here, since it is not obvious that the optimal coupling of lifted measures is induced by a map (and furthermore we would need to know the optimal lifts). Even if it were, that would not be enough for our purposes — we would want to associate a unique mass r to each initial point x , and map that to another unique r' and x' . Instead we turn to Theorem 26, and Theorem 27.

Suppose the μ_0, μ_1 are such that $\mu_i \ll \eta_i$, so that $\mu_i^\perp = 0$ in Theorem 27. This happens, for instance, when the conditions of Proposition 29 are satisfied along the geodesic between them. The optimal η for Theorem 26 is supported on a map T , and thus an optimal γ_η for Theorem 24 is supported on the assignment

$$(x_0, r_0(x_0)) \rightarrow (T(x_0), r_1(T(x_0)))$$

which associates to each x_0 a unique mass $r_0(x_0) = \sqrt{\sigma_0(x_0)}$ and maps it to another unique location and mass $r_1(T(x_0)) = \sqrt{\sigma_1(T(x_0))}$. This is much stronger than γ_η being induced by a map on \mathfrak{C} .

Now, define the operator $S_t^{\mathfrak{C}}[z_0, \dots, z_N]$ to be the Riemannian cubic interpolant in \mathfrak{C} of the points z_0, \dots, z_N . We define *transport splines over WFR* by the following procedure

1. For each i solve (5.4) between μ_i and μ_{i+1} to obtain the optimal coupling $\eta_{i \rightarrow i+1}$, and thus the map $T_{i \rightarrow i+1}$. Let $T_i = T_{0 \rightarrow 1} \circ \dots \circ T_{i-1 \rightarrow i}$.
2. For each x , compute the sequence of masses $r_i(x) = d\mu_i(T_i(x))$.
3. For each x , form a path $X_t(x)$ interpolating the $x_i = T_i(x)$, and a mass path $R_t(x)$ interpolating the masses $\sqrt{r_i(x)}$.
4. Define the interpolating curve by $d\mu_t(x) = R_t^2(x)d((X_t)_\# \mu_0)$ by pushing masses along the curve X_t

We have left the interpolating strategy in step 3 general for two reasons. First, though ideally one would interpolate the masses $\sqrt{\sigma_i(x_i)}$, this is not feasible since two adjacent couplings $\eta_{i-1,i}$ and $\eta_{i,i+1}$ may not have the same i -th marginal, so σ_i is not well-defined. Second, though ideally for any sequence $(x_i, r_i)_i$ one would interpolate using Riemannian cubics on \mathfrak{C} , these are difficult to compute; unlike the Euclidean case, there appears to be no closed formula. We suggest then interpolating the position and mass separately. For the positions x_i we use Euclidean cubics; for the masses we may interpolate $(\sqrt{r_i})$ using a cubic spline, which will then be positive once we square it, or we may use techniques for approximating curvature-minimizing interpolants that are constrained to be positive.

Appendix A

Details for Numerical Experiments

In this section we provide further details for the experiments in this work.

A.1 Figure 3-1

In this figure, we set five Gaussians as our interpolation knots, alternating between

$$\mathcal{N}\left(\begin{bmatrix} 7(k-1) \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\right) \quad \text{for } k \text{ odd}$$

and

$$\mathcal{N}\left(\begin{bmatrix} 7(k-1) \\ 7 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}\right) \quad \text{for } k \text{ even,}$$

where $k = 1, \dots, 5$.

To determine the linear and cubic spline interpolations we first computed the optimal transport maps between the neighboring Gaussians. The closed-form formula for the Monge map from $\mathcal{N}(\mu_1, \Sigma_1)$ to $\mathcal{N}(\mu_2, \Sigma_2)$ is

$$T(x) = \mu_2 + A(x - \mu_1), \quad A = \Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}.$$

The gray lines in both figures show the trajectories of individual sample points along

our interpolations. To draw them, we obtained a sample X_0 from the Gaussian at time $t = 0$, repeatedly applied the Monge maps between successive Gaussians in time, and fit a piecewise linear or natural cubic spline through these points as described in Section 3.4.

Since the maps between successive Gaussians are linear and the formula for the linear or natural cubic spline is linear in its knots, the value of the spline $S_t(X_0)$ interpolation at time t is linear in X_0 . Hence, given the covariance of the Gaussian at time $t = 0$, we used this linear map S_t to compute the covariance of the interpolated Gaussian at time t . Likewise, by taking a linear or cubic spline through the means of the Gaussians at the knot points, we obtained the means of the interpolated Gaussians at any given time. Using this information, we plotted the interpolated Gaussians at the halfway points between the knots for both the linear and cubic spline interpolations.

A.2 Figure 3-5

Here we give more details on the thin-plate spline interpolation leading to Figure 3-5. The data is a representation of the temperature at various weather stations throughout California on June 1 of each year in a thirty year period. That is, we consider the *distribution* of temperatures recorded on each of June 1, 1981, June 1, 1982, \dots , June 1, 2010, and we model this distribution as Gaussian (characterized by its mean and standard deviation). This data is processed and released each decade by the NOAA NCEI [4]. We interpolate these measures using our transport spline technique, obtaining Gaussian measures at each point in California. The left side of Figure 3-5 summarizes these measures by their quantiles, while the right side illustrates the behavior of our method as we sample increasingly many weather stations. The median temperature in the top left quantile plot is taken to be equal to the mean temperature due to our assumption that the temperature distribution is Gaussian at every location. Though there are 484 stations in the NOAA dataset, we used substantially fewer to better capture the convergence of our method.

A.3 Figure 3-3

To simulate the n -body trajectories, we used the Python `nBody` simulator by Cabrera & Li, which can be accessed at <https://github.com/GabrielSCabrera/nBody>.

We created 15 smaller bodies, each of mass 5×10^9 and radius 1. Each body was initialized with a position x and a velocity v drawn randomly according to

$$x \sim \mathcal{N}\left(\begin{bmatrix} 100 \\ 100 \end{bmatrix}, \begin{bmatrix} 30 & 0 \\ 0 & 20 \end{bmatrix}\right), \quad v \sim \mathcal{N}\left(\begin{bmatrix} 10 \\ -20 \end{bmatrix}, \begin{bmatrix} 20 & 0 \\ 0 & 10 \end{bmatrix}\right).$$

In addition, we also created one larger body, with mass 10^{11} and radius 10, initialized at the origin with no initial velocity.

We simulated the trajectories of the planets for 5 seconds sampled every 0.02 seconds. We took the positions of the bodies at 5 evenly spaced times as the knots for our interpolation. In order to solve the matching problem between planets at neighboring knot times, we placed a uniform empirical distribution over the planets at both times and used the Python Optimal Transport (POT) library function `ot.emd` to compute the Monge map between these two distributions. We checked post process that the Monge maps computed were indeed valid matchings (i.e. permutation matrices).

Given the Monge maps between knots, we applied Algorithm 1 to interpolate the empirical distributions of the bodies using cubic splines. Note that in our cubic spline reconstruction, it is possible to observe mistakes in the matching, i.e., the Monge map may not necessarily map a body at one time to the same body at a future time. Such mismatches seem unavoidable without using a more sophisticated method which takes into account the physical model in the simulation.

Appendix B

Details on Natural Cubic Splines

In the main text we require some technical results on natural cubic splines, which we collect here.

We are given times $0 = t_0 < t_1 < \dots < t_N = 1$ and corresponding points $(x_{t_0}, x_{t_1}, \dots, x_{t_N})$ in \mathbb{R}^d . Our goal is to construct a piecewise cubic polynomial interpolation $y : [0, 1] \rightarrow \mathbb{R}^d$ which is \mathcal{C}^2 smooth.

We parametrize y in the following way: for each $i \in [N]$ and for $t \in [t_{i-1}, t_i]$, we set $y(t) = y_i(t)$, where

$$y_i(t) = a_i (t - t_{i-1})^3 + b_i (t - t_{i-1})^2 + c_i (t - t_{i-1}) + d_i$$

Computing derivatives,

$$\begin{aligned}x_{t_{i-1}} &= y_i(t_{i-1}) = d_i, \\x_{t_i} &= y_i(t_i) = a_i \delta_i^3 + \frac{m_i}{2} \delta_i^2 + c_i \delta_i + d_i, \\ \dot{y}_i(t_{i-1}) &= c_i, \\ \dot{y}_i(t_i) &= 3a_i \delta_i^2 + m_i \delta_i + c_i, \\ \ddot{y}_i(t_{i-1}) &= m_i, \\ \ddot{y}_i(t_i) &= 6a_i \delta_i + m_i\end{aligned}$$

where define $\delta_i := t_i - t_{i-1}$ and $m_i := 2b_i$ (and anticipating the natural boundary

condition, which asserts $\ddot{y}(0) = \ddot{y}(1) = 0$, we make the convention $m_{N+1} := 0$). Using continuity of the first and second derivatives of y at the knots, we solve for the coefficients of the polynomial y_i in terms of the variables m and x :

$$\begin{aligned} a_i &= \frac{m_{i+1} - m_i}{6\delta_i}, \\ b_i &= \frac{m_i}{2}, \\ c_i &= \frac{x_{t_i} - x_{t_{i-1}}}{\delta_i} - \frac{m_{i+1} + 2m_i}{6} \delta_i, \\ d_i &= x_{t_{i-1}} \end{aligned}$$

Therefore, it suffices to work with the variables m .

If we plug these equations back into the continuity condition for the first derivative at the knot, after some algebra we obtain the equations

$$6\Delta_i = \delta_i m_i + 2(\delta_i + \delta_{i+1})m_{i+1} + \delta_{i+2}m_{i+2}, \quad i = 1, \dots, N-1,$$

where we have defined the quantities

$$\Delta_i := \frac{x_{t_{i+1}} - x_{t_i}}{\delta_{i+1}} - \frac{x_{t_i} - x_{t_{i-1}}}{\delta_i}$$

a proxy for the second derivative of the data points.

We can express these equations in matrix form (including also the natural boundary condition $m_1 = 0$):

$$\underbrace{\begin{bmatrix} 2(\delta_1 + \delta_2) & \delta_2 & & & \\ \delta_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \delta_{N-1} & \\ & & \delta_{N-1} & 2(\delta_{N-1} + \delta_N) & \\ & & & & \end{bmatrix}}_{:=\mathbf{T}} m = 6\Delta$$

The matrix \mathbf{T} above is a symmetric tridiagonal matrix of size $N-1$.¹ To obtain

¹To be precise, we should write this as the block matrix equation $(\mathbf{T} \otimes I_d)m = 6\Delta$.

bounds on m , we will study the inverse \mathbf{T}^{-1} of \mathbf{T} .

Lemma 10. *Assume that for each $i \in [N]$, we have $\alpha\delta \leq t_i - t_{i-1} \leq \delta$. Then, we have the entrywise bound*

$$\left| (\mathbf{T}^{-1})_{i,j} \right| \leq \frac{1}{4\alpha^2 (1 + \alpha)^{|i-j|-1}} \frac{1}{\delta}, \quad i, j \in [N - 1]$$

Proof. We write $\mathbf{T} = \mathbf{B} + \mathbf{D}$, where

$$\mathbf{B} := \begin{bmatrix} 0 & \delta_2 & & & \\ \delta_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \delta_{N-1} & \\ & & & \delta_{N-1} & 0 \end{bmatrix},$$

$$\mathbf{D} := 2 \operatorname{diag}(\delta_1 + \delta_2, \dots, \delta_{N-1} + \delta_N)$$

Therefore,

$$\begin{aligned} \mathbf{T}^{-1} &= (\mathbf{B} + \mathbf{D})^{-1} \\ &= \mathbf{D}^{-1/2} (\mathbf{I}_{N-1} + \mathbf{D}^{-1/2} \mathbf{B} \mathbf{D}^{-1/2})^{-1} \mathbf{D}^{-1/2} \\ &= \sum_{k=0}^{\infty} (-1)^k \mathbf{D}^{-1/2} \underbrace{(\mathbf{D}^{-1/2} \mathbf{B} \mathbf{D}^{-1/2})^k}_{:=\mathbf{M}} \mathbf{D}^{-1/2} \end{aligned}$$

The matrix \mathbf{M} is

$$\mathbf{M} = \begin{bmatrix} 0 & \gamma_2 & & & \\ \gamma_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \gamma_{N-1} & \\ & & & \gamma_{N-1} & 0 \end{bmatrix}$$

where we set

$$\gamma_i := \frac{\delta_i}{2\sqrt{(\delta_{i-1} + \delta_i)(\delta_i + \delta_{i+1})}} \leq \frac{1}{2(1 + \alpha)}.$$

Since \mathbf{M} has non-negative entries, we have the entrywise bound

$$\mathbf{M}^k \leq \frac{1}{\{2(1+\alpha)\}^k} \underbrace{\begin{bmatrix} 0 & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & & 1 \\ & & & 1 & 0 \end{bmatrix}}_{:=\mathbf{A}}^k$$

The matrix \mathbf{A} is the adjacency matrix of the path graph on $\{1, \dots, N-1\}$, so $(\mathbf{A}^k)_{i,j}$ is the number of paths from i to j of length k . We can trivially bound this number by $2^k \mathbb{1}_{|i-j| \leq k}$. From this we deduce the entrywise bound

$$(\mathbf{M}^k)_{i,j} \leq \frac{1}{(1+\alpha)^k} \mathbb{1}_{|i-j| \leq k}$$

Therefore,

$$\begin{aligned} |(\mathbf{T}^{-1})_{i,j}| &\leq \sum_{k=0}^{\infty} \frac{1}{2\sqrt{(\delta_i + \delta_{i+1})(\delta_j + \delta_{j+1})}} \frac{\mathbb{1}_{|i-j| \leq k}}{(1+\alpha)^k} \\ &\leq \frac{1}{4\alpha\delta} \sum_{k=|i-j|}^{\infty} \frac{1}{(1+\alpha)^k} = \frac{1}{4\alpha^2(1+\alpha)^{|i-j|-1}} \frac{1}{\delta} \end{aligned}$$

□

Bibliography

- [1] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1961–1971, 2017.
- [2] Luigi Ambrosio and Nicola Gigli. *A User’s Guide to Optimal Transport*, pages 1–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
- [3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
- [4] Anthony Arguez, Imke Durre, Scott Applequist, Mike Squires, Russell Vose, Xungang Yin, and Rocky Bilotta. Noaa’s u.s. climate normals (1981-2010) [daily], 2010.
- [5] Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. Bayesian learning with Wasserstein barycenters. *arXiv e-prints*, 2018.
- [6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [7] Jean-David Benamou, Thomas O. Gallouët, and François-Xavier Vialard. Second-order models for optimal transport and cubic splines on the Wasserstein space. *Found. Comput. Math.*, 19(5):1113–1143, 2019.
- [8] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures-Wasserstein distance between positive definite matrices. *Expo. Math.*, 37(2):165–191, 2019.
- [9] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. Henri Poincaré Probab. Stat.*, 53(1):1–26, 2017.
- [10] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electron. J. Stat.*, 12(2):2253–2289, 2018.

- [11] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1 – 26, 2017.
- [12] Garrett Birkhoff and Carl de Boor. Error bounds for spline interpolation. *J. Math. Mech.*, 13:827–835, 1964.
- [13] Geir Bogfjellmo, Klas Modin, and Olivier Verdier. Numerical algorithm for c2-splines on symmetric spaces. *SIAM Journal on Numerical Analysis*, 56, 03 2017.
- [14] Andrea Braides. Chapter 2 a handbook of γ -convergence. volume 3 of *Handbook of Differential Equations: Stationary Partial Differential Equations*, pages 101–213. North-Holland, 2006.
- [15] Yongxin Chen, Giovanni Conforti, and Tryphon T. Georgiou. Measure-valued spline curves: an optimal transport viewpoint. *SIAM J. Math. Anal.*, 50(6):5947–5968, 2018.
- [16] Yongxin Chen, Giovanni Conforti, Tryphon T Georgiou, and Luigia Ripani. Multi-marginal schrödinger bridges. In *International Conference on Geometric Science of Information*, pages 725–732. Springer, 2019.
- [17] Sinho Chewi, Julien Clancy, Thibaut Le Gouic, Philippe Rigollet, George Stepaniants, and Austin J. Stromme. Fast and smooth interpolation on wasserstein space, 2020.
- [18] Sinho Chewi, Julien Clancy, Thibaut Le Gouic, Philippe Rigollet, George Stepaniants, and Austin Stromme. Fast and smooth interpolation on wasserstein space. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3061–3069. PMLR, 13–15 Apr 2021.
- [19] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1276–1304. PMLR, 2020.
- [20] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher-Rao metrics. *Found. Comput. Math.*, 18(1):1–44, 2018.
- [21] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Beijing, China, 2014.

- [22] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [23] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.
- [24] Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *Journal of Machine Learning Research*, 19(41):1–42, 2018.
- [25] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2010. Corrected reprint of the second (1998) edition.
- [26] Manfredo P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.
- [27] Manfredo P. do Carmo. *Differential geometry of curves & surfaces*. Dover Publications, Inc., Mineola, NY, 2016.
- [28] Nicola Gigli. Second order analysis on $(\mathcal{P}_2(M), W_2)$. *Mem. Amer. Math. Soc.*, 216(1018):xii+154, 2012.
- [29] Pierre-Yves Gousenbourger, Estelle Massart, and P.-A. Absil. Data fitting on manifolds with composite Bézier-like curves and blended cubic splines. *J. Math. Imaging Vision*, 61(5):645–671, 2019.
- [30] A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In Sebastien Ourselin, Daniel C. Alexander, Carl-Fredrik Westin, and M. Jorge Cardoso, editors, *Information Processing in Medical Imaging*, pages 261–272, Cham, 2015. Springer International Publishing.
- [31] Charles A. Hall and W. Weston Meyer. Optimal error bounds for cubic spline interpolation. *J. Approximation Theory*, 16(2):105–122, 1976.
- [32] A. Karimi, L. Ripani, and T. T. Georgiou. Statistical learning in wasserstein space. *IEEE Control Systems Letters*, 5(3):899–904, 2021.
- [33] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

- [34] Soheil Kolouri and Gustavo 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 (CVPR)*, June 2015.
- [35] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12):1117 – 1164, 2016.
- [36] Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *arXiv e-prints*, 2019.
- [37] Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.
- [38] Christian Léonard. From the Schrödinger problem to the Monge-Kantorovich problem. *J. Funct. Anal.*, 262(4):1879–1920, 2012.
- [39] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.*, 34(4):1533–1574, 2014.
- [40] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: the Hellinger-Kantorovich distance and geodesic curves. *SIAM J. Math. Anal.*, 48(4):2869–2911, 2016.
- [41] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Invent. Math.*, 211(3):969–1117, 2018.
- [42] Stefania Maniglia. Probabilistic representation and uniqueness results for measure-valued solutions of transport equations. *Journal de Mathématiques Pures et Appliquées*, 87(6):601–626, 2007.
- [43] Estelle Massart, Julien M. Hendrickx, and P.-A. Absil. Curvature of the manifold of fixed-rank positive-semidefinite matrices endowed with the Bures-Wasserstein metric. In *Geometric science of information*, volume 11712 of *Lecture Notes in Comput. Sci.*, pages 739–748. Springer, Cham, 2019.
- [44] Klas Modin. Geometry of matrix decompositions seen through optimal transport and information geometry. *J. Geom. Mech.*, 9(3):335–390, 2017.
- [45] T Muthukumar. Introduction to γ -convergence.
- [46] Lyle Noakes, Greg Heinzinger, and Brad Paden. Cubic splines on curved spaces. 6(4):465–473.
- [47] Victor M. Panaretos and Yoav Zemel. Amplitude and phase variation of point processes. *Ann. Statist.*, 44(2):771–812, 2016.

- [48] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- [49] Tomasz Popiel and Lyle Noakes. Bézier curves and c_2 interpolation in riemannian manifolds. *Journal of Approximation Theory*, 148(2):111–127, 2007.
- [50] R. Michael Range. *Holomorphic functions and integral representations in several complex variables*, volume 108 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1986.
- [51] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe’er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. Science forum: The human cell atlas. *eLife*, 6:e27041, 2017.
- [52] Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Inf. Inference*, 8(4):691–717, 2019.
- [53] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.
- [54] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928 – 943.e22, 2019.
- [55] E. Schrödinger. Über die umkehrung der naturgesetze. *Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math.*, 144:144–153, 1931.
- [56] E. Schrödinger. Sur la théorie relativiste de l’électron et l’interprétation de la mécanique quantique. *Ann. Inst. H. Poincaré*, 2(4):269–310, 1932.
- [57] Amit Singer. From graph to manifold Laplacian: the convergence rate. *Appl. Comput. Harmon. Anal.*, 21(1):128–134, 2006.

- [58] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
- [59] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.
- [60] Grace Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics, 1990.
- [61] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648, 2019.