Inverse Statistical Problems

Luca Raffo
EPFL, Institute of Mathematics - luca.raffo@epfl.ch
Capital Fund Management - luca.raffo@cfm.com

July 2025

Introduction

In statistical physics, *forward problems* consist in predicting macroscopic behavior of a system starting from a known microscopic description of it. For example, given the system's Hamiltonian, one can compute average quantities by evaluating expectations with respect to the corresponding Boltzmann distribution.

Inverse statistical problems reverse this perspective: from empirical data, one aims to infer the underlying interaction parameters. In a more general scenario, not even the model is fixed, and we could use some principle to assume a reasonable one (e.g. maximum entropy principle with fixed first two moments).

A prototypical example is the inverse Ising problem: given access to samples $\sigma^{(1)}, \ldots, \sigma^{(n)} \in \{-1, +1\}^d$ drawn from an unknown distribution of the form

$$P(\sigma) \propto \exp\left(\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\right),$$

the goal is to estimate the couplings J_{ij} and external fields h_i.

In this manuscript we will start by giving some theoretical results that will be used throughout. Then we will address inverse statistical problems by describing a selection of methods, both in the continuous and discrete cases, together with the numerical experiments.

Contents

Chapter 1

Framework

This chapter sets the theoretical foundations and notations that will be used throughout the manuscript. We begin by reviewing the forward perspective of statistical mechanics, where one computes observables from a known microscopic model. Then, we shift to the inverse setting, where the objective is to reconstruct the unknown parameters of a model from empirical data.

The general framework we adopt is probabilistic: given a probability distribution P_θ over a configuration space (e.g., spin configurations $\{-1, +1\}^d$ or continuous states in $\mathbb{R}^d$), we aim to describe both the forward problem (computing expectations with respect to P_θ), and the inverse problem (estimating the structure or parameters that generated P_θ).

Furthermore, we fix the notation and conventions used throughout the rest of the manuscript, such as the meaning of empirical distributions, expectation operators, and the form of energy-based models.

1.1 Forward computations

In the forward setting of statistical mechanics, the goal is to compute macroscopic observables from a known microscopic model, described by an Hamiltonian (or energy function) H_θ. These computations are typically performed with respect to a probability distribution of the Boltzmann–Gibbs type, which is induced by the Hamiltonian over the configuration space.

In particular, we are often interested in evaluating expectations of certain observables under this distribution, which takes the form

$$P_\theta(x) = \frac{1}{Z(\theta)} e^{-H_\theta(x)},$$

where $H_\theta(x)$ is the energy function parametrized by θ, and $Z(\theta)$ is the normalizing constant, also known as the *partition function*, or *Z function.*

1.1.1 Z function

The partition function

$$Z(\theta) = \int_{\Omega_\theta} e^{-H_\theta(x)}\, d\mu(x)$$

(the measure $\mu(x)$ could be for instance the Lebesgue measure for continuous models, or the counting measure for the discrete ones) plays a central role in statistical mechanics: it ensures the normalization of the Boltzmann–Gibbs distribution and encodes the thermodynamic properties of the system. However, in most nontrivial models, computing $Z(\theta)$ is a highly nontrivial task.

Example 1.1.1 (Ising model). In the classical Ising model on d spins, the configuration space is $\{-1,+1\}^d$, and the Hamiltonian is given by

$$H_{J,h}(\sigma) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i.$$

The corresponding partition function reads

$$Z(J,h) = \sum_{\sigma\in\{-1,+1\}^d} \exp\left(\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\right).$$

This sum runs over 2^d configurations, which becomes computationally intractable already for moderate values of d, due to the exponential growth of the state space.

The intractability of $Z(\theta)$ has practical consequences. For instance, computing expectations such as the magnetization

$$\mathbb{E}_{P_\theta}[\sigma_i] = \frac{1}{Z(\theta)} \sum_\sigma \sigma_i e^{-H_\theta(\sigma)},$$

or pairwise correlations $\mathbb{E}_{P_\theta}[\sigma_i\sigma_j]$ becomes unfeasible in closed form.

To address this issue, one typically resorts to approximate methods, the most important of which are MCMC techniques. These methods allow to sample approximately from the Boltzmann distribution without computing $Z(\theta)$ explicitly. Once samples $\sigma^{(1)}, \dots, \sigma^{(n)} \sim P_\theta$ are obtained, empirical estimates can be formed:

$$\mathbb{E}_{P_\theta}[\sigma_i] \approx \frac{1}{n}\sum_{k=1}^{n}\sigma_i^{(k)},\ \mathbb{E}_{P_\theta}[\sigma_i\sigma_j] \approx \frac{1}{n}\sum_{k=1}^{n}\sigma_i^{(k)}\sigma_j^{(k)}.$$

Even with this method, for big models the approach is too expensive in terms of computing time.

1.2 Inverse problems

Inverse statistical problems reverse the forward paradigm: rather than computing macroscopic observables from a known microscopic model, the objective is to infer the microscopic parameters that best explain the observed macroscopic behavior. Given access to empirical observations drawn from an unknown distribution P, one aims to estimate the parameters θ of a model P_θ such that $P_\theta \approx P$.

This setting is common across statistical physics, machine learning, and disordered systems. Such problems arise when trying to infer interaction parameters in systems where only macroscopic quantities (e.g., magnetizations or correlations) are experimentally accessible.

1.2.1 Maximum entropy models

The maximum entropy (MaxEnt) principle provides a natural and principled way to select a probability distribution consistent with given empirical constraints, while making no further analytical assumptions.

Suppose we are given empirical expectations of some observables $f_1, \dots, f_m$, i.e.,

$$\mathbb{E}_P[f_j] \approx \hat{f}_j \quad \text{for } j = 1, \dots, m.$$

Then, among all distributions Q absolutely continuous with respect to a reference measure μ and satisfying these constraints,

$$\mathcal{Q} := \left\{ Q \ll \mu \,\middle|\, \int_{\Omega_Q} f_j(x)\, dQ(x) = \hat{f}_j \text{ for all } j \right\},$$

the maximum entropy principle selects the distribution Q^* which maximizes the Shannon entropy:

$$Q^* = \arg\max_{Q \in \mathcal{Q}} H(Q), \quad \text{where } H(Q) = -\int_{\Omega_Q} \log\left(\frac{dQ}{d\mu}(x)\right) dQ(x).$$

This constrained optimization problem can be solved via Lagrange multipliers, and the solution Q^* takes the exponential (Boltzmann-like) form:

$$\frac{dQ^*}{d\mu}(x) = \frac{1}{Z(\theta)} \exp\left(\sum_{j=1}^{m} \theta_j f_j(x)\right),$$

for suitable parameters $\theta_1, \ldots, \theta_m$ enforcing the moment constraints. The corresponding partition function $Z(\theta)$ ensures normalization:

$$Z(\theta) = \int_{\Omega_\theta} \exp\left(\sum_{j=1}^{m} \theta_j f_j(x)\right) d\mu(x).$$

This construction leads to a flexible and general family of models—*exponential families*—which play a central role in inverse problems. Importantly, many classical models from statistical physics (e.g., the Ising model) are special cases of this construction, with μ being the counting measure.

1.2.2 The inverse Ising problem

A prototypical example of an inverse problem is the *inverse Ising problem*: given a collection of samples $\sigma^{(1)}, \ldots, \sigma^{(n)} \in \{-1, +1\}^d$ drawn from an unknown discrete model.

In this setting, the configuration space is finite and the reference measure μ is the uniform counting measure on $\{-1, +1\}^d$, so integrals reduce to finite sums.

A natural approach, grounded in the maximum entropy principle, is to consider the empirical averages of the sufficient statistics, i.e., the empirical magnetizations

$$\hat{m}_i := \frac{1}{n}\sum_{k=1}^{n} \sigma_i^{(k)}$$

and the empirical correlations

$$\hat{c}_{ij} := \frac{1}{n}\sum_{k=1}^{n} \sigma_i^{(k)}\sigma_j^{(k)}.$$

Then, one looks for the distribution over $\{-1,+1\}^d$ that has maximum entropy among all those matching these empirical moments. This leads precisely to the Boltzmann–Gibbs distribution of the Ising form.

Formally, the model can be written as

$$P_\theta(\sigma) = \frac{1}{Z(\theta)} \exp\left(\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\right), \quad \sigma \in \{-1,+1\}^d,$$

with parameters $\theta = (J,h)$, and partition function

$$Z(\theta) = \sum_{\sigma\in\{-1,+1\}^d} \exp\left(\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\right) = \int_{\Omega_\theta} \exp\left(-H_\theta(x)\right) d\mu(x).$$

The parameters J_{ij} and h_i can be estimated by requiring that the model expectations match the empirical ones:

$$\mathbb{E}_{P_\theta}[\sigma_i] = \hat{m}_i, \quad \mathbb{E}_{P_\theta}[\sigma_i\sigma_j] = \hat{c}_{ij},$$

which corresponds to a moment-matching condition. One can perform maximum likelihood estimation, which leads to these conditions at the optimum.

The gradient ascent steps for maximizing the log-likelihood are:

$$\begin{aligned} J_{ij}^{(t+1)} &= J_{ij}^{(t)} + \eta\left(\hat{c}_{ij} - \mathbb{E}_{P_\theta}[\sigma_i\sigma_j]\right), \\ h_i^{(t+1)} &= h_i^{(t)} + \eta\left(\hat{m}_i - \mathbb{E}_{P_\theta}[\sigma_i]\right), \end{aligned}$$

where $\eta > 0$ is the learning rate and the parameters are initialized in any way. These updates increase the log-likelihood by aligning the model statistics with the empirical ones.

However, this approach requires computing model expectations under P_θ, which involve the partition function $Z(\theta)$ or its derivatives. Since this is intractable for large d, approximate methods must be employed. These include minimum probability flow, score matching, Langevin matching, pseudolikelihood maximization, and interaction screening. We will explore these techniques in the next chapters.

Chapter 2

Methods

Now that we have shown that gradient ascent is unwise, we move towards a series of different methods to perform inverse statistical problems.

The general framework is as follows. We assume we are given i.i.d. samples $x^{(1)}, \ldots, x^{(n)} \sim \hat{P}$, drawn from an unknown probability distribution $P_\theta \propto \exp(-H_\theta(x))$. The aim is to estimate the unknown parameters θ of the energy function $H_\theta(x)$, given only the samples.

1. In the *discrete setting*, the space is finite: $x \in \{-1, +1\}^d$, and the model is an Ising model of the form

$$P_{J,h}(\sigma) \propto \exp\left(\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\right).$$

   We assume the data are drawn from such a Boltzmann distribution, and our goal is to infer the couplings J_{ij} and external fields h_i.

2. In the *continuous setting*, the state space is $\mathbb{R}^d$, and the distribution takes the form

$$P_\theta(x) \propto \exp(-H_\theta(x)),$$

   where H_θ is a differentiable potential depending on parameters θ. Here too, the goal is to estimate θ from empirical samples.

The next sections will present various inference techniques adapted to this frameworks.

These methods have in common the idea of doing the inference without the need of computing the normalization Z.

2.1 Minimum probability flow

2.1.1 Discrete case

2.1.2 Continuous case

2.2 Score matching

2.2.1 Continuous case

2.2.2 Connection with MPF

2.2.3 Discrete case

2.3 Langevin matching

2.3.1 Continuous case

2.4 Pseudolikelihood maximization

2.5 Interaction screening

Chapter 3

Experiments

3.1 Complete graph

3.2 Regular graph

3.3 Sparse graph

Chapter 4

Appendix

4.1 Wasserstein spaces

This section's aim is to define Wasserstein spaces and show their properties.

We assume that the reader is already familiar with optimal transport, which in turn assumes a solid knowledge of measure theory and basic functional analysis. An intuitive understanding of Riemannian geometry could help as well, but is by no means required.

4.1.1 Informal introduction to metric geometry

Metric geometry is the field of mathematics that studies abstraction of key ideas from differential geometry, relying only on the intrinsic notion of distance.

Let us fix a metric space $(\mathcal{S}, d)$. Suppose we have enough regularity to be able to define paths $\omega : I \subseteq \mathbb{R} \to \mathcal{S}$, and their lengths $L(\omega)$ in the usual way.

As in the familiar Euclidean settings, two paths $\omega_1 : I_1 \to \mathcal{S}$ and $\omega_2 : I_2 \to \mathcal{S}$ are *equivalent* if there exists a continuous, non-decreasing and surjective function $\phi : I_1 \to I_2$ such that $\omega_1 = \omega_2 \circ \phi$. In this case, ω_2 is called a *reparametrization* of ω_1 (and vice-versa by symmetry), and it is trivial to check that $L(\omega_1) = L(\omega_2)$.

We say that a path $\omega : [a, b] \to \mathcal{S}$ has *constant speed* if for all $a \leq s \leq t \leq b$,

$$L(\omega_{[s,t]}) = \frac{t-s}{b-a} L(\omega),$$

and we have that any rectifiable path $\omega : [a, b] \to \mathcal{S}$ has a constant-speed reparametrization $\bar{\omega} : [0, 1] \to \mathcal{S}$ by usual multivariate calculus.

For fixed $x_0, x_1 \in \mathcal{S}$, a *geodesic* is defined as a path $\omega : [0, 1] \to \mathcal{S}$ such that

$$d(x_0, x_1) = L(\omega).$$

A metric space $(\mathcal{S}, d)$ is said to be a *geodesic space* if for any $x_0, x_1 \in \mathcal{S}$ we can find a (constant speed) geodesic.

In a geodesic space, given $x_0, x_1 \in \mathcal{S}$, we can consider the constant speed geodesic $\omega : [0, 1] \to \mathcal{S}$ such that $\omega(0) = x_0, \omega(1) = x_1$ and define the *midpoint* between x_0 and x_1 to be $x_{\frac{1}{2}} := \omega(0.5)$.

Let us denote with λ the Lebesgue measure on $\mathbb{R}^d$.

Notably, if we take $(\mathcal{S}, d) = (\mathcal{P}_2^{ac}(\lambda), L_2(\lambda))$ the space of absolutely continuous measures with finite second moment with the metric induced by the Lebesgue L_2 norm, we see that geodesics represent *teleportation*, i.e. the vertical interpolation, as shown below.

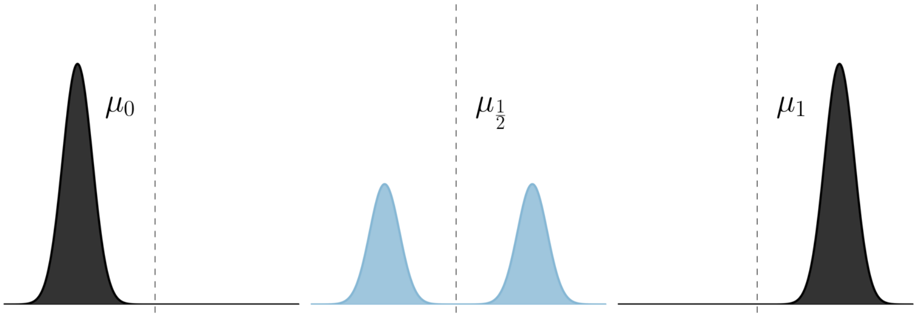

Figure 4.1: Geodesic in $(\mathcal{P}_2^{ac}(\lambda), L_2(\lambda))$.

The natural horizontal interpolation will be achieved in the next sections by $(\mathcal{S}, d) = (\mathcal{P}_2^{ac}(\lambda), \mathcal{W}_2)$, as shown below.

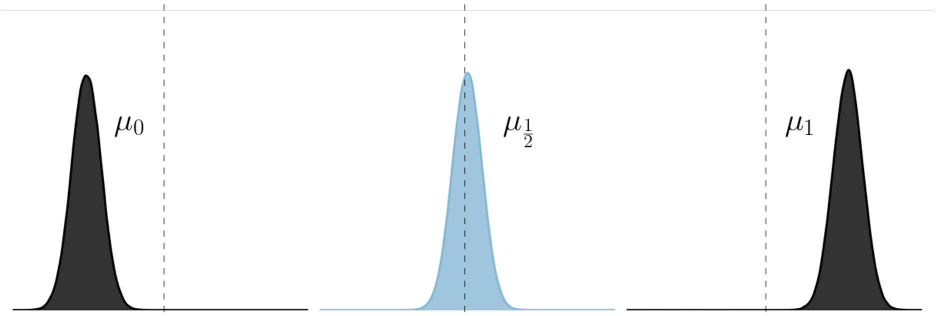

Figure 4.2: Geodesic in $(\mathcal{P}_2^{ac}(\lambda), \mathcal{W}_2)$.

4.1.2 Hints of optimal transport

We are going to denote with $\mathcal{P}$ the set of probability measures on $\mathbb{R}^d$ and with $\mathcal{P}_2$ the subset of probability measures with finite second moment.

Definition 4.1.1. Given $\mu_0 \in \mathcal{P}$ and $T : \mathbb{R}^d \to \mathbb{R}$, we define the *push forward measure*

$$T\#\mu_0 := \mu_0(T^{-1}(A)), \text{ for any } A \subseteq \mathbb{R}^d.$$

Remarkably, $\mu_1 = T\#\mu_0$ if and only if $\int_{\mathbb{R}^d} \phi \, d\mu_1 = \int_{\mathbb{R}^d} \phi \circ T \, d\mu_0$ for any $\phi : \mathbb{R}^d \to \mathbb{R}$ measurable and bounded, as shown in [3].

We have the following theorem which is useful in computations.

Theorem 4.1.2. Let $\mu_0, \mu_1 \in \mathcal{P}$, with $\mu_0, \mu_1 \ll \lambda$, where λ is the Lebesgue measure on $\mathbb{R}^d$. If $f = \frac{d\mu_0}{d\lambda}$ and $g = \frac{d\mu_1}{d\lambda}$ are the Radon-Nykodim derivatives, and $T : \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism such that $\mu_1 = T\#\mu_0$, we have that

$$f(x) = |\det J_{T(x)}| g(T(x)).$$

Proof. It follows from the standard change of variables formula. Indeed, for any $\phi : \mathbb{R}^d \to \mathbb{R}$ measurable and bounded,

$$\begin{aligned}\int_{\mathbb{R}^d} \phi \circ T(x) f(x) \, dx &= \int_{\mathbb{R}^d} \phi(y) g(y) \, dy \\ &= \int_{\mathbb{R}^d} \phi \circ T(x) g(T(x)) |\det J_{T(x)}| \, dx,\end{aligned}$$

which concludes because ϕ was arbitrary and T was bijective. □

Definition 4.1.3. We will call *canonical projections* the functions

$$\pi_X : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, \quad \pi_X[(x, y)] = x$$
$$\pi_Y : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, \quad \pi_Y[(x, y)] = y$$

Sometimes we will abuse of this notation without specifying the projections' domain and codomain, but it will always be clear from the context.

Definition 4.1.4. Given $\mu_0 \in \mathcal{P}$ and $T : \mathbb{R}^d \to \mathbb{R}$, we denote the set of *couplings* as

$$\Gamma(\mu_0, \mu_1) := \{\gamma \in \mathcal{P} \times \mathcal{P} : \ \pi_X\#\gamma = \mu_0, \, \pi_Y\#\gamma = \mu_1\}.$$

We have the machinery to define the fundamental optimization problems.

Definition 4.1.5. Given $\mu_0, \mu_1 \in \mathcal{P}$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we define the *Monge's problem* as

$$(\text{MP}) = \inf_{T:\mathbb{R}^d \to \mathbb{R}} \left\{ \int_{\mathbb{R}^d} c(T(x), x) \, d\mu_0(x) : T\#\mu_0 = \mu_1 \right\}$$

This formally translates that we aim to move a probability measure to another one by minimizing a given cost function. We can relax the need of a transport map by only requiring a restriction in terms of couplings.

Definition 4.1.6. Given $\mu_0, \mu_1 \in \mathcal{P}$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we define the *Kantorovich's problem* as

$$(\text{KP}) = \inf_{\gamma \in \mathcal{P} \times \mathcal{P}} \left\{ \int_{\mathbb{R}^d} c(x, y) \, d\gamma(x, y) : \gamma \in \Gamma(\mu_0, \mu_1) \right\}.$$

Remarkably the set of couplings is never empty: it is enough to consider $\mu_0 \otimes \mu_1 \in \mathcal{P} \times \mathcal{P}$. Moreover, due to Prokhorov and Banach-Alaoglu's theorems, the infimum is actually reached, besides very pathological situations [3].
Furthermore any transport map T such that $T\#\mu_0 = \mu_1$ induces a coupling $\gamma_T := (id, T)\#\mu_0 \in \Gamma(\mu_0, \mu_1)$, as can be checked into [3].

In the following we are going to work only with $c(x, y) = \|x - y\|^2$, as this choice is the most natural and the most studied, and furthermore we can anticipate that this choice will enable us to define the tangent space (in a fixed point of our Wasserstein space) as a Hilbert space. Accordingly to this choice, to avoid issues with $+\infty$, we will restrict ourselves to $\mathcal{P}_2$.

We have the important theorem, due to Brenier.

Theorem 4.1.7. Given $\mu_0 \in \mathcal{P}_2^{ac}(\lambda)$ and $\mu_1 \in \mathcal{P}_2$, then there exists a unique optimizer $\bar{\gamma}$ in (KP). In addition, there exists $T = \nabla\phi$, with $\phi : \mathbb{R}^d \to \mathbb{R}$ convex, such that $\bar{\gamma} = (id, T)\#\mu_0$ and T is the unique optimizer in (MP).

4.1.3 Wasserstein geometry

In this section we are going to consider the metric induced by (KP) on the space of measures and study its properties. In particular, we will show that the geometry induced by this metric lifts the geometry of the underlying space, as desired.

It turns out that this space also has a natural differential structure that leads not only to defining geodesics, but also to notions of tangent spaces and gradients. These, in turn, allow for the definition of Wasserstein gradient flows, which are tools of fundamental importance in statistical applications.

As hinted in the previous section, the inf in (KP) is really a min, besides pathological situations. Recalling that we restricted to the quadratic cost, we can thus define the following.

Definition 4.1.8. Given two probability measures $\mu_0, \mu_1 \in \mathcal{P}_2$, we define the *Wasserstein distance* between them as

$$\mathcal{W}_2(\mu_0, \mu_1) := \min_{\gamma} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, d\gamma(x, y) \mid \gamma \in \Gamma(\mu_0, \mu_1) \right)^{\frac{1}{2}}.$$

This is indeed a metric, as the next result shows.

Theorem 4.1.9. $\mathcal{W}_2 : \mathcal{P}_2 \times \mathcal{P}_2 \to \mathbb{R}$ is a metric on $\mathcal{P}_2$.

The proof can be found in [1]. Remarkably, when the optimal coupling is induced (uniquely) by an optimal map, the distance simplifies to

$$\mathcal{W}(\mu_0, \mu_1) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|T_{\mu_0 \to \mu_1}(x) - x\|_2^2 \, d\mu_0(x) \right)^{\frac{1}{2}}.$$

From now on we are going to stick with $\mathcal{P}_2^{ac}(\lambda)$, the family of measure with finite second moment and absolutely continuous with respect to Lebesgue, so that we can always assume the existence of such representation, thanks to Brenier's theorem.

Recalling our definitions from the first section, we can show that $(\mathcal{P}_2, \mathcal{W}_2)$ is a geodesic space by exhibiting geodesics between any two fixed $\mu_0, \mu_1 \in \mathcal{P}_2^{ac}(\lambda)$.

Proposition 4.1.10. Given any $\mu_0, \mu_1 \in \mathcal{P}_2$, the constant speed geodesic with respect to the Wasserstein distance is $\mu_t := T_t \# \mu_0$, where $T_t(x) := (1-t)x + tT_{\mu_0 \to \mu_1}(x)$.

The proof can be found in [1].

Moreover, we can see that the underlying space's geometry is preserved by looking at the isometry $(x, \|\cdot\|) \mapsto (\delta_x, \mathcal{W}_2)$:

Proposition 4.1.11. Given any $x_0, x_1 \in \mathbb{R}^d$, we have that

$$\|x_1 - x_0\|_2 = \mathcal{W}_2(\delta_{x_0}, \delta_{x_1}).$$

Proof. Obviously we notice that $\Gamma(\delta_{x_0}, \delta_{x_1}) = \delta_{(x_0, x_1)}$, and the cost induced by the coupling (which is actually induced by the transport map that sends $x \mapsto x_1, \ \forall x \in \mathbb{R}^d$), is

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, d\delta_{(x_0, x_1)} = \|x_0 - x_1\|^2,$$

which concludes our argument. □

Moreover, the transport structure of the Wasserstein space induces not only a metric space, but a manifold-like geometry, referred to in the literature as *Otto calculus*. In order to fix ideas, let us take as a toy example the usual Riemannian geometry framework on $\mathbb{S}^2$.

Example 4.1.12. Let us consider a generic $x_0 \in \mathbb{S}^2$. Now, for any $x_1 \in \mathbb{S}^2$ (not equal and not antipodal to x_0), there exists a unique constant speed geodesic ω, i.e. the smaller section of the maximal circumference through x_0 and x_1. By smoothness, the vector $\omega'(0)$ which tells a moving particle starting in x_0 where to (infinitesimally) move if it wants to follow ω, is well defined and unique.

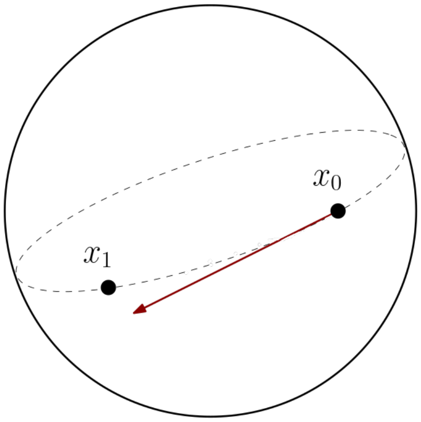

Figure 4.3: The vector $\omega'(0)$ is in red.

Now, for fixed x_0, if we take the linear combination of the vectors $\omega'(0)$ as x_1 varies in $\mathbb{S}^2$, we obtain a plane. We denote such plane as the tangent space $\mathcal{T}_{x_0}\mathbb{S}^2$. We are going to use the same exact idea to lift the concept of tangent space to our measures space.

Under our assumptions, for fixed $\mu_0 \in \mathcal{P}_2$, we can identify uniquely any $\mu_1 \in \mathcal{P}_2$ with $T_{\mu_0 \to \mu_1}$. Furthermore, since we showed that $(\mathcal{P}_2, \mathcal{W}_2)$ is a geodesic space, we are motivated to define a *tangent space* like structure in the following way accordingly to our example:

Definition 4.1.13. Given $\mu_0 \in \mathcal{P}_2$, we define the *tangent space* at μ_0 as

$$\mathcal{T}_{\mu_0}\mathcal{P}_2 := \overline{\{\zeta(T_{\mu_0 \to \mu_1} - id) : \mu_1 \in \mathcal{P}_2;\, \zeta > 0\}}^{L^2(\mu_0)}.$$

The geometric intuition is straightforward: in this metric space, the role of the tangent vector is played by $\frac{dT_t}{dt} = \frac{d}{dt}\left[(1-t)id + tT_{\mu_0 \to \mu_1}\right] = T_{\mu_0 \to \mu_1} - id$. Then, by our assumption on finiteness of second moments, every such map satisfies $T_{\mu_0 \to \mu_1} - id \in L_2(\mu_0)$. We hence take the closure with respect to $\| \cdot \|_{L_2(\mu_0)}$.

Though not obvious, an equivalent (and often useful) definition is the following:

$$\mathcal{T}_{\mu_0}\mathcal{P}_2 = \overline{\{\nabla\phi |\, \phi : \mathbb{R}^d \to \mathbb{R} \text{ compactly supported and smooth}\}}^{L_2(\mu_0)}. \tag{4.1}$$

which is given in [2]. As a subset of $L_2(\mu_0)$, the tangent space inherits the inner product:

$$\langle f, g\rangle_{\mu_0} = \int_{\mathbb{R}^d} f(x)g(x)\, d\mu_0(x),\ \ f, g \in L_2(\mu_0),$$

and this makes it clear why we chose to work with $\mathcal{W}_2$ instead of any other $\mathcal{W}_p, p \in [1, +\infty]$. Though not obvious from the definition, $\mathcal{T}_{\mu_0}\mathcal{P}_2$ is a linear space, as shown in [1]. Motivated by this differential structure, we can define the following tools:

Definition 4.1.14. Given $\mu_0 \in \mathcal{P}_2$ we define a formal *exponential map* as

$$exp_{\mu_0} : \mathcal{T}_{\mu_0}\mathcal{P}_2 \to \mathcal{P}_2, \qquad exp_{\mu_0}(T) = (T + id)\#\mu_0.$$

Moreover, we define a *logarithm map* as its left inverse, projecting onto $\mathcal{T}_{\mu_0}\mathcal{P}_2$

In particular if $\mu_0 \ll \lambda$, Brenier's theorem yields that

$$log_{\mu_0}(\mu_1) = T_{\mu_0\to\mu_1} - id.$$

In general, for fixed $\mu_0 \in \mathcal{P}_2$, the exponential map takes a generic transformation as input and interprets it as a *tangent vector* to μ_0, and by our identification $\mu_1 \longleftrightarrow T_{\mu_0\to\mu_1}$ it outputs the unique measure that induces the constant speed geodesic that produces that tangent vector. The logarithm map does the inverse.

4.1.4 Evolution of measures

Now that we have described the geometry of this space, we are interested in studying the evolution of probability measures when a family of time-dependent vector fields is acting on the underlying space. More precisely let $v_t : \mathbb{R}^d \to \mathbb{R}^d$, where $t \geq 0$ represents time. The motion of particles in the underlying space is described by the following Ordinary Differential Equation (ODE):

$$\dot{X}_t = v_t(X_t), \quad t \geq 0, \tag{4.2}$$

where X_t represents the position of a particle at time t.

The distribution μ_t of X_t solving (4.2) evolves over time according to the action of the vector fied, and in the context of this setup, we will show that this evolution is governed by the *continuity equation*, which expresses the conservation of probability as the underlying particles move through the space.

Preliminarly, we define $\partial_t \mu_t$ as the measure that satisfies $\int_{\mathbb{R}^d} g \, d(\partial_t \mu_t) = \partial_t \mathbb{E}[g(X_t)]$ for any test function $g : \mathbb{R}^d \to \mathbb{R}$ (where test means *smooth and compactly supported*).

Theorem 4.1.15. Let $(v_t)_{t \geq 0} : \mathbb{R}^d \to \mathbb{R}^d$ be a time dependent vector field with $v_t \in L_1(\mathbb{R}^d)$ for any $t \geq 0$, and suppose that particles evolve according to (4.2). Then $X_t \sim \mu_t$, where μ_t satisfies

$$\int_{\mathbb{R}^d} g \, d(\partial \mu_t) = \int_{\mathbb{R}^d} \langle \nabla g, v_t \rangle_2 \, d\mu_t, \tag{4.3}$$

for every test function g.

The proof can be found in [1].

Whenever $d\mu_t = f_t \, d\lambda$, with $f_t \in C^1(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$, then (4.3) is equivalent to

$$\partial_t f_t + \langle \nabla, (f_t v_t) \rangle_2 = 0 \tag{4.4}$$

in weak sense. To see this, fix a test function g, and notice that

$$\begin{aligned} \int_{\mathbb{R}^d} g \, d(\partial_t \mu_t)(x) &= \partial_t \int_{\mathbb{R}^d} g(x) f_t(x) \, d(x) \quad \text{by definition} \\ &= \int_{\mathbb{R}^d} g(x) \partial_t f_t(x) \, d(x). \quad \text{by dominated convergence} \end{aligned}$$

On the other side,

$$\begin{aligned} \int_{\mathbb{R}^d} \langle \nabla g(x), v_t(x) \rangle_2 \, d\mu_t(x) &= \int_{\mathbb{R}^d} \langle \nabla g(x), v_t(x) \rangle_2 f_t(x) \, dx \quad \text{by definition} \\ &= -\int_{\mathbb{R}^d} g(x) \langle \nabla, (f_t(x) v_t(x)) \rangle_2 \, dx, \quad \text{integration by parts} \end{aligned}$$

which implies, by identification, that $\partial_t f_t + \langle \nabla, (f_t v_t) \rangle_2 = 0$.

In this flavor, up to defining $\langle \nabla, (\mu_t v_t) \rangle_2$ as the distribution that satisfies

$$\int_{\mathbb{R}^d} g(x) \, d(\langle \nabla, \mu_t v_t \rangle_2)(x) = -\int_{\mathbb{R}^d} \langle \nabla g(x), v_t(x) \rangle_2 \, d\mu_t(x),$$

for any test function g; even when μ_t does not admit density with respect to the Lebesgue measure, we say that (4.3) is equivalent to the weak generalized continuity equation (or weak continuity equation for brevity)

$$\partial_t \mu_t + \langle \nabla, (\mu_t v_t) \rangle_2 = 0. \tag{4.5}$$

Every *nice* curve of probability measures can be interpreted as a fluid moving along a time varying vector field. And by *nice* we mean:

Definition 4.1.16. A curve $t \mapsto \mu_t \in \mathcal{P}_2$ is said to be *absolutely continuous* if at every time, the metric derivative is finite, i.e., if

$$\text{for all } t, \ |\dot{\mu}|(t) := \lim_{s \to t} \frac{\mathcal{W}_2(\mu_s, \mu_t)}{|s-t|} < +\infty.$$

Theorem 4.1.17. Let $t \mapsto \mu_t$ be an absolutely continuous curve of measures. Then:

1. For any vector field $(\tilde{v})$ that satisfies the weak continuity equation, we have
$$|\dot{\mu}_t| \leq \|\tilde{v}_t\|_{L_2(\mu_t)}.$$
2. Conversely, there exists a unique choice of vector field $(v_t)_{t \geq 0}$ that satisfies the weak continuity equation and
$$\|v_t\|_{L_2(\mu_t)} \leq |\dot{\mu}_t|.$$

Moreover, $v_t = \nabla \psi_t$ for $\psi : \mathbb{R}^d \to \mathbb{R}$ and

$$v_t = \lim_{\delta \to 0} \frac{T_{\mu_t \to \mu_{t+\delta}} - id}{\delta}.$$

For a proof, see [1].

The theorem just states that starting from an (absolutely continous) curve $t \mapsto \mu_t$ we can identify any infinitesimal displacement as a constant speed geodesic between μ_t and $\mu_{t+\delta}$ and accordingly find uniquely the vector field v_t that induces it.

4.1.5 First variation of functionals

The goal of this and the following subsections is to understand how functionals $\mathcal{F} : \mathcal{P}_2 \to \mathbb{R},\ \mu_t \mapsto \mathcal{F}(\mu_t)$ evolve as the underlying space is subject to a family of time dependent vector fields. This will lead us to a notion of gradient flow in $(\mathcal{P}_2, \mathcal{W}_2)$, which will be our main tool for applications.

This subsection is specifically devoted to an informal description of the tools required to work with functionals. Although its abstract nature may momentarily disrupt the flow of the exposition, we have chosen to include it here rather than in the preliminaries, so as to provide sufficient motivation drawn from the surrounding context.

We want to work our way towards a rigorous definition of *differential operator* associated to $\mathcal{F}$.

The first obstacle comes from the non-linearity of probability measures (the sum of two probability measures is no longer a measure), so that the direct differentiation does not make sense. Indeed, we typically understand differentiation of a smooth function f on a Euclidean space as:

$$f(x+h) - f(x) = [(\delta f)(x)](h) + o(h), \qquad h \to 0,\ x, h \in \mathbb{R}^d,$$

where $[(\delta f)(x)]$ is a linear and bounded functional on $\mathbb{R}^d$, i.e., a matrix.

However, say the functional $\mathcal{F}$ admits an extension over the space of *signed measures* $\mathcal{M}$ (on $\mathbb{R}^d$). This is a linear space, so that differentation is more easily understood in the usual way, in that the (extended) functional $\mathcal{F}$ is differentiable if there exists a continuous linear functional $[\delta\mathcal{F}(\mu)]$ on $\mathcal{M}$ such that:

$$\mathcal{F}(\mu + \epsilon\chi) - \mathcal{F}(\mu) = \epsilon[\delta\mathcal{F}(\mu)](\chi) + o(\epsilon), \qquad \mu, \chi \in \mathcal{M},\ \epsilon \to 0. \tag{4.6}$$

It turns out that every such continuous linear functional has a particular form. This result, known as *Kantorovich-Rubinstein duality*, states that the dual space of $\mathcal{M}$ can the identified with the space of bounded continuous functions:

$$\mathcal{M}^* \simeq C_b(\mathbb{R}^d),$$

in the sense that for any linear functional G acting on measures defined on $\mathcal{M}$, there exists $g \in C_b(\mathbb{R}^d)$ such that:

$$G(\chi) = \int_{\mathbb{R}^d} g \, d\chi, \qquad \forall \chi \in \mathcal{M}.$$

Therefore, with a slight abuse of notation, i.e. identifying the map $[\delta\mathcal{F}(\mu)]$ with its Rieszs representative living in $C_b(\mathbb{R}^d)$, we can rewrite (4.6) as:

$$\mathcal{F}(\mu + \epsilon\chi) - \mathcal{F}(\mu) = \epsilon \int_{\mathbb{R}^d} [\delta\mathcal{F}(\mu)] \, d\chi + o(\epsilon), \qquad \mu, \chi \in \mathcal{M}, \ \epsilon \to 0. \tag{4.7}$$

4.1.6 Wasserstein gradient flows

Coming back to probability measures, by Theorem 4.1.17, we know that given a regular enough flow μ_t we can associate to it a unique vector field v_t such that (4.3) (a.k.a. the weak continuity equation) holds. Furthermore we can write in weak sense

$$\mu_t = \mu_0 + t\,\partial_t\mu_t + o(t), \quad t \to 0,$$

and by substituting this into (4.7), we get

$$\lim_{t \to 0} \frac{\mathcal{F}(\mu_t) - \mathcal{F}(\mu_0)}{t} = \int_{\mathbb{R}^d} [\delta\mathcal{F}(\mu_0)] \, d(\partial_t\mu_t),$$

Now, assuming vanishing boundary conditions (which will be the case in functional of interest), we get:

$$\begin{aligned}\lim_{t \to 0} \frac{\mathcal{F}(\mu_t) - \mathcal{F}(\mu_0)}{t} &= \int_{\mathbb{R}^d} [\delta\mathcal{F}(\mu_0)] \, d(\partial_t\mu_t) \\ &= \int_{\mathbb{R}^d} \langle (\nabla[\delta\mathcal{F}(\mu_0)])(x), v_t(x) \rangle_2 \, d\mu_t(x) = \langle (\nabla[\delta\mathcal{F}(\mu_0)]), v_t \rangle_{L_2(\mu_t)},\end{aligned}$$

by the same steps of Theorem 4.1.15.

Moreover, since $\nabla[\delta\mathcal{F}(\mu)]$ is the gradient of a bounded function, by the characterization in (4.1), we know that $\nabla[\delta\mathcal{F}(\mu_0)] \in \mathcal{T}_{\mu_0}\mathcal{P}_2$.

We have informally proven the following[1]:

[1] For a rigorous proof, take a look at [1].

Theorem 4.1.18. Let $\mathcal{F} : \mathcal{P}_2 \to \mathbb{R}$ be a functional with bounded first variation. Then, the Wasserstein gradient of $\mathcal{F}$ is the vector field defined by:

$$\nabla_{\mathcal{W}} \mathcal{F}(\mu_0) = \nabla[\delta\mathcal{F}(\mu_0)],$$

where $[\delta\mathcal{F}(\mu_0)] \in C_b(\mathbb{R}^d)$ is tacitly the Rieszs representative of the first variation of $\mathcal{F}$ at μ_0, while ∇ is the usual Euclidean gradient.

Example 4.1.19. Given a potential $V : \mathbb{R}^d \to \mathbb{R}$, we can define the *potential energy* as

$$\mathcal{V}(\mu) := \int_{\mathbb{R}^d} V \, d\mu, \quad \text{for any } \mu \in \mathcal{P}_2.$$

Then,

$$\partial_t \mathcal{V}(\mu_t) = \int_{\mathbb{R}^d} V \, d(\partial_t \mu_t),$$

and thus we can identify $\delta\mathcal{V}(\mu) = V$, for any $\mu \in \mathcal{P}_2$. Therefore,

$$\nabla_{\mathcal{W}_2} \mathcal{V}(\mu) = \nabla V, \quad \text{for any } \mu \in \mathcal{P}_2.$$

Example 4.1.20. Given $\mu \in \mathcal{P}_2^{ac}$, with $d\mu = f \, d\lambda$, we can define the *entropy functional* as

$$\mathrm{Ent}(\mu) := \int_{\mathbb{R}^d} f \log(f) \, d\lambda.$$

Then,

$$\partial_t \mathrm{Ent}(\mu_t) = \int_{\mathbb{R}^d} (\partial_t f_t \log(f_t) + \partial_t f_t) \, d\lambda = \int_{\mathbb{R}^d} \partial_t f_t (\log(f_t) + 1) \, d\lambda,$$

and therefore we can identify $\delta\mathrm{Ent}(\mu) = \log(f) + 1$. Then we can write

$$\nabla_{\mathcal{W}_2} \mathrm{Ent}(\mu) = \nabla \log f.$$

Now, Theorem 4.1.18 is showing that the first variation of $\mathcal{F}$ naturally leads to a gradient structure in the Wasserstein space, indeed we can expand $\mathcal{F}(\mu_t)$ to the first order:

$$\mathcal{F}(\mu_{t+h}) = \mathcal{F}(\mu_t) + h\langle \nabla_{\mathcal{W}} \mathcal{F}(\mu_t), v_t \rangle_{L_2(\mu_t)} + o(h), \quad h \to 0.$$

We can now finally define the Wasserstein gradient flow of a functional.

Informally, a gradient flow in the Wasserstein space is a curve of measures $(\mu_t)_{t \geq 0}$ such that the tangent vector to the curve at t equals $-\nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)$. Recalling that the tangent vector governs the evolution of $(\mu_t)_{t \geq 0}$ via the continuity equation (4.3), we arrive at the following definition.

Definition 4.1.21. Let $\mathcal{F} : \mathcal{P}_2 \to \mathbb{R}$ a functional. Then $(\mu_t)_{t\geq 0}$ is called the *Wasserstein gradient flow* of $\mathcal{F}$ if it solves the following PDE in weak sense:

$$\partial_t \mu_t = \langle \nabla, (\mu_t \nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)) \rangle_2. \tag{4.8}$$

Unsurprisingly, this gradient flow yields a principled approach for dynamically, smoothly evolving a probability measure in the Wasserstein space, with the aim of minimizing the objective functional $\mathcal{F}$:

$$\partial_t \mathcal{F}(\mu_t) = \langle \nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t), v_t \rangle_{L_2(\mu_t)} = -\|\nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)\|^2_{L_2(\mu_t)}.$$

Clearly, in order to have theoretical and quantitative guarantees about the convergence to a minima we need to address the convexity of the functional of interest. This will be studied in the next section for a specific case of interest.

4.1.7 KL divergence

We start by properly defining our notion of divergence between measures.

Definition 4.1.22. Given two probability measures $\mu, \pi \ll \lambda$ on $\mathbb{R}^d$, with $d\mu = g d\lambda$, $d\pi = f d\lambda$, the *Kullback-Leibler divergence* is defined as:

$$\mathcal{D}_{KL}(\mu \| \pi) := \int_{\mathbb{R}^d} \log\left(\frac{g(x)}{f(x)}\right) g(x)\, dx = \int_{\mathbb{R}^d} \log\left(\frac{g}{f}\right) g\, d\lambda.$$

The KL divergence quantifies how much the measure μ differs from the target measure π. Let us start by stating a natural but nonetheless important property of our chosen divergence.

Theorem 4.1.23. Given two probability measures $\mu, \pi \ll \lambda$ on $\mathbb{R}^d$, with $d\mu = g d\lambda$, $d\pi = f d\lambda$, we have that $\mathcal{D}_{KL}(\mu\|\pi) \geq 0$, and $\mathcal{D}_{KL}(\mu\|\pi) = 0$ if and only if $f = g$ $\lambda-$a.e.

The proof follows from Jensen's inequality.

Unfortunately, in general $\mathcal{D}_{KL}(\nu\|\pi) \neq \mathcal{D}_{KL}(\pi\|\nu)$, which already shows that the KL divergence is not a metric.

From now on, we are going to focus on the situation in which $\pi = f\, d\lambda$

with $f(x) \propto e^{-V(x)}$. This is very common in applications such as Bayesian statistics, complex systems and statistical mechanics.

As we hinted in the introduction, the optimization problem

$$\mu^* = \arg \min_{\mu \in \mathcal{P}_2^{ac}(\lambda)} \mathcal{D}_{KL}(\mu \| \pi) \tag{4.9}$$

where $d\mu = g d\lambda$, $d\pi = f d\lambda = k\tilde{f} d\lambda = k d\tilde{\pi}$ (and we have direct access to $\tilde{f}$), does not require us to compute the normalizing constant k. Indeed, a direct calculation yields

$$\mathcal{D}_{\mathrm{KL}}(\mu \| \pi) = \int_{\mathbb{R}^d} \log\left(\frac{g}{k\tilde{f}}\right) g \, d\lambda = \int_{\mathbb{R}^d} \left[\log\left(\frac{g}{\tilde{f}}\right) - \log(k)\right] g \, d\lambda.$$

Separating the terms:

$$\mathcal{D}_{\mathrm{KL}}(\mu \| \pi) = \int_{\mathbb{R}^d} \log\left(\frac{g}{\tilde{f}}\right) g \, d\lambda - \log(k) \int_{\mathbb{R}^d} g \, d\lambda,$$

but $\int_{\mathbb{R}^d} g \, d\lambda = 1$, and thus:

$$\mathcal{D}_{\mathrm{KL}}(\mu \| \pi) = \mathcal{D}_{\mathrm{KL}}(\mu \| \tilde{\pi}) - \log(k)$$

which implies

$$\arg \min_{\mu \in \mathcal{P}_2^{ac}(\lambda)} \mathcal{D}_{KL}(\mu \| \pi) = \arg \min_{\mu \in \mathcal{P}_2^{ac}(\lambda)} \mathcal{D}_{KL}(\mu \| \tilde{\pi}),$$

as wanted to prove.

We will show that whenever V is convex, then $\mathcal{D}_{KL}(\cdot \| \pi)$ is *geodesically convex* (in $(\mathcal{P}_2^{ac}(\lambda), \mathcal{W}_2)$): a property that brings important results, as the next subsection will show.

4.1.8 Geodesic convexity of KL divergence

Let us recall some general knowledge about convex functions from analysis.

Definition 4.1.24. A function $V : \mathbb{R}^d \to \mathbb{R}$ is said to be *convex* if

$$V((1-t)x_0 + tx_1) \leq (1-t)V(x_0) + tV(x_1), \qquad \forall x_0, x_1 \in \mathbb{R}^d, \ \forall t \in [0,1].$$

In short, a function is convex if for any two fixed points in the domain, if we project the segment joining them to the epigraph the curve we obtain lies uniformly below the convex interpolation of the images of the point.

We can strength this definition to requiring a *uniform rate of convexity*:

Definition 4.1.25. A function $V : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is said to be *α-convex* (or *α-strongly convex*) for some $\alpha > 0$ if

$$V((1-t)x_0+tx_1) \leq (1-t)V(x_0)+tV(x_1)-\frac{\alpha}{2}t(1-t)\|x_1-x_0\|^2, \qquad \forall x_0, x_1 \in \mathbb{R}^d,\ \forall t \in [0,1].$$

We have the following useful characterization.

Proposition 4.1.26. Given a function $V : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we have

$$\begin{aligned} V \text{ is } \alpha\text{−convex } &\iff V - \alpha\|\cdot\|^2 \text{ is convex} \\ &\iff D^2V \geq \alpha Id_d. \end{aligned}$$

The proof follows from standard multivariate calculus.

In order to extend the concept of convexity to functionals we have to use the same strategy of the first section: instead of segments, we look at geodesics.

Definition 4.1.27. Let $(\mathcal{X}, d)$ be a geodesic space. A functional $\mathcal{F} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be *geodesically convex* (or *displacement convex*) if for every pair of points $x_0, x_1 \in \mathcal{X}$, and every constant-speed geodesic $(x_t)_{t\in[0,1]}$ joining them, the map $t \mapsto V(x_t)$ is convex. That is,

$$V(x_t) \leq (1-t)V(x_0) + tV(x_1), \qquad \forall t \in [0,1].$$

Definition 4.1.28. Let $(\mathcal{X}, d)$ be a geodesic space. A functional $\mathcal{F} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be *α-geodesically convex* for some $\alpha > 0$ if for every pair of points $x_0, x_1 \in \mathcal{X}$, and every constant-speed geodesic $(x_t)_{t\in[0,1]}$ joining them, the map $t \mapsto V(x_t)$ satisfies

$$V(x_t) \leq (1-t)V(x_0) + tV(x_1) - \frac{\alpha}{2}t(1-t)\|x_1 - x_0\|^2, \qquad \forall t \in [0,1].$$

For us, the geodesic space upon which the functionals are acting will be $(\mathcal{P}_2, \mathcal{W}_2)$.

We can already show that the entropy functional, introduced in Example 4.1.19, is geodesically convex.

Proposition 4.1.29. $\mathrm{Ent}(\mu)$ is geodesically convex.

Proof. Let $\mu_0 = f_0 \, d\lambda$ and $\mu_1 = f_1 \, d\lambda$ be two absolutely continuous probability measures in $\mathcal{P}_2$. Consider the Wasserstein geodesic $(\mu_t)_{t\in[0,1]}$ defined by

$$\mu_t := ((1-t)\mathrm{id} + tT_{\mu_0\to\mu_1})_{\#}\mu_0.$$

Clearly $\mu_t = f_t \, d\lambda$ is absolutely continuous for each $t \in [0,1]$. Using the change of variables formula from Theorem 4.1.2, the density f_t satisfies:

$$f_t((1-t)id + tT_{\mu_0\to\mu_1}))(x) = \frac{f_0(x)}{\det((1-t)Id_d + t\nabla T_{\mu_0\to\mu_1})(x))},$$

and we can define $y = g(x) := ((1-t)id + tT_{\mu_0\to\mu_1})(x)$.

Clearly id is differentiable and bijective, and since $T_{\mu_0\to\mu_1}$ is the gradient of a convex function by Brenier's, it is also differentiable almost everywhere, thanks to Alexandrov theorem [3]. On top of that, as a corollary to Brenier's, it can be shown that the inverse of $T_{\mu_0\to\mu_1}$ is well defined almost surely, so that it is bijective.

Finally, the weighted sum of two (almost everywhere) differentiable and bijective functions is itself (almost everywhere) differentiable and bijective, so that our change of variables $f_t(g(x)) = \frac{f_0(x)}{\det(\nabla g(x))}$ is well justified.

The entropy at time t is:

$$\begin{aligned}
\mathrm{Ent}(\mu_t) &= \int_{\mathbb{R}^d} f_t(y) \log f_t(y) \, dy \\
&= \int_{\mathbb{R}^d} f_t(g(x)) \log f_t(g(x)) \det(\nabla g(x)) \, dx \\
&= \int_{\mathbb{R}^d} \frac{f_0(x)}{\det(\nabla g(x))} \log \left(\frac{f_0(x)}{\det(\nabla g(x))} \right) \det(\nabla g(x)) \, dx \\
&= \int_{\mathbb{R}^d} f_0(x) \log \left(\frac{f_0(x)}{\det((1-t)Id_d + t\nabla T_{\mu_0\to\mu_1})(x))} \right) dx,
\end{aligned}$$

This can be split as:

$$\mathrm{Ent}(\mu_t) = \int_{\mathbb{R}^d} f_0(x) \log f_0(x) \, dx - \int_{\mathbb{R}^d} f_0(x) \log \det((1-t)Id_d + t\nabla T_{\mu_0\to\mu_1})(x)) \, dx.$$

The first term is constant in t, and the second term is concave in t, since $A \mapsto \log\det A$ is concave on the space of positive definite matrices, and $t \mapsto (1-t)id + t\nabla T_{\mu_0\to\mu_1})(x)$ is affine in t.

Therefore, $t \mapsto \mathrm{Ent}(\mu_t)$ is convex, and we conclude:

$$\mathrm{Ent}(\mu_t) \leq (1-t)\mathrm{Ent}(\mu_0) + t\mathrm{Ent}(\mu_1),$$

proving that the entropy functional is geodesically convex. □

Additionally, we have an important result regarding the potential energy functional $\mathcal{V}(\mu) = \int_{\mathbb{R}^d} V\, d\mu$ defined in Example 4.1.20.

Theorem 4.1.30. V is $\alpha-$convex if and only if $\mathcal{V}$ is $\alpha-$geodesically convex.

Proof. ($\Rightarrow$) We claim that if $\mu_0, \mu_1 \in \mathcal{P}_2$, and $\bar{\gamma} \in \Gamma(\mu_0, \mu_1)$ is the optimal coupling between the two, then $t \mapsto \mu_t$ is convex, where μ_t is the constant speed geodesic.

$$\begin{aligned}\int_{\mathbb{R}^d} V\, d\mu_t &= \int_{\mathbb{R}^d\times\mathbb{R}^d} V((1-t)x_0 + tx_1)\, d\bar{\gamma}(x_0,x_1)\\ &\leq \int_{\mathbb{R}^d\times\mathbb{R}^d} \left[(1-t)V(x_0) + tV(x_1) - \frac{\alpha}{2}t(1-t)\|x_1-x_0\|^2\right] d\bar{\gamma}(x_0,x_1)\\ &= (1-t)\int_{\mathbb{R}^d} V\, d\mu_0 + t\int_{\mathbb{R}^d} V\, d\mu_1 - \frac{\alpha}{2}t(1-t)\mathcal{W}_2^2(\mu_0,\mu_1).\end{aligned}$$

($\Leftarrow$) For the other direction, it is sufficient to apply the $\alpha-$geodesically convexity of $\mathcal{V}$ to δ_{x_0} and δ_{x_1} to get the $\alpha-$convexity of V. □

We can finally show the main result for this subsection.

Theorem 4.1.31. If $\pi = f\, d\lambda$, with $f = ke^{-V}$ for some normalizing constant k, and $V : \mathbb{R}^d \to \mathbb{R}$ is convex, then $\mathcal{D}_{KL}(\cdot\|\pi)$ is geodesically convex. Accordingly, if V is $\alpha-$convex, then $\mathcal{D}_{KL}(\cdot\|\pi)$ is $\alpha-$geodesically convex.

Proof. Let $\pi = f\, d\lambda$ with $f = ke^{-V}$ for some normalizing constant $k > 0$, and assume $V : \mathbb{R}^d \to \mathbb{R}$ is convex. For any absolutely continuous probability measure $\mu \ll \lambda$ with density $g = \frac{d\mu}{d\lambda}$, we can write the Kullback–Leibler divergence as:

$$\mathcal{D}_{KL}(\mu\|\pi) = \int_{\mathbb{R}^d} g\log\left(\frac{g}{f}\right)\, d\lambda = \int_{\mathbb{R}^d} g\log g\, d\lambda + \int_{\mathbb{R}^d} Vg\, d\lambda + \log k.$$

That is,

$$\mathcal{D}_{KL}(\mu\|\pi) = \mathrm{Ent}(\mu) + \mathcal{V}(\mu) + \log k,$$

where $\mathrm{Ent}(\mu)$ is the (negative) entropy relative to the Lebesgue measure; and $\mathcal{V}(\mu) := \int V \, d\mu$ is the potential energy. Now observe the following:

1. The entropy functional $\mu \mapsto \mathrm{Ent}(\mu)$ is geodesically convex by Proposition 4.1.29.
2. From Theorem 4.1.30, if V is convex, then $\mathcal{V}(\mu)$ is also geodesically convex.

Since the sum of two geodesically convex functionals remains geodesically convex, we conclude that $\mathcal{D}_{KL}(\cdot\|\pi)$ is geodesically convex.

Moreover, if V is α-convex, then $\mathcal{V}$ is α-geodesically convex (again by Theorem 4.1.30), and therefore $\mathcal{D}_{KL}(\cdot\|\pi)$ is α-geodesically convex as well. □

The main point of this section is that if the family $\mathcal{Q}$ is geodesically convex (i.e. whenever $\mu_0, \mu_1 \in \mathcal{Q}$, then the constant speed geodesic between them is also in $\mathcal{Q}$), and $\mathcal{V}$ is $\alpha-$geodesically convex then the solution to (4.9) (in which we optmize restricted to $\mathcal{Q} \subseteq \mathcal{P}_2^{ac}(\lambda)$) is unique.

To see this, suppose for the sake of contradiction that there exist two distinct minimizers $\mu_0^*, \mu_1^* \in \mathcal{Q}$ such that $\mathcal{V}(\mu_0^*) = \mathcal{V}(\mu_1^*) = \inf_{\mu\in\mathcal{Q}} \mathcal{V}(\mu)$. Since $\mathcal{Q}$ is geodesically convex, the constant-speed geodesic $(\mu_t^*)_{t\in[0,1]}$ connecting μ_0^* and μ_1^* is contained in $\mathcal{Q}$.

By α-geodesic convexity of $\mathcal{V}$, we have for all $t \in (0,1)$:

$$\mathcal{V}(\mu_t^*) \leq (1-t)\mathcal{V}(\mu_0^*) + t\mathcal{V}(\mu_1^*) - \frac{\alpha}{2} t(1-t) W_2^2(\mu_0^*, \mu_1^*).$$

Since both μ_0^* and μ_1^* are minimizers, the right-hand side equals $\inf_{\mu\in\mathcal{Q}} \mathcal{V}(\mu)$ minus a strictly positive term (unless $\mu_0^* = \mu_1^*$). Therefore:

$$\mathcal{V}(\mu_t^*) < \inf_{\mu\in\mathcal{Q}} \mathcal{V}(\mu),$$

which contradicts the minimality of μ_0^* and μ_1^*.

We conclude that the minimizer is unique.

Recall that by what we computed in the previous section (examples 4.1.19, 4.1.20), we can write the Wasserstein gradient flow as

$$\nabla_{\mathcal{W}_2} \mathcal{D}_{KL}(\cdot \| \pi) = \nabla V + \nabla \log f. \tag{4.10}$$

On top of uniqueness of the optimizer in (4.9), we have the Poljak–Łojasiewicz inequality that gives us a rate of convergence to the optimal solution when the optimization is addressed via Wasserstein gradient flows.

Theorem 4.1.32. Let $\pi \propto \exp(-V)$ be a density on $\mathbb{R}^d$, where V is α-convex. Let $\mathcal{Q} \subseteq \mathcal{P}_2^{ac}(\lambda)(\mathbb{R}^d)$ be geodesically convex. Then, the Wasserstein gradient flow $(\mu_t)_{t \geq 0}$ of $\mathrm{KL}(\cdot \,\|\, \pi)$ constrained to lie in $\mathcal{Q}$ satisfies

$$\mathcal{D}_{KL}(\mu_t \| \pi) - \mathcal{D}_{KL}(\mu^* \| \pi) \leq e^{-2\alpha t}[\mathcal{D}_{KL}(\mu_0 \| \pi) - \mathcal{D}_{KL}(\mu^* \| \pi)].$$

4.2 RKHS

An RKHS (*reproducing kernel hilbert space*) is an Hilbert space of functions $\mathcal{H}$ which has some interesting properties. We will see how we can leverage on these spaces to define the MMD distance, upon which we will build the optimization in Section 2.3.

4.2.1 Basics and representer theorem

In this subsection we review the main definitions of RKHS theory and state the important *representer theorem.*

Definition 4.2.1. A Hilbert space $\mathcal{H}$ is a complete, possibly infinite-dimensional linear space endowed with a inner product.

In the following we will only work with functions defined on $\mathbb{R}^d$. Essentially, a Hilbert space lets us apply concepts from finite-dimensional linear algebra to infinite-dimensional spaces.

A norm in $\mathcal{H}$ can be naturally defined from the given inner product, as $\|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{H}}}$. Our norm will always be assumed to be the one arising from the latter. Furthermore, we always assume that $\mathcal{H}$ is separable (contains a countable dense subset) so that $\mathcal{H}$ has a countable orthonormal basis[2].

The main tool when dealing with these spaces is the *Rieszs representaiton theorem.*

Theorem 4.2.2 (Riesz Representation Theorem I)**.** Let $\mathcal{H}$ be a Hilbert space over $\mathbb{R}$, and let $\Lambda : \mathcal{H} \to \mathbb{R}$ be a continuous linear functional. Then, there exists a unique element $g \in \mathcal{H}$ such that for all $f \in \mathcal{H}$:

$$\Lambda(f) = \langle f, g\rangle_{\mathcal{H}}.$$

Moreover, $\|\Lambda\| = \|g\|_{\mathcal{H}}$.

This result allows us to represent any continuous linear functional on $\mathcal{H}$ as an inner product with a unique element in $\mathcal{H}$.

[2] This is just a technical condition used in the proof of the *representer theorem.*

Theorem 4.2.3 (Riesz Representation Theorem II)**.** *Let $\mathcal{H}$ be a Hilbert space, and let $h \in \mathcal{H}$. Then:*

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, h \rangle_{\mathcal{H}} = \|h\|_{\mathcal{H}}.$$

That is, the operator norm of the linear functional $f \mapsto \langle f, h \rangle_{\mathcal{H}}$ is equal to the norm of h in $\mathcal{H}$.

Example 4.2.4*.* Let us consider the class of square integrable functions on the interval $[a, b]$, denoted by $L^2[a, b]$. We define the inner product as

$$\langle f, g \rangle = \int_a^b f(x) g(x)\, dx$$

with associated norm

$$\|f\| = \left(\int_a^b f^2(x)\, dx \right)^{1/2}$$

It can be checked that this space is complete, so it is a Hilbert space. However, there is one problem with the functions in this space. Consider trying to evaluate the function $f(x)$ at the point $x = k$. There exists a function g in the space defined as follows:

$$g(x) = \begin{cases} c & \text{if } x = k \\ f(x) & \text{otherwise.} \end{cases}$$

Because it differs from f only at one point, g is clearly still square-integrable, and moreover, $\|f - g\| = 0$. In fact, we can set the constant c (or, more generally, the value of $g(x)$ at any finite number of points) to an arbitrary real value. What this means is that a condition on the integrability of the function is not strong enough to guarantee that we can use it predictively, since prediction requires evaluating the function at a particular data value. This characteristic is what will differentiate reproducing kernel Hilbert spaces from ordinary Hilbert spaces, as we discuss in the following.

Definition 4.2.5. An *evaluation functional* over a Hilbert space of functions $\mathcal{H}$ is a linear functional

$$\mathcal{F}_t : \mathcal{H} \to \mathbb{R}$$

that evaluates each function in the space at the point t, or

$$\mathcal{F}_t[f] = f(t) \quad \text{for all } f \in \mathcal{H}.$$

Definition 4.2.6. A Hilbert space $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) if the evaluation functionals are bounded, i.e., if for all t there exists some $M > 0$ such that

$$|\mathcal{F}_t[f]| = |f(t)| \leq M\|f\|_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H},$$

which means that the norm topology is stronger than the topology induced by pointwise convergence.

While this condition might seem obscure or specific, it is actually quite general and is the weakest possible condition that ensures us both the existence of an inner product and the ability to evaluate each function in the space at every point in the domain.

In practice, it is difficult to work with this definition directly. We would like to establish an equivalent notion that is more useful in practice. To do this, we will need the *reproducing kernel* from which the reproducing kernel Hilbert space takes its name.

Firstly, from the definition of the reproducing kernel Hilbert space, we get that evaluations are linear and bounded functionals, hence we can apply the Riesz representation theorem.

Proposition 4.2.7. If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists a function $k_t \in \mathcal{H}$ (called the *representer of* t) with the *representing property*

$$\mathcal{F}_t[f] = \langle k_t, f\rangle_{\mathcal{H}} = f(t) \quad \text{for all } f \in \mathcal{H}.$$

This allows us to represent our linear evaluation functional by taking the inner product with an element of $\mathcal{H}$. Since k_t is a function in $\mathcal{H}$, by the representing property, for each $x \in X$ we can write

$$k_t(x) = \langle k_t, k_x\rangle_{\mathcal{H}}.$$

We take this to be the definition of reproducing kernel in $\mathcal{H}$.

Definition 4.2.8. The *reproducing kernel of* $\mathcal{H}$ is a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, defined by

$$k(t, x) := k_t(x).$$

Definition 4.2.9. A function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *reproducing kernel* if it is symmetric, i.e. $k(x, y) = k(y, x)$, and positive definite:

$$\sum_{i,j=1}^{n} c_i c_j k(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, \ldots, t_n \in X$ and $c_1, \ldots, c_n \in \mathbb{R}$.

Having this general notion of a reproducing kernel is important because it allows us to define an RKHS in terms of its reproducing kernel, rather than attempting to derive the kernel from the definition of the function space directly. The following theorem formally establishes the relationship between the RKHS and a reproducing kernel.

Proposition 4.2.10. A RKHS defines a corresponding reproducing kernel. Conversely, a reproducing kernel defines a unique RKHS.

Let us move to the main result for this subsection. Let us suppose having N data points $(x_i, y_i)_{i=1,\ldots,N}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We want to find the function in $\mathcal{H}$ that best interpoles the data while not being too complex, which formally is translated into:

$$f^* := \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} l(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

for a loss function $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and a constant $\lambda \in \mathbb{R}$. Optimizing through a vector space with an infinite number of dimension is a priori undoable on a computer, but we have the following theorem, known as *representer theorem* which allows us to move directly to a finite dimension optimization problem.

Theorem 4.2.11. The minimizer over the RKHS $\mathcal{H}$, of the regularized empirical loss functional

$$f^* := \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} l(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

can be represented by the expression

$$f^*(x) = \sum_{i=1}^{N} c_i K(x_i, x),$$

for some N-tuple $(c_1, \ldots, c_N) \in \mathbb{R}^N$. Hence, minimizing over the (possibly infinite-dimensional) Hilbert space boils down to minimizing over $\mathbb{R}^{N \times d}$.

For the proof we suggest reading [4].

As a fimal remark, we show some examples of commonly used RKHS.

1. *Linear kernel* $k(x, x') = x \cdot x'$
2. *Gaussian kernel* $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{\sigma^2}\right), \quad \sigma > 0$
3. *Polynomial kernel* $k(x, x') = (x \cdot x' + 1)^d, \quad d \in \mathbb{N}$

4.2.2 MMD

We now introduce the *Maximum Mean Discrepancy* (MMD), a distance between probability measures defined via RKHS embeddings.

Definition 4.2.12. Let $\mathcal{H}$ be a RKHS over $\mathbb{R}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The *Maximum Mean Discrepancy* (MMD) between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as

$$\mathrm{MMD}_{\mathcal{H}}(\mu, \nu) := \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}_{x\sim\mu}[f(x)] - \mathbb{E}_{y\sim\nu}[f(y)]\right).$$

This quantity measures how well functions in $\mathcal{H}$ can distinguish between μ and ν. Crucially, it admits a closed-form expression due to the RKHS structure.

Proposition 4.2.13. Let $\mu_k := \mathbb{E}_{x\sim\mu}[k(x, \cdot)] \in \mathcal{H}$ denote the *mean embedding* of μ in the RKHS $\mathcal{H}$. Then,

$$\mathrm{MMD}_{\mathcal{H}}(\mu, \nu) = \|\mu_k - \nu_k\|_{\mathcal{H}}.$$

Proof. For any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$, we have:

$$\mathbb{E}_{x\sim\mu}[f(x)] = \langle f, \mu_k \rangle_{\mathcal{H}}, \quad \mathbb{E}_{y\sim\nu}[f(y)] = \langle f, \nu_k \rangle_{\mathcal{H}}.$$

Hence,

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\langle f, \mu_k - \nu_k \rangle_{\mathcal{H}}\right) = \|\mu_k - \nu_k\|_{\mathcal{H}},$$

by the Riesz representation theorem. □

The MMD can thus be computed explicitly using the kernel function:

$$\mathrm{MMD}^2_{\mathcal{H}}(\mu, \nu) = \mathbb{E}_{x,x' \sim \mu}[k(x, x')] + \mathbb{E}_{y,y' \sim \nu}[k(y, y')] - 2\mathbb{E}_{x \sim \mu, y \sim \nu}[k(x, y)].$$

Regarding empirical estimation, if $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{m} \sum_{j=1}^m \delta_{y_j}$, the empirical (biased) estimator of MMD^2 is:

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j) + \frac{1}{m^2} \sum_{i,j=1}^m k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(x_i, y_j).$$

4.2.3 Spectral regualarized MMD

To complete.

Bibliography

[1] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.

[2] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport, 2024.

[3] Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows.* EMS Press, Berlin, second edition edition, 2023.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.