GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN
IN PUBLICA COMMODA
SEIT 1737

Estimation of Wasserstein Barycenters with Unimodal Densities

Bachelor Thesis in Mathematics
at the University of Göttingen

by

Janek Große

Advisor

Prof. Dr. Axel Munk

Second advisor

Dr. Housen Li

Date of submission

16 July 2024

Contents

1 Introduction

Imagine that for n different persons we have measured the concentration of some drug in the blood over time after taking a pill. This then gives us n time series that we would expect to look somewhat alike. Our task is now to find one time series that represents the development in a typical person's body. In other words, we have to find the ”average” of these n time series. Now, there are many ways to compute averages of functions but the most common ones (such as the L_2-Barycenter) might not always deliver the results one is looking for (see the figure below). This is why in this thesis we will focus on the Wasserstein distance, a metric that comes from the theory of Optimal Transport and that has many useful properties. The Wasserstein distance is a metric defined on probability measures which is why we will interpret our time series as probability density functions (pdfs). Which standardization technique we choose to turn these functions into pdfs will influence the result and will be one of the points of discussion in the following chapters.

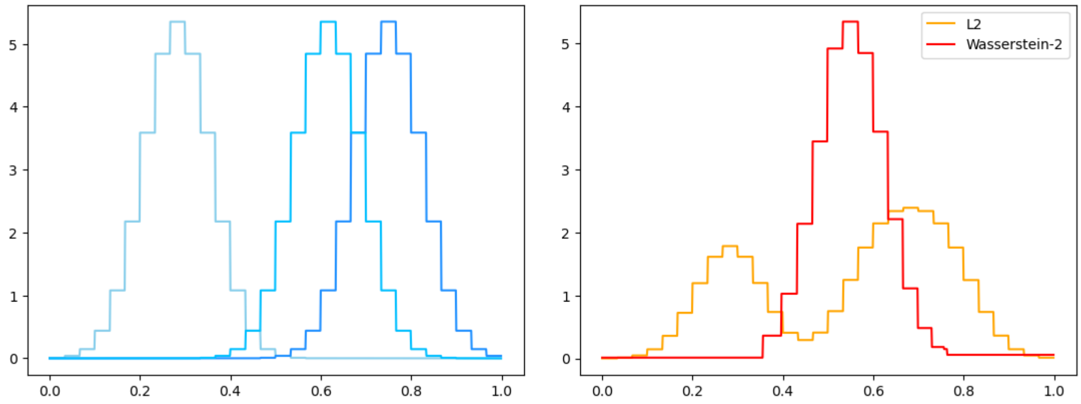

Figure 1: On the left side we see the (discretized) pdfs of three Gaussians and on the right side we can see their L_2-Barycenter as well as their Wasserstein-2-Barycenter. Arguably, the Wasserstein-Barycenter represents the input distributions better.

In the figure above we can see that one problem with L_2-Barycenters seems to be that they do not conserve unimodality well; the three Gaussians are all unimodal and yet their L_2-Barycenter has two local maxima. The goal of this thesis is to investigate the behaviour of Wasserstein-Barycenters if we input unimodal densities, and, in a second step, to compute the Wasserstein-Barycenter with unimodality imposed as a constraint. It turns out that, contrary to what you might guess from looking at Figure 1, in general Wasserstein-Barycenters do not conserve unimodality either (though there are special cases where they do). With a bit of ”reverse engineering” (see chapter 3) it is easy to come up with unimodal functions whose W_2-Barycenter is not unimodal:

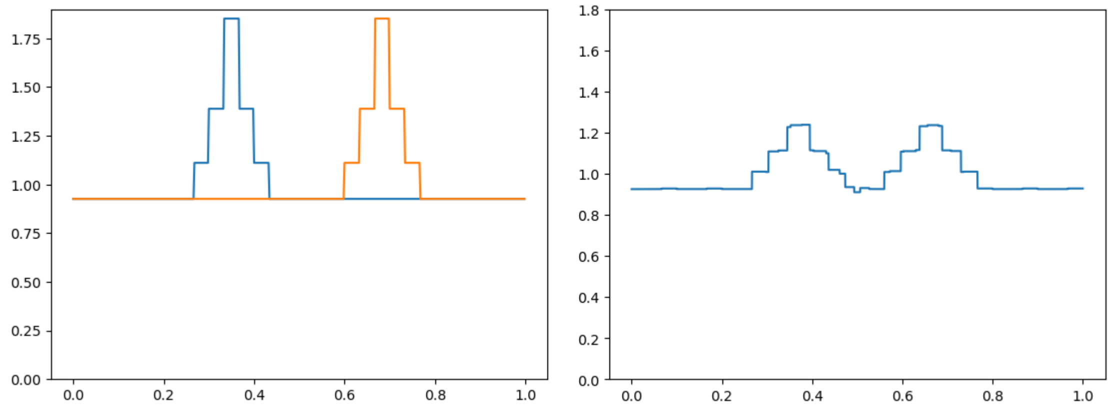

Figure 2: Now we have two densities as our input (left picture) that are both unimodal, but the Wasserstein-2-Barycenter (graph on the right) is not unimodal. The difference to Figure 1, as we will see later on, lies in the fact that here we have shifted the densities up.

We will start with a quick introduction to Optimal Transport, leading up to the results that we are going to need in the second part of the thesis, where we will then investigate Wasserstein Barycenters with imposed unimodality constraints. In addition to theoretical results, we will also show how to implement different versions of unimodal Wasserstein Barycenters and compare them to each other.

2 Theoretical backgrounds

In this section I will give a quick introduction to the mathematical concepts the project is based on. The first part will be about Optimal Transport and Wasserstein Barycenters and then I will give a short overview of some of the results in Convex Optimization that are going to be applied to our specific problem in the second part.

2.1 Optimal Transport

This part is mostly based on [San15] and [PZ20] with the later parts also using [PC19].

2.1.1 Monge and Kantorovich problems

Optimal Transport is an area of mathematics that was first investigated by the French mathematician Gaspard Monge in 1781. Imagine that there is a big pile of sand and, in some distance, there is a hole of equal volume in the ground. Your task is to fill the hole with the sand. You know that it costs you an effort of $c(x, y)$ to transport a grain of sand from a point x in the sandpile to a point y in the hole and your task is, of course, to move the pile of sand into the hole with the least effort possible. Now, mathematically, the sand pile as well as the hole are probability distributions μ and ν on abstract spaces X and Y (in our example $X = Y = \mathbb{R}^2$) and the minimal cost of moving one to the other will give us some kind of "distance" between the two. Since we want to find the best way of moving the pile into the hole, we first have to find a mathematical way to describe what exactly we mean by a "movement". It will be a transport map $T : X \to Y$ that takes a point x in the sand pile and returns the point in the hole to which x is transported. Furthermore, we need to guarantee that the hole is actually filled if we apply T and that the volume is conserved (sand cannot blow up in volume all of a sudden). This can be done by demanding that $\nu(A) = \mu(T^{-1}(A))$ for all measurable $A \subset Y$ (and T also has to be measurable). In this case we call ν the pushforward of μ under T. The cost of moving a grain of sand from x to $T(x)$ is $c(x, T(x))$ and therefore the cost associated with the whole transport map T is $\int_X c(x, T(x))d\mu(x)$. It will be assumed in this chapter that X and Y are separable Banach spaces (even though some of the definitions also work for more general spaces). We now have all the ingredients to state the Monge Optimal Transport problem:

Definition 2.1 Let $\mu \in P(X)$ and $\nu \in P(Y)$ probability measures on X and Y. For a cost function $c : X \times Y \to \mathbb{R}_{\geq 0}$ the optimization problem

$$\inf\{\int_X c(x, T(x))\ d\mu(x) \quad | \quad T_{\#}\mu = \nu\}.$$

is known as the **Monge Optimal Transport problem**.

Now, while there is a rich theory about the Monge OT problem, it also comes with a couple of problems. Most importantly, there is no guarantee that feasible transport maps even exist. If, for example, the first measure (the sandpile) has its mass concentrated at finitely many points and the other one has a density, then there can be no feasible transport map. Even if both measures are discrete but the second one has its mass distributed among a bigger number of points, the feasible set will be empty. To overcome these problems, Soviet mathematician Leonid Kantorovich slightly adapted Monge's approach in the 20th century. Instead of demanding that all the mass at $x \in X$ be transferred to the same point $T(x) \in Y$, he allowed for mass to be split up and to be sent to different points. We can think of this as a two-dimensional or a two-step approach: At each point x in X, there is a different measure μ_x which indicates how the sand sitting at x is split up and where in the hole it is sent to. If we do this for all x, we end up with a probability measure π on the product space $X \times Y$ that has μ and ν as its marginals, i.e.

$$\pi(A \times Y) = \mu(A) \quad \forall A \subset X \text{ measurable},$$

$$\pi(X \times B) = \nu(B) \quad \forall B \subset Y \text{ measurable}.$$

We denote the set of all measures satisfying these conditions as $\Gamma(\mu, \nu)$. Note that this set is never empty because it always contains the product measure $\mu \otimes \nu$. Any $\pi \in \Gamma(\mu, \nu)$ is also referred to as a **transport plan**: For $A \subset X$ and $B \subset Y$ we can interpret $\pi(A \times B)$ as the amount of mass going from any point in A to any point in B. Therefore, the first condition above reads as "the amount of mass that leaves A has to be exactly the amount of mass that sits at A" and the second condition reads as "the amount of mass ending up in B has to fill the volume of the hole at B exactly". Hence, these conditions can be thought of as guaranteeing that no mass can be created or destroyed during the transport. Similarly to before we can now define the Kantorovich Optimal Transport problem.

Definition 2.2 Let $\mu \in P(X)$ and $\nu \in P(Y)$ again be probability measures on X and Y. For a cost function $c : X \times Y \to \mathbb{R}_{\geq 0}$ the optimization problem

$$\inf_{\pi \in \Gamma(\mu,\nu)} \int_{X \times Y} c(x, y) \; d\pi(x, y).$$

is known as the **Kantorovich Optimal Transport problem**.

Not only do feasible plans exist now, but it can also be shown that there is always an optimal plan in the Kantorovich problem if we assume c to be continuous (see chapter 1.1 in [PZ20]).

2.1.2 Wasserstein distance

Returning to our initial intuition of measuring the distance between μ and ν, now assume that $X = Y$. Remember that we assumed that they are separable Banach spaces, i.e. they are equipped with a norm. Then the Kantorovich OT problem gives rise to a new family of metrics on a subset of $P(X)$, the so called Wasserstein distance. The subset we can define the Wasserstein metric on is introduced next.

Definition 2.3 Let $p \geq 1$. The Wasserstein space $W_p(X)$ is defined as

$$W_p(X) = \{\mu \in P(X) \mid \int_X ||x||^p \; d\mu(x) < \infty\}.$$

Now we are ready to define the distance between two measures that the Kantorovich OT problem gives us.

Definition 2.4 Let again $p \geq 1$. For $\mu, \nu \in W_p(X)$ the Wasserstein distance between them is defined as the minimal value in the Kantorovich OT problem applied to the cost function $c(x, y) = d(x, y) = ||x - y||$:

$$W_p(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \left(\int_{X \times Y} d(x, y)^p \; d\pi(x, y) \right)^{\frac{1}{p}}.$$

Theorem 2.5 *The Wasserstein distance as just defined is indeed a metric on $W_p(X)$.*

A proof can be found in chapter 7 of [Vil03]. After this general introduction to OT, we now narrow our focus down to the theory of cases that we are going to need in Section 2 of the thesis.

2.1.3 OT in one dimension

Now, while the Kantorovich problem always admits a solution, it is in general not straightforward to find this solution (nor the corresponding minimal score, i.e. the Wasserstein distance). If we however assume that $X = \mathbb{R}$, then there is a closed form representation of the Wasserstein distance:

Theorem 2.6 *If $X = \mathbb{R}$, then*

$$W_p(\mu, \nu) = \left(\int_0^1 |F^{-1}(x) - G^{-1}(x)|^p dx \right)^{\frac{1}{p}},$$

where F^{-1} and G^{-1} are the quantile functions corresponding to μ and ν, respectively.

For proofs of this theorem as well as the following corollary see chapter 1.5 in [PZ20].

While this is already a useful theorem which will be the basis for our work in section 2, it can be improved even further if $p = 1$:

Corollary 2.7 *If $p = 1$, then the formula can be simplified even further to give us*

$$W_1(\mu, \nu) = \int_{\mathbb{R}} |F(x) - G(x)| dx,$$

replacing the quantile function with the cumulative distribution function (cdf).

2.1.4 Wasserstein Barycenters

Now recall that our initial problem was to find the (constrained) 'average' of the given measures, i.e. one measure which best approximates all the other measures. The concept of a 'Barycenter', also referred to as a 'Fréchet Mean', is the natural way to formalize this average (see section 3.1.4 in [PZ20] and section 9.2 in [PC19] for this chapter).

Definition 2.8 Let (X, d) be a metric space, $x_1, x_2, ..., x_n \in X$ and let $q > 0$. If a solution to the minimization problem

$$\min_{x \in X} \frac{1}{n} \sum_{i=1}^{n} d(x_i, x)^q$$

exists, then the minimizer is called the **Barycenter** of $x_1, x_2, ..., x_n$. In this thesis we will of course consider the Wasserstein distance $d = W_p(\cdot, \cdot)$ and the parameter q from the definition above will be set to p from our respective Wasserstein space.

We will now have a closer look at the 1D-Barycenters of the two most common cases ($p = 1$ and $p = 2$) that are also going to be the focus later on in section 2. It turns out that in these cases we can find a solution to the unconstrained Barycenter problem in closed form.

Lemma 2.9 *If X is a Hilbert space (and $d(x, y) = ||x - y||$, $q = 2$), then the Barycenter of $x_1, x_2, ..., x_n \in X$ is $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.*

Proof Since we do not only need the result later on but also a part of the proof, we will have a look at it here. Let's denote the objective in the definition by g, i.e. $g(x) = \frac{1}{n} \sum_{i=1}^{n} ||x_i - x||^2$. We now compute

$$\begin{aligned}
g(x) &= \frac{1}{n}\sum_{i=1}^{n}\langle x_i - x, x_i - x\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle x, x\rangle - \frac{2}{n}\sum_{i=1}^{n}\langle x_i, x\rangle + \frac{1}{n}\sum_{i=1}^{n}\langle x_i, x_i\rangle \\
&= \langle x, x\rangle - 2\,\langle \frac{1}{n}\sum_{i=1}^{n} x_i, x\rangle + \langle \frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} x_i\rangle - \langle \frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} x_i\rangle + \frac{1}{n}\sum_{i=1}^{n}\langle x_i, x_i\rangle \\
&= \langle x - \frac{1}{n}\sum_{i=1}^{n} x_i, x - \frac{1}{n}\sum_{i=1}^{n} x_i\rangle - \langle \frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} x_i\rangle + \frac{1}{n}\sum_{i=1}^{n}\langle x_i, x_i\rangle \\
&= ||x - \overline{x}||^2 + c
\end{aligned}$$

where in the last line we abbreviated the last two terms of the line before with the constant c (as they do not depend on x). In this form it is now obvious that the unique minimizer of g is $\overline{x}$. □

Now recall that Theorem 2.6 gave us a representation of the Wasserstein distance as the usual L_p-distance between the quantile functions corresponding to μ and ν. However, from Functional Analysis we know that if (and only if) $p = 2$, then L_p is actually a Hilbert space, which immediately gives us the following result:

Corollary 2.10 *For $\mu_1, \mu_2, ..., \mu_n \in W_2(\mathbb{R})$, their 2-Barycenter is the measure μ with quantile function $F_\mu^{-1} = \frac{1}{n}\sum_{i=1}^{n} F_{\mu_i}^{-1}$.*

Now, for the case $p = 1$ we can use Corollary 2.7 to rewrite our objective from Definition 2.8 to

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} |F_\mu(x) - F_{\mu_i}(x)|dx = \frac{1}{n}\int_{\mathbb{R}}\sum_{i=1}^{n} |F_\mu(x) - F_{\mu_i}(x)|dx$$

An integral will certainly be minimized if the pointwise minimization of the integrand is admissible. We will show that this is the case here and start with a well-known result:

Lemma 2.11 *Let $x_1, ..., x_n$ be real numbers. Then the expression $\sum_{i=1}^{n} |x - x_i|$ is minimized by any number x satisfying $x = median(x_1, ..., x_n)$.*

If we now set $F_\mu(x) = median(F_{\mu_1}(x), ..., F_{\mu_n}(x))$ pointwise, then our Lemma tells us that this function minimizes the integrand pointwise. To make this definition of G unique we take the mean of the two midpoints for n even. Now, if all the F_{μ_i}'s are cdfs, then their pointwise median will again be a cdf (i.e. increasing, right-continuous and with limits 0 and 1 at $-\infty$ and ∞), giving rise to the following result:

Corollary 2.12 *For $\mu_1, \mu_2, ..., \mu_n \in W_1(\mathbb{R})$, their 1-Barycenter is any measure μ whose cumulative distribution function satisfies $F_\mu = median(F_{\mu_1}, ..., F_{\mu_n})$.*

Remark We could have used the more general representation given by Theorem 2.6 with $p = 1$ and then would have gotten the same result as Corollary 2.12, only with all cdfs replaced by quantile functions. In general however cdfs are slightly easier to work with (especially if we are actually interested in the corresponding density).

2.2 Densities, Cdfs and Quantile functions

Remember that our problem was about measures with unimodal densities (exact definition below). The formulas we have seen in the last chapter however all involved either cdfs or quantile functions. This is why in this section we will see which kinds of cdfs and quantile functions correspond to a unimodal pdf. Note that even though the Wasserstein distance is technically only defined on probability measures, we will also be referring to the Wasserstein distance between density functions or cumulative distribution functions as long as it is clear what we are talking about; the Wasserstein distance between pdfs f_1, f_2 is for example defined as the Wasserstein distance between the measures μ_1, μ_2, where $\mu_i(A) = \int_A f_i(x)\ dx$ for all measurable A. We will begin this section by defining some basic terminology which is somewhat self-explanatory but which will come up often:

Definition 2.13 We say that a function $f : (a, b) \to \mathbb{R}$, where a and b may assume the values $-\infty$ and ∞ respectively, is **unimodal** with mode $M \in (a, b)$ if it is monotonically increasing on $(a, M]$ and monotonically decreasing on $[M, b)$. We say that f is **convex-concave** with mode $M \in (a, b)$ if it is convex on $(a, M]$ and concave on $[M, b)$. Analogously, we define what it means for a function to be **concave-convex**.

The reason why we define *convex-concave* is that if f is unimodal, then the corresponding cdf F is convex-concave. As we have already seen in the first section, cdfs and also quantile functions play an important role in computing the Wasserstein distance. This is why we write down these results in our first theorem which is preceded by a quick and well-known lemma (see Proposition 1.1.10 (f) in [NP18] for example).

Lemma 2.14 *Let $g : (a, b) \to (c, d)$ be an increasing and invertible function. Then g is convex if and only if g^{-1} is concave. This statement still holds if the intervals are closed on one or both ends.*

Now we are ready to state and prove the theorem relating pdfs, cdfs and quantile functions in our setting:

Theorem 2.15 *Let $F : \mathbb{R} \to [0,1]$ be a distribution function and $M \in \mathbb{R}$. Then the following statements are equivalent:*

$$(i)\ F \text{ has a density } f \text{ that is unimodal with mode } M \tag{1}$$

$$(ii)\ F \text{ is convex-concave with mode } M \text{ and continuous at } M \tag{2}$$

$$(iii)\ F^{-1} \text{ is concave-convex with mode } F(M) \text{ and does not have a plateau at } F(M) \tag{3}$$

Remark We only need the requirements (continuity/no plateau) in (ii) and (iii) at one single point as we will see in the proof that convexity/concavity already takes care of it everywhere else.

Remark Note that the density in (i) is of course not unique, as we can choose any value on null sets. The statement therefore reads as "F has a density that we can choose to be unimodal with mode M".

Proof The proof is best done by showing $(i) \Leftrightarrow (ii)$ and $(ii) \Leftrightarrow (iii)$. The first part is also proved in [Unk15].

$(i) \Rightarrow (ii)$ is a well-known result from basic analysis (the integral of an increasing function is convex).

For $(ii) \Rightarrow (i)$ we note that convexity as well as concavity on open intervals implies continuity (Thm 1.1.2 in [NP18]). By Theorem 1.5.2 in that same book we get absolute continuity of F on on $(-\infty, M)$ and (M, ∞). Hence, F has a density f_- on $(-\infty, M)$ as well as a density f_+ on (M, ∞). The assumed continuity at M implies that we can combine these densities to one common density f that takes any value at M. Now, since the left derivatives (or right derivatives) of F exist everywhere and are non-decreasing where F is convex and non-increasing where F is concave (Theorem 1.4.2 in [NP18]), we get that this f can be chosen unimodal with mode M.

To prove $(ii) \Leftrightarrow (iii)$ we want to use Lemma 2.14. However, we are not directly able to do that as quantile functions and cdfs are generally not everywhere inverse to each other. Assuming (ii), we have already established that this implies that F is continuous (convexity/concavity implies continuity up to M/starting from M and the continuity at M is assumed). The convexity on $(-\infty, M]$ also implies that F is either strictly increasing on all of $(-\infty, M]$ or there is a point $r \in (-\infty, M]$ such that $F = 0$ on $(-\infty, r)$ and F is strictly increasing on $[r, M]$. In other words, F possibly starts off at 0 for some time, but once it leaves 0, it is strictly monotone up until M. This is easy to see, as if there were two intervals on which F is constant but with different values, then we could easily find a chord that connects them and lies below the graph, thus violating the convexity. Hence,

F is invertible on $(r, M]$ (r might be $-\infty$ now). Therefore, we can apply Lemma 2.14 and get that F^{-1} is concave on $(F(r), F(M)] = (0, F(M)]$. Similarly, due to concavity, F will be strictly increasing on $[M, \infty)$ whenever it takes a value other than 1. This shows that F^{-1} will be convex on $[F(M), 1)$, again making use of Lemma 2.14. The fact that F^{-1} does not have a plateau at $F(M)$ follows directly from the continuity of F at M which we assumed.

$(iii) \Rightarrow (ii)$ works exactly the same way. Again we argue that concavity/convexity implies continuity of F^{-1}. Just as before, it also implies that F^{-1} is strictly monotone except possibly around the mode (the right end of the concave interval or the left end of the convex interval). This is however ruled out as by assumption there is no plateau at $F(M)$. Hence, we get that F^{-1} is invertible on $(0, F(M)]$ as well as on $[F(M), 1)$ and by Lemma 2.14 F is convex-concave with mode M. The continuity at M follows directly from the fact that F^{-1} has no plateau at $F(M)$. □

2.3 Convex Optimization

As the project requires a significant amount of convex optimization, in this section I will briefly mention the concepts that we are going to use later on (the last section was technically already part of it). It should not come as a surprise that we will need convex optimization as the problem of finding the OT Barycenter in general is a nested optimization problem; computing the objective of our optimization problem is again an optimization problem. This section follows along the lines of [BV04]. We'll start by defining a *convex optimization problem*:

Definition 2.16 For convex functions $f_0, f_1, ..., f_n : \mathbb{R}^m \to \mathbb{R}$, $A \in \mathbb{R}^{p \times m}$ and $b \in \mathbb{R}^p$, the minimization problem

$$\begin{aligned} \text{Minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \;\; i = 1, ..., n \\ & Ax = b \end{aligned}$$

is called a **convex optimization problem**.

The big advantage of convex optimization problems over general (non-convex) optimization problems is that every local minimum is immediately also a global minimum: Suppose that x is feasible and minimizes f on $B_r(x)$ for some $r > 0$, i.e. x is a local minimum. Now assume that x is not a global minimum. Then there is a feasible y outside of that ball, such that $f_0(y) < f_0(x)$. Consider the convex combination $z = (1 - \lambda)x + \lambda y$ for $\lambda = \frac{r}{2||y-x||}$. Note that then $\lambda < 1$ as y is outside of $B_r(x)$, i.e. $||y - x|| \geq r$. Since x and y are feasible, any convex combination of them is again feasible (as the inequality constraint functions are convex and the equality constraints linear). Now, by design we have $||z - x|| = \frac{r}{2} < r$ and therefore $z \in B_r(x)$. However, by convexity of the objective we have $f_0(z) \leq (1 - \lambda) f_0(x) + \lambda f_0(y) < f_0(x)$, where the last step follows from the fact that $f_0(y) < f_0(x)$. This however is a contradiction to x being locally optimal, as we have found $z \in B_r(x)$ with an even better score. Therefore, any local minimum is also a global minimum in convex optimization problems.

2.3.1 Linear Programming

Now, while convex optimization problems are easier to handle than general optimization problems, there are further classifications into even more convenient classes of problems. Perhaps the simplest form of convex optimization problems are *linear programs*:

Definition 2.17 A convex optimization problem is called a linear program if $f_0, f_1, ..., f_n$ are all affine (i.e. linear with an added constant term). This means that it can be stated in the following form:

$$\begin{aligned} \text{Minimize} \quad & c^T x \\ \text{subject to} \quad & Gx \leq h \\ & Ax = b \end{aligned}$$

for $c \in \mathbb{R}^m$, $G \in \mathbb{R}^{n \times m}$ and $h \in \mathbb{R}^n$.

In other words, now we require not only the equality constraints to be (affine) linear, but also the inequality constraints as well as the objective function. Linear programs have the advantage that there are very fast solvers available. In many cases, the optimization problem one is facing does not seem to be a linear program at first sight, but can be rearranged and reformulated so that the resulting (equivalent) problem is a linear program. That will be the case later on in this thesis with the following problem (see chapter 6.1.1 in [BV04]):

Example 2.18 *Consider the optimization problem*

$$\begin{aligned} \underset{x}{\textit{Minimize}} \quad & ||x - u||_1 \\ \textit{subject to} \quad & Gx \leq h \\ & Ax = b \end{aligned}$$

where $u \in \mathbb{R}^m$ is some constant. While the constraints are all in linear programming form, the objective is not and therefore the problem does not look like a linear program.

We can however do some steps to transform it into one. First, we introduce another m additional variables $t \in \mathbb{R}^m$ that will represent the scores. Our objective will be $c^T \begin{pmatrix} t \\ x \end{pmatrix}$ with $c_1 = ... = c_m = 1$ and $c_{m+1} = ... = c_{2m} = 0$. If we add the constraints $|x_i - u_i| \leq t_i$ then this new problem is equivalent to the original one:

$$\begin{aligned} \underset{t,x}{\mathit{Minimize}} \quad & c^T \begin{pmatrix} t \\ x \end{pmatrix} \\ \mathit{subject\ to} \quad & Gx \leq h \\ & |x_i - u_i| \leq \ t_i, i = 1, ..., m \\ & Ax = b \end{aligned}$$

We are now minimizing the sum of the best upper bounds of $|x_i - u_i|$ which is of course equivalent to minimizing $\sum |x_i - u_i|$ directly. The advantage is that we now managed to shift the absolute value into the constraints where we can split $|x_i - u_i| \leq t_i$ up into $x_i - u_i \leq t_i$ and $x_i - u_i \geq -t_i$, and thus get rid of the absolute value. Writing this into matrices, our final problem in linear programming standard form is

$$\begin{aligned} \underset{t,x}{\mathit{Minimize}} \quad & c^T \begin{pmatrix} t \\ x \end{pmatrix} \\ \mathit{subject\ to} \quad & Ax = b \\ & \begin{pmatrix} -I & I \\ -I & -I \\ 0 & G \end{pmatrix} \begin{pmatrix} t \\ x \end{pmatrix} \leq \begin{pmatrix} u \\ -u \\ h \end{pmatrix} \end{aligned}$$

where I is the $m \times m$ identity matrix.

2.3.2 Quadratic Programming

Another class of convex optimization problems are *Quadratic programs.* As we will encounter one of them in the next section, I will briefly introduce them now:

Definition 2.19 A quadratic program is a convex program of the form

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2} x^T P x + q^T x + r \\ \text{subject to} \quad & Gx \leq h \\ & Ax = b \end{aligned}$$

for $q \in \mathbb{R}^m$, $r \in \mathbb{R}$ and $P \in \mathbb{R}^{m \times m}$ symmetric and positive semi-definite.

Remark Note that P being symmetric is not a restriction. If we have a general matrix Q we can just replace it by the symmetric matrix $P = \frac{1}{2}(Q + Q^T)$. Since

$$\frac{1}{2}x^t Px = \frac{1}{4}(x^T Qx + x^T Q^T x) = \frac{1}{4}(x^T Qx + (Qx)^T x) = \frac{1}{4}(x^T Qx + (x^T Qx)^T) = \frac{1}{2}x^T Qx,$$

both matrices yield the same score. The last step follows from the fact that in $\mathbb{R}$ taking the transpose doesn't change the value.

Remark Sometimes the condition that P must be positive semi-definite is skipped in the definition of a quadratic program. This would, however, in general lead to a non-convex program which is why we keep this constraint.
To see that positive semi-definiteness of P guarantees convexity of a quadratic program, we set $LHS = ((1-\lambda)x + \lambda y)^T P((1-\lambda)x + \lambda y) - (1-\lambda)x^T Px - \lambda y^T Py$ for $\lambda \in [0,1]$ and $x, y \in \mathbb{R}^m$. If we can show that this expression is less than or equal to zero then we've shown that the quadratic part of the objective is convex. Since the linear and constant parts are certainly convex, this would then prove that the whole objective is convex. And indeed we can compute

$$\begin{aligned} LHS &= (1-\lambda)^2 x^T Px + \lambda^2 y^T Py + \lambda(1-\lambda)(x^T Py + y^T Px) - (1-\lambda)x^T Px - \lambda y^T Py \\ &= -\lambda(1-\lambda)\Big(x^T Px + y^T Py - x^T Py - y^T Px\Big) \\ &= -\lambda(1-\lambda)(x-y)^T P(x-y) \\ &\leq 0 \end{aligned}$$

where in the last line we used the positive semi-definiteness of P.

Remark Note that any linear program is also a quadratic program with $P = 0$, i.e. the class of linear programs is contained inside the class of quadratic programs.

3 Wasserstein Barycenters with unimodality constraints

Now we turn the attention to our specific case. The setup will be as follows: We are given n functions $f_1, f_2, ..., f_n$ that are each defined at equidistant fixed points $x_1, x_2, ..., x_m$ (think of the results from an experiment). We will then continue these function to all of $\mathbb{R}$ by setting $f_i(x) = f_i(x_j)$ whenever $x \in (x_{j-1}, x_j)$ for $j = 1, ..., m$. We are also going to assume that $x_m = 1$ and we set $x_0 = 0$, i.e. $x_j = \frac{j}{m}$ and $f_i = 0$ outside of $[0, 1]$. We will then standardize these functions by dividing each of the f_i's by $\frac{1}{m}\sum_{j=1}^{m} f_i(x_j)$ for $i = 1, ..., n$ to ensure that they integrate to 1 (more on standardization in chapter 3.3.2). The density functions obtained by these operations will again be called f_i's.

Our task is now to find their Wasserstein Barycenter under the constraint that this Barycenter must have a unimodal density. This means that we have to minimize

$$\frac{1}{n}\sum_{i=1}^{n} W_p^p(\mu, \mu_i)$$

over all measures μ that correspond to a unimodal density function.

3.1 The W_2-case

For $p = 2$, Theorem 2.6 gives us the following representation of our objective function. We need to minimize

$$\frac{1}{n}\sum_{i=1}^{n} \int_0^1 (F^{-1}(x) - F_i^{-1}(x))^2 dx, \tag{$\star$}$$

over the set of all quantile functions F^{-1} that correspond to a unimodal probability density function. We already know when a quantile function corresponds to a unimodal pdf from Theorem 2.15, but let me make a quick remark about quantile functions in our setup. Working with them can at times be slightly annoying, because - even though the notation suggests otherwise - quantile functions are in general not the exact inverse of a cdf, since a cdf might not be invertible. In our setup however, we get the following result:

Remark If f is a unimodal pdf, then the corresponding quantile function is continuous and strictly increasing (and therefore invertible) on $(0, 1)$.

The reasoning is the same as in the proof of Theorem 2.15: The unimodality implies that the pdf is strictly positive everywhere except possibly at the very start and the very end of the interval $[0, 1]$. This means that the corresponding cdf is strictly increasing - and hence invertible - whenever it takes values other than 0 or 1. Therefore, in our setup, the quantile function is the "real" inverse of the cdf on $(0, 1)$ - and not only the pseudo-inverse.

Now we are ready to prove our first result about Wasserstein-2-Barycenters in this section.

Theorem 3.1 *Let $M \in \mathbb{R}$ and let $f_1, f_2, ..., f_n$ be unimodal densities with mode M. Then their unconstrained Wasserstein-2-Barycenter is again unimodal with mode M.*

PROOF From Corollary 2.10 we know that $F^{-1} = \frac{1}{n}\sum_{i=1}^{n} F_i^{-1}$ is the quantile function of the Barycenter. Now we just need to apply the previous Theorem 2.15 a couple of times: Since all $f_i's$ are unimodal with mode M, implication $(i) \Rightarrow (iii)$ gives us that all $F_i^{-1}{}'s$ are concave on $(0, F(M)]$ and convex on $[F(M), 1)$. Since linear combinations of convex (concave) functions are again convex (concave), it follows that F^{-1} is again concave-convex with mode $F(M)$. Using $(iii) \Rightarrow (i)$ in the theorem, we get that the Barycenter again has a unimodal density. □

This is a useful theorem as we now know that the 2-Barycenter conserves unimodality at a fixed mode. If the $f_i's$ however are not unimodal or even unimodal with different modes, then we've seen in the introduction that this result does not hold anymore. Indeed, a counterexample is easy to construct: You just have to find two concave-convex quantile functions whose sum is not concave-convex (see Example 3.2). In the proof of Lemma 2.9 we have seen that our objective $(\star)$ only differs by a constant from the expression

$$\int_0^1 (F^{-1}(x) - \frac{1}{n}\sum_{i=1}^{n} F_i^{-1}(x))^2 dx$$

.

This insight simplifies things as we do not have to deal with n different inputs anymore but only the "average" of the quantile functions of the input. In other words, if we set $G^{-1}(x) := \frac{1}{n}\sum_{i=1}^{n} F_i^{-1}(x)$, then our problem becomes the L_2-minimization problem

$$\min \int_0^1 (F^{-1}(x) - G^{-1}(x))^2 dx$$

over the set of all concave-convex quantile functions G^{-1}. This means that what we're doing is we calculate the unconstrained W_2-Barycenter and then find the "unimodal measure" (i.e. the measure with a unimodal density) that best approximates this unconstrained Barycenter in W_2. Before actually computing the Barycenter let me make one last remark about the existence and uniqueness of our solution:

Remark In the next chapter we will see that we have at most $n \cdot m$ possible modes and our solution will be continuous and piecewise linear with up to $n \cdot m$ points of non-differentiability. Now, for fixed mode M the set of all concave-convex quantile functions of that form (continuous and piecewise linear) with mode M is convex and closed. Since squaring is strictly convex and integrating conserves convexity as well, our objective is also strictly convex. As our objective is continuous and the feasible set is closed and bounded (in our setup all quantile functions only map into $[0, 1]$), a solution exists. Strict convexity then gives us that for fixed mode M there exists exactly one solution to our problem. We can therefore say that a solution to our problem exists and that there are at most $n \cdot m$ of them. Indeed, as we will see later on, the solution is not necessarily unique (i.e. there are cases where we get the same optimal score for different modes).

3.1.1 Computing the Barycenter

I want to start with a note on the shape which our solution will take (see chapter 4.3.3 in [PZ20]). Recall that our input data will be defined only at finitely many fixed timestamps $x_1, x_2, ..., x_m$ and we will continue these m points to a density function that takes the value it takes at x_j on all of $(x_{j-1}, x_j]$ (i.e. our pdf will be a step function). Now, it would be natural to expect the result to be of the same form. If we were to write a program to solve our problem it would even be a canonical approach to look for the m values $f(x_1), f(x_2), ..., f(x_m)$ that meet the unimodality conditions and that give the best score. However, it turns out that this approach will not only complicate things, but it will also prevent us from finding the real solution: We just saw in the last section that the solution to our problem has quantile function $F^{-1} = \frac{1}{n}\sum_{i=1}^{n} F_i^{-1}$ in some cases. Now, if f_i is a step function then the corresponding cdf F_i will be piecewise linear with points of non-differentiability at $x_1, x_2, ..., x_m$. Therefore the corresponding quantile function will also be piecewise linear with non-differentiabilities at $F_i(x_1), F_i(x_2), ..., F_i(x_m)$. If we now sum all the quantile functions for $i = 1, ..., n$, this will give us a piecewise linear function with up to $n \cdot m$ points of non-differentiability. Converting this back to a cdf and then to a pdf, we find that the Barycenter has a density of a different shape than our input functions; it has $n \cdot m$ jumps and the locations of the jumps have little to do with the locations of the original $x_i's$. It is therefore crucial to not restrict ourselves too early.

Having established that, our input now does not consist of n functions anymore but only of a single quantile function that is piecewise linear with $n \cdot m$ points of non-differentiability. These points are $F_1(x_1), F_1(x_2), ..., F_1(x_m), F_2(x_1), ..., F_2(x_m), ...F_n(x_1), ..., F_n(x_m)$. Let's sort this list, remove the duplicates (such that they are all at least ϵ apart) and rename the ordered elements $y_1, y_2, ..., y_l$, i.e. $y_i \leq y_{i+1}\ \forall i$. We also set $y_0 = 0$ and $y_{l+1} = 1$. Recall that our objective was $\int_0^1 (F^{-1}(y) - G^{-1}(y))^2\, dy$, where G^{-1} was known and F^{-1} had to be concave-convex. Knowing that both of them are piecewise linear with non-differentiabilities at $y_1, y_2, ..., y_l$, we can now write

$$G^{-1}(y) = G^{-1}(y_k) + \frac{(y - y_k)}{y_{k+1} - y_k} \cdot (G^{-1}(y_{k+1}) - G^{-1}(y_k))$$

for $y \in [y_k, y_{k+1})$. The same equation holds if we replace G^{-1} by F^{-1}. To keep the following expression as simple as possible, we set $c_k := F^{-1}(y_k) - G^{-1}(y_k)$ for $k = 1, ..., l$ and $c_0 = c_{l+1} = 0$. Now we can rewrite the objective to

$$\begin{aligned}
\int_0^1 (F^{-1}(y) - G^{-1}(y))^2 dy &= \sum_{k=0}^{l} \int_{y_k}^{y_{k+1}} \left(c_k + \frac{(y - y_k)}{y_{k+1} - y_k} \cdot (c_{k+1} - c_k) \right)^2 dy \\
&= \sum_{k=0}^{l} \frac{1}{(y_{k+1} - y_k)^2} \int_{y_k}^{y_{k+1}} \left((c_{k+1} - c_k) \cdot y + c_k y_{k+1} - c_{k+1} y_k \right)^2 dy \\
&= \sum_{k=0}^{l} \frac{1}{(y_{k+1} - y_k)^2} \left[\frac{1}{3} (c_{k+1} - c_k)^2 (y_{k+1}^3 - y_k^3) \right. \\
&+ (c_{k+1} - c_k)(c_k y_{k+1} - c_{k+1} y_k)(y_{k+1}^2 - y_k^2) \\
&+ \left. (c_k y_{k+1} - c_{k+1} y_k)^2 (y_{k+1} - y_k) \right] \\
&= \sum_{k=0}^{l} \frac{y_{k+1} - y_k}{3} \cdot (c_{k+1}^2 + c_{k+1} c_k + c_k^2)
\end{aligned}$$

where the last step involved nothing but multiplying out everything and collecting the terms (start by applying the identity $(a - b)^3 = (a - b)(a^2 + ab + b^2)$ as well as the second binomial formula to cancel one of the $y_{k+1} - y_k$'s in the denominator and then get rid of all the brackets to see that a lot of things cancel and that there are two more factors of $y_{k+1} - y_k$ in the numerator). The last expression is a quadratic in $c = (c_1, ..., c_l)$. We now want to get this objective into standard form for quadratic programming $\frac{1}{2} x^t Q x + x^t q$ but let me first make a quick remark.

Remark Another approach would have been to "discretize" the quantile functions and only consider their values at $y_1, ..., y_m$. This would have saved us the calculation involving the integral above but we wouldn't have gotten the mixed term $c_{k+1} c_k$ in the sum, i.e. this approach would have slightly changed the problem.

Back to our quadratic program, we set $q = 0 \in \mathbb{R}^l$ as there are no linear terms in the objective and define $Q \in \mathbb{R}^{l \times l}$ as follows: $Q_{i,i} = \frac{2}{3}(y_{i+1} - y_{i-1})$ for $i = 1, ..., l$ and $Q_{i,i+1} = Q_{i+1,i} = \frac{1}{3}(y_{i+1} - y_i)$ for $i = 1, ..., l - 1$. All other entries are set to zero. Note that we defined the edges $(c_0, y_0, c_{l+1}, y_{l+1})$ so that all of these expressions are well defined, the optimization however is of course only over $c_1, ..., c_l$.

Let's have a look at the constraints we need to impose on c. We want F^{-1} to be concave-convex which means that we will need some mode $M \in \{1, ..., l\}$. We are going to fix this mode for now and then run our program for all l possible modes. Comparing the scores will then yield the final result. Now, for our function F^{-1} to be concave up to y_M, we need

$$\frac{F^{-1}(y_{k+1}) - F^{-1}(y_k)}{y_{k+1} - y_k} \leq \frac{F^{-1}(y_k) - F^{-1}(y_{k-1})}{y_k - y_{k-1}}$$

for all $k < M$. For $k > M$ we want F^{-1} to be convex which means that we get the same inequality, only with $\geq$ instead of $\leq$. Translating this to use the $c's$ instead of the $F^{-1\prime}s$ we get the constraints

$$\frac{c_{k+1} - c_k + G^{-1}(y_{k+1}) - G^{-1}(y_k)}{y_{k+1} - y_k} - \frac{c_k - c_{k-1} + G^{-1}(y_k) - G^{-1}(y_{k-1})}{y_k - y_{k-1}} \leq 0$$

With these constraints we ensure that our function is concave-convex.

Now we have a convex quadratic program in standard form that we can solve using any quadratic programming solver. I will be using the qp_solvers library in Python [CAB+24] (see Appendix B for an implementation). The first step of the implementation is to convert the input functions $f_1, ..., f_n$ into one quantile function G^{-1} and the last step will be to convert the result of the quadratic program (which is a quantile function) back into a pdf. To do this, we first go from pdf to cdf by adding the values using numpy.cumsum and then from cdf to quantile function by keeping track of the locations of all points of non-differentiability of the quantile function and inverting the cdf between these points.

3.1.2 Observations and Outlook

We'll walk through an example to see what happens during the calculation:

Example 3.2 *Our two input density fuctions are both Gaussians shifted up a bit (see Figure below). The first one has its mode at about 0.23 and the second one has its mode at 0.77. Of course, to make them probability densities again, we divided both by their respective integral after the up-shift.*

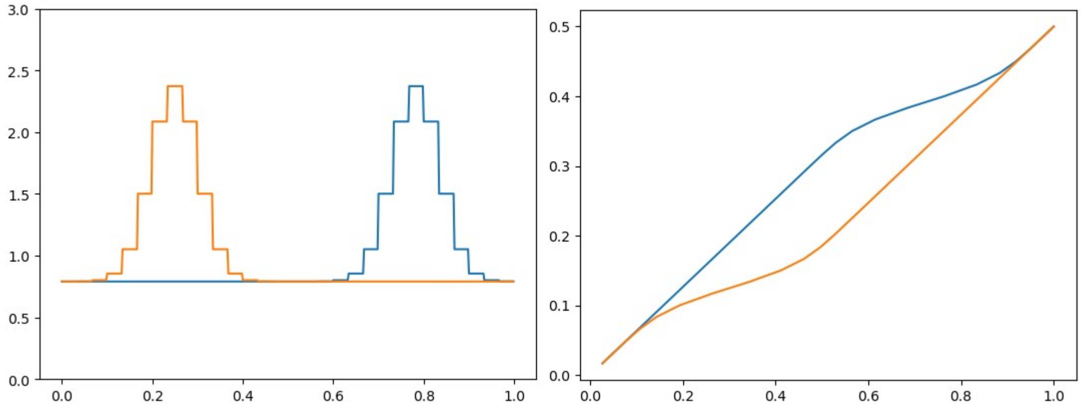

Figure 3: On the left we see the pdfs and on the right the corresponding quantile functions.

Since both densities are unimodal, the quantile functions are both concave up to their respective mode and convex from then onwards. However, if we now take the average of the two at each point, the next figure shows that this new quantile function is not concave-convex anymore:

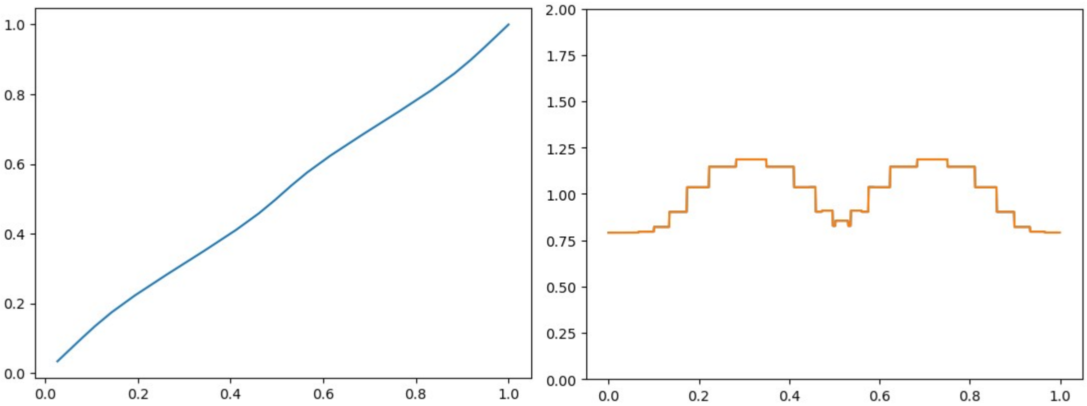

Figure 4: The quantile function of the Barycenter (left side) is not concave-convex which leads to a non-unimodal density function (right side).

We now use our quadratic programming solver to impose concave-convexity on the quantile function:

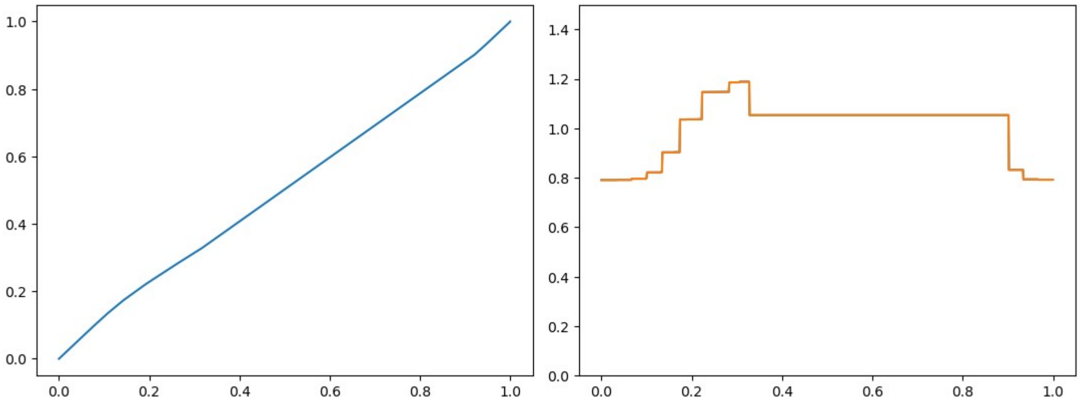

Figure 5: On the left side we can see the optimal solution among all concave-convex quantile functions which our program gives us and on the right side the corresponding pdf.

Now, this distribution resembles the first input distribution up to a certain point and then just goes flat for a long time. What surprises is that the solution seems pretty asymmetric for two very symmetric input densities; it would have been natural to expect a mode at around 0.5. A glance into the optimal scores which our program returns for each mode y_j for $j \in \{1, ..., l\}$ however tells us that the best solution when the mode is at 0.5 is significantly worse than the optimal solution which our program returned and which has its mode at around 0.3. Our intuition that the solution to a symmetric problem has *to be symmetric again however is not wrong: Remember that we only concluded that a unique minimizer exists for any fixed mode. This still leaves the possibility of solutions with a different mode but with the same score and indeed, another look into the l best scores reveals that there is another mode, at around 0.7, which gives the same score as the perfect solution that had its mode at 0.3. Plotting this other solution, it turns out that the corresponding pdf has pretty much the same density as the first solution, but mirrored at the line $x = 0.5$:*

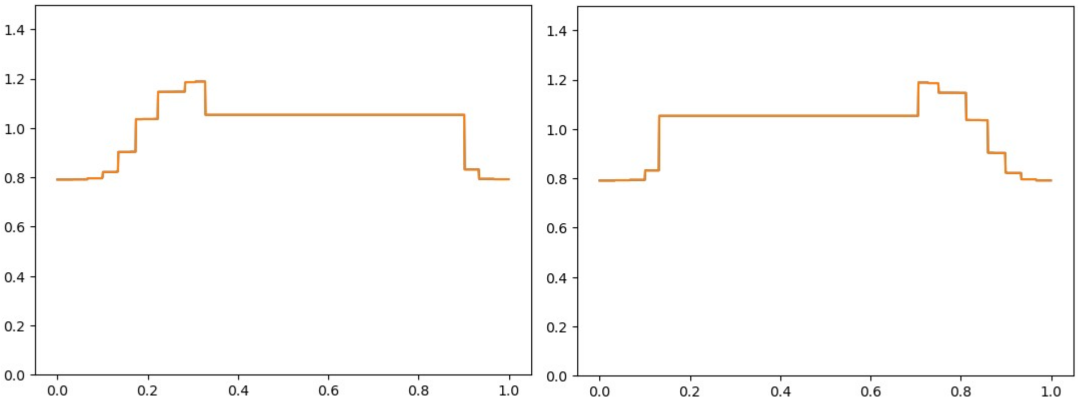

Figure 6: The two solutions to our problem

Now, while at the end we were able to restore the symmetry in the solutions from a mathematical point of view, the resulting densities still might not seem aesthetically satisfying to some, as any single solution (and we will certainly have to pick one for most applications) continues to be highly asymmetric. If we look at the solution with mode 0.5 however, we get a density that consists of a flat line for most of the time and therefore does not represent the input distributions well either (see Figure 7 below). An approach that could solve our problem in many cases is to change the way in which we standardize our input distributions. Remember that the (unshifted) Gaussians we saw in the Introduction did not have this problem; their Barycenter had a mode that was right in the center of the input densities' modes. The difference between shifted Gaussians and the normal Gaussians from the Introduction is that the normal Gaussians are basically zero outside of their mode ± a few standard deviations. Whenever a density is zero, then that will result in a jump in the quantile function which means that it will have no weight in the integral over that quantile function. This means that only the actual "hill" really counts towards the integral.

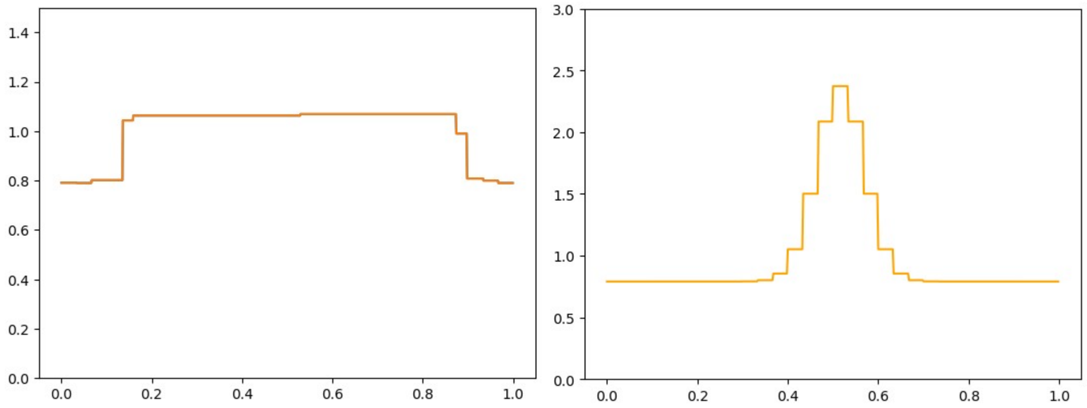

Figure 7: On the left we see the resulting pdf if we set the mode to 0.5 and on the right side we see the result with our new standardization technique that includes a shift along the y-axis.

We will therefore consider the following standardization procedure that is especially useful if we want to focus on the hills and not give too much attention to whatever happens elsewhere. Note that this is an assumption which is well in line with the general setup; if we want to find the best unimodal approximation in W_2, then that means that we expect a hill somewhere and probably care more about this part of the solution than the other parts. Note also that the following standardization procedure is not always a good choice; for a perfect uniform distribution for example it is not even well-defined.

As a first step, we subtract from each f_i its minimal value; this ensures that the new minimal value is zero. As a positive side effect it also means that we can now process inputs with negative values. As a second step, we then still have to divide each f_i by its integral. After running the program, our result should then be multiplied by the average of the numbers we divided by and, in a second step, we have to add the average number we subtracted in the beginning. Obviously, addition of or multiplication with a (positive) constant has no influence on whether the density is unimodal or not. The result of this new standardization applied to the distributions from Example 3.2 can be seen in the figure above. We will talk about standardization again in section 3.3.2.

3.2 The W_1-case

Dealing with the case $p = 1$ has some advantages and some disadvantages compared to the case $p = 2$. One advantage is that Corollary 2.7 gives us a representation of the Wasserstein-1-distance using cumulative distribution functions rather than quantile functions which are somewhat easier to work with. Using this, our optimization problem is to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}}|F(x)-F_i(x)|dx,$$

over all distribution functions F that correspond to a unimodal pdf. Recall that Theorem 2.15 gave us a characterization of cdfs that have a unimodal pdf. In short, we are minimizing over convex-concave cdfs. One of the disadvantages is the following: If we fix a mode M, then the set of all piecewise linear convex-concave distribution functions with mode M is convex, closed and bounded (again our cdfs only take non-zero values on $[0,1]$). It is easy to see that the objective is also continuous and convex since the absolute value is convex and integrating - just as before - conserves convexity. However, unlike for the case $p=2$, we don't have strict convexity anymore. This means that for a fixed mode M, a solution still exists but it might not be unique. Again, we will see that in our setup we can bound the possible number of modes which then gives us the following result:

Remark A solution to our problem exists but it is generally not unique. Indeed, there may be infinitely many solutions.

This result should not come as a surprise as we already saw in Corollary 2.12 that the unconstrained 1-Barycenter is the measure whose cdf is the pointwise median of the input cdfs - and the median is not necessarily unique! A quick example where this happens:

Example 3.3 *Using the fact that any monotone rearrangement is optimal in W_1, we can easily construct a case that has infinitely many solutions. In fact, if the input densities separate the mass in the way they are doing it in the figure below, then any distribution that has all its mass in between will be a solution to the Barycenter optimization problem.*

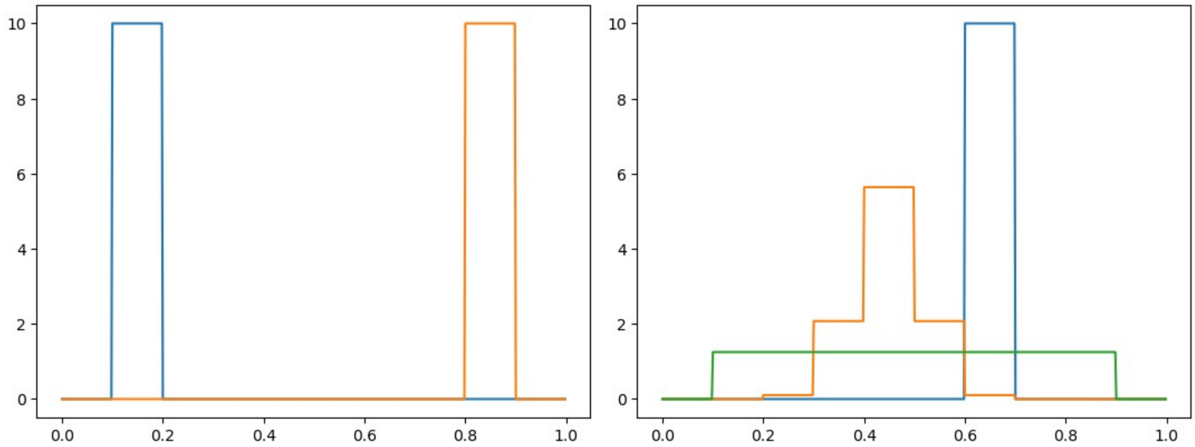

Figure 8: Any of the distributions on the right are (unimodal) W_1-Barycenters of the two input distributions on the left

Another thing we do not get for $p = 1$ is a result similar to Theorem 3.1; in W_1 it can happen that the input distributions are all unimodal with the same mode and yet their Barycenter is not (better: none of the Barycenters is), as the following example shows. This is essentially due to the fact that - unlike the mean - the median of convex functions is not necessarily convex again.

Example 3.4 *If we have 3 input distributions then the unconstrained Barycenter is unique and to calculate it we just have to take the pointwise median of the cdfs. In the left picture we see three unimodal pdfs with the same mode at 0.6 and in the second picture the corresponding cdfs. The median cdf is the blue one up until 0.5, the green one between 0.5 and 0.6, then the orange one and in the end the blue one again. Following this median in the second picture, it is obvious that this is not the course of a convex-concave function and therefore the Barycenter's density is not unimodal (picture 3).*

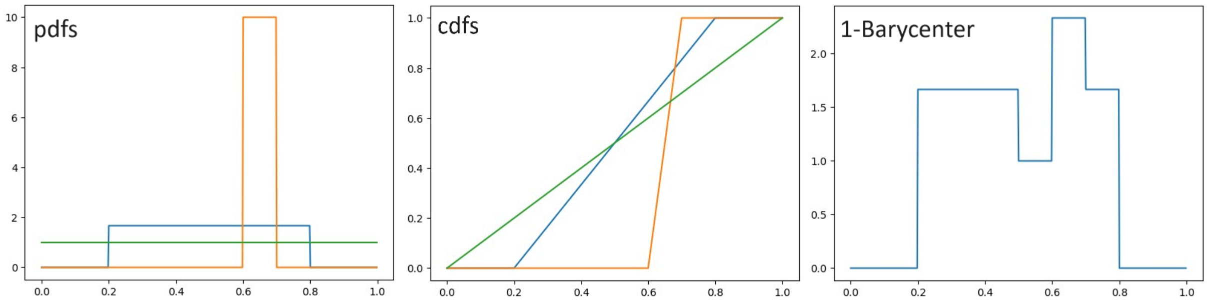

Figure 9: All input distributions are unimodal with the same mode and yet their W_1-Barycenter is not.

3.2.1 Computing the Barycenter

After these first results, we will now begin to calculate the Barycenter. To do this we start with a quick lemma that tells us how to calculate one of the integrals we are about to encounter:

Lemma 3.5 *Let $m \in \mathbb{N}$ and $a, b \in \mathbb{R}$ such that $a \neq b$. Then*

$$2m \int_0^{\frac{1}{m}} |a + m \cdot (b - a) \cdot x| dx = \frac{b \cdot |b| - a \cdot |a|}{b - a} = \begin{cases} b + a & \textit{if } a, b > 0 \\ -(b + a) & \textit{if } a, b \leq 0 \\ \frac{b^2 + a^2}{b - a} & \textit{if } a \leq 0, b > 0 \\ -\frac{b^2 + a^2}{b - a} & \textit{if } a > 0, b \leq 0 \end{cases}$$

This expression is continuous in (a, b) on all of $\mathbb{R}^2$ if we set it to $2|a|$ on the diagonal $a = b$. If we approximate it with $|a| + |b|$ then the error is zero whenever $sgn(a) = sgn(b)$ and $|\frac{2ab}{b-a}|$ else.

Remark Note that the error given by the approximation cannot blow up in size as it only applies whenever a and b have a different sign and hence the denominator $|b-a|$ can't get close to zero unless both a and b are close to zero in which case the numerator $|2ab|$ approaches zero even faster.

PROOF With the substitution $y = a + m \cdot (b-a) \cdot x$ and the fact that $(\frac{y|y|}{2})' = |y|$ we compute

$$2m \int_0^{\frac{1}{m}} |a + m(b-a)x| dx = \frac{2m}{m(b-a)} \int_a^b |y| dy = \frac{2}{b-a} (\frac{b|b|}{2} - \frac{a|a|}{2}) = \frac{b|b| - a|a|}{b-a}.$$

If $a = b$ then the integrand is just the constant $|a|$ and hence the integral evaluates to $\frac{|a|}{m}$ which, with the factor of $2m$ in front means the final result is $2|a|$. Now we calculate the error if we approximate the obtained expression by $|a| + |b|$. If a and b have the same sign, then $\frac{b|b|-a|a|}{b-a} = |b+a| = |b| + |a|$ and hence the error is 0. If they do not have the same sign (WLOG $a < 0 < b$), then $\frac{b^2+a^2}{b-a} - (|a| + |b|) = \frac{b^2+a^2}{b-a} - (b-a) = \frac{2ab}{b-a}$ and hence the error is $|\frac{2ab}{b-a}|$. □

Now we return to our objective. Recall that our input functions f_i are step functions with jumps at $x_1, , , , , x_m$ which means that their respective cdfs F_i will be continuous piecewise linear functions with points of non-differentiability at $x_1, x_2, ..., x_m$. We know that at these points we have $F_i(x_k) = \frac{1}{m} \sum_{j=1}^k f(x_j)$ and using linear interpolation we get the representation

$$F_i(x) = F_i(x_k) + m \cdot (x - x_k) \cdot (F_i(x_{k+1}) - F_i(x_k))$$

whenever $x \in (x_k, x_{k+1}]$. Remember that $x_0 = 0$ and of course $F(x_0) = 0$. Setting $c_{i,k} = F(x_k) - F_i(x_k)$, we can now write

$$\begin{aligned}
\int_0^1 |F(x) - F_i(x)| dy &= \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} |c_{i,k} + m(x - x_k)(c_{i,k+1} - c_{i,k})| dx \\
&= \sum_{k=0}^{m-1} \int_0^{\frac{1}{m}} |c_{i,k} + mx(c_{i,k+1} - c_{i,k})| dx \\
&= \sum_{k=0}^{m-1} \frac{c_{i,k+1} \cdot |c_{i,k+1}| - c_{i,k} \cdot |c_{i,k}|}{2m \cdot (c_{i,k+1} - c_{i,k})} \\
&\approx \sum_{k=0}^{m-1} \frac{|c_{i,k+1}| + |c_{i,k}|}{2m}.
\end{aligned}$$

In the second step we made the simple substitution $x \to x + x_k$, in the third step we used Lemma 3.5 and in the fourth step the approximation mentioned in that same Lemma. Using that $c_{i,0} = c_{i,m} = 0$, we can simplify the last expression even further to get

$$\frac{1}{m} \sum_{k=1}^{m-1} |c_{i,k}|.$$

Remark Another way to reach this result (instead of the approximation we used) would have been to only consider the values assumed at x_k for $k = 0, ..., m$, i.e. to discretize the cdfs.

To get our objective, we now have to sum our obtained expression over $i = 1, ..., n$. Skipping the constant $\frac{1}{nm}$, we therefore get

$$\sum_{i=1}^{n} \sum_{k=1}^{m-1} |F(x_k) - F_i(x_k)|.$$

Now, this almost has the form of the objective in Example 2.18, only that here we are dealing with a double sum. We will now rephrase our objective and constraints so that we can apply this example. For convenience, I will copy the statement from Example 2.18 into a Lemma:

Lemma 3.6 *The minimization problem*

$$\begin{aligned} \underset{y}{\text{Minimize}} \quad & ||y-u||_1 \\ \text{subject to} \quad & Gy \leq h \\ & Ay = b \end{aligned}$$

for $u \in \mathbb{R}^M,\ G \in \mathbb{R}^{N \times M},\ h \in \mathbb{R}^N,\ A \in \mathbb{R}^{P \times M},\ b \in \mathbb{R}^P$ *is equivalent to the linear program*

$$\begin{aligned} \underset{t,y}{\text{Minimize}} \quad & c^T \begin{pmatrix} t \\ y \end{pmatrix} \\ \text{subject to} \quad & Ay = b, \\ & \begin{pmatrix} -I & I \\ -I & -I \\ 0 & G \end{pmatrix} \begin{pmatrix} t \\ y \end{pmatrix} \leq \begin{pmatrix} u \\ -u \\ h \end{pmatrix} \end{aligned}$$

where I *is the* $M \times M$*-Identity matrix and* $c \in \mathbb{R}^{2M}$ *such that* $c_1 = ... = c_M = 1$, $c_{M+1} = ... = c_{2M} = 0$

Note that I renamed some of the variables from Example 2.18 to not confuse them with the variables we have already established in this section. Using the variable names from the lemma above we set $M = n(m-1)$ and define $u \in \mathbb{R}^M$ as follows: $u_{(k-1)n+i} = F_i(x_k)$ for $i = 1, ..., n$ and $k = 1, ..., m-1$. We are therefore optimizing over an $n \cdot (m-1)$-dimensional vector y which has the additional equality constraints $y_1 = y_2 = ... = y_n,\ y_{n+1} = ... = y_{2n},\ ..., y_{(m-2)n+1} = ... = y_{(m-1)n}$. We will write these constraints into the equality matrix A (and the vector b will be zero). The inequality constraints will have to ensure that our solution is convex-concave. This means that we need to fill G and h such that $y_{i\cdot n} - y_{i\cdot(n-1)} \leq y_{(i+1)\cdot n} - y_{i\cdot n}$ before the mode and the same equation with a "$\geq$" after the mode. This way our problem now has the form of the Lemma above and we can therefore transform it into a linear program in standard form that can be solved with any linear programming solver (see Appendix B for an implementation).

3.2.2 Unimodalizing the Barycenter

Remember that for the case $p = 2$ one of our main results was that the Barycenter with imposed unimodality constraints was actually just the unimodalized unconstrained Barycenter, i.e. we could calculate the unconstrained Barycenter and then find the best unimodal fit in W_2 to this single measure. In formulas, this can be expressed as

$$\arg\min_{\mu \in U} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu_i, \mu) = \arg\min_{\mu \in U} W_2^2(\mu, \nu)$$

where ν is the unconstrained Barycenter and U the set of measures with unimodal density functions. The question that naturally arises is whether we can get the same (or a similar) result for the case $p = 1$. Can we just take the median F_{med} of the cumulative distribution functions and then find the function that best approximates F_{med} among all convex-concave cdfs? The first problem that comes up is that generally none of these terms is uniquely defined; recall that in Example 3.3 we saw how there can be many unconstrained Barycenters to a set of input distributions. Unimodalizing these Barycenters will in turn result in many different unimodalized Barycenters. Our question should therefore be rephrased to "Is every unimodalized unconstrained Barycenter necessarily also a solution to the original problem, i.e. does it minimize the Barycenter functional among all unimodal pdfs?". We are now going to investigate this question.

We already know how to calculate the Barycenter with imposed unimodality constraints from the last section. In a similar way we can get an algorithm that returns the median of the input cdfs and then unimodalizes it. Taking the median is an easy thing (though it is of course not uniquely defined) and unimodalizing this median works in a similar way as before, again making use of Lemma 3.6: M will be $m-1$ and $u_k = F_{med}(x_k)$. We will not need any equality constraints this time (i.e. $A = b = 0$) and the inequality constraints (encoded in G and h) will have to ensure convex-concavity, just as before.

It turns out that in general, the answer to our question is ”no”, i.e. it is not true that any unimodalized unconstrained Barycenter is necessarily also a solution to the unimodality-constrained Barycenter problem, as the following example shows.

Example 3.7 *We are going to reuse Example 3.3 and choose the in some sense canonical median which takes the mean of the two midpoints for even n. This gives us the following unconstrained Barycenter:*

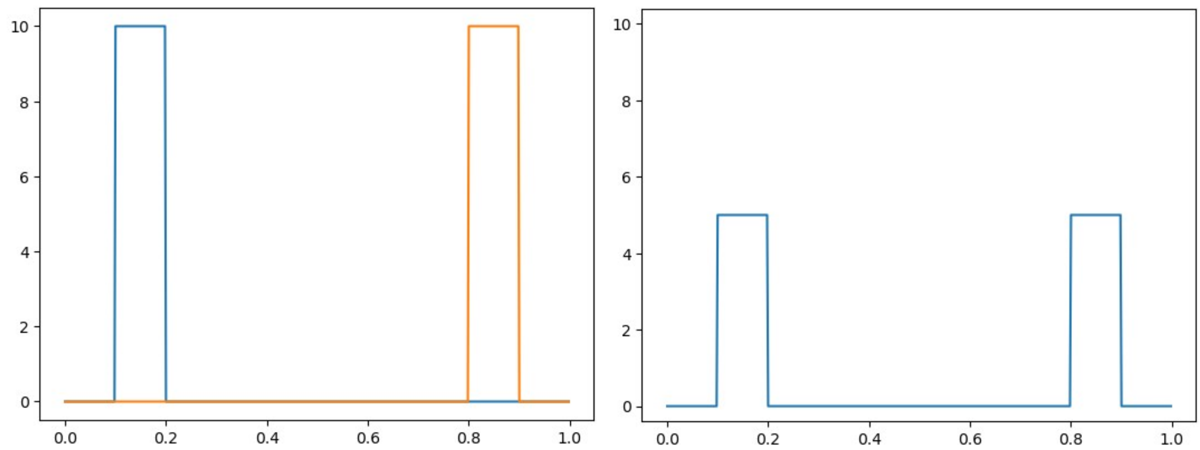

Figure 10: If we take the pointwise median of the cdfs of the two distributions on the left and then convert it back to a pdf, we get a Barycenter that is not unimodal (picture on the right).

We now compute the best unimodal approximation in W_1 to this Barycenter, using the algorithm described just before this example.

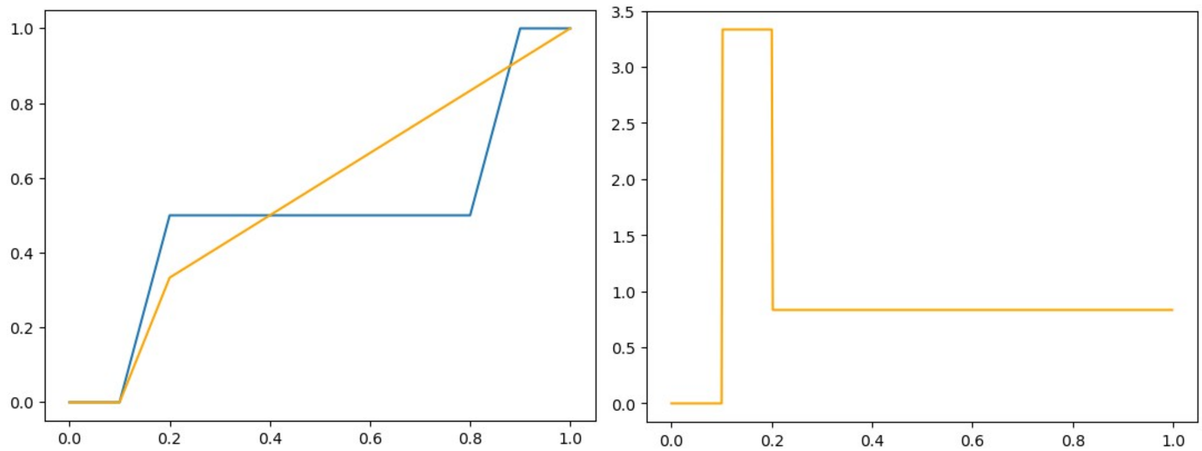

Figure 11: Left: The best convex-concave fit to the cdf of the unconstrained Barycenter (blue) is given by the orange curve (note that this might not be unique). On the right: The pdf of this unimodalized unconstrained Barycenter.

Now, this pdf has mass to the right of both input distributions (i.e. at $x > 0.9$) and will therefore intuitively not be a solution of the unimodality-constrained Barycenter problem. We could calculate the perfect score of this problem with our program and check that the distribution we just obtained gives a score strictly higher than that but here this is not even necessary. In fact, we can easily find a median of the input cdfs that translates to a unimodal pdf (see Example 3.3), while the cdf of our unimodalized unconstrained Barycenter (Figure 11 above) certainly is not a median of the input cdfs as it does not assume the value 1 on $(0.9, 1)$ whereas both input cdfs do. However, if it is not a median of the input cdfs but another convex-concave cdf is, then that other cdf will certainly give a better score.

At first, one might think that this result might be due to our approximation of the continuous distribution by a discrete one. This however is not the case as a quick manual calculation shows. In addition, it still serves as a counterexample for big values of m where the effects of discretization grow smaller and smaller (have a look at the simulation in the Appendix A). We have now learned that, in theory, we get two different ways to compute a unimodal W_1-Barycenter. However, it turns out that in practice the solutions to both problems tend to behave very similarly most of the time. In the Appendix A I included a simulation that investigates how big the difference typically is. There is a difference in runtime as the unimodality-constrained Barycenter needs m l_1-norm minimization programs with about $n \cdot m$ variables each whereas unimodalizing the unconstrained Barycenter is slightly less costly with m such programs that have m variables each. Since linear programs can be solved very fast however, this is hardly a concern unless we are dealing with very big values for n (and also m).

3.3 Applying the results and discussion

We have now seen how to compute a Barycenter that we want to ensure to be unimodal using the W_1- as well as the W_2-metric. This chapter will be dedicated to discussing which of the programs should be used depending on the input data (or the expected shape of the output) and how exactly it should be applied.

3.3.1 W_1 vs W_2

To demonstrate the advantages and disadvantages of the different p's, we will start with an artificial application but one that is designed in a way that it resembles a typical application case:

Our $n = 15$ input functions are constructed in the following way. All errors mentioned are i.i.d normally distributed with a mean of 0, the standard deviations vary. All input functions start as Gaussians. The mean is set to 0.3 + a small error for 9 of them and to 0.6 + a small error for the other 6. The standard deviation is set to 0.05 + a small error. In a second step, all of the Gaussians are shifted up by 1 + a small error. Lastly, we make all of them gradually decrease some time after their respective maximum (at 0.3 or 0.7). The result can be seen in the figure below. Note that all of these choices are of course somewhat random but I believe that the obtained input functions represent a typical application (e.g. from an experiment); given that all inputs are unimodal (except for small errors) we will certainly want our result to be unimodal as well.

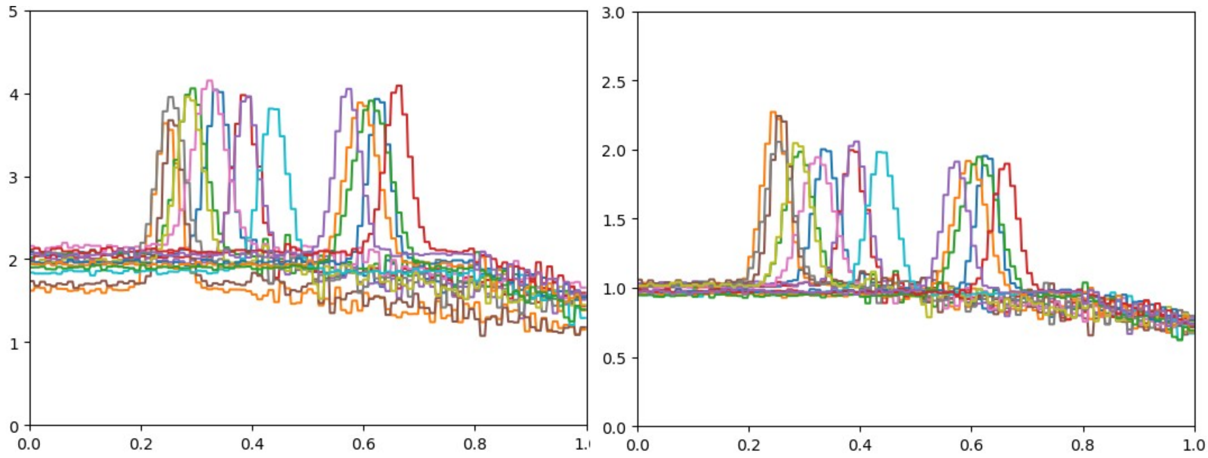

Figure 12: Our input functions before (left) and after (right) "basic" standardization, i.e. ensuring that they integrate to 1.

We now calculate the unimodality-constrained Barycenters for both the W_1- and the W_2-metric. During the standardization process we kept track of the numbers we divided each pdf by. In the end we multiply the result by the mean of these numbers (since we are talking about multiplication the geometric mean might be a better choice than the arithmetic mean), giving us the following results:

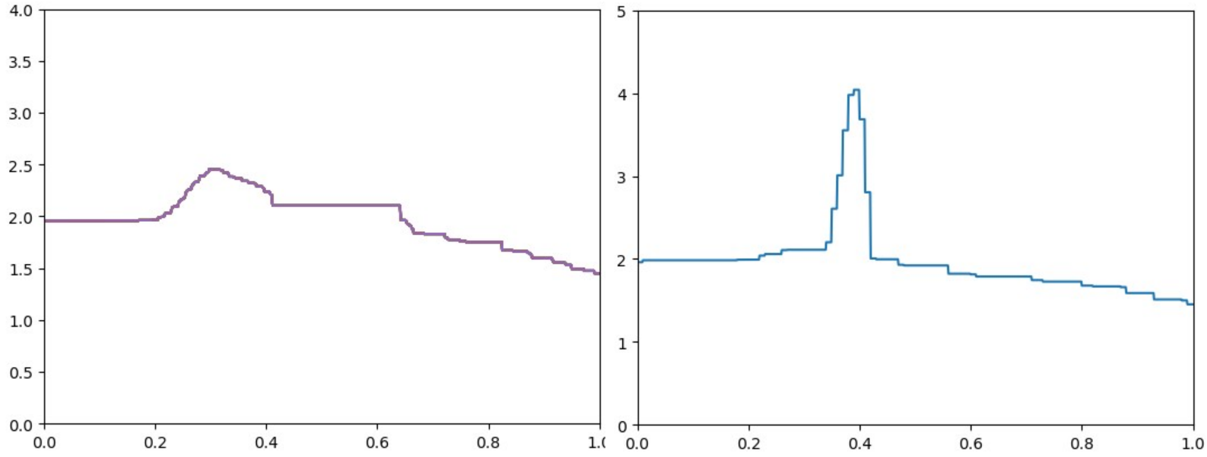

Figure 13: The unimodalized Barycenter for W_2 (left) and W_1 (right).

The most striking difference at first sight is the shape of the hill; in W_2 the hill is flat and wide, in W_1 it is high and narrow. Looking back at the input distributions, the hill of the W_1-Barycenter certainly is a better fit. In general, the W_2-Barycenter tends to smoothen things and make extreme features less extreme. This might sometimes be desirable but here it is too much and it just makes the Barycenter's hill look very different from the original distributions' hills. To further highlight the unsmooth nature of the 1-Barycenter and the positive side effect which the unimodalization has on it, we also calculate the unconstrained Barycenters:

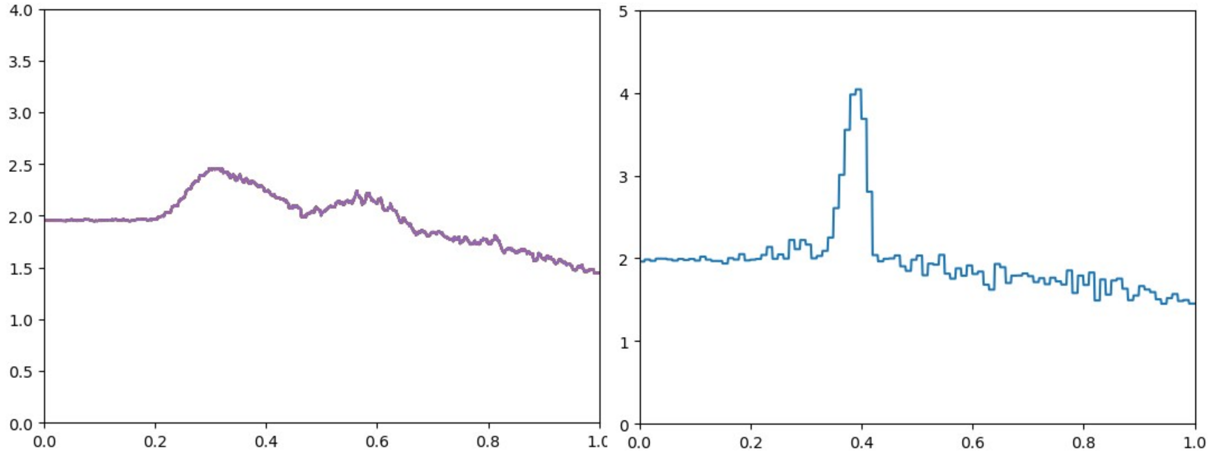

Figure 14: The unconstrained Barycenter for W_2 (left) and W_1 (right).

Another difference that becomes obvious after looking at the unconstrained Barycenters is that, on the whole, the W_1-Barycenter is already unimodal without even adding the unimodality constraints whereas the W_2-Barycenter is not. The distributions with their peak at 0.6 are seen as outliers by the program using the W_1-distance and at first do not seem to have a big influence on the result. In the W_2-program on the other hand they even get an own small hill. This is because, in general, the W_1-distance, just as the L_1-distance, is a lot more robust, i.e. not as susceptible to outliers. It is however not entirely true that the second group of distributions (with their peak at 0.6) has no influence on the result in W_1. Paying attention to the location of the peak, we see that it is almost at 0.4 for the W_1-Barycenter, which is slightly more to the right than for the big chunk of the input distributions and the W_2-Barycenter. This means that the outliers do have an effect on the location of the peak in W_1 in this example.

In summary we can say that if the data is prone to outliers that behave very differently than the rest, then the W_1-Barycenter might be the better choice as a couple of outliers do not have as much of an impact on it as on the W_2-Barycenter. If, on the other hand, all the input functions behave in a similar way, the W_2-Barycenter might be the better choice as it tends to yield the smoother solution. This argument however is a bit weaker than usually as the unimodality constraints make even the W_1-solution quite smooth. This is why, in general, I would suggest the unimodalized W_1-Barycenter for typical applications (e.g. in cases where we expect a single clearly distinguishable hill).

3.3.2 Standardization

So far, we have standardized our input distributions by dividing each of them by their integral. This was essentially so that we can interpret them as densities (else we wouldn't have been able to use the Wasserstein distance on them). This is the most basic and the only absolutely necessary form of standardization. There are however more standardization techniques, each of which has its advantages and its disadvantages. One inherent disadvantage of standardization itself is that it erases information. It is therefore crucial to not overuse it, however appealing the results might look like; the more standardization we use, the further the solution we get is from the true solution. The methods I'm proposing here are all designed to put more importance on the "hill" which we expect to exist in our Barycenter as well as in the input densities in our setup (if we didn't expect a clearly distinguishable hill then there would be no point in using a program that is designed to guarantee unimodality).

One standardization technique I've already briefly mentioned in Example 3.2 is a shift along the y-axis. We could either subtract from each input density its own minimum or we subtract the global minimum from each of the f_i's. Subtracting the global minimum will of course have less of an impact and can be considered a moderate form of standardization. The effect of this method is that from each point we are taking away an equal amount of mass, resulting in even more mass percentage wise at the points that had a lot of mass in the first place. These points are of course the location of the hill and thus we give more importance to the hill than without the shift.

Another possible - though unusual - form of standardization that might be useful in certain cases in our context is a shift along the x-axis. We could take the average of the x-locations of the maxima of the input densities (the assumption here is that the majority of the input densities already have a clearly distinguishable hill) and then shift each density along the x-axis such that it has its mode at this "average mode". If we assume that all input densities are more or less unimodal then this procedure would ensure that we do not need to impose any constraints on the W_2-Barycenter (see Theorem 3.1). This can be seen as a two-step approach: First we determine the location of the hill and then, in a second step, we determine the shape of the hill. A big disadvantage is that we will either need to invent data at the edges or shorten our interval. This method is therefore only recommendable if it is very obvious what happens outside of the hill and we only really care about the location as well as the shape of that hill.

In conclusion, both techniques increase the importance of the hill, but to varying degrees. While the shift along the y-axis as proposed above gives more relative weight to the hill, for the shift along the x-axis we manually determine the location of the mode and then need to delete (or invent) data outside of the area with the hill.

4 Conclusion

In this thesis we have dealt with the problem of finding the average or a typical representative of functions obtained from an experiment in the case that we suspect that this average must be unimodal. To do so, we used the Wasserstein distances W_1 and W_2 that come from the field of Optimal Transport. Since the Wasserstein distance is a distance between probability measures, in a first step we converted the more or less random input functions into probability density functions. This meant that we needed to somehow standardize these functions and we discussed different approaches to it; from the minimalistic standardization that only guarantees that each function integrates to 1 to more advanced shifts along the axes.

In addition to theoretical results which showed some inherent differences between the cases W_1 and W_2, we have also seen how to compute the unimodal Barycenters using linear/quadratic programming and implemented them. One question we investigated in more detail was whether it makes a difference if we calculate the unimodality-constrained Barycenter or just calculate the unconstrained Barycenter and then find the best unimodal fit to that. While in W_2 the formulas directly showed that there is no difference, in W_1 this was not the case anymore. A simulation however showed that the difference tends to be very small even in this case.

We discussed which of the programs should be used depending on the exact application case and came to the conclusion that in a typical application the W_1-program delivers better results (see figure below). Apart from guaranteeing one peak in the big picture, the unimodalization also serves to smooth the result by disallowing even small peaks.

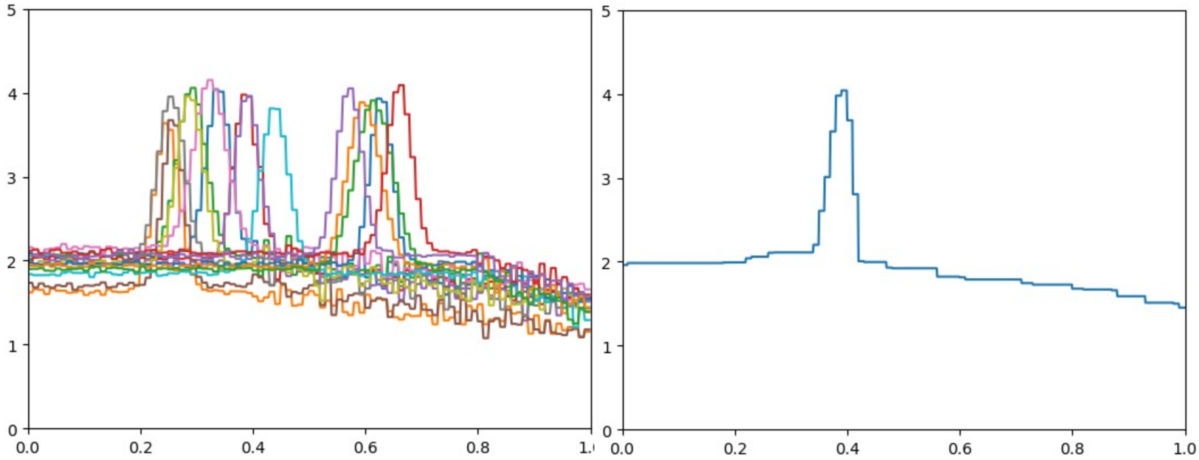

Figure 15: The unimodalized W_1-Barycenter (right) of the input functions (left) from section 3.3.1.

The work could be continued by investigating the unimodalized W_p-Barycenter for non-standard choices of p (i.e. $p \notin \{1, 2\}$). For $p \in (1, 2)$ for example, we might get a result that somewhat combines the advantages of the two cases we dealt with. While this will again be a convex problem (even a strictly convex one), it will probably not boil down to a standard linear/quadratic program; at least not with the approach we chose in this thesis.

5 References

[BV04] Stephen Boyd and Lieven Vandenberghe. Convex Optimization. Cambridge University Press, 2004.

[CAB$^+$24] Stéphane Caron, Daniel Arnström, Suraj Bonagiri, Antoine Dechaume, Nikolai Flowers, Adam Heins, Takuma Ishikawa, Dustin Kenefake, Giacomo Mazzamuto, Donato Meoli, Brendan O'Donoghue, Adam A. Oppenheimer, Abhishek Pandala, Juan José Quiroz Omaña, Nikitas Rontsis, Paarth Shah, Samuel St-Jean, Nicola Vitucci, Soeren Wolfers, Fengyu Yang, @bdelhaisse, @MeindertHH, @rimaddo, @urob, and @shaoanlu. qpsolvers: Quadratic Programming Solvers in Python, March 2024.

[NP18] Constantin P. Niculescu and Lars-Erik Persson. Convex Functions and Their Applications: A Contemporary Approach. Springer, 2018.

[PC19] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[PZ20] Victor M. Panaretos and Yoav Zemel. An Invitation to Statistics in Wasserstein Space. Springer, 2020.

[San15] Filippo Santambrogio. Optimal Transport for Applied Mathematicians. Birkhäuser, 2015.

[Unk15] Unknown. Unimodality and the dip statistic. `https://math.mit.edu/~rmd/465/unimod-dip-bds.pdf`, 2015. Unpublished handout, Massachusetts Institute of Technology.

[Vil03] C. Villani. Topics in Optimal Transportation. American Mathematical Society, 2003.

A Simulation complementing section 3.2.2

In section 3.2.2 we investigated whether the equation

$$\arg\min_{\mu\in U}\frac{1}{n}\sum_{i=1}^{n}W_p^p(\mu_i,\mu)=\arg\min_{\mu\in U}W_p^p(\mu,\nu), \tag{$\star$}$$

which proved to be true for $p=2$, also holds for $p=1$ (in the equation above ν was the unconstrained p-Barycenter and U the set of measures with unimodal density functions). In Example 3.7 we saw that this is generally not the case. In this section we will run a simulation and compare the results for both sides of the equation above to see how different they really are. In chapter 3 we derived algorithms that allowed us to formulate both optimization problems as a group of linear programs. We therefore have all the prerequisites we need to compute the solutions and it is also easy to compute their W_1-distance. The only question that remains is how we choose the input distributions. The first choice we have to make is which values for n and m we want to use. In this simulation, we choose three different values for m and three different values for n to be able to see how the magnitude of these two variables influences the results. The next question is what kind of distributions we want to have. We certainly want to have different kinds of distributions but at the same time we also want to use distributions where it is reasonable to look for a unimodal Barycenter. To do that, each of the input distributions will be a convex combination of a uniform distribution and a Gaussian with random weights. The Gaussian will have a random mean between 0 and 1 and also a random standard deviation. In the end we add error terms. That way, we include many different distributions, ranging from a uniform distribution (mass is equally distributed everywhere) to the other extreme (there is only mass at one mth of the points). The results are displayed in the table below.

m \ n	4	10	30
5	0.01080	0.00474	0.00184
20	0.00609	0.00208	0.00092
50	0.00287	0.00090	0.00039

Table 1: The table displays the average W_1-distance between the LHS and the RHS of ($\star$) for different values of n and m. Each pair (n, m) has been tested 1000 times with input distributions as specified above.

There are two things that stick out. First of all, all these numbers are pretty low. As a comparison, the difference between two Gaussians with std=0.1 and means $\frac{1}{100}$ apart is about 0.01, i.e. as high as the lowest value in the table. The other observation one can make from the data is that the difference seems to go down as n and m increase. This is however not uniformly true, since another quick computation shows that for any values of n and m, however high, we can find input distributions for which the results of the two sides of ($\star$) are at least 0.1 apart in W_1. To do this we just reuse Example 3.7 by putting all mass to the interval $[\frac{5}{m}, \frac{6}{m})$ for half of the distributions and to $[\frac{m-6}{m}, \frac{m-5}{m})$ for the other half (the exact numbers don't matter too much). With these configurations the lowest value we get for $m < 300$ is 0.096 and there is no downwards trend to be observed as m increases.

B Implementation in Python

In addition to the code for the simulation, I also attached the code for the unimodality-constrained Barycenters in W_1 and W_2 in a separate zip-folder.

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich diese Bachelorarbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe und dass ich diese Arbeit nicht bereits zur Erlangung eines akademischen Grades eingereicht habe.

Göttingen, 16. Juli 2024