Chapter 1

Numerical considerations

1.1 Introduction

1.1.1 A preliminary note

In this vignette, we carefully follow the work of Philippe Rigollet and Jonathan Weed in their paper [6]. We explore the computational estimator they propose in section 2.2 of their paper and reiterate some of their findings, aswell as give some new insights.

Also note that some parts in this document don't have enough context to be comprehensible - this is a preliminary version and is merely lazily copied together from my master's thesis.

1.1.2 Setting and basic notation

In our setting, we assume the data to be modeled by the equation

$$Y_i = m(X_i) + \varepsilon_i$$

for $i = 1, ..., n$, where $m : [0,1] \to [-V, V]$ is an unknown, non-decreasing function. We do not make any continuity assumptions on m, however, we assume that $V > 1$ is known and finite.

We observe the data in the form of multisets $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_n\}$. Data is observed in an unordered manner, i.e., we cannot tell which value X_i belongs to which value Y_i. The design points $\{X_1, ..., X_n\}$ are assumed to be non-random and fixed.

We assume that the error variables $\varepsilon_1, ..., \varepsilon_n$ are independent, identically distributed and sub-exponential.

The class of non-decreasing functions from $[0,1]$ to $[-V, V]$ will be referred to as $\mathcal{F}_V$.

For a function f, we will denote by π_f the pushforward of the empirical measure on $\{X_1, ..., X_n\}$ through f, i.e.,

$$\pi_f := f_* \left(\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \right) = \frac{1}{n} \sum_{i=1}^{n} \delta_{f(X_i)}.$$

The symbol $\hat{\pi}$ will be reserved for the empirical measure on our multiset of observations $\{Y_1, ..., Y_n\}$, i.e.,

$$\hat{\pi} := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}.$$

In two-dimensional settings, π^1 and π^2 will refer to the projection maps on the first or second component, respectively.
We will frequently refer to a collection of constants as C, where C might change from line to line.
The symbol $\mathcal{D}$ will typically refer to some distribution that is known but not explicitly specified.
The terms "Bernoulli noise" or the symbol Ber will always refer to random variables with $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = \frac{1}{2}$. We realize that this is non-standard terminology, as usually, one would expect to have $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}$ (or, more generally, some non-symmetric version). However, with our definition, the distribution has a variance of 1. Furthermore, as will be pointed out in the following section, isotonic regression problems for non-centered noise can always be rewritten in a form with centered noise.

Definition 1.1.1. We call

$$\hat{m} \in \operatorname*{argmin}_{g \in \mathcal{F}_V} \mathrm{W}_1(\pi_g * \mathcal{D}, \hat{\pi})$$

a *minimum Wasserstein deconvolution estimator.* [6]

Theorem 1.1.1. *Let μ, ν be probability measures on $\mathcal{B}(\mathbb{R})$ with inverse distribution functions F^{-1} and G^{-1}, respectively. If μ and ν have finite p-th moment, then*

$$W_p^p(\mu, \nu) = \int_0^1 |F^{-1}(x) - G^{-1}(x)|^p \,\mathrm{d}x.$$

1.1.3 Some preliminary propositions and lemmata

Theorem 1.1.2. *Let μ, ν be probability measures on $\mathcal{B}(\mathbb{R})$ with inverse distribution functions F^{-1} and G^{-1}, respectively. If μ and ν have finite p-th moment, then*

$$W_p^p(\mu, \nu) = \int_0^1 |F^{-1}(x) - G^{-1}(x)|^p \,\mathrm{d}x.$$

Lemma 1.1.3. *Let μ, ν and ρ be Borel-measures with finite p-th moment. It holds that*

$$\mathrm{W}_p(\mu, \nu) \leq \mathrm{W}_p(\mu, \rho) + \mathrm{W}_p(\rho, \nu).$$

Lemma 1.1.4. *With μ, ν as in Theorem 1.1.2, it holds that*

$$\mathrm{W}_1(\mu * \mathrm{Ber}, \nu * \mathrm{Ber}) \leq \mathrm{W}_1(\mu, \nu).$$

Theorem 1.1.5. *The estimator defined in Definition 1.1.1 fulfills the inequality*

$$\sup_{m \in \mathcal{F}_V} \left(\mathbb{E}\|m - \hat{m}\|_p^p\right)^{1/p} \leq Cp\frac{\log\log n}{\log n}$$

for all $1 \leq p < \infty$ and $n \geq \max\{4V^2, e^e\}$.

Proposition 1.1.6. *The estimator defined in Definition 1.1.1 fulfills the inequality*

$$\sup_{m\in\mathcal{F}_V} \left(\mathbb{E}\|m-\hat{m}\|_p^p\right)^{1/p} \leq \left(\frac{(2V)^{p-1}(V+1)(4V+2)}{\sqrt{n}}\right)^{1/p}$$

for all $p \geq 1$.

Proposition 1.1.7. *Let* M_V *be the space of probability measures supported on* $[-V,V]$. *The map*

$$\begin{aligned} \mathrm{W}_p^p(\cdot,\nu) : M_V &\to \mathbb{R} \\ \mu &\mapsto \mathrm{W}_p(\cdot,\nu) \end{aligned}$$

is convex.

1.2 A computationally efficient variation of the estimator

1.2.1 First steps

For a generic noise distribution $\mathcal{D}$, the problem of calculating an explicit $\hat{m}$ that fulfills

$$\hat{m} \in \operatorname*{argmin}_{g\in\mathcal{F}_V} \mathrm{W}_p(\pi_g * \mathcal{D}, \hat{\pi})$$

does not seem easily approachable via numerical methods at first. Luckily for us, though, Rigollet and Weed show that solutions to a discretized variation of the minimization problem are sufficiently close to solutions of the original problem. We will now examine the problems one faces when trying to calculate $\hat{m}$ for a generic distribution $\mathcal{D}$ and see if those problems are present in our special case $\mathcal{D} = \mathrm{Ber}$.

First off all, there is some nuisance associated with handling measures of the form $\pi_g = \sum_{i=1}^n \frac{1}{n}\mathbb{1}_{g(X_i)}$. Put $a_1 := g(X_1), ..., a_n := g(X_n)$, and consider the cumulative distribution function of π_g. By changing the values of the a_i, one controls where the $\frac{1}{n}$ jumps are in the cumulative distribution function.
Our goal is to work in a slightly different setting: We want to fix the values of the a_i, and instead control the magnitudes of the jumps in the cumulative distribution function. Instead of considering measures of the form $\pi_g = \sum_{i=1}^n \frac{1}{n}\mathbb{1}_{g(X_i)}$ where we control the value of $g(X_i)$, we consider measures of the form $\mu = \sum_{i=1}^N \mu_i \mathbb{1}_{a_i}$ where we control the values of the μ_i and fix a_i beforehand.
One advantage of this approach is immediately obvious: We can freely choose the number of a_i's, i.e., we can choose N. Since the calculations that will follow can get quite resource-intensive, the ability to reduce the dimension of the space where we search for solutions will turn out to be very useful.
Put $A := \{a_1, ..., a_N\}$ as an equidistant set and let μ be supported on A. The remaining problems all depend on the nature of $\mathcal{D}$. If $\mathcal{D}$ is not discrete, $\mu * \mathcal{D}$ will not be discrete. If $\mathcal{D}$ has unbounded support, $\mu * \mathcal{D}$ will have unbounded support. Neither of those problems concerns our case $\mathcal{D} = \mathrm{Ber}$, though it is still interesting to see how this can be handled in the general case.

1.2.2 The new estimator

Let $A = \{a_1, ..., a_N\}$ be a discrete, equidistant set with $a_1 \leq -V$, $a_N \geq V$ and $a_i < a_j$ for $i < j$, where N is some function of n. Denote by Π^A the projection that maps each point in $\mathbb{R}$ to its nearest neighbor in A, i.e.,

$$\Pi^A : \mathbb{R} \to A$$
$$x \mapsto \begin{cases} a_1 & \text{if } x \leq \frac{a_1+a_2}{2}, \\ a_i & \text{if } \frac{a_{i-1}+a_i}{2} < x \leq \frac{a_i+a_{i+1}}{2}, i = 2, ..., N-1, \\ a_N & \text{if } \frac{a_{N-1}+a_N}{2} < x. \end{cases}$$

Let M_V be the class of all Borel-measures supported on $[-V, V]$, and $M_{A,V}$ be the class of all Borel-measures supported on $A_V := [-V, V] \cap A$. For continuous noise distributions $\mathcal{D}$, take

$$\hat{\mu} \in \operatorname*{argmin}_{\mu \in M_{A,V}} \mathrm{W}_1(\Pi^A_*(\mu * \mathcal{D}), \hat{\pi}), \tag{1.1}$$

where $\Pi^A_*(\mu * \mathcal{D})$ denotes the push-forward of $\mu * \mathcal{D}$ through Π^A_*. For discrete noise distributions $\mathcal{D}$, it is sufficient to choose

$$\hat{\mu} \in \operatorname*{argmin}_{\mu \in M_{A,V}} \mathrm{W}_1(\mu * \mathcal{D}, \hat{\pi}).$$

Put

$$\hat{g}(x_i) = F^{-1}\left(\frac{i}{n}\right), i = 1, ..., n, \tag{1.2}$$

where F^{-1} denotes the inverse distribution function of $\hat{\mu}$.

The goal now is to show that $\mathbb{E}[\mathrm{W}_p(\pi_{\hat{g}} * \mathcal{D}, \hat{\pi})]$ is close to $\mathbb{E}[\mathrm{W}_p(\pi_{\hat{m}} * \mathcal{D}, \hat{\pi})]$. A proof for the general case can be found in [6, Proposition 2].

Of the two problems mentioned in the introduction, we've already seen how non-discreteness of $\mathcal{D}$ can be circumvented: By pushing $\mu * \mathcal{D}$ through Π_A, we again end up with a discrete measure.
For sub-exponential $\mathcal{D}$ with unbounded support, Rigollet and Weed show that is suffices to choose A in a clever way to handle this obstacle: They choose $A = \{a_1, ..., a_N\}$, such that with increasing N, the distance between individual elements gets smaller, but the overall "width" of A is unbounded and of order $\log(n)$. Details can be found in [6, chapter 2.2].

1.2.3 A proof for $\mathcal{D} = \mathrm{Ber}$

We will show that

$$\mathrm{W}_1(\pi_{\hat{g}} * \mathrm{Ber}, \hat{\pi}) \leq \mathrm{W}_1(\pi_{\hat{m}} * \mathrm{Ber}, \hat{\pi}) + \frac{C}{\sqrt{n}},$$

which gives us that

$$\sup_{m \in \mathcal{F}_V} \mathbb{E}\|m - \hat{g}\|_1$$

converges with the same rate as

$$\sup_{m \in \mathcal{F}_V} \mathbb{E}\|m - \hat{m}\|_1.$$

To this end, we need to prove some lemmata. We assume that A is an equidistant set on $[-V-1, V+1]$ with $N = \lceil \sqrt{n} \rceil$. The following lemma is the bounded support equivalent to [6, Lemmata 5 & 6]:

Lemma 1.2.1. *Let ν be a Borel-measure supported on $[-V-1, V+1]$. Then, for*

$$\nu' := \Pi^A_*(\nu)$$

it holds that

$$\mathrm{W}_1(\nu, \nu') \leq \frac{V+1}{\sqrt{n}-1}.$$

Proof. Put

$$\gamma := \sum_{i=1}^{N} \nu\Big|_{A_i} \times \delta_{\{a_i\}},$$

where we put $A_i = \left(\Pi^A\right)^{-1}(\{a_i\})$ for $i = 1, ..., N$, i.e. the A_i are the disjoint subsets of $[-V-1, V+1]$ that are projected onto the respective values a_i by Π^A. Clearly, γ is a coupling between ν and ν'. Because A is equidistant on $[-V-1, V+1]$, the maximum distance between points from A_i and the corresponding a_i is $\frac{2V+2}{2(N-1)} \leq \frac{V+1}{\sqrt{n}-1}$. Thus,

$$\mathrm{W}_1(\nu, \nu') \leq \int |x-y| \,\mathrm{d}\gamma \leq \int \frac{V+1}{\sqrt{n}-1} \,\mathrm{d}\gamma = \frac{V+1}{\sqrt{n}-1}.$$

□

The next result corresponds to [6, Lemma 7].

Lemma 1.2.2. *Let μ be supported on $[-V, V]$ with inverse distribution function F^{-1}, and let $g \in \mathcal{F}_V$ satisfy*

$$g(X_i) = F^{-1}\left(\frac{i}{n}\right).$$

Then,

$$\mathrm{W}_1(\mu, \pi_g) \leq \frac{2V}{n}.$$

Proof. Denote by G^{-1} the inverse distribution function of π_g. By definition of π_g we have

$$G^{-1}(x) = F^{-1}\left(\frac{i}{n}\right) \text{ for } \frac{i-1}{n} < x \leq \frac{i}{n}.$$

Thus, by Theorem 1.1.2, it holds that

$$\begin{aligned}
\mathrm{W}_1(\mu, \pi_g) &= \int \left|F^{-1}(x) - G^{-1}(x)\right| \mathrm{d}x \\
&= \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} \left|F^{-1}(x) - G^{-1}(x)\right| \mathrm{d}x \\
&= \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} \left|F^{-1}(x) - F^{-1}\left(\frac{i}{n}\right)\right| \mathrm{d}x \\
&\leq \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} \left|F^{-1}\left(\frac{i-1}{n}\right) - F^{-1}\left(\frac{i}{n}\right)\right| \mathrm{d}x \\
&= \sum_{i=1}^{n} \frac{1}{n} \left(F^{-1}\left(\frac{i}{n}\right) - F^{-1}\left(\frac{i-1}{n}\right)\right) \\
&\overset{\text{teleskope}}{=} \frac{1}{n} \left(F^{-1}(1) - F^{-1}(0)\right) \\
&\leq \frac{2V}{n}.
\end{aligned}$$

□

Similarly to [6, Proposition 2], one can use the previously established bounds to show the following proposition:

Proposition 1.2.3. *The estimator from equation 1.2 fulfills the equation*

$$\mathrm{W}_1(\pi_{\hat{g}} * \mathrm{Ber}, \hat{\pi}) \leq \mathrm{W}_1(\pi_{\hat{m}} * \mathrm{Ber}, \hat{\pi}) + \frac{C}{\sqrt{n}}$$

for some $C > 0$.

Proof. The claimed inequality can be achieved by multiple applications of the triangle inequality (prop. 1.1.3) and the previous lemmata.
Denote by $\hat{\mu}$ the estimator defined in 1.1, define

$$\hat{\nu} := \operatorname*{argmin}_{\nu \in M_V} \mathrm{W}_1(\nu * \mathrm{Ber}, \hat{\pi})$$

and put $\hat{\nu}' := \Pi_*^A(\nu)$.
By the triangle inequality,

$$\mathrm{W}_1(\pi_{\hat{g}} * \mathrm{Ber}, \hat{\pi}) \leq \mathrm{W}_1(\hat{\mu} * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\pi_{\hat{g}} * \mathrm{Ber}, \hat{\mu} * \mathrm{Ber}).$$

By Lemma 1.1.4,

$$\mathrm{W}_1(\hat{\mu} * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\pi_{\hat{g}} * \mathrm{Ber}, \hat{\mu} * \mathrm{Ber}) \leq \mathrm{W}_1(\hat{\mu} * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\pi_{\hat{g}}, \hat{\mu}),$$

and by Lemma 1.2.2,

$$\mathrm{W}_1(\hat{\mu} * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\pi_{\hat{g}}, \hat{\mu}) \leq \mathrm{W}_1(\hat{\mu} * \mathrm{Ber}, \hat{\pi}) + \frac{C}{n}.$$

By the triangle inequality,

$$\mathrm{W}_1(\hat{\mu} * \mathrm{Ber}, \hat{\pi}) \leq \mathrm{W}_1(\Pi_*^A(\hat{\mu} * \mathrm{Ber}), \hat{\pi}) + \mathrm{W}_1(\Pi_*^A(\hat{\mu} * \mathrm{Ber}), \hat{\mu} * \mathrm{Ber}).$$

By Lemma 1.2.1,

$$\mathrm{W}_1(\Pi^A_*(\hat{\mu} * \mathrm{Ber}), \hat{\pi}) + \mathrm{W}_1(\Pi^A_*(\hat{\mu} * \mathrm{Ber}), \hat{\mu} * \mathrm{Ber}) \leq \mathrm{W}_1(\Pi^A_*(\hat{\mu} * \mathrm{Ber}), \hat{\pi}) + \frac{C}{\sqrt{n}}.$$

By optimality of $\hat{\mu}$,

$$\mathrm{W}_1(\Pi^A_*(\hat{\mu} * \mathrm{Ber}), \hat{\pi}) \leq \mathrm{W}_1(\Pi^A_*(\hat{\nu}' * \mathrm{Ber}), \hat{\pi}),$$

and by the triangle inequality,

$$\mathrm{W}_1(\Pi^A_*(\hat{\nu}' * \mathrm{Ber}), \hat{\pi}) \leq \mathrm{W}_1(\hat{\nu}' * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\Pi^A_*(\hat{\nu}' * \mathrm{Ber}), \hat{\nu}' * \mathrm{Ber}).$$

By Lemma 1.2.1,

$$\mathrm{W}_1(\hat{\nu}' * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\Pi^A_*(\hat{\nu}' * \mathrm{Ber}), \hat{\nu}' * \mathrm{Ber}) \leq \mathrm{W}_1(\hat{\nu}' * \mathrm{Ber}, \hat{\pi}) + \frac{C}{\sqrt{n}}.$$

By the triangle inequality,

$$\mathrm{W}_1(\hat{\nu}' * \mathrm{Ber}, \hat{\pi}) \leq \mathrm{W}_1(\hat{\nu} * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\hat{\nu}' * \mathrm{Ber}, \hat{\nu} * \mathrm{Ber})$$

by Lemma 1.1.4,

$$\mathrm{W}_1(\hat{\nu} * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\hat{\nu}' * \mathrm{Ber}, \hat{\nu} * \mathrm{Ber}) \leq \mathrm{W}_1(\hat{\nu} * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\hat{\nu}', \hat{\nu})$$

and by Lemma 1.2.1 / definition of $\hat{\nu}'$,

$$\mathrm{W}_1(\hat{\nu} * \mathrm{Ber}, \hat{\pi}) + \mathrm{W}_1(\hat{\nu}', \hat{\nu}) \leq \mathrm{W}_1(\hat{\nu} * \mathrm{Ber}, \hat{\pi}) + \frac{C}{\sqrt{n}}.$$

Finally, by optimality of $\hat{\nu}$, we have

$$\mathrm{W}_1(\hat{\nu} * \mathrm{Ber}, \hat{\pi}) \leq \mathrm{W}_1(\pi_{\hat{f}} * \mathrm{Ber}, \hat{\pi}),$$

and thus overall

$$\mathrm{W}_1(\pi_{\hat{g}} * \mathrm{Ber}, \hat{\pi}) \leq \mathrm{W}_1(\pi_{\hat{f}} * \mathrm{Ber}, \hat{\pi}) + \frac{C}{\sqrt{n}},$$

where $C \leq \frac{2V}{\sqrt{n}} + 3\frac{(V+1)\sqrt{n}}{\sqrt{n}-1}$. □

Remark 1.2.4. By replacing $\hat{m}$ with $\hat{g}$ in the proofs of Theorem 1.1.5 or Proposition 1.1.6, one easily gets the same bounds of convergence for $\hat{g}$ as we have for $\hat{m}$. The first objective in these proofs is bounding $\mathrm{W}(\hat{g}, \hat{\pi})$ by $\mathrm{W}(\hat{g} * \mathrm{Ber}, \hat{\pi})$, which in fact does not use any properties of $\hat{g}$ at all. At that point, we can use Proposition 1.2.3 to bound $\mathrm{W}(\hat{g} * \mathrm{Ber}, \hat{\pi})$ by $\mathrm{W}(\pi_{\hat{m}} * \mathrm{Ber}, \hat{\pi})$, which lets us continue as before.

1.3 Details on minimizing the objective function

In this section, we will examine the different steps needed to actually produce a working program that calculates the minimum Wasserstein estimator. We will do this for the general case where $\mathcal{D}$ is not necessarily a Bernoulli distribution.

We denote the measure associated with $\mathcal{D}$ by $P_{\mathcal{D}}$.

Suppose $A = \{a_1, ..., a_N\}$ is an equidistant set and put $A_V := A \cap V = \{a_1^V, ..., a_K^V\}$. Consider the following calculation, where $h = \frac{a_2 - a_1}{2}$ and $j \in \{2, ..., N-1\}$:

$$\begin{aligned}\Pi_A(\mu * \mathcal{D})(\{a_j\}) &= \int\int \mathbb{1}_{(a_j-h,a_j+h]}(x+y)\,\mathrm{d}P_{\mathcal{D}}(x)\,\mathrm{d}\mu(y)\\ &= \int\int_{a_j-h-y}^{a_j+h-y} \mathrm{d}P_{\mathcal{D}}(x)\,\mathrm{d}\mu(y)\\ &= \int P_{\mathcal{D}}(a_j-h-y, a_j+h-y]\,\mathrm{d}\mu(y)\\ &= \sum_{i=1}^{K} \mu(\{a_i^V\})P_{\mathcal{D}}(a_j-h-a_i, a_j+h-a_i].\end{aligned}$$

For $j = 1$, this reads

$$\begin{aligned}\Pi_A(\mu * \mathcal{D})(\{a_j\}) &= \int\int \mathbb{1}_{(-\infty,a_j+h]}(x+y)\,\mathrm{d}P_{\mathcal{D}}(x)\,\mathrm{d}\mu(y)\\ &= \sum_{i=1}^{K} \mu(\{a_i^V\})P_{\mathcal{D}}(-\infty, a_j+h-a_i],\end{aligned}$$

and for $j = N$ this reads

$$\begin{aligned}\Pi_A(\mu * \mathcal{D})(\{a_j\}) &= \int\int \mathbb{1}_{(a_j-h,\infty)}(x+y)\,\mathrm{d}P_{\mathcal{D}}(x)\,\mathrm{d}\mu(y)\\ &= \sum_{i=1}^{K} \mu(\{a_i^V\})P_{\mathcal{D}}(a_j-h-a_i, \infty).\end{aligned}$$

Defining

$$\bar{\mu} := \begin{pmatrix}\mu(\{a_1\})\\ \vdots\\ \mu(\{a_K\})\end{pmatrix}$$

offers the convenient representation

$$\begin{pmatrix}\Pi_A(\mu * \mathcal{D})(\{a_1\})\\ \vdots\\ \Pi_A(\mu * \mathcal{D})(\{a_K\})\end{pmatrix} = \overline{P_{\mathcal{D}}} \cdot \bar{\mu},$$

where $\overline{P_{\mathcal{D}}}^{i,j} = P_{\mathcal{D}}(a_j-h-a_i, a_j+h-a_i]$ for $j = 2, ..., N-1$, $\overline{P_{\mathcal{D}}}^{i,j} = P_{\mathcal{D}}(-\infty, a_j+h-a_i]$ for $j = 1$ and $\overline{P_{\mathcal{D}}}^{i,j} = P_{\mathcal{D}}(a_j-h-a_i, \infty)$ for $j = N$. This is where it becomes very advantageous to have fixed $\{a_1, ..., a_N\}$: The $N \times K$-matrix $\overline{P_{\mathcal{D}}}$ does not depend on μ, and we can consider $\overline{P_{\mathcal{D}}} \cdot \bar{\mu}$ as a (linear and therefore) differentiable function of $\bar{\mu} \in \Delta^{N-1}$, where $\Delta^{N-1} := \{(x_1, ..., x_N) \in \mathbb{R}^N : \sum_{i=1}^{N} x_i = 1, x_i \geq 0\}$ denotes the probability simplex. If we now identify

$$\begin{pmatrix}\Pi_A(\mu * \mathcal{D})(\{a_1\})\\ \vdots\\ \Pi_A(\mu * \mathcal{D})(\{a_K\})\end{pmatrix}$$

with the measure $\Pi_A(\mu * \mathcal{D})$, we can, by extension, consider

$$\mathrm{W}_1(\overline{P_{\mathcal{D}}}\bar{\mu}, \hat{\pi}) := \mathrm{W}_1(\Pi_A(\mu * \mathcal{D}), \hat{\pi})$$

as a function of $\bar{\mu}$. We have reformulated the problem of finding

$$\operatorname*{argmin}_{\mu \in M_{A,V}} \mathrm{W}_1(\Pi_*^A(\mu * \mathcal{D}), \hat{\pi})$$

to finding

$$\operatorname*{argmin}_{\bar{\mu} \in \Delta^{N-1}} \mathrm{W}_1(\overline{P_{\mathcal{D}}}\bar{\mu}, \hat{\pi}).$$

It is straightforward to see that this is a convex problem: First of all, Δ^{N-1} is convex. By Proposition 1.1.7, $\bar{\mu} \mapsto \mathrm{W}_1(\bar{\mu}, \nu)$ is convex. The composition of a convex function with an affine function is still convex (see e.g. [1]), thus

$$\bar{\mu} \mapsto \mathrm{W}_1(\overline{P_{\mathcal{D}}}\bar{\mu}, \hat{\pi})$$

is convex.

Rigollet and Weed claim "subgradients [of this map] can be obtained by standard methods in computational optimal transport [5]", though we struggle to see what precisely is meant by that.
The approach we chose is the following: Instead of trying to solve the minimization problem

$$\operatorname*{argmin}_{\bar{\mu} \in \Delta^{N-1}} \mathrm{W}_1(\overline{P_{\mathcal{D}}}\bar{\mu}, \hat{\pi}),$$

we try to solve an entropically regularized approximation of that problem. The details for this approach are established by Peyré and Cuturi in [5].

For two discrete measures $\mu = \sum_{i=1}^N u_i \delta_{X_i}$ and $\nu = \sum_{j=1}^n v_j \delta_{Y_i}$, a coupling between μ and ν can be represented as a matrix $\gamma = (\gamma_{i,j})_{i=1,\dots,N; j=1,\dots,n}$. For such a matrix, define

$$H(\gamma) := -\sum_{i,j} \gamma_{i,j}(\log(\gamma_{i,j}) - 1).$$

We define the the *entropically regularized Wasserstein distance* between μ and ν as

$$\begin{aligned} L_p^\varepsilon(\mu, \nu) &:= \min_{\gamma \in \Gamma(\mu,\nu)} \int |x - y|^p \,\mathrm{d}\gamma - \varepsilon H(\gamma) \\ &= \min_{\gamma \in \Gamma(\mu,\nu)} \sum_{i,j} \gamma_{i,j} \left[|x_i - y_j|^p - \varepsilon(\log(\gamma_{i,j}) - 1) \right]. \end{aligned}$$

For the remainder of this segment, we will consider the support of μ and ν to be fixed, i.e., $(X_1, ..., X_N)$ and $(Y_1, ..., Y_n)$ are considered fixed and we consider $\mu = \mu^u$ and $\nu = \nu^u$ to be functions of the weight vectors $(u_1, ..., u_N) \in \Delta^{N-1}$ and $(v_1, ..., v_n) \in \Delta^{n-1}$. This in turn lets us consider $L_p^\varepsilon(\mu^u, \nu^v)$ as a function of u and v.

In [5, Proposition 4.1] it is shown that

$$L_p^\varepsilon(\mu^u,\nu^v) \overset{\varepsilon\to 0}{\to} \mathrm{W}_p^p(\mu^u,\nu^v),$$

and since $\Delta^{N-1}\times\Delta^{n-1}$ is compact, this convergence is uniform on that domain. In [5, Proposition 4.6], Peyré and Cuturi give a formula for the gradient of L_p^ε at (μ^u,ν^v) (i.e., a gradient with respect to u and v, since we consider the support of μ and ν fixed). In [5, Remark 4.20], they display an iterative algorithm to simultaneously calculate the value and the gradient of $L_p^\varepsilon(\cdot,\cdot)$ at a point (μ^u,ν^v), the *Sinkhorn algorithm.*
Putting $\nu^v := \hat\pi$, an application of the chain rule gives us

$$\nabla^T L_2\left(\Pi_A(\mu * \mathcal{D}),\hat\pi\right) = \left(\nabla^T L_2(\cdot,\hat\pi)\right)\Big|_{\overline{P_\mathcal{D}}\bar\mu} \cdot \overline{P_\mathcal{D}}.$$

Our practical approach for minimizing $L_1^\varepsilon(\overline{P_\mathcal{D}}\cdot,\hat\pi)$ is to apply a gradient descent approach leveraging the aforementioned formula. This can be summed up in the following algorithm:

```
Input:
Observations $Y = (Y_1,...,Y_n)$.
Noise distribution $\mathcal{D}$.

Calulate starting parameters:
Choose suitable domain $A = \{a_1,...,a_N\}$ (depends on $\mathcal{D}$).
Put $\bar\mu_c = (\frac{1}{K},...,\frac{1}{K})$.
Put $\hat\pi = (\frac{1}{n},...,\frac{1}{n})$.
Put $\gamma = 1$.
Calculate $\overline{P_\mathcal{D}}$.
Calculate $C = (C_{i,j})_{i=1,...,N;j=1,...,n}$ with $C_{i,j} = |X_i - Y_j|$.
Choose $\varepsilon$ = small.
Put $K = \exp(\varepsilon C)$ [exp component-wise, not matrix-exponential]

Repeat:
Do Sinkhorn algorithm at $(\overline{P_\mathcal{D}}\bar\mu_c,\nu)$ with matrix $K$.
Use output to update $L_1^\varepsilon(\overline{P_\mathcal{D}}\bar\mu_c,\hat\pi)$ and $\nabla L_1^\varepsilon(\cdot,\hat\pi)\Big|_{\overline{P_\mathcal{D}}\bar\mu_c}$.
Update $\gamma$.
Update $\bar\mu_l = \bar\mu_c$
Update $\nabla L_2(\cdot,\hat\pi)\Big|_{\overline{P_\mathcal{D}}\bar\mu_l} = \nabla L_2(\cdot,\hat\pi)\Big|_{\overline{P_\mathcal{D}}\bar\mu_c}$
Update $\bar\mu_c = \bar\mu_c - \gamma\left(\overline{P_\mathcal{D}}^T\nabla L_2(\cdot,\hat\pi)\Big|_{\overline{P_\mathcal{D}}\bar\mu_c}\right)$.
Project $\bar\mu_c$ back onto probability simplex.
Until cancelation criterion dependent on $L_1^\varepsilon(\overline{P_\mathcal{D}}\bar\mu_c,\hat\pi)$.

Output:
$\bar\mu_l$.
```

There are several common numerical ways to update γ, e.g. the Barzilai–Borwein

[2] method:

$$\gamma = \frac{\left| \left(\nabla^T L_1(\cdot, \hat{\pi}) \Big|_{\overline{P_{\mathcal{D}}}\bar{\mu}_c} - \nabla^T L_1(\cdot, \hat{\pi}) \Big|_{\overline{P_{\mathcal{D}}}\bar{\mu}_l} \right) \overline{P_{\mathcal{D}}} \cdot (\bar{\mu}_c - \bar{\mu}_l) \right|}{\left|\left| \left(\nabla^T L_1(\cdot, \hat{\pi}) \Big|_{\overline{P_{\mathcal{D}}}\bar{\mu}_c} - \nabla^T L_1(\cdot, \hat{\pi}) \Big|_{\overline{P_{\mathcal{D}}}\bar{\mu}_l} \right) \overline{P_{\mathcal{D}}} \right|\right|_2^2}.$$

This obviously only works from the second step onward, as one needs both $\bar{\mu}_c$ and $\bar{\mu}_l$ (indices c and l denote "current step" and "last step").

As a convergence criterion, one could choose that consecutive values of L_1^ε are sufficiently close, i.e.,

$$||L_1^\varepsilon(\overline{P_{\mathcal{D}}}\bar{\mu}_c, \hat{\pi}) - L_1^\varepsilon(\overline{P_{\mathcal{D}}}\bar{\mu}_l, \hat{\pi})|| \leq \text{tolerance}.$$

For two vectors a, b and a matrix K, the Sinkhorn algorithm in its simplest form, as described in [5, p. 63], reads

```
Input:
Vectors a, b.
Matrix K.
Regularization constant $\varepsilon$.

Start parameters:
$v = (1, .., 1)$, same length as $b$.

Repeat:
$u = \frac{a}{Kv}$ [division is component-wise]
$v = \frac{b}{K^T u}$
Until: $||diag(u)Kv - a||$ and $||diag(v)K^T u - b||$ are small.

Output:
Gradients $f = \varepsilon \log(u)$ and $g = \varepsilon \log(v)$.
Value $L_1^\varepsilon = < f, a > + < g, b > -\varepsilon u^T K v$.
```

We also implemented a variation of this algorithm from [5, Remark 4.21] that is better conditioned, but unfortunately also noticeably slower.

Due to constraints on calculation time, the gradients obtained via the Sinkhorn algorithm are generally not very exact. Therefore, $\bar{\mu}$ is generally not on the probability simplex anymore after being updated. We deal with this problem by projecting $\bar{\mu}$ back onto the probability simplex after every update. There are a few possible ways to do that, and we choose to do a least-squares projection on the probability simplex. To this end, we employ an implementation from Fritz Schelten's bachelor thesis [7] of an algorithm proposed by Laurent Condat [3].

Finally one has to recover a monotone function from $\bar{\mu}_l = (\bar{\mu}^1, ..., \bar{\mu}^K)$. As mentioned in Lemma 1.2.2, this can be achieved by putting

$$\hat{g}(X_i) = a_{k_i}^V, \quad k_i := \min\{k : \sum_{j=1}^{k} \bar{\mu}^j \geq \frac{i}{n}\}.$$

We conclude this section with a final note on the convergence of the gradient descent method. Common assumptions on the objective function that guarantee the convergence of the gradient descent method to a global minimum are convexity and Lipschitz continuity of the first derivative [4]. Clearly,

$$\sum_{i,j} \gamma_{i,j} \left[|x_i - y_j|^p - \varepsilon(\log(\gamma_{i,j}) - 1) \right]$$

is convex with respect to γ, since log is concave, which ensures that finding

$$\operatorname*{argmin}_{\gamma \in \Gamma(\mu,\nu)} \sum_{i,j} \gamma_{i,j} \left[|x_i - y_j|^p - \varepsilon(\log(\gamma_{i,j}) - 1) \right]$$

is a convex problem. In [5, Proposition 4.6.] it is stated, that

$$L_1^\varepsilon(\mu^u, \nu^v) = \min_{\gamma \in \Gamma(\mu,\nu)} \sum_{i,j} \gamma_{i,j} \left[|x_i - y_j|^p - \varepsilon(\log(\gamma_{i,j}) - 1) \right]$$

is also convex with respect to u and v. We are not sure whether its gradient is Lipschitz.

1.4 Heuristic estimation by value-matching

Using the representation of W_p distances via inverse distribution functions, we can express the problem of finding

$$\operatorname*{argmin}_{g \in \mathcal{F}_V} \mathrm{W}_2^2(\pi_g * \mathrm{Ber}, \hat{\pi})$$

in a very explicit manner. Consider the support vector $\mu^* \in \mathbb{R}^{2n}$ of a measure of the form $\pi_g * \mathrm{Ber}$: If $\mu \in \mathbb{R}^n$ is the support vector of π_g, μ^* can easily be computed from μ by concatenating $\mu + 1$ and $\mu - 1$ and sorting the entries of the resulting vector in an increasing order. Using Theorem 1.1.2, we can then calculate that it holds that

$$\mathrm{W}_p^p(\pi_g * \mathrm{Ber}, \hat{\pi}) = \sum_{k=1}^{n} |\pi_k - \mu_{2k-1}^*|^p + |\pi_k - \mu_{2k}^*|^p.$$

Thus, the problem of finding the argmin can be formulated as the following optimization problem:

For a given vector $\pi \in \mathbb{R}^n$ with increasing entries, find $\mu \in \mathbb{R}^n$ that minimizes

$$r(\mu) := \sum_{k=1}^{n} |\pi_k - \mu_{2k-1}^*|^p + |\pi_k - \mu_{2k}^*|^p,$$

where μ_k^* is the order statistic on $\{\mu_1 - 1, ..., \mu_n - 1, \mu_1 + 1, ..., \mu_n + 1\}$.

We suspect that the following might give an be an explicit solution to the aforementioned optimization problem:

Conjecture 1.4.1. Let S_n be the set of permutations of n elements. Let M be the set of all empirical measures on n or fewer points. It might hold that

$$\min_{\mu \in M} \mathrm{W}_p^p(\mu * \mathrm{Ber}, \pi) = \min_{\sigma \in S_n} \sum_{i=1}^{n} \left| \frac{|\pi_i - \pi_{\sigma(i)}| - 2}{2} \right|^p .$$

A minimizing μ can easily be recovered from a minimizing σ by putting $\mu_k = \frac{\pi_k + \pi_{\sigma(k)}}{2}$ for $k = 1, ..., n$. Computing a minimizing σ via brute force is computationally quite expensive, though, as S_n has $n!$ elements.

This conjecture is based on the following heuristic:
Write

$$\hat{\pi} = \sum_{i=1}^{n} \frac{1}{n} \delta_{\{Y_i\}} = \sum_{i=1}^{n} \frac{1}{2n} \delta_{\{Y_i\}} + \frac{1}{2n} \delta_{\{Y_i\}}$$

and consider this as an empirical measure on $2n$ points. Convolving point measures with Bernoulli noise gives us a measure of the form

$$\delta_{\{a\}} * \mathrm{Ber} = \frac{1}{2} \delta_{\{a-1\}} + \frac{1}{2} \delta_{\{a+1\}}.$$

If we have a measure of the from $\frac{1}{2n}\delta_{\{Y_1\}} + \frac{1}{2n}\delta_{\{Y_2\}}$ such that $|Y_1 - Y_2| \approx 2$, then putting a point measure right in the middle of Y_1 and Y_2 and convolving it with the Bernoulli distribution will generate a measure that is close in Wasserstein distance to $\frac{1}{2n}\delta_{\{Y_1\}} + \frac{1}{2n}\delta_{\{Y_2\}}$:

$$\mathrm{W}_p \left(\frac{1}{n} \delta_{\frac{Y_1 + Y_2}{2}} * \mathrm{Ber},\ \frac{1}{2n} \delta_{\{Y_1\}} + \frac{1}{2n} \delta_{\{Y_2\}} \right) \approx 0.$$

For a given

$$\hat{\pi} = \sum_{i=1}^{n} \frac{1}{2n} \delta_{\{Y_i\}} + \frac{1}{2n} \delta_{\{Y_i\}},$$

the idea is now to match the points $Y_1, ..., Y_n$ into pairs $(Y_1, Y_{\sigma(1)}), ..., (Y_{1n}, Y_{\sigma(n)})$ such that the distance of the points in each pair is as close to 2 as possible.

One could get the idea that a collection of such pairs fulfills a certain symmetry relation, namely that if the pair (Y_i, Y_j) is part of an optimal matching, then (Y_j, Y_i) is also part of the same matching. This is false, as the following remark shows.

Remark 1.4.1. Let $H := \{\sigma \in S_n | \sigma \text{ is representable by disjoint 1- and 2-cycles}\}$ be the set of permutations of n elements, which are involutions. It is insufficient to restrict the search for a minimizing σ to H, i.e. it holds that

$$\min_{\sigma \in H} \sum_{i=1}^{n} \left| \frac{\pi_i - \pi_{\sigma(i)} + 2}{2} \right|^p \neq \min_{\sigma \in H} \sum_{i=1}^{n} \left| \frac{\pi_i - \pi_{\sigma(i)} + 2}{2} \right|^p .$$

Proof. Consider

$$\hat{\pi} = \frac{1}{3} \delta_{\{0\}} + \frac{1}{3} \delta_{\{\frac{3}{2}\}} + \frac{1}{3} \delta_{\{\frac{5}{2}\}}.$$

An optimal pairing using only involutions is given by $(0, \frac{3}{2}), (\frac{3}{2}, 0), (\frac{5}{2}, \frac{5}{2})$. Putting $\mu_1 = \frac{1}{3}\delta_{\{\frac{3}{4}\}} + \frac{1}{3}\delta_{\{\frac{3}{4}\}} + \frac{1}{3}\delta_{\{\frac{5}{2}\}}$ gives us

$$\mathrm{W}_1(\mu_1 * \mathrm{Ber}, \hat{\pi}) = \frac{1}{3}\frac{1}{4} + \frac{1}{3}\frac{1}{4} + \frac{1}{3} \cdot 2 = \frac{10}{12}.$$

On optimal pairing using arbitrary permutations is given by $(0, \frac{3}{2}), (\frac{3}{2}, \frac{5}{2}), (\frac{5}{2}, 0)$. Putting $\mu_2 = \frac{1}{3}\delta_{\{\frac{3}{4}\}} + \frac{1}{3}\delta_{\{2\}} + \frac{1}{3}\delta_{\{\frac{5}{4}\}}$ gives us

$$\mathrm{W}_1(\mu_2 * \mathrm{Ber}, \hat{\pi}) = \frac{1}{3}\frac{1}{4} + \frac{1}{3}\frac{1}{4} + \frac{1}{3} \cdot 1 = \frac{6}{12},$$

which of course is smaller than $\frac{10}{12}$. □

For large n, we do not have a smart way of calculating an optimal permutation σ. Instead, though, we observed that an approximate version of this approach performs reasonably well:

Consider the vector $Y = (Y_1, ..., Y_n)$ which denotes a random permutation of the data $\{Y_1, ..., Y_n\}$. Put $A_1 := \{Y_1, ..., Y_n\}$ and put $Y_{j_1} = \operatorname{argmin}_{Y_j \in A_1} ||Y_1 - Y_j| - 2|$. Now inductively set $A_{i+1} = A_i \backslash \{Y_i\}$ and $Y_{j_i} := \operatorname{argmin}_{Y_j \in A_i} ||Y_1 - Y_j| - 2|$. We then put

$$\mu_{\text{match}} := \sum_{i=1}^{n} \frac{1}{n} \delta_{\frac{Y_i + Y_{j_i}}{2}}$$

and denote by $\tilde{g}$ the cumulative distribution function of μ_{match}, and the estimator

$$g_{\text{match}} := \max\{\min\{\tilde{g}, V\}, -V\}.$$

The computational complexity of this process is of order $O(n^2)$. Considering the simplicity of this approach, the performance of this estimator in our simulations was quite impressive. While for low n, the matching estimator was generally outperformed by the gradient descent approach, the results for high n were generally comparable in our simulations. Especially for high n, the matching approach method required significantly less computation time than the gradient descent approach.

Bibliography

[1] Daniel Fischer (https://math.stackexchange.com/users/83702/daniel-fischer). *How do I prove that the composition of an affine function preserves convexity?* Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/529052 (version: 2013-10-16). eprint: `https://math.stackexchange.com/q/529052`. URL: `https://math.stackexchange.com/q/529052`.

[2] Jonathan Barzilai and Jonathan M Borwein. “Two-point step size gradient methods”. In: *IMA journal of numerical analysis* 8.1 (1988), pp. 141–148.

[3] Laurent Condat. “Fast projection onto the simplex and the l1 ball”. In: *Mathematical Programming* 158.1-2 (2016), pp. 575–585.

[4] Robert M. Gower. *Convergence Theorems for Gradient Descent.* 2018. URL: `https://perso.telecom-paristech.fr/rgower/pdf/M2_statistique_optimisation/grad_conv.pdf`.

[5] Gabriel Peyré, Marco Cuturi, et al. “Computational optimal transport”. In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[6] Philippe Rigollet and Jonathan Weed. *Uncoupled isotonic regression via minimum Wasserstein deconvolution.* 2018. eprint: `arXiv:1806.10648`.

[7] Fritz Schelten. “Efficient ℓ_1-Regularization in High-Dimensional Spaces”. B.S. Thesis. Heidelberg: University of Heidelberg, 2019.