Entropic optimal transport is maximum-likelihood deconvolution

Citation: Rigollet, Philippe and Jonathan Weed. "Entropic optimal transport is maximum-likelihood deconvolution." Comptes Rendus Mathematique 356, 11-12 (November 2018): 1228-1235 © 2018 Académie des sciences

As Published: http://dx.doi.org/10.1016/j.crma.2018.10.010

Publisher: Elsevier BV

Persistent URL: https://hdl.handle.net/1721.1/126692

Version: Original manuscript: author's manuscript prior to formal peer review

Entropic optimal transport is maximum-likelihood deconvolution

PHILIPPE RIGOLLET[*] AND JONATHAN WEED[†]

Massachusetts Institute of Technology

Abstract. We give a statistical interpretation of entropic optimal transport by showing that performing maximum-likelihood estimation for Gaussian deconvolution corresponds to calculating a projection with respect to the entropic optimal transport distance. This structural result gives theoretical support for the wide adoption of these tools in the machine learning community.

Key words and phrases: Entropy, optimal transport, deconvolution.

1. INTRODUCTION

Optimal transport is a fundamental notion with arising in several branches of mathematics, including probability, analysis and statistics. More recently, it has found new applications in computational domains such as machine learning and image processing [ACB17, BvdPPH11, CFTR17, RTG00, SDGP+15, JSCG16, MJ15, RCP16]. This newfound utility was largely fueled by algorithmic advances allowing optimal transport distances to be computed quickly between large scale discrete distributions [PC17]. At the heart of these algorithmic techniques is the idea of entropic penalization, which has been leveraged to obtain near-linear-time approximation schemes for optimal transport distances [Wil69, Cut13, AWR17]. Hereafter, we refer to this technique as *entropic optimal transport*. This line of well established computational research stands in sharp contrast with our *statistical* understanding of regularization for optimal transport, which is still in its infancy [FHN+18]. Entropic regularization has been shown to play a central statistical role in a variety of problems related to model selection [JRT08, Rig12, RT11, RT12, DT08, DT12b, DT12a] but our knowledge of its effect on optimal transport is currently limited to experimental evidence [Cut13, PC17] without theoretical support.

In this note, we give a statistical interpretation of entropic optimal transport, showing that under some modeling assumptions, it corresponds to the objective function in maximum-likelihood estimation for deconvolution problems involving additive Gaussian noise. This interpretation provides a first indication that optimal transport problems where data is subject to Gaussian observation error should be handled with entropic regularization. Moreover, our results indicate

[*] This work was supported by NSF DMS-1712596, NSF TRIPODS-1740751, ONR N00014-17-1-2147, the Chan Zuckerberg Initiative DAF 2018-182642, and the MIT Skoltech Seed Fund.

[†] This work was supported in part by NSF Graduate Research Fellowship DGE-1122374.

that in the same context, a relaxed version of optimal transport should be preferred, as it is equivalent to maximum-likelihood estimation even in absence of said modeling assumptions.

2. ENTROPIC OPTIMAL TRANSPORT

Throughout we denote by γ a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ and by $\|\cdot\|$ the Euclidean norm over $\mathbb{R}^d$. Given such a measure, we denote by $\pi_X\gamma$ and $\pi_Y\gamma$ the two measures on $\mathbb{R}^d$ obtained by projecting onto the first and second component, respectively. Given probability measures μ and ν on $\mathbb{R}^d$, define

$$\mathcal{M}(\mu,\nu) := \{\gamma : \pi_X\gamma = \mu, \pi_Y\gamma = \nu\} \quad \text{and} \quad \mathcal{M}(\mu) := \{\gamma : \pi_Y\gamma = \mu\}\,.$$

We also recall the definition of Kullback-Leibler (KL) divergence between probability measures μ and ν:

$$D(\mu\|\,\nu) = \begin{cases} \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu, & \text{if } \mu \ll \nu \\ \infty, & \text{otherwise.} \end{cases}$$

Definition 1. *The* entropic optimal transport distance *between μ and ν is*

$$W_{\sigma^2}(\mu,\nu) := \min_{\gamma\in\mathcal{M}(\mu,\nu)} \int \frac{1}{2}\|x-y\|^2\,\mathrm{d}\gamma(x,y) + \sigma^2 I(\gamma)\,, \tag{1}$$

where $I(\gamma)$ is the mutual information *defined by*

$$I(\gamma) := D(\gamma\,\|\,\pi_X\gamma\otimes\pi_Y\gamma)\,.$$

Note that when $\sigma = 0$, this corresponds the squared 2-Wasserstein distance between probability measures over $\mathbb{R}^d$ [San15]. When $\sigma > 0$, W_{σ^2} no longer satisfies the axioms of a (squared) distance, but it still possesses useful distance-like properties [Cut13]. We employ the term “distance” for all values of σ for terminological consistency. When μ and ν are discrete measures, the minimizer of (1) is also discrete and agrees with the minimizer of

$$\min_{\gamma\in\mathcal{M}(\mu,\nu)} \int \frac{1}{2}\|x-y\|^2\,\mathrm{d}\gamma(x,y) - \sigma^2 H(\gamma)\,, \tag{2}$$

where H is the standard Shannon entropy:

$$H(\gamma) := \sum_{ij} \gamma_{ij}\log\frac{1}{\gamma_{ij}}\,,$$

where $\gamma_{ij} := \gamma(x_i, y_j)$ for $(x_i, y_j) \in \operatorname{supp}(\gamma)$. Note that definition (2) is the one proposed by [Cut13] for discrete measures, whereas definition (1) corresponds to the appropriate generalization studied in [GCPB16] for measures that are not necessarily discrete. It is not hard to check that $W_{\sigma^2}(\mu,\nu) < \infty$ for any pair (μ,ν) possessing finite second moments, since the independent coupling $\mu\otimes\nu$ in the minimization that appears in (1) leads to a finite objective value.

A maximum-likelihood interpretation of entropic optimal transport is already known in the context of a large-deviation principle for Brownian motion [Léo14].

In this context, given two distributions μ and ν (viewed as the positions of particles at times $t = 0$ and $t = 1$), Schrödinger [Sch32] gave a heuristic argument motivated by statistical physics establishing that the law of independent particles undergoing Brownian motion conditioned on having initial and final distributions μ and ν respectively is induced by the solution to the optimal transport problem with entropic regularization [Léo14]. While suggestive, this interpretation does not hold any immediate implications for estimation problems where only data is available rather than distributions μ and ν. Below, we introduce the classical deconvolution model for corrupted observations and show that in this context, entropic optimal transport is precisely the maximum-likelihood estimator.

3. DECONVOLUTION

Let $\mathcal{P}$ be a given family of probability distributions over $\mathbb{R}^d$ with finite second moments and let P^* be an unknown distribution, also with finite second moment. The deconvolution problems consists in estimating P^* on the basis of corrupted observations $Y_1, \dots, Y_n$, where

$$Y_i = X_i + Z_i\,, \qquad i = 1, \dots, n\,, \tag{3}$$

and the errors $Z_1, \dots, Z_n$ are independent of $X_1, \dots, X_n$. For identifiability purposes, the random variables $\{Z_i\}$ are assumed to be independent copies of a random variable Z with known distribution: $Z \sim \mathcal{N}(0, \sigma^2)$ where the variance σ^2 is known.

In this context, the distribution of Y_i admits a density $\varphi_\sigma \star \mathrm{d}P^*$ with respect to λ where, for any $P \in \mathcal{P}$, we define

$$\varphi_\sigma \star \mathrm{d}P(y) = \int \varphi_\sigma(y - x)\,\mathrm{d}P(x) \tag{4}$$

and φ_σ denotes the density of $Z \sim \mathcal{N}(0, \sigma^2)$. Under these assumptions, we call (3) the *Gaussian deconvolution model.*

Deconvolution is a classical question of nonparametric statistics [CH88, Fan91, CCDM11] and is core to mixture models [Lin95] as well as statistical models with measurement errors [CRSC06]. As such, it has received significant attention from the statistics literature. More recently, it was shown that deconvolution has strong methodological and mathematical connections to optimal transport in the context of a problem known as uncoupled regression [RW18].

A natural candidate to estimate P^* is the maximum-likelihood estimator (MLE) $\hat{P}$ defined by

$$\hat{P} = \operatorname*{argmax}_{P \in \mathcal{P}} \sum_{i=1}^{n} \log \varphi_\sigma \star \mathrm{d}P(Y_i)\,, \tag{5}$$

The statistical properties of the MLE are well known and have been established under general conditions on the class $\mathcal{P}$ [BD06, GN16]. In the next section, we show that entropic optimal transport is in fact implementing $\hat{P}$.

4. ENTROPIC OPTIMAL TRANSPORT IS MAXIMUM-LIKELIHOOD DECONVOLUTION

In this section, we adopt the Gaussian deconvolution model (3) of the previous section. The extension to other distributions for the corruption errors $\{Z_i\}$ is postponed to Section 5.

Our main result involves families of distributions satisfying a particular closure condition.

DEFINITION 2. *A class $\mathcal{P}$ of probability measures is said to be* closed under domination *if $Q \ll P$ for some $P \in \mathcal{P}$ implies that $Q \in \mathcal{P}$.*

Many families are closed under domination. For example the class of all measures, the class of all measures absolutely continuous with respect to some reference measure σ, the class of discrete measures, and the set of measures whose support is finite or contains at most k points all possess this property. A class $\mathcal{P}$ of probability measures may always be augmented to be closed under domination by adding to it the set of probability measures $\bigcup_{P\in\mathcal{P}}\{Q : Q \ll P\}$. The extension to families not closed under domination is considered in Section 5.

We are now in a position to state our main result: a structural representation of the maximum-likelihood estimator $\hat{P}$ in terms of entropic optimal transport.

THEOREM 1. *Let $\mathcal{P}$ be a class of probability measures that is closed under domination and assume the Gaussian deconvolution model (3). Then the maximum-likelihood estimator $\hat{P}$ over $\mathcal{P}$ defined in (5) satisfies:*

$$\hat{P} = \operatorname*{argmin}_{P\in\mathcal{P}} W_{\sigma^2}\Big(P, \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}\Big)\,.$$

In other words, the maximum-likelihood estimator $\hat{P}$ is the projection of the empirical measure $\frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$ onto $\mathcal{P}$ with respect to the entropic optimal transport distance W_{σ^2}.

The projection estimator $\operatorname{argmin}_{P\in\mathcal{P}} W_{\sigma^2}\Big(P, \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}\Big)$ has been employed in the machine learning community [MMC16, GPC18] as a smoothed version of a minimum Kantorovich distance estimator [BBR06] more suitable for optimization. Theorem 1 shows that this estimator has a statistical interpretation in addition to its computational benefits.

As noted above, in the special case when $\sigma^2 = 0$, the quantity W_{σ^2} reduces to the squared 2-Wasserstein distance W. In the context of Gaussian mixture models, when $\mathcal{P}$ is the class of probability distributions supported on at most k points, solving $\operatorname{argmin}_{P\in\mathcal{P}} W\Big(P, \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}\Big)$ corresponds to performing a "hard" clustering of the data by minimizing the k-means objective. It is known, however, that hard k-means clustering does not lead to consistent estimation of the centroids in a mixture of Gaussians model, whereas consistent estimation *can* be achieved with the MLE, which induces a relaxed "soft" clustering [KMN97]. Theorem 1 implies that replacing W by W_{σ^2} precisely corresponds to this relaxation.

PROOF. Write for simplicity $\ell_P := \log \varphi_\sigma \star \mathrm{d}P$. By (4), we have

$$\ell_P(y_i) = C + \log \int \exp\big(-\frac{1}{2\sigma^2}\|x - y_i\|^2\big)\,\mathrm{d}P(x)\,,$$

where C is a constant not depending on y_i or P. The Gibbs variational principle [Cat04, Equation (5.2.1)] then implies that

$$\ell_P(y_i) = C - \frac{1}{\sigma^2}\min_{Q_i}\Big\{\frac{1}{2}\mathop{\mathbb{E}}_{x\sim Q_i}\|x - y_i\|^2 + \sigma^2 D(Q_i\,\|\,P)\Big\}\,,$$

where the minimization is taken over probability measures on $\mathbb{R}^d$. Then, by definition of the MLE, we have

$$\hat{P} = \operatorname*{argmin}_{P \in \mathcal{P}} \min_{Q_1,\dots,Q_n} \frac{1}{n}\sum_{i=1}^n \left[\frac{1}{2}\mathop{\mathbb{E}}_{x\sim Q_i}\|x-y_i\|^2 + \sigma^2 D(Q_i \,\|\, P)\right].$$

Next, given any set of n distributions $\{Q_1,\dots,Q_n\}$ on $\mathbb{R}^d$, we can define the joint probability measure $\bar{\gamma}$ on $\mathbb{R}^d \times \{y_1,\dots,y_n\}$ by

$$\bar{\gamma} := \frac{1}{n}\sum_{i=1}^n Q_i \otimes \delta_{y_i}\,.$$

Note that $\pi_Y\gamma = U$, the uniform distribution on $\{y_1,\dots,y_n\}$. Conversely, for any joint probability measure $\bar{\gamma}$ on $\mathbb{R}^d\times\mathbb{R}^d$ satisfying $\pi_Y\gamma = U$, we can decompose γ as $\frac{1}{n}\sum_{i=1}^n Q_i\otimes\delta_{y_i}$ for some $Q_1,\dots,Q_n$. This bijection between sets of n probability measures $\{Q_1,\dots,Q_n\}$ on $\mathbb{R}^d$ and joint measures $\bar{\gamma}\in\mathcal{M}(U)$ satisfies the equality

$$\frac{1}{n}\sum_{i=1}^n \left[\frac{1}{2}\mathop{\mathbb{E}}_{x\sim Q_i}\|x-y_i\|^2 + \sigma^2 D(Q_i \,\|\, P)\right] = \frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim\bar{\gamma}}\|x-y\|^2 + \sigma^2 D(\bar{\gamma}\,\|\,P\otimes U)\,.$$

We can therefore leverage this bijection to rewrite $\hat{P}$ as

$$\hat{P} = \operatorname*{argmin}_{P\in\mathcal{P}} \min_{\gamma\in\mathcal{M}(U)} \left\{\frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim\gamma}\|x-y\|^2 + \sigma^2 D(\gamma\,\|\,P\otimes U)\right\}.$$

By Lemma 1,

$$D(\gamma\,\|\,P\otimes U) = I(\gamma) + D(\pi_X\gamma\|P)\,,$$

where we have used that $D(\pi_Y\gamma\|U)=0$ for any $\gamma\in\mathcal{M}(U)$. We obtain

$$\hat{P} = \operatorname*{argmin}_{P\in\mathcal{P}} V(P)\,,$$

where

$$V(P) := \min_{\gamma\in\mathcal{M}(U)}\left\{\frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim\gamma}\|x-y\|^2 + \sigma^2 I(\gamma) + \sigma^2 D(\pi_X\gamma\|P)\right\} \tag{6}$$

Next, for any $P\in\mathcal{P}$, denote by γ_P the coupling that achieves the minimum in the definition (1) of $W_{\sigma^2}(P,U)$ and observe that since $D(\pi_X\gamma_P\|P)=0$, we get

$$V(P) \le \frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim\gamma_P}\|x-y\|^2 + \sigma^2 I(\gamma_P) = W_{\sigma^2}(P,U) < \infty\,,$$

since P has finite second moment. We now show that the functions $P\mapsto V(P)$ and $P\mapsto W_{\sigma^2}(P,U)$ achieve their minimum over $\mathcal{P}$ at the same P. To that end, observe that since $V(P)<\infty$, the minimum in the definition (6) of V may be restricted to couplings γ such that $\pi_X\gamma \ll P$, since otherwise $D(\pi_X\gamma\|P)$ is infinite. Thus,

$$
\begin{aligned}
(7)\qquad \min_{P\in\mathcal{P}} V(P) &= \min_{P\in\mathcal{P}} \min_{\substack{\gamma\in\mathcal{M}(U)\\ \pi_X\gamma\ll P}} \left\{\frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim\gamma}\|x-y\|^2 + \sigma^2 I(\gamma) + \sigma^2 D(\pi_X\gamma\|P)\right\}\\
&\geq \min_{P\in\mathcal{P}} \min_{\substack{\gamma\in\mathcal{M}(U)\\ \pi_X\gamma\ll P}} \left\{\frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim\gamma}\|x-y\|^2 + \sigma^2 I(\gamma)\right\}\\
&= \min_{P\in\mathcal{P}} \min_{\gamma\in\mathcal{M}(P,U)} \left\{\frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim\gamma}\|x-y\|^2 + \sigma^2 I(\gamma)\right\}\\
&= \min_{P\in\mathcal{P}} W_{\sigma^2}(P,U)\,,
\end{aligned}
$$

where in the inequality we used that $D(\pi_X\gamma\|P)\geq 0$ and in the second equality we used that $\mathcal{P}$ is closed under domination. In light of the previous two displays, we conclude that $\hat{P} = \operatorname{argmin}_{P\in\mathcal{P}} W_{\sigma^2}\left(P, \frac{1}{n}\sum_{i=1}^n \delta_{y_i}\right)$, as claimed.

□

5. EXTENSIONS

5.1 Relaxed transport

Traditional parametric classes of distributions $\mathcal{P}$ are often not closed under domination. For example, the one-dimensional scale/location family with template density φ with respect to the Lebesgue measure Leb on $\mathbb{R}$ is defined by

$$
\mathcal{P} = \left\{P \,:\, \frac{\mathrm{d}P}{\mathrm{dLeb}}(\cdot) = \frac{1}{\tau}\varphi\big(\frac{\cdot-\mu}{\tau}\big)\,, \mu\in\mathbb{R}, \tau>0\right\}.
$$

Clearly $\mathcal{P}$ is not closed under domination. In such cases, Theorem 1 can fail to hold as illustrated by Proposition 1 below. However, it follows from (7) in the proof of Theorem 1 that the following representation for the MLE *always* holds:

$$
\hat{P} = \operatorname*{argmin}_{P\in\mathcal{P}} W^{\mathsf{rel}}_{\sigma^2}\big(P, \frac{1}{n}\sum_{i=1}^n \delta_{y_i}\big)\,,
$$

where $W^{\mathsf{rel}}_{\sigma^2}$ denotes the *relaxed entropic optimal transport* distance defined for any probability measures μ,ν by

$$
W^{\mathsf{rel}}_{\sigma^2}(\mu,\nu) = \min_{\substack{\gamma\in\mathcal{M}(\nu)\\ \pi_X\gamma\ll\mu}} \int \frac{1}{2}\|x-y\|^2\,\mathrm{d}\gamma(x,y) + \sigma^2\big[I(\gamma) + D(\pi_X\gamma\,\|\,\mu)\big]\,.
$$

This result indicates that it may be preferable to use relaxed transport in statistical contexts. Relaxing the marginal constraints in the optimal transport problem is an idea which has attracted significant recent interest [FZM$^+$15, LMS18, CPSV16] after it was first formally proposed in [Ben03] under the name "unbalanced transport" to generalize optimal transport to apply to nonnegative measures with different total mass. Relaxed optimal transport has since been used to improve robustness to sampling noise in statistical applications [SST$^+$17].

We now exhibit a simple example of a class $\mathcal{P}$ that is not closed under domination and for which Theorem 1 fails to hold. For any $\sigma > 0$, let $\mathcal{P} = \{P_1, P_2\}$ where P_1 and P_2 are two probability measures on the real line defined respectively by

$$P_1 := \frac{1}{2}(\delta_0 + \delta_{4\sigma}) \quad \text{and} \quad P_2 := \frac{1}{2}(\delta_{2\sigma} + \delta_{6\sigma})$$

Let $X \sim P_1$ and let $Y = X + Z$ where $Z \sim \mathcal{N}(0, \sigma^2)$ is independent of X.

Proposition 1. *With probability at least* .15*, we have*

$$P_1 = \operatorname*{argmin}_{P \in \mathcal{P}} W_{\sigma^2}(P, \delta_Y)\,, \quad \textit{and} \quad P_2 = \operatorname*{argmax}_{P \in \mathcal{P}} \log \varphi_\sigma \star \mathrm{d}P(Y)\,.$$

In other words, the maximum-likelihood estimator and the projection with respect to the entropic optimal transport distance do not agree.

Proof. By rescaling, it suffices to consider the case $\sigma^2 = 1$. For each $P \in \mathcal{P}$, the set $\mathcal{M}(P, \delta_Y)$ contains only the independent coupling $P \otimes \delta_Y$, for which the mutual information vanishes. Therefore,

$$W_{\sigma^2}(P_1, \delta_Y) = W_0(P_1, \delta_Y) = \frac{1}{4}(Y^2 + (Y-4)^2) \qquad W_{\sigma^2}(P_2, \delta_Y) = \frac{1}{4}((Y-2)^2 + (Y-6)^2)\,,$$

so that P_1 is the unique minimizer over $\mathcal{P}$ of $W_{\sigma^2}(P, \delta_Y)$ as long as $Y < 3$.

On the other hand, P_2 is the unique maximum-likelihood estimator if

$$e^{-Y^2/2} + e^{-(Y-4)^2/2} < e^{-(Y-2)^2/2} + e^{-(Y-6)^2/2}\,,$$

and it can be checked that this condition holds on the interval $[1.01, 3)$.

Therefore, if $Y \in [1.01, 3)$, then the claimed situation occurs. To conclude the proof, it suffices to observe that $P_1(1.01 \le Y < 3) \ge .5\mathbb{P}(1.01 \le |Z| \le 2.99) \ge .15$. □

5.2 General noise distribution

In the is section, we raise the question of non-Gaussian deconvolution that arises when the errors $\{Z_i\}$ are not Gaussian. It turns out that a simple modification of our argument can be made to accommodate any noise distribution that admits a density f with respect to the Lebesgue measure on $\mathbb{R}^d$. In this context, the MLE takes the form

$$\hat{P} = \operatorname*{argmax}_{P \in \mathcal{P}} \sum_{i=1}^{n} \log f \star \mathrm{d}P(Y_i) \tag{8}$$

The use of the squared Euclidean norm $\frac{1}{2\sigma^2}\|x - y\|^2$ in the objective (1) is tailored to Gaussian errors and may be replaced with $-\log f(x-y)$. After rescaling by σ^2, we define

$$W_f(\mu, \nu) := \min_{\gamma \in \mathcal{M}(\mu, \nu)} -\int \log f(x-y)\,\mathrm{d}\gamma(x,y) + I(\gamma)\,.$$

We assume in what follows that $\log f(\cdot - y) \in L_1(P)$ for all $y \in \mathbb{R}^d$ and $P \in \mathcal{P}$. The following proposition is stated without proof as it follows from exactly the same arguments as Theorem 1.

Proposition 2. *Let $\mathcal{P}$ be a class of probability measures that is closed under domination and assume the deconvolution model* (3) *where Z_i has density f with respect to the Lebesgue measure. Then the maximum-likelihood estimator $\hat{P}$ over $\mathcal{P}$ defined in* (8) *satisfies:*

$$\hat{P} = \operatorname*{argmin}_{P \in \mathcal{P}} W_f\Big(P, \frac{1}{n}\sum_{i=1}^n \delta_{y_i}\Big)\,.$$

In other words, the maximum-likelihood estimator $\hat{P}$ is the projection of the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ onto $\mathcal{P}$ with respect to the entropic optimal transport distance W_f.

In the case where $f(z) \propto \exp(-\|z\|_p^p)$, the cost $-\log f(x-y)$ corresponds to the ℓ_p metric arising in the definition of the p-Wasserstein distance. Another intriguing example is the cost $-\log\cos^2(\|x-y\| \wedge \pi/2)$, which appears in the definition of the Wasserstein-Fisher-Rao [CPSV16] or Hellinger-Kantorovich [LMS18] distance between positive measures. These formulations differ from ours in that they consider a version of relaxed transport which allows for misspecification of both marginals; nevertheless, our analysis suggests that inference involving these distances is likely to be robust to convolutional noise Z supported on the Euclidean ball of radius $\pi/2$ around the origin with density

$$f(z) \propto \cos^2(\|z\|)\mathbb{1}\big(\|z\| \leq \frac{\pi}{2}\big)\,.$$

The Wasserstein-Fisher-Rao/Hellinger-Kantorovich distance is motivated by a dynamic formulation of unbalanced transport and it is unclear whether the above noise distribution plays a special role in the context of deconvolution.

6. ADDITIONAL LEMMAS

Lemma 1. *Let γ be a measure on $\mathbb{R}^d \times \mathbb{R}^d$, and let α and β be probability measures on $\mathbb{R}^d$. Then*

$$D(\gamma \,\|\, \alpha \otimes \beta) = I(\gamma) + D(\pi_X\gamma \,\|\, \alpha) + D(\pi_Y\gamma \,\|\, \beta)\,.$$

Proof. We assume $\gamma \ll \pi_X\gamma \otimes \pi_Y\gamma \ll \alpha \otimes \beta$, since otherwise both sides are infinite. Under this condition, we have

$$\begin{aligned}
D(\gamma \,\|\, \alpha \otimes \beta) &= \int \log \frac{\mathrm{d}\gamma}{\mathrm{d}\alpha \mathrm{d}\beta}(x,y)\,\mathrm{d}\gamma(x,y) \\
&= \int \log \frac{\mathrm{d}\gamma}{\mathrm{d}\pi_X\gamma \mathrm{d}\pi_Y\gamma}(x,y)\,\mathrm{d}\gamma(x,y) + \int \log \frac{\mathrm{d}\pi_X\gamma \mathrm{d}\pi_Y\gamma}{\mathrm{d}\alpha\mathrm{d}\beta}(x,y)\,\mathrm{d}\gamma(x,y) \\
&= I(\gamma) + \int \log \frac{\mathrm{d}\pi_X\gamma}{\mathrm{d}\alpha}(x,y)\,\mathrm{d}\gamma(x,y) + \int \log \frac{\mathrm{d}\pi_Y\gamma}{\mathrm{d}\beta}(x,y)\,\mathrm{d}\gamma(x,y) \\
&= I(\gamma) + \int \log \frac{\mathrm{d}\pi_X\gamma}{\mathrm{d}\alpha}(x)\,\mathrm{d}\pi_X\gamma(x) + \int \log \frac{\mathrm{d}\pi_Y\gamma}{\mathrm{d}\beta}(y)\,\mathrm{d}\pi_Y\gamma(y) \\
&= I(\gamma) + D(\pi_X\gamma \,\|\, \alpha) + D(\pi_Y\gamma \,\|\, \beta)\,.
\end{aligned}$$

□

REFERENCES

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[AWR17] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1961–1971, 2017.

[BBR06] F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statist. Probab. Lett.*, 76(12):1298–1302, 2006.

[BD06] P. J. Bickel and K. A. Doksum. *Mathematical statistics: basic ideas and selected topics*, volume 1. Updated printing. Prentice-Hall, 2 edition, 2006.

[Ben03] J.-D. Benamou. Numerical resolution of an “unbalanced” mass transport problem. *M2AN Math. Model. Numer. Anal.*, 37(5):851–868, 2003.

[BvdPPH11] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Trans. Graph.*, 30(6):158:1–158:12, 2011.

[Cat04] O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[CCDM11] C. Caillerie, F. Chazal, J. Dedecker, and B. Michel. Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Stat.*, 5:1394–1423, 2011.

[CFTR17] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865, 2017.

[CH88] R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

[CPSV16] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016.

[CRSC06] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement error in nonlinear models*, volume 105 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2006. A modern perspective.

[Cut13] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2292–2300, 2013.

[DT08] A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1):39–61, 2008.

[DT12a] A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012.

[DT12b] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci. (to appear). arXiv:0903.1223*, 2012.

[Fan91] J. Fan. On the estimation of quadratic functionals. *Ann. Statist.*, 19(3):1273–1294, 1991.

[FHN$^+$18] A. Forrow, J.-C. Hütter, M. Nitzan, G. Schiebinger, P. Rigollet, and J. Weed. Statistical Optimal Transport via Geodesic Hubs. *ArXiv:1806.07348*, 2018.

[FZM$^+$15] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. A. Poggio. Learning with a Wasserstein loss. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2053–2061, 2015.

[GCPB16] A. Genevay, M. Cuturi, G. Peyré, and F. R. Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3432–3440, 2016.

[GN16] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016.

[GPC18] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In A. J. Storkey and F. Pérez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 2018.

[JRT08] A. Juditsky, P. Rigollet, and A. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008.

[JSCG16] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 181–189, 2016.

[KMN97] M. J. Kearns, Y. Mansour, and A. Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In D. Geiger and P. P. Shenoy, editors, *UAI '97: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Brown University, Providence, Rhode Island, USA, August 1-3,* *1997*, pages 282–293. Morgan Kaufmann, 1997.

[Léo14] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.*, 34(4):1533–1574, 2014.

[Lin95] B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume Volume 5 of *Regional Conference Series in Probability and Statistics.* Institute of Mathematical Statistics and American Statistical Association, Haywood CA and Alexandria VA, 1995.

[LMS18] M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Invent. Math.*, 211(3):969–1117, 2018.

[MJ15] J. Mueller and T. S. Jaakkola. Principal differences analysis: Interpretable characterization of differences between distributions. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1702–1710, 2015.

[MMC16] G. Montavon, K. Müller, and M. Cuturi. Wasserstein training of restricted boltzmann machines. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3711–3719, 2016.

[PC17] G. Peyré and M. Cuturi. Computational optimal transport. Technical report, 2017.

[RCP16] A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed wasserstein loss. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 630–638. JMLR.org, 2016.

[Rig12] P. Rigollet. Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 2012.

[RT11] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011.

[RT12] P. Rigollet and A. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

[RTG00] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[RW18] P. Rigollet and J. Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *arXiv:1806.10648*, 06 2018.

[San15] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

[Sch32] E. Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Ann. Inst. H. Poincaré*, 2(4):269–310, 1932.

[SDGP$^+$15] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[SST$^+$17] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, S. Liu, S. Lin, P. Berube, L. Lee, et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *bioRxiv*, page 191056, 2017.

[Wil69] A. G. Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, 3(1):108–126, 1969.

Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(rigollet@math.mit.edu)

Jonathan Weed
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(jweed@mit.edu)