An Alternative to the Log-Likelihood with Entropic Optimal Transport

Candidate Number: 1073087

University of Oxford

A dissertation submitted for the degree of

MSc in Mathematical Sciences - OMMS

Trinity 2023

Word count: 7117 (TeXcount)

Abstract

One of the most important question in statistics is parameter estimation. Although minimizing the likelihood loss l is the de-facto general purpose method, being the most efficient in the classical setting (achieving the Cramer-Rao bound etc.), statistical methods are in the perpetual need of being adapted and tailored to new data regime, which might necessitate specific properties (robustness, expressivity, computational efficiency etc.). In this essay, we consider the entropic optimal transport (EOT) loss $L(\theta)$ and its associated estimator (EOTE). This method has been popularized quite recently, with the advent of new computational methods in the field of optimal transport (Sinkhorn algorithm [PC19a] etc.). In models admitting an additive noise structure (such as Gaussian Mixture Models), it has been shown that both estimators recover the true parameter (([Men+20], [RW18]), but with L being a better optimization objective than l, avoiding bad local optima and with faster convergence. An appealing property of L, over the classical method, is that it seems to be naturally more robust. This has been described empirically by [Men+20]. What is more, a *semi-dual* formulation of EOT loss can be understood as an 'adversarial' estimator, thus improving confidence in the robustness claims. In this essay, we work toward a theoretical framework to justify these potential gains, backed with experimental results. In particular, we lead a sensitivity analysis on model misspecification for mixture models, which is then applied to a simple Gaussian Mixture Model with two symmetric component. Finally, we use semiparametric theory to obtain influence functions for the semi-dual of the EOT loss, for possible further study.

Contents

Introduction

The tools of optimal transport have recently found many applications in areas such as computer vision, statistics or machine learning, following breakthrough work of in early 2010 [Cut13]. In particular, entropic regularization enables efficient computations with tractable convergence (see the thorough work of [PC19a]), as exemplified by Sinkhorn's algorithm, that can be easily implemented and adapted to statistical procedures such as the EM algorithm [Men+20], for estimation purposes.

Aim of the paper

In this paper we focus on the associated Entropic Optimal Transport Estimator (EOTE). This estimator essentially satisfies a minimal distance/projection type condition. While the main focus in the literature has been on the computation side, and on the properties of the associated distances defined between measures (such as Wasserstein distances), the statistical properties of this estimator in itself, in the asymtotic/non-asymptotic regime or in misspecified models, are neither well understood nor sufficiently researched. The objective here is to close this gap, while leading a comparison with the reference Maximum Likelihood Estimator (MLE). We will assume familiarity with parametric estimation theory (else, see [Del]).

Outline and Contributions

Based on the recent work of [Men+20] and [RW18], it is proved that in some settings, both estimators are comparable and recover the same optimal parameter. The models introduced in each respective work are:

- (Q1) Model the observations Y by the joint probability Q^θ of (X, Y) where X is a latent variable which distribution μ_X is known:

  $$dQ^\theta(x, y) = e^{-g^\theta(x,y)} d\mu_X(x) d\nu(y) = q^\theta(x, y) d\mu_X(x) d\nu(y).$$

  Then it is proved that the population/empirical EOT loss always dominates the log-likelihood loss. Moreover, when the model is well-specified, both losses are minimized in the population regime at the true parameter: $l(\theta^*) = L(\theta^*)$.

- (Q2) Model the observations Y as follows:

  $$Y = X_\theta + Z_{\sigma^2},$$

  where $Z_{\sigma^2} \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise, and $X_\theta \sim P_\theta \in \mathcal{P}$ where $\mathcal{P}$ is a family of measures closed under domination. Then the associated MLE and EOTE agree even in the finite regime.

In order to obtain an effective method to compute these estimations, we will also work to adapt the Expectation-Maximization (EM) algorithm to its EOT counterpart, thanks to the Sinkhorn algorithm. Further, as presented in [Men+20], the

Sinkhorn EM further shows improvements over the instabilities of the classical EM algorithm, requiring both less iterations to converge, and less random initializations before getting a satisfying result. In this work, the EOT loss also shows better theoretical properties for convergence, reinforcing the belief it is worth studying.

Our original contribution is mainly the sensitivity analysis on model misspecification, where we conclude that the EOTE should indeed be more robust. This lies outside of the setting of (Q1) and (Q2), as we seek clear distinct behaviours between both estimators, favoring one to the other. An alternative representation of the EOT loss, the semi-dual formulation, is a path to explaining this robustness claims, especially in the context of model misspecification. As introduced by [Men], it consists in:

$$\theta_{EOT} = \arg\max_{\theta} \min_{f} KL(\alpha_\theta \parallel \alpha_{\theta,f}) + \mathbb{E}_{Y\sim\beta}(\log \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\alpha_{\theta,f}(Y+Z)),$$

where Y is the noisy observed variable, $X \sim \alpha_\theta$ is the latent variable we model, and $\alpha_{\theta,f}$ is its f-tilting. Indeed, this semi-dual formulation shows, among other things, that the loss is able to account for models outside of the parametric family, thanks to the tilting (in the case of a GMM, this allows fixed mixture weights to vary). Thus we will lead our study in the simplest of such setting:

$$Y = X + Z = \alpha\mathcal{N}(\theta, 1) + (1-\alpha)\mathcal{N}(-\theta, 1).$$

We will enforce misspecification in the mixing parameter: $\alpha_{\text{true}} = \alpha + \varepsilon$. Using an approach introduced by [Gus96], we will devise formulas to approximate and compare sensitivity of both the MLE and the EOTE, and conclude that the EOTE should indeed be more robust. **Simulations on synthetic data confirm the derived formula for the sensitivity of the maximum likelihood estimator is true**.

Shortfalls

Our approach suffers from a few shortfalls. First, theoretically, as this subject of research is still in its infancy. We did not have the time to fully lead a statistical analysis of this estimator in the context of mixture models (consistency, regularity, asymptotical linearity...), with misspecification or not. Moreover, concerning the semi-dual approach, we had to use a further simplification letting us consider the usual parametric estimation theory, instead of going through semiparametrics. We will, however, give directions for future steps and, in particular, a theorem to obtain the influence function for a semi-parametric m-estimator, which applies to the semi-dual representation of EOT.

Second, on the practical viewpoint, even if the obtained formulas in the population regime are verified to be accurate, in practice the simulations using the different EM algorithms on synthetic data show no real difference between the MLE and the EOTE. It is possible that the simplifications needed to use parametric methods for the semi-dual formulation implicitly assume statistical properties, like infinite data, that give a virtual, unobtainable edge to the EOTE.

1 Parametric Estimation

Parametric statistical models at hand will be described by families of distributions identified by subsets of $\mathbb{R}^q$: $\mathcal{P}_\Theta = \{P_\theta,\ \theta \in \Theta \subset \mathbb{R}^q\}$.

When the true distribution of the data lies in the parametric family, we say that the model is **well specified**. Otherwise it is **misspecified**.

Let g be a function from Θ to $\mathbb{R}^d$. We denote by $\hat{g}$ an estimator for $g(\theta)$. When $g = Id$, we will just write $\hat{\theta} := \hat{g}$. When the statistical model describes n data points we may write $\hat{g}_n$ instead of $\hat{g}$.

1.1 Maximum-Likelihood Estimator (MLE)

Suppose the parametric family is dominated by a measure ν, and write p_θ the densities of the distributions in the family at hand. Let $l_\theta(Z_1, \cdots, Z_n) = \sum_{i=1}^n \log p_\theta(Z_i)$. Then the MLE is an M-estimator associated with the above objective function. Thus

$$\hat{\theta}_n^{\text{MLE}} = \arg\max_\theta l_\theta. \tag{1}$$

It is well known that when the model is well-specified, this estimator $\hat{\theta}_n$ is consistent, regular, asymptotically normal, and, when unbiased, reaches Cramer-Rao bound (efficient in a reasonable class of estimators). It has thus always been the preferred method for parameter estimation. Moreover, it admits the interesting property that it minimizes the KL divergence (a well-known notion of 'distance' between measures) between the true distribution of the data P and the distributions in the model:

$$\theta_{\text{MLE}} = \arg\min_\theta KL(P \parallel P_\theta) = \arg\min_\theta \int_{\mathcal{X}} \log\left(\frac{dP}{dP_\theta}(x)\right) dP(x).$$

This seems reassuring in terms of the robustness of the MLE, for instance in the case of model misspecification, where it is also known that the MLE holds pretty well as an estimator, conserving its nicest statistical properties such as linearity etc [Whi82]. However, the KL divergence is a rather bad notion of distance. First it is not even a metric, as it is not symmetric. But more importantly, it also also fails on natural examples; suffices to find two "close" measures such that one does not dominate the other:

$$KL(\delta_{-\varepsilon} \parallel \delta_\varepsilon) = +\infty.$$

This is rather unfortunate, and one could wonder if other approaches could resolve such a failure. To our greatest delight, this happens to be the case, and one should immediately think of the theory of optimal transport: we know how to metrize spaces of measures.

1.2 Entropic Optimal Transport Estimator (EOTE)

The theory of optimal transport comes to our rescue. It is strongly recommended to read Appendix B if one is not familiar with it, and one may also begin with Appendix A for some probabilistic tools. For now, take $\mathcal{X}, \mathcal{Y}$ measurable spaces, denote by $\mathcal{P}(\mathcal{X})$ the set of probabilities on one such space, denote by $\Pi_X : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{P}(\mathcal{X})$ the projection operator, and write $\mathcal{M}(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \Pi_X \gamma = \mu,\ \Pi_Y \gamma = \nu\}$. Introduce the Wasserstein metric distance between measure:

$$W_p(\mu, \nu) = \inf_{\gamma \in \mathcal{M}(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} \|x - y\|^p d\gamma(x, y) \right)^{1/p}. \tag{2}$$

One should read the appendix to be convinced that this, indeed, defines a distance. Notice how, now, $W_p(\delta_{-\varepsilon}, \delta_\varepsilon) = 2\varepsilon$. One could not have hoped better. This leaves us with an immediate estimation method; $\hat{\theta}_{\text{OT}} = \arg\min_\theta W_p(P, P_\theta)$
Unfortunately, inspecting the discrete formulation of this estimation method yields a sad conclusion:

$$\hat{\theta}_n = \arg\min_\theta W_p(\hat{P}, \hat{P}_\theta) = \arg\min_\theta \inf_{P\mathbf{1}=\hat{G}, P^T\mathbf{1}=\hat{F}_\theta} \sum_{1\le i,j\le n} (x_i - y_j)^p P_{ij}.$$

This is a linear program, and best known approaches (network simplex, $O(n^3 \log n)$ [BLO]) are way too slow. The solution is to add an entropic term to the transport problem (entropic regularization):

$$W_{p,\sigma^2}(\mu, \nu) = \inf_{\Pi_X\gamma=\mu, \Pi_Y\gamma=\nu} \int_{\mathcal{X}\times\mathcal{Y}} \|x - y\|^p d\gamma(x, y) + \boldsymbol{\sigma^2 KL(\gamma \parallel \mu \otimes \nu)}, \tag{3}$$

and to then solve the estimation problem:

$$\hat{\theta}_{\text{EOT}} = \arg\min_\theta W_{p,\sigma^2}(P, P_\theta). \tag{4}$$

Entropy acts as a smoothing, and enables very fast optimization methods (**Sinkhorn**, see Appendix B). It also smooths the optimal transport plan γ_{σ^2}:

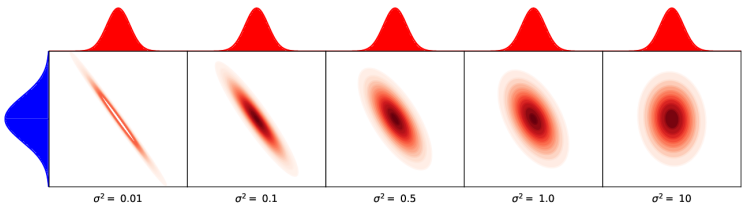

Figure 1: EOT coupling between two Gaussians [Jan+20]

Remark how $\gamma_{\sigma^2} \xrightarrow{\varepsilon\to\infty} \mu \otimes \nu$. Indeed, this does not define a distance anymore, and $\arg\min_\nu W_\varepsilon(\mu, \nu) \neq \mu$ (we can still recover a distance as explained in [PC19a]). This is actually not an issue. The objective is not solely to benefit from computations speed up, while minimizing approximation error by letting $\varepsilon \to 0$. The problem is also of its own interest; it is entropic regularization that provides a bridge with MLE, and good statistical properties, such as robustness.

2 Comparison between MLE and EOT

Let us introduce this well known lemma, that will be of great use:

Lemma 2.1. *(Donsker and Varadhan's variational formula)* Let h be a continuous L^1 function. Then for distributions P and Q:

$$\log \mathbb{E}_P(\exp(h)) = \sup_{Q \ll P} \mathbb{E}_Q(h) - KL(Q \parallel P),$$

equality iff $dQ = \frac{\exp(h)}{\mathbb{E}_P(\exp(h))} dP$.

We now present the ground breaking work of both [Men+20] and [RW18], showing the intricate relationship between MLE and EOT.

2.1 (Q1) $(X, Y) \sim Q^\theta$, Population Regime

As introduced by [Men+20]. Model the observations Y by the joint probability (X, Y) where X is a latent variable. Suppose the distribution of the latent variable is already known; for instance, in a GMM, this means we already know the mixture weight. We are in the following parametric estimation setting:

$$dQ^\theta(x, y) = e^{-g^\theta(x,y)} d\mu_X(x) d\nu(y) = q^\theta(x, y) d\mu_X(x) d\nu(y). \tag{5}$$

The distribution of X is known to be μ_X, ν is a suitable base measure for values of Y, and we require $q^\theta(x, y)d\nu(y)$ to be a probability measure on values of Y (which is equivalent as stating that $\Pi_X Q^\theta = \mu_X$).

This covers many settings. If we have a mixture model, we can have $X \sim \frac{1}{n}\sum \delta_{x_i}$ and $g^\theta \propto (y - x)^T \Sigma_x^{-1} (y - x)$. Then $Y|X \sim \mathcal{N}(\theta_X, \Sigma_X)$, with ν is the Lebesgue measure. Actually, this parametrization with g^θ is flexible enough to model all the parameters of the GMM (we can have $Y|X \sim \mathcal{N}(\mu^\theta(X), \sigma^\theta(X))$ instead of $Y|X \sim \mathcal{N}(X, \sigma^2)$). Thus indeed μ_X is only a guess on mixing weights.

Definition 1. *Define the population regime/empirical log-likelihood $l, \hat{l}$ to be:*

$$l(\theta) = -\mathbb{E}_{Y \sim Q_Y^{\theta^*}} \log q^\theta(Y), \qquad \hat{l}(\theta) = -\mathbb{E}_{Y \sim U_Y} \log q^\theta(Y),$$

where $U_Y = \frac{1}{n}\sum \delta_{y_i}$ is the empirical distribution. Likewise, define the EOT-loss:

$$L(\theta) = W_\theta(\mu_X, Q_Y^{\theta^*}), \qquad \hat{L}(\theta) = W_\theta(\mu_X, U_Y).$$

Theorem 2.2. *(Equivalence in Q1) In the setting of Q1, the EOT loss always dominates the log-likelihood loss:*

$$l(\theta) \leq L(\theta) \qquad \hat{l}(\theta) \leq \hat{L}(\theta)$$

Moreover, whenever the model is well-specified, with the true distribution being identified by θ^*, both population loss are minimized at θ^*:

$$L(\theta^*) = l(\theta^*).$$

Thus the two associated estimator consistently recover the same parameter in the population regime. As a side note, the results obtained here are also of interest for the computation of the EM algorithm, as described in Appendix D.2.

Proof. **Log-Likelihood Loss** $l(\theta), \hat{l}(\theta)$

By writing $q^\theta(x, y) = q^\theta(x|y)q^\theta(y)$, write

$$F_y(\mu, \theta) = \log q^\theta(y) - KL(\mu \parallel Q^\theta(\cdot|y)),$$

where $Q^\theta(\cdot|y)$ is a kernel of conditional probabilities induced by any joint probability over (X, Y) where we only require $Y \sim \mu_Y$. F_y is alike the F-Functional introduced in Appendix D.2. *The following work actually also proves the relations between the F-Functional and the EM algorithms.* Anyway, by taking P a joint probability between (X, Y) we can confidently write

$$\log q^\theta(Y) = \max_{P \sim (X,Y)} F_Y(P(\cdot, Y), \theta) = F_Y(Q^\theta(\cdot|Y), \theta).$$

Having replaced μ by P a joint probability on (X, Y) with $Y \sim \mu_Y$, we are able to switch supremum and integration;

$$\mathbb{E}_{Y \sim \mu_Y} \log q^\theta(Y) = \max_{P \sim (X,Y)} \mathbb{E}_{Y \sim \mu_Y} F_Y(P(\cdot|Y), \theta).$$

Now when $\mu_Y = Q_Y^{\theta^*}$ or $\mu_Y = \frac{1}{n}\sum \delta_{y_i}$ we recover a formula for either the log-likelihood in the population regime $l(\theta)$ or the empirical one $\hat{l}(\theta)$.

EOT Loss $L(\theta), \hat{L}(\theta)$

This time see that:

$$\sup_\mu F_y(\mu, \theta) = \inf_\mu g^\theta(x, y) + KL(\mu \parallel \mu_X).$$

In the same way just introduce a joint distribution P such that $Y \sim \mu_Y$ and now:

$$\mathbb{E}_Y \; F_Y(P(\cdot|Y), \theta) = -\mathbb{E}_{X,Y} \; g^\theta(X, Y) - KL(P(\cdot|Y) \parallel \mu_X).$$

Thus

$$-\sup_P \mathbb{E}_Y F_Y(P(\cdot|Y),\theta) \leq -\sup_{P\in\mathcal{M}(\mu_X,\mu_Y)} \mathbb{E}_Y F_Y(P(\cdot|Y),\theta),$$
$$-\sup_P \mathbb{E}_Y F_Y(P(\cdot|Y),\theta) \leq \inf_{P\in\mathcal{M}(\mu_X,\mu_Y)} \mathbb{E}_{X,Y}\ g^\theta(X,Y) + KL(P(\cdot|Y) \parallel \mu_X),$$

i.e

$$-\mathbb{E}_{Y\sim\mu_Y} \log q^\theta(Y) \leq W_\theta(\mu_X,\mu_Y).$$

We immediately get the following inequalities for the relevant distributions μ_Y:

$$l(\theta) \leq L(\theta) \qquad \hat{l}(\theta) \leq \hat{L}(\theta)$$

Moreover, when $(X,Y) \sim Q^{\theta^*}$, $l(\theta^*)$ is attained at $P = Q^{\theta^*} \in \mathcal{M}(\mu_X, Q_Y^{\theta^*})$. So

$$L(\theta^*) = -\sup_{P\in\mathcal{M}(\mu_X,Q_Y^{\theta^*})} \mathbb{E}_Y F_Y(P(\cdot|Y),\theta^*) \leq -\mathbb{E}_{Y\sim Q_Y^{\theta^*}} F_Y(Q^{\theta^*}(\cdot|Y),\theta^*) = l(\theta^*),$$

and in fact:

$$L(\theta^*) = l(\theta^*).$$

□

2.2 (Q2) Closed Under Domination

As introduced by [RW18], the setting is different. Suppose we work on Euclidean space: $\mathcal{X},\mathcal{Y},\mathcal{Z} = \mathbb{R}^d$. Model our observations Y the following way:

$$Y = X + Z_c, \tag{6}$$

where X is a random variable whose distribution belongs to a (non necessarily parametric) family of measures $\mathcal{P}$ closed under domination (Definition 2), and Z_c is a noise which distribution admits c as a density (e.g $c(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\|x\|^2/2\sigma^2}$ for a Gaussian noise with variance σ^2). Here, the maximum likelihood method is directly used on the family of distribution modeling X rather than $(Y|X)$. We must know the noise distribution.

Definition 2. *(Closed under domination) A family of measures $\mathcal{P}$ is said to be closed under domination if for any measure Q, $Q \ll P$ for some $P \in \mathcal{P}$ implies $Q \in \mathcal{P}$.*

This accommodates for many examples; the class of all measures dominated by the Lebesgue measure, the class of discrete measures, the class of measure with finite support or with at most k points in its support (ex: finite GMM). However, this is still pretty restrictive. Consider for instance a very usual parametric family one might consider: $\{\mathcal{N}(\mu,\sigma^2)\}_{\mu,\sigma\in\mathbb{R}}$. It is definitely not closed under domination.

The main theorem is the following:

Theorem 2.3. *(EOT is Noise Deconvolution [RW18]) Consider the previous model described in Equation 6. Let the distribution of X lie in a family $\mathcal{P}$ closed under domination. Then the MLE $\hat{P}$ satisfies*

$$\hat{P} = \arg\min_{P\in\mathcal{P}} W_c(P, \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}) = \arg\max_{P\in\mathcal{P}} \sum_{i=1}^{n} \log c * dP(y_i), \tag{7}$$

where $*$ denotes convolution. Thus EOTE and MLE agree, even in the finite data regime.

Proof. Write $l_P := \log c * dP$. Applying the convolution to a data point yields

$$l_P(y_i) = \log \int_{\mathcal{X}} c(x - y_i) dP(x).$$

The lemma 2.1 implies that

$$l_P(y_i) = \min_{Q_i} \mathbb{E}_{X\sim Q_i} c(X - y_i) + KL(Q_i \parallel P)$$

By definition the MLE is

$$\hat{P} = \arg\min_{P\in\mathcal{P}} \min_{Q_1,\cdots,Q_n} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X\sim Q_i} c(X - y_i) + KL(Q_i \parallel P).$$

Define U to be $\frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$ the empirical distribution of the observations. Remark that any joint probability measure $\overline{\gamma}$ on $\mathcal{X}\times\mathcal{Y}$ satisfying $\overline{\gamma}\in\mathcal{M}(U)$ (i.e $\Pi_Y\overline{\gamma} = U$) can be uniquely written (easily provable by properties of Dirac measures)

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} Q_i \otimes \delta_{y_i}.$$

Use this bijection to rewrite our MLE as

$$\hat{P} = \arg\min_{P\in\mathcal{P}} \min_{\gamma\in\mathcal{M}(U)} \mathbb{E}_{(X,Y)\sim\gamma} c(X - Y) + KL(\gamma \parallel P \otimes U).$$

But remark that $KL(\gamma \parallel P\otimes U) = KL(\gamma \parallel \Pi_X\gamma \otimes U) + KL(\Pi_X\gamma \parallel P)$, since of course $KL(\Pi_Y\gamma \parallel U) = 0$. Thus we rewrite the MLE equation again as

$$\begin{aligned}\hat{P} &= \arg\min_{P\in\mathcal{P}} \min_{\gamma\in\mathcal{M}(U)} \mathbb{E}_{(X,Y)\sim\gamma} c(X - Y) + KL(\gamma \parallel \Pi_X\gamma\otimes U) + KL(\Pi_X\gamma \parallel P)\\ &= \arg\min_{P\in\mathcal{P}} V(P).\end{aligned}$$

Now, for any P, consider γ_P the coupling that achieves the minimum in $W(P, U)$ (existence and uniqueness proved in Appendix B). Then, $KL(\Pi_X\gamma \parallel P) = 0$ so

$$V(P) \leq \mathbb{E}_{(X,Y)\sim\gamma} c(X-Y) + KL(\gamma \parallel \Pi_X\gamma \otimes U) = W_c(P,U).$$

We now show that we have the reverse equality for the respective minimum of these values, thus showing equality between MLE and EOTE. To this end, observe that $V(P) < +\infty$ implies $\Pi_X\gamma \ll P$, for the divergence not to be infinite. Thus there must be $\Pi_X\gamma \in \mathcal{P}$ by our closed under domination assumption. This simple fact will lead to the conclusion:

$$\begin{aligned}
\min_{P\in\mathcal{P}} V(P) &= \min_{P\in\mathcal{P}} \min_{\gamma\in\mathcal{M}(U),\Pi_X\gamma\in\mathcal{P}} \mathbb{E}_{(X,Y)\sim\gamma} c(X-Y) + KL(\gamma \parallel \Pi_X\gamma \otimes U) + KL(\Pi_X\gamma \parallel P) \\
&\geq \min_{P\in\mathcal{P}} \min_{\gamma\in\mathcal{M}(U),\Pi_X\gamma\in\mathcal{P}} \mathbb{E}_{(X,Y)\sim\gamma} c(X-Y) + KL(\gamma \parallel \Pi_X\gamma \otimes U) \\
&\geq \min_{P\in\mathcal{P}} \min_{\gamma\in\mathcal{M}(P,U)} \mathbb{E}_{(X,Y)\sim\gamma} c(X-Y) + KL(\gamma \parallel \Pi_X\gamma \otimes U) \\
&\geq \min_{P\in\mathcal{P}} W_c(P,U).
\end{aligned}$$

This ends the proof. □

2.3 Discussion

Let us recapitulate our findings.

(Q1) is a more general setting that encompasses usual mixture models up to known fixed mixing weights. Guarantees of equivalence only exist in population regime and in well-specified models. Moreover, empirical study [Men+20] shows that EOT loss behaves better than MLE loss: sharper hessian at true parameter ($\nabla^2 l(\theta^*) \preceq \nabla^2 L(\theta^*)$), fewer bad local optima, and faster convergence using EM algorithm.

(Q2) is somewhat restrictive, even if it works in finite sample regime. It only models distribution of X rather than (X,Y). Noise σ^2 must be known beforehand to calibrate W_{σ^2} the EOT loss. Plus, the closed under domination assumption is very restrictive; it is not satisfied by $\{\mathcal{N}(\theta,\sigma^2)\}_{\theta,\sigma\in\mathbb{R}^2}$ for instance.

Even if these results are interesting by themselves, it is hard to improve them further or tie them together. Rather, we will use them as a proxy as to know where not to look. Now that we are set to begin our robustness study, we want to investigate a data regime with a model such as the resulting estimators show different behaviours. In our case, we would like to focus on the robustness property of the EOTE, and will look at model misspecification, in a non closed under domination family.

3 Robustness Considerations

As thoroughly explained in Appendix C (Theorem C.5), and first pointed out by [Men], the EOT loss admits the following semi-dual formulation:

$$\theta_{EOT} = \arg\max_{\theta} \min_{f} KL(\alpha_\theta \parallel \alpha_{\theta,f}) + \mathbb{E}_{Y\sim\beta}(\log L * \alpha_{\theta,f}(Y)), \tag{8}$$

where $d\alpha_{\theta,f} = \frac{e^f}{\mathbb{E}_{\alpha_\theta}(e^f(X))} d\alpha_\theta$ is the f-tilting, L is such that it admits $e^{-c(x)}$ as a density against the Lebesgue measure (this is alike the noise descirbed in the previous section), where we assume that the cost c verifies $c(x,y) = c(x-y)$, and $*$ denotes the convolution operator. There are some immediate remarks:

- (Adversarial Interpretation) The method tries to make the data look implausible, while tilting not too far away from a distribution in the model.
- (Variance Regularization) This can be thought as variance regularization. Approximate $\log(x) \approx x - \frac{x^2}{2}$. Then

$$\begin{aligned} KL(\alpha_\theta \parallel \alpha_{\theta,f}) &= \log \mathbb{E}_\theta e^f - \int \log\left(e^{f(x)} d\alpha_\theta(x)\right) \\ &\approx \mathbb{E}_\theta e^f - \frac{1}{2}\mathbb{E}_\theta(e^f)^2 - \mathbb{E}_\theta e^f + \frac{1}{2}\mathbb{E}_\theta(e^{2f}) \\ &\approx \frac{1}{2} Var_\theta(e^f). \end{aligned}$$

- (Robustness) Tilting suggests to study model misspecification.
- (Semi parametric estimation) We should appeal to the theory of semiparametrics since the optimization is made on (θ, f), where f can be interpreted as a nuisance parameter, and θ the parameter of interest. However, here it does not work the usual way: the estimation equation involves computation of a sup inf instead of a sup sup.

As we have stated all throughout, our strategy to prove these very serious claims of robustness is to work on the sensitivity on model misspecification of both MLE and EOTE. Since it is easier, we will be working on parametric estimation instead of semiparametrics, by considering that the infimum over the dual potential f is described by a function of θ. As we will see, this is enough to obtain sensible results.

4 Sensitivity Analysis: Model Misspecification

4.1 Simplified Setting: Symmetric GMM with Two Components

We can focus on a mixture of two Gaussians and see how the adversarial approach compares to the MLE alternative. Suppose

$$X \sim \alpha_\theta = \alpha^* \delta_{\theta^*} + (1-\alpha^*)\delta_{-\theta^*},$$

and $c(x,y) = \|x-y\|^2/2\sigma^2$ so that $Z \sim p_z \propto e^{-c(x)}$ and

$$Y = X + Z \sim \alpha^* \mathcal{N}(\theta^*, \sigma^2) + (1-\alpha^*)\mathcal{N}(-\theta^*, \sigma^2).$$

The question is what happens in model misspecification, where

$$X \sim \alpha \mathcal{N}(\theta^*, \sigma^2) + (1-\alpha)\mathcal{N}(-\theta^*, \sigma^2),$$

and $\alpha^* = \alpha + \varepsilon$.

There are many questions to ask. What happens in finite sample regime, inference, out of sample performance etc. On our part, we will focus on a sensitivity analysis on model misspecification, in the population regime, i.e study the value of $\hat{\theta}_{EOT/MLE}(\varepsilon)$ or $\hat{\theta}'_{EOT/MLE}(0)$. As suggested by the semi-dual formulation, the EOT approach is designed to be the most robust.

4.2 Sensitivity Analysis on M-Estimators of Mixture Models

We follow the strategy presented in [Gus96].

4.2.1 Setting and Notations

We look at the following mixture model.

- We observe a variable Y, modeled by specifying distributions of $Y|X$ and X, with X the unobservable parameter (mixing distribution).
- Parametrization $\theta = (\theta_1, \theta_2)$ such that $Y|X \sim F_{\theta_1,X}$ and $Z \sim G_{\theta_2}$.

Here we investigate what happens when our model correctly specifies the conditional distribution $Y|X$ but not the mixing distribution of X. Specifically we look at what happens when the true distribution is 'ε-away' from a member of the parametric family. [CGT17] introduces interesting work regarding what we could use as a distance to characterize this 'ε-away', but here we will simply work with what [Gus96] calls an 'ε-contamination' (which is the same idea for sensitivity analysis introduced by influence functions, in the usual context of robust estimation):

Definition 3. *(ε-contamination) G^ε is an ε contamination of G_{θ^*} if there exists a (necessarily unique) distribution $\tilde{G}$ s.t*

$$G^\varepsilon = (1-\varepsilon)G_{\theta^*} + \varepsilon\tilde{G}.$$

We will then assume that

$$X \sim G^\varepsilon$$

$\tilde{G}$ can be any distribution s.t the estimation method still finds $F_{\theta_1^*}$ as the optimal solution. Typically it is required that $\tilde{G} \in \Lambda(G_{\theta^*})$, which is the class of distributions sharing the same first two moments as G_{θ^*} [Gus96]. But in our case we will see that θ_1 really belongs to a singleton, so there are no such considerations. [dit] uses a Dirac as $\tilde{G}$ and ties sensitivity analysis of leave one out with first order approximation of bias for log likelihood estimation; the analysis is extremely similar even though it does not cover the case of mixture models, and is, as we will point out, a special case of the method here.

Notations Denote the perturbation distribution by $\tilde{G}$. Denote the m-estimator by m. Let the (conditional) m-score functions be $s(\theta, x) = \mathbb{E}(\frac{\partial m}{\partial \theta}(Y,\theta)\ |X = x) = \int_{\mathcal{Y}} \frac{\partial m}{\partial \theta}(y,\theta)dF_{\theta^*,x}(y)$. Write the information matrix $I_m(\theta) = \mathbb{E}(-\frac{\partial^2 m}{\partial\theta\theta^T}(Y,\theta))$. The m-estimation procedure computes

$$\arg\max_\theta\ \mathbb{E}(m(Y,\theta)). \tag{9}$$

Theorem 4.1. *(Sensitivity of parametric m-estimators, mixture models) Keep previous notations and setting. Suppose the information matrix I_m is everywhere invertible. Then, for all $\theta \in \Theta$, there exists a neighbourhood U of 0 and V of θ s.t there exists a smooth function $\theta : U \to \mathbb{R}$ verifying*

$$\theta(\varepsilon) = \arg\max_\theta\ \mathbb{E}(m(Y,\theta)).$$

Moreover we can compute its derivative at zero to be exactly

$$\theta'(0) = I_m^{-1}(\theta^*)\int s(\theta^*, z)d\tilde{G}(x), \tag{10}$$

where $\theta^* = \theta(0)$ is the parameter identifying truth, recovered without perturbation.

Proof. For an 'ε-contamination' of the mixing distribution, the estimation θ_ε must verify

$$\frac{\partial}{\partial\theta}\mathbb{E}\left(m(Y,\theta)\right) = \frac{\partial}{\partial\theta}\mathbb{E}_{G^\varepsilon}\mathbb{E}_{F_{\theta^*,X}}\left(m(Y,\theta)|X\right)\Big|_{\theta_\varepsilon} = \mathbb{E}_{G^\varepsilon}s(\theta,X)\Big|_{\theta_\varepsilon} = 0.$$

Denote by $g : (\varepsilon,\theta) \mapsto \mathbb{E}_{G^\varepsilon}s(\theta,X)\Big|_\theta$ the above function, and by θ^* the unique solution at $\varepsilon = 0$ of the above problem, s.t $g(0,\theta^*) = 0$. This is well defined since the problem is correctly specified in the usual statistical framework. See that

$$\frac{\partial g}{\partial\theta}(0,\theta^*) = \mathbb{E}\left(\frac{\partial^2 m}{\partial\theta\theta^T}(Y,\theta)\right) = -I_m(\theta^*),$$

which we suppose to be invertible. Thus the implicit function theorem assures us that for ε in a small neighbourhood U of zero we have a differentiable function θ such that

$$g(\varepsilon, \theta(\varepsilon)) = 0.$$

Now we will find a formula for $\theta'(0)$, which is our desired approximation result (impossible to come up with an exact formula since we we are manipulating unknown distributions). Compute

$$\begin{aligned}\frac{\partial g}{\partial \varepsilon}(\varepsilon, \theta) &= \frac{\partial}{\partial \varepsilon}(1-\varepsilon)\int s(\theta(\varepsilon), x) dG_{\theta^*}(x) + \epsilon \int s(\theta(\varepsilon), x) d\tilde{G}(x) \\ &= \frac{\partial}{\partial \varepsilon}(1-\varepsilon)A(\varepsilon) + \varepsilon B(\varepsilon) \\ &= -A(\varepsilon) + (1-\varepsilon)A'(\varepsilon) + B(\varepsilon) + \varepsilon B'(\varepsilon).\end{aligned}$$

See how

$$A'(\varepsilon) = \int \frac{\partial m}{\partial \theta \theta^T}(\theta(\varepsilon), x)\theta'(\varepsilon) dG_{\theta^*}(x) = -I_m(\theta(\varepsilon))\theta'(\varepsilon).$$

Moreover, $A(0) = g(0, \theta^*) = 0$ so at $\varepsilon = 0$:

$$\begin{aligned}\frac{\partial g}{\partial \varepsilon}(\varepsilon, \theta) &= A'(0) + B(0) - A(0) \\ &= -I_m(\theta^*)\theta'(0) + \int s(\theta^*, x) d\tilde{G}(x) \\ &= 0.\end{aligned}$$

Thus finally

$$\theta'(0) = I_m^{-1}(\theta^*) \int s(\theta^*, x) d\tilde{G}(x).$$

□

Corollary 4.1.1. *Remark that when we are not working with a mixture model, such as in [dit], we recover the (very) famous formula for the influence function of the m-estimator by simply considering $\tilde{G} = \delta_x$; then*

$$\theta'_x(0) = IF(\hat{\theta}, G_{\theta^*}, \delta_x) = I_m^{-1}(\theta^*) s_\theta(\theta^*, x). \tag{11}$$

We can easily deduce the asymptotic variance of the estimator for instance:

$$Var(\hat{\theta}) = Var(\theta'_X(0)) = I_m^{-1}(\theta^*)\mathbb{E}(s_\theta(\theta^*, X) s_\theta(\theta^*, X)^T) I_m^{-1}.$$

4.2.2 Log-Likelihood

Now let us put this into practice with a Gaussian mixture, and the m-estimator being the log-likelihood. Take

$$F_{\theta^*,x}(y) = F_x(y) \sim \mathcal{N}(x, 1).$$

Here, see how the distribution of $Y|X$ is not parametrized. Thus we do not require that $G^\varepsilon \in \Lambda(G_\theta)$ (i.e matching moments), since no matter the distribution of Z the optimal parameter found by the model will (obviously) always result in $F = F_{\theta^*,x}$. Without loss of generality, suppose $\alpha^* \geq 1/2$ and take

$$\begin{aligned} G^{\varepsilon'} &= (\alpha^* + \varepsilon')\delta_{\theta^*} + (1 - \alpha^* - \varepsilon')\delta_{-\theta^*} \\ G_{\theta^*} &= \alpha^*\delta_{\theta^*} + (1 - \alpha^*)\delta_{-\theta^*} \\ \tilde{G} &= \delta_{\theta^*}, \end{aligned}$$

So that we have

$$G^{\varepsilon'} = (1 - \varepsilon)G_{\theta^*} + \varepsilon\tilde{G}, \tag{12}$$

with $\varepsilon' = \varepsilon(1-\alpha^*)$. In the case where $\alpha^* < 1/2$, simply exchange θ^* by $-\theta^*$, which is a convenient symmetry. The likelihood function we are working with is thus defined as follows

$$p_\theta(y) = d\left(\alpha^*\mathcal{N}(\theta, 1) + (1 - \alpha^*)\mathcal{N}(-\theta, 1)\right)(y),$$

where, as we set out to study, the misspecification is on the mixing weight α^*; Y is indeed distributed as

$$Y \sim (\alpha^* + \varepsilon')\mathcal{N}(\theta^*, 1) + (1 - \alpha^* - \varepsilon')\mathcal{N}(-\theta^*, 1),$$

which we will denote by $Y \sim (F, G^\varepsilon)$. The m-estimator at hand is $l_\theta = \log p_\theta$. Thus

$$\theta(\varepsilon') = \arg\max_\theta \mathbb{E}l_\theta(Y),$$

where we rather have θ a function of ε' for the natural misspecification we introduced in the Gaussian mixture (Equation 12). Since $\frac{\partial\theta}{\partial\epsilon} = (1 - \alpha^*)\frac{\partial\theta}{\partial\epsilon'}$, we can directly compute our final result using Theorem 4.1:

$$\theta'(0) = \frac{1}{(1 - \alpha^*)} I^{-1}(\theta^*)s(\theta^*, \theta^*),$$

with $I^{-1}(\theta^*) = -\left(\mathbb{E}\frac{\partial^2}{\partial\theta^2}l_{\theta^*}(Y)\right)^{-1}$ and $s(\theta^*, \theta^*) = \mathbb{E}\left(\frac{\partial}{\partial\theta}l_{\theta^*}(Y)|Z = \theta^*\right)$:

$$\theta'(0) = \frac{-1}{(1 - \alpha^*)}\left(\mathbb{E}\frac{\partial^2}{\partial\theta^2}l_{\theta^*}(Y)\right)^{-1}\int_{\mathbb{R}}\frac{\partial}{\partial\theta}l_{\theta^*}(y)d\mathcal{N}(\theta^*, 1)(y). \tag{13}$$

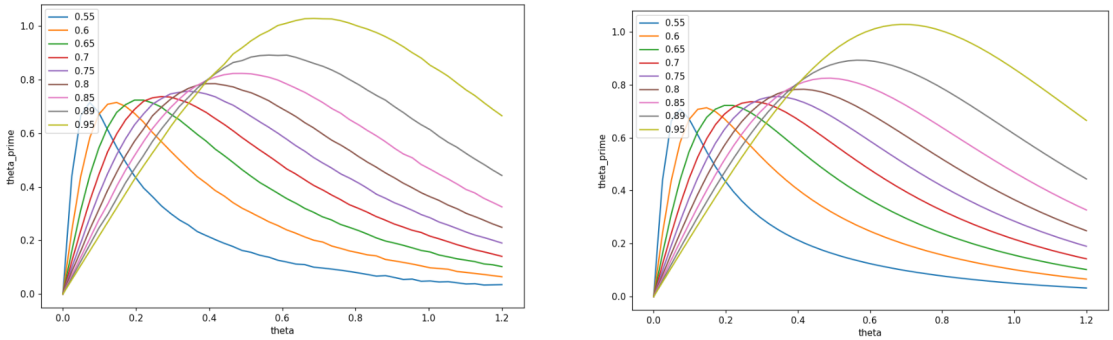

Figure 2: Numerical simulations (left) vs theoretical formula (right)

Experimentation Results Confirms the truthfulness of the formula we have found.

4.2.3 Semi-dual EOTE as a Parametric M-Estimator

Remember we have the semiparametric semi-dual EOT estimator (Equation 8):

$$\begin{aligned}\theta_{EOT} &= \arg\max_{\theta}\min_{f} \mathbb{E}G(\theta, f, Y) \\ &= \arg\max_{\theta}\min_{f} \mathbb{E}_{Y\sim\beta}\left(KL(\alpha_\theta \parallel \alpha_{\theta,f}) + \log L * \alpha_{\theta,f}(Y)\right).\end{aligned}$$

This is hard to manipulate, because of the nuisance parameter f. Instead of appealing to semiparametrics, let us just consider that the infimum is reached on an almost everywhere smooth function of θ, such that we obtain the following parametric m-estimator:

$$\theta_{EOT} = \arg\max_{\theta} \mathbb{E}m(\theta, f(\theta), Y).$$

Now, $m : (y, \theta) \mapsto m(\theta, f(\theta, y))$ is a perfectly good m-estimator to which we can apply our sensitivity analysis. Consider the same mixture setting

$$Y|X \sim \mathcal{N}(X, 1), \quad X \sim G_{\theta^*} = \alpha^*\delta_{\theta^*} + (1-\alpha^*)\delta_{-\theta^*},$$

and perturbation G_θ^ε as before. Assuming I_m invertible we directly obtain the sensitivity formula from Theorem 4.1:

$$\theta'(0) = (1-\alpha^*)^{-1} I_m^{-1}(\theta^*) \int s(\theta^*, x) d\tilde{G}(x). \tag{14}$$

Since $\tilde{G}(x) = \delta_{\theta^*}$, we have $\theta'(0) = (1-\alpha^*)^{-1} I_m^{-1}(\theta^*) s(\theta^*, \theta^*)$. In other words

$$\begin{aligned}\theta'(0) &= (1-\alpha^*)^{-1} I_m^{-1}(\theta^*) \mathbb{E}(s(\theta^*, Y)|X = \theta^*) \\ &= (1-\alpha^*)^{-1} I_m^{-1}(\theta^*) \\ &\quad \times \left[\frac{\partial}{\partial\theta}(KL(\alpha_\theta \parallel \alpha_{\theta,f(\theta)})) + \int_{\mathcal{Y}} \left(\frac{\partial}{\partial\theta} \log \int_{\mathcal{X}} e^{-c(y-x)} d\alpha_{\theta,f(\theta)}(x)\right) d\mathcal{N}(\theta^*, 1)(y)\right].\end{aligned}$$

Let us expand everything to get to some nice formulas. We will separate work between the information matrix, and the score function (terms between brackets).

Score function Write

$$\alpha_{\theta,f(\theta)} = h(\theta)\delta_{\theta^*} + (1-h(\theta))\delta_{-\theta^*}, \text{ with } \; h(\theta) = \frac{\alpha^* e^{f(\theta)}}{\alpha^* e^{f(\theta)} + (1-\alpha^*)e^{f(-\theta)}}.$$

Then

$$\begin{aligned}\frac{\partial}{\partial\theta}(KL(\alpha_\theta \parallel \alpha_{\theta,f(\theta)})) &= \frac{\partial}{\partial\theta}\alpha^* \log\frac{\alpha^*}{h(\theta)} + (1-\alpha^*)\log\frac{1-\alpha^*}{1-h(\theta)} \\ &= \frac{(\alpha^* - h(\theta))h'(\theta)}{(h(\theta)-1)h(\theta)}.\end{aligned}$$

Further defining c to be $x \mapsto \|x\|^2 + \log(\sqrt{2\pi})$:

$$\begin{aligned}\left[\cdots\right] = \Bigg[&\frac{(\alpha^* - h(\theta))h'(\theta)}{(h(\theta)-1)h(\theta)} \\ &+ \int_{\mathcal{Y}} \left(\frac{\partial}{\partial\theta}\log h(\theta)e^{-c(y-\theta)} + (1-h(\theta))e^{-c(y+\theta)}\right) d\mathcal{N}(\theta^*,1)(y)\Bigg] \\ = \Bigg[&\frac{(\alpha^* - h(\theta))h'(\theta)}{(h(\theta)-1)h(\theta)} \\ &+ \int_{\mathcal{Y}} \left(\frac{h'(\theta)(e^{-c(y-\theta)} - e^{-c(y+\theta)})}{h(\theta)e^{-c(y-\theta)} + (1-h(\theta))e^{-c(y+\theta)}}\right) d\mathcal{N}(\theta^*,1)(y) \\ &+ \int_{\mathcal{Y}} \left(\frac{-2h(\theta)((y-\theta)e^{-c(y-\theta)} - (y+\theta)e^{-c(y+\theta)})}{h(\theta)e^{-c(y-\theta)} + (1-h(\theta))e^{-c(y+\theta)}}\right) d\mathcal{N}(\theta^*,1)(y) \\ &+ \int_{\mathcal{Y}} \left(\frac{-2(y+\theta)e^{-c(y+\theta)}}{h(\theta)e^{-c(y-\theta)} + (1-h(\theta))e^{-c(y+\theta)}}\right) d\mathcal{N}(\theta^*,1)(y)\Bigg].\end{aligned}$$

Information matrix We simply compute

$$\begin{aligned}I_m(\theta) = &\frac{\partial^2}{\partial\theta^2}(KL(\alpha_\theta \parallel \alpha_{\theta,f(\theta)})) \\ &+ \int_{\mathcal{Y}} \left(\frac{\partial^2}{\partial\theta^2}\log\left(h(\theta)e^{-c(y-\theta)} + (1-h(\theta))e^{-c(y+\theta)}\right)\right) dG_{\theta^*}(y),\end{aligned}$$

where $\dfrac{\partial^2}{\partial\theta^2}(KL(\alpha_\theta \parallel \alpha_{\theta,f(\theta)})) =$

$$\frac{(h(\theta)-\alpha^*)(h(\theta)-1)h'(\theta)^2 + (h(\theta)-\alpha)h(\theta)h'(\theta)^2 + [(\alpha - h(\theta))h''(\theta) - h'(\theta)^2](h(\theta)-1)h(\theta)}{(h(\theta)-1)^2h(\theta)^2},$$

$$
\begin{aligned}
&\text{and } \frac{\partial^2}{\partial\theta^2}\log\left(h(\theta)e^{-c(y-\theta)}+(1-h(\theta))e^{-c(y+\theta)}\right)= \\
&-\left(\left((\theta-y)\,h(\theta)-h'(\theta)\right)e^{-c(y+\theta)}+\left(-(\theta+y)\,(h(\theta)-1)+h'(\theta)\right)e^{-c(y-\theta)}\right) \\
&\quad\times\left((\theta-y)\,(h(\theta)-1)\,e^{-c(y-\theta)}-(\theta+y)\,h(\theta)e^{-c(y+\theta)}+e^{-c(y-\theta)}h'(\theta)-e^{-c(y+\theta)}h'(\theta)\right) \\
&+\left((h(\theta)-1)\,e^{-c(y-\theta)}-h(\theta)e^{-c(y+\theta)}\right) \\
&\quad\times\Bigg(-(\theta-y)\left((\theta+y)\,(h(\theta)-1)-h'(\theta)\right)e^{-c(y-\theta)} \\
&\qquad+(\theta+y)\left((\theta-y)\,h(\theta)-h'(\theta)\right)e^{-c(y+\theta)} \\
&\qquad+\left((\theta-y)\,h'(\theta)+h(\theta)-h''(\theta)\right)e^{-c(y+\theta)} \\
&\qquad+\left(-(\theta+y)\,h'(\theta)-h(\theta)+h''(\theta)+\right)e^{-c(y-\theta)}\Bigg) \\
&\times\frac{1}{\left((h(\theta)-1)\,e^{-c(y-\theta)}-h(\theta)e^{-c(y+\theta)}\right)^2},
\end{aligned}
$$

which entirely determines the value of $\theta'(0)$. However, we have not been able to determine analytical solutions to $h(\theta)$. Many various techniques were tried (hand calculation, symbolical computations), never with success. Thus, in the following, computing the formula will involve numerical integration.

Experimentation Results Confirm the truthfulness of the formula we have found.

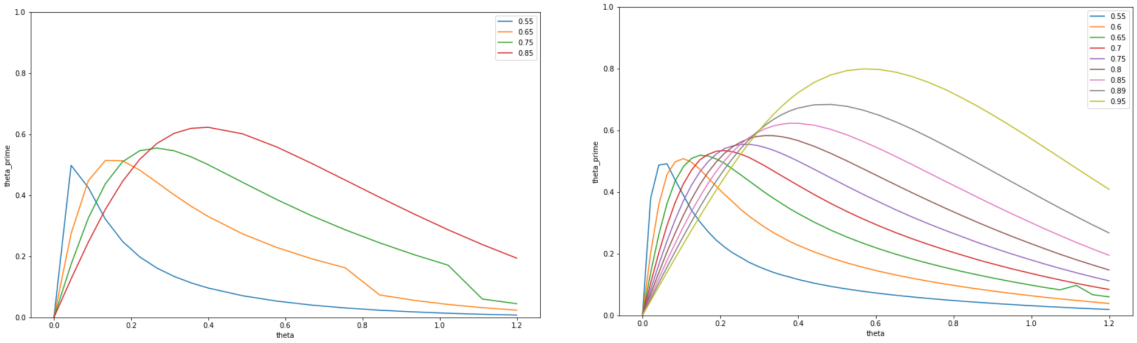

Figure 3: Numerical simulations (left) vs theoretical formula (right)

4.3 Comparison Between EOT and MLE

Compare $\theta'_{EOT}(0), \theta'_{MLE}(0)$ for different (α, θ), on the misspecified model

$$\alpha\mathcal{N}(\theta^*, 1) + (1-\alpha)\mathcal{N}(-\theta^*, 1),$$

with the true data being distributed with the mixture weights specified by $\alpha^* = \alpha + \varepsilon$. In a general fashion, we always have:

$$\theta'_{EOT}(0) \leq \theta'_{MLE}(0), \tag{15}$$

which theoretically proves that the EOT method is indeed more robust in this setting.

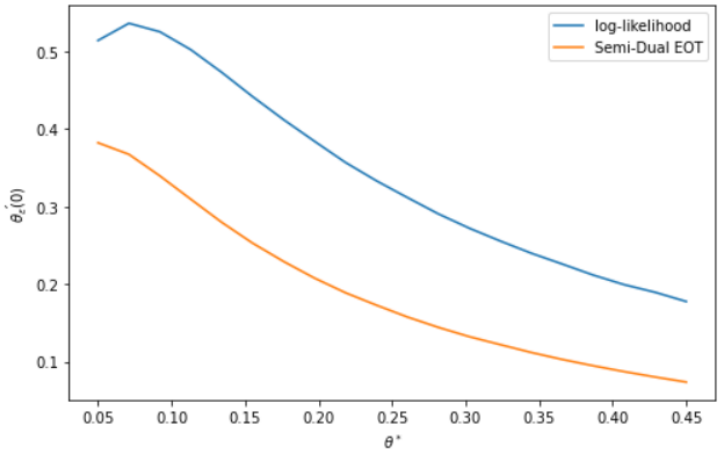

Figure 4: $\alpha^* = 0.55$

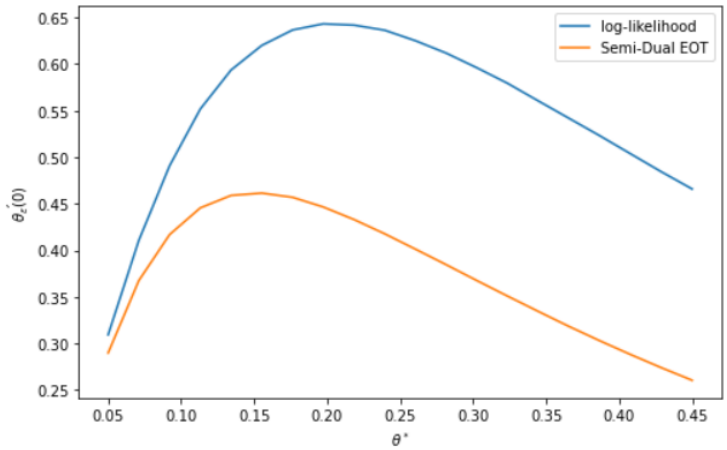

Figure 5: $\alpha^* = 0.65$

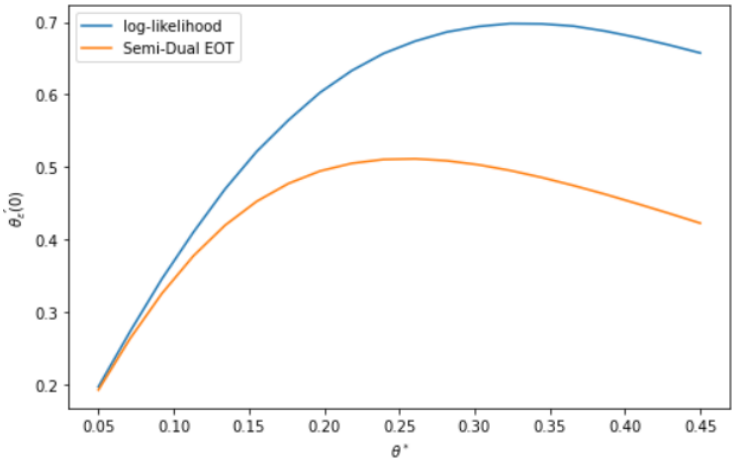

Figure 6: $\alpha^* = 0.75$

4.4 Simulations with Classical EM and Sinkhorn EM

Please refer to Appendix D to understand how the EM algorithms are implemented, and the update formulas for a mixture of symmetric Gaussians.

Because of numerical imprecision, we do not plot the estimated derivative of the estimator, but simply the estimator resulting from the ε-tilting, and we denote it by θ_ε. By 'true', on the plots, we mean the value of $\theta(\varepsilon)$ found by each misspecified model, when using `scipy` optimization procedures to get the exact results on our Gaussian mixture models. By 'EM' we mean running the Classical EM algorithm ('true' thus means the true curve for log-likelihood misspecification), and idem with Sinkhorn EM (see D for algorithm).

In order to have sensible results, for each EM simulation we choose the best estimation out of 5 runs, each randomly initialized close to the true computed value of $\theta_{MLE/EOT}(\varepsilon)$, with $\theta_{init} = (1 + 0.2\text{random}(-1, 1))\theta_{MLE/EOT}(\varepsilon)$.

We use $N = 3000$ points for each simulations, 10 EM iterations as prescribed in [XHM16] (and verified empirically, there is no need for more with 2-GMM), and compute a total of $M = 100$ estimation per (α, θ) couple with different randomly generated data. We plot the true and simulated curves for 20 equally spaced θ^* between 0.05 and 0.45 and for three different alphas. Each data point of the EM or Sinkhorn algorithms include a vertical error bar that indicate $\text{mean}(\hat{\theta}_n) \pm \frac{\text{std}(\hat{\theta}_n)}{\sqrt{M}}$. Misspecification is $\varepsilon = 0.05$, chosen so that $\theta_\varepsilon - \theta^* > \text{std}(\hat{\theta}_n)$.

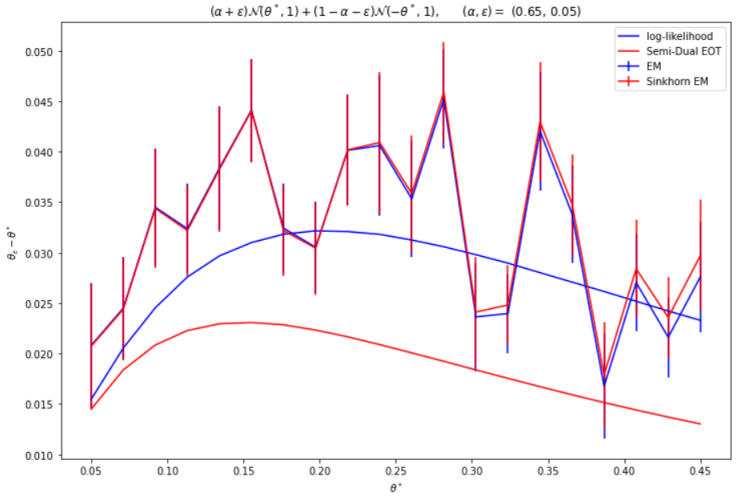

Figure 7: $\alpha^* = 0.65$

We can see that the MLE theoretical curve coincides with the simulation, which is reassuring. However, Sinkhorn EM produces approximately the same values as the EM algorithm. This has been consistently the case during many hours of different simulations and hyper-parameters selection. This odd behaviour needs further study. It is possible that the simplifications needed to use parametric methods for the semi-dual formulation implicitly assume statistical properties, like infinite data, that give a

virtual, unobtainable edge to the EOTE. It is also possible that numerical subtleties in the implementation explain such a behaviour, but it is yet to be fully elucidated. Anyway, this is encouraging, and **the MLE is well fitted.**

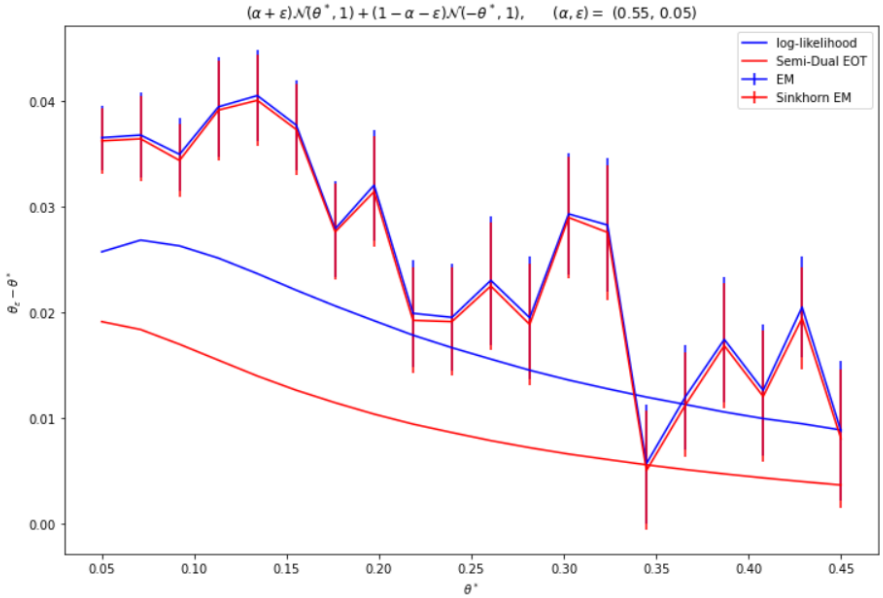

Figure 8: $\alpha^* = 0.55$

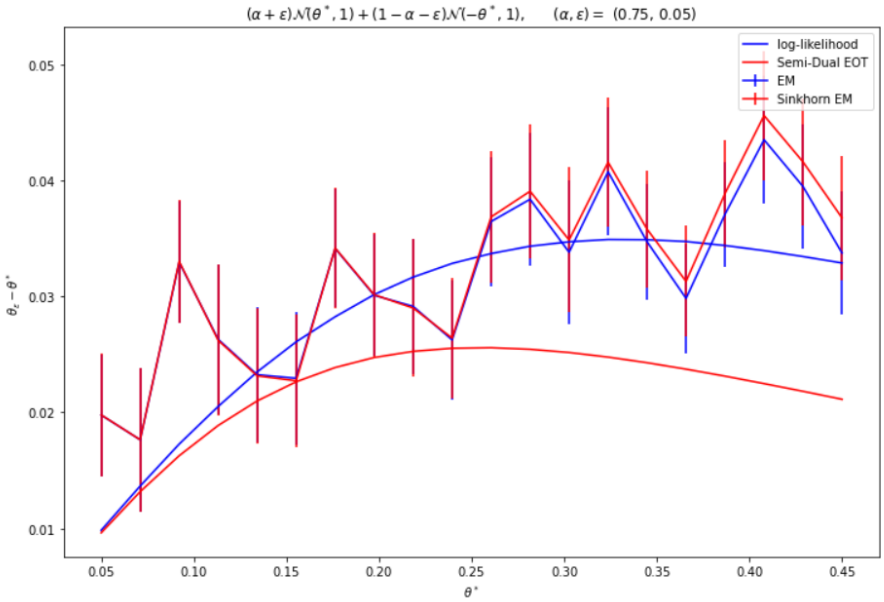

Figure 9: $\alpha^* = 0.75$

5 Semiparametrics for Semi-Dual EOT

Remember the Equation 8 for the semi-dual EOT:

$$\theta_{EOT} = \arg\max_{\theta} \min_{f} KL(\alpha_\theta \parallel \alpha_{\theta,f}) + \mathbb{E}_{Y\sim\beta}(\log L * \alpha_{\theta,f}(Y))$$

Working directly with this formulation is harder. Here, f can be interpreted as a nuisance parameter, embedded in a space of possibly infinite dimension, and θ is the parameter of interest. One can take a look at [Tsi07] and [Kos07] for a good introduction to the theory of semiparametric estimators. However, here it does not go the usual way: the estimation equation involves the computation of a sup inf instead of a sup sup. But this is not necessarily an obstacle. As a direction for further result, we propose the following (applicable) theorem, presented in [Kos07]. Refer to the latter for full necessary hypothesis.

5.1 A Formula for Further Studies

Theorem 5.1. *(Influence function for semiparametric M-Estimator [Kos07]) Write $m(\theta, f)$ the semiparametric estimator. Use subscripts to denote derivative, bracket $[A]$ for Frechet derivative along A in nuisance space. Suppose $(\hat{\theta}_n, \hat{f}_n)$ satisfy Equation 21.6 and conditions A1-A4 hold. Then*

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) = &-\sqrt{n}\mathbb{E}\left(m_{11}(\theta_0 f_0) - m_{21}(\theta_0, f_0)[A^*]\right)^{-1} \\ &\times \mathbb{E}_n(m_1(\theta_0, f_0) - m_2(\theta_0, f_0)[A^*]) + o_P(1)\end{aligned}$$

where A^* verifies a projection criterion accommodating for the $\min_f$ in the semi-dual equation 8. Thus the estimator is asymptotically linear with mean 0, and we easily deduce variance etc. This could also be an entry point to compare with the parametric results, and study misspecification.

References

[Bil68] P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1968. ISBN: 9780471072423. URL: https://books.google.co.uk/books?id=O9oQAQAAIAAJ.

[BLO] ADAM BLOCK. *OPTIMAL TRANSPORT FOR DISCRETE DISTRIBUTIONS*. URL: https://patrikgerber.github.io/assets/pdf/Discrete_Optimal_Transport.pdf.

[Bon] Thomas Bonald. *Expectation-maximization for the gaussian mixture model*. URL: https://perso.telecom-paristech.fr/bonald/documents/gmm.pdf.

[CGT17] Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. *Optimal transport for Gaussian mixture models*. 2017. DOI: 10.48550/ARXIV.1710.07876. URL: https://arxiv.org/abs/1710.07876.

[CP15] Marco Cuturi and Gabriel Peyre. *A Smoothed Dual Approach for Variational Wasserstein Problems*. 2015. arXiv: 1503.02533 [stat.ML].

[CP18] Marco Cuturi and Gabriel Peyre. "Semi-dual Regularized Optimal Transport". In: *CoRR* abs/1811.05527 (2018). arXiv: 1811.05527. URL: http://arxiv.org/abs/1811.05527.

[Cut13] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

[Del] George Deligiannidis. *Foundations of Statistical Inference*. URL: https://www.stats.ox.ac.uk/~deligian/pdf/sb21/sb21_notes.pdf.

[dit] Francois ditraglia. *Lecture Notes for Econ722*. URL: http://ditraglia.com/econ722/main.pdf.

[Gus96] Paul Gustafson. "The Effect of Mixing-Distribution Misspecification in Conjugate Mixture Models". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 24.3 (1996), pp. 307–318. ISSN: 03195724. URL: http://www.jstor.org/stable/3315741 (visited on 03/02/2023).

[Jan+20] Hicham Janati et al. "Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed Form". In: *arXiv: Statistics Theory* (2020).

[Kos07] M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, 2007. ISBN: 9780387749785. URL: https://books.google.co.uk/books?id=FXaYjTQIUZ8C.

[Men] Gonzalo Mena. *Parameter estimation in deconvolution models using optimal transport: a note*.

[Men+20] Gonzalo Mena et al. *Sinkhorn EM: An Expectation-Maximization algorithm based on entropic optimal transport*. 2020. arXiv: 2006.16548 [stat.ML].

[PC19a] Gabriel Peyre and Marco Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.

[PC19b] Gabriel Peyre and Marco Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.

[RW18] Philippe Rigollet and Jonathan Weed. *Entropic optimal transport is maximum-likelihood deconvolution*. 2018. arXiv: 1809.05572 [math.ST].

[Sch] Martin Schweizer. *Weak convergence of probability measures*. URL: https://www2.math.ethz.ch/education/bachelor/lectures/fs2014/math/bmsc/weak-conv.pdf.

[Tho] Matthew Thorpe. *Introduction to Optimal Transport*. URL: https://www.damtp.cam.ac.uk/research/cia/files/teaching/Optimal_Transport_Notes.pdf.

[Tsi07] A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2007. ISBN: 9780387373454. URL: https://books.google.co.uk/books?id=xqZFi2EMB40C.

[Whi82] Halbert White. "Maximum Likelihood Estimation of Misspecified Models". In: *Econometrica* 50.1 (1982), pp. 1–25. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912526 (visited on 03/02/2023).

[XHM16] Ji Xu, Daniel Hsu, and Arian Maleki. *Global analysis of Expectation Maximization for mixtures of two Gaussians*. 2016. arXiv: 1608.07630 [math.ST].

A Notations and Preliminaries

$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ will be denoting some arbitrary measurable space with the convenient topology (Polish space: separable topological space which can be metrized using a distance which makes it a complete metric space).

The set of measures on each space will be denoted $\mathcal{P}(\mathcal{X})$, we'll usually write $\mu, \nu, \eta \in \mathcal{P}(\mathcal{X})$. The letter P in $P \in \mathcal{P}(\mathcal{X})$ will specifically designate a probability measure.

Denote by $X \sim G$ the fact that G is the distribution of the random variable X. With f a measurable function, $\mathbb{E}_G(f(X)) = \int_{\mathcal{X}} f(x)dG(x)$ denotes the expectation of the random variable $f(X)$ with respect to the distribution G (we may mix the distribution with its associated probability measure). We may also unequivocally write $\mathbb{E}_{X\sim G}$, or $\mathbb{E}_{X\sim g}$ if G admits a density g against another measure (usually the Lebesgue measure).

Types of convergence; $\mu_n \rightharpoonup \mu$ denotes convergence in law, $\mu_n \xrightarrow{\mathcal{P}} \mu$ convergence in probability, and $\mu_n \xrightarrow{a.s} \mu$ a.s convergence.

Let $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a joint probability measure. Denote by $\Pi_X\gamma \in \mathcal{P}(\mathcal{X})$ its projection on the first variable. Likewise, define $\Pi_Y\gamma \in \mathcal{P}(\mathcal{Y})$.

Define

$$\mathcal{M}(\mu, \nu) := \{\gamma \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : \Pi_X\gamma = \mu, \Pi_Y\gamma = \nu\}, \quad \mathcal{M}(\nu) := \{\gamma \in \mathcal{X} \times \mathcal{Y} : \Pi_Y\gamma = \nu\}.$$

The pushforward of measure μ by a measurable function T is $T_{\#}\mu = \nu$ and is such that $\forall B$ measurable, $\nu(B) = \mu(T^{-1}(B))$. When μ, ν absolutely continuous against the Lebesgue measure in Euclidean space, we have $\mu(x) = \nu(T(x))|\det(T'(x))|$.

We will sometimes talk of disintegration of measures, in particular in relation to conditional expectation.

Let f be a continuous bounded function. An f-tilting parametrizes an exponential tilting, that is, a change of measure defined by the random variable $L = \frac{e^{f(X)}}{\mathbb{E}_X(e^{f(X)})}$.

Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. We say that $\mu \ll \nu$ (ν dominate μ) whenever for all measurable sets $B \subset \mathcal{X}$, $\nu(B) = 0 \Rightarrow \mu(B) = 0$. Denote by KL the Kullback-Leiber divergence:

$$KL(\mu \parallel \nu) = \begin{cases} E_\mu(\log(\frac{d\mu}{d\nu})) & \text{if } \mu \ll \nu \\ +\infty & \text{else} \end{cases}$$

B Theory of Optimal Transport

The goal of this section is to introduce the relevant tools to rightfully construct our alternative estimator. We will burn through the usual presentation and history of optimal transport to get to our means, while trying to stay pedagogic. This is why we introduce Monge's formulation as we feel this is an easier way into understanding the motivation of the theory. We will also focus on Sinkhorn algorithm, which will give a procedure to actually compute the EOT estimator.

B.1 Optimal Transport

Most of the content here is inspired by [PC19b], [PC19a], [Tho].

B.1.1 Monge Formulation

The initial setting for optimal transport is the following. Imagine you are given a starting measure μ and a final measure ν, such that you wish to transport one to the other. For instance we could consider that μ models the distribution in space of a pile of sand, and ν models the distribution in space of a corresponding storage space such that their volume is the same. We also specify a cost function c such that $c(x, y)$ models the cost of transporting x to y. Then the optimal transport problem consists in finding the best transport plan such that the total cost is minimized. This very simple example is also why this is sometimes referred as the earth mover's problem.

Let us define the problem formally.

Definition 4. *(Monge Formulation for Optimal Transport) Given $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the Monge formulation for optimal transport consists in determining the following:*

$$\mathbb{M}(\mu, \nu) = \inf_{\text{measurable } T} \int_{\mathcal{X}} c(x, T(x)) d\mu(x), \quad T_{\#}\mu = \nu. \tag{16}$$

Example B.1. *(Matching Problem) An interesting example is the matching problem. Suppose μ is an empirical measure admitting n uniformly weighted points in its support; $\mu \propto \sum_i \delta_{x_i}$. The mapping T must then be one-one; $\nu = \sum_i \delta_{y_i}$, and T is described by a permutation $\sigma \in \mathcal{S}_n$. The cost can be described by $C \in \mathcal{M}_{n,n}(\mathbb{R})$ where C_{ij} is the cost of transporting an element of mass from point i to j. The optimal transport problem becomes equivalent to:*

$$\min_{\sigma \in \mathcal{S}_n} \sum_{i}^{n} C_{i\sigma(i)}. \tag{17}$$

One can directly see the computational downside; solving this with a simple algorithm checking all permutations would run in $O(n!)$... which is clearly prohibitive. It is thus important to enforce additional properties to devise effective approaches. In particular, in this problem, if we suppose that $C_{ij} = h(x_i - y_j)$ where h is convex (e.g $h(x) = x^2$),

then one can prove that the permutation has to identically map the sorted sequences $(x_i)_i$, $(y_i)_i$ one to another; thus describing a procedure with much better complexity $O(n \log n)$.

In the previous example, one can convince himself that the minimal cost $\mathbb{M}(\mu, \nu)$ seems to provide a good notion of distance between measure. As it happens, this is indeed the case, and one should not forget it from now on. Let us now give an important existence theorem. It is hard to prove and requires to build many more tools, but we give it here none the less, for clarity.

Theorem B.2. *(Brenier Theorem) Assume $\mathcal{X}, \mathcal{Y} = \mathbb{R}^d$, $c(x, y) = \|x - y\|^2$ and μ is dominated by the Lebesgue measure. Then, there exists a unique optimal map T solving the OT problem, both in its Monge formulation and its Kantorovitch relaxation (that we introduce later). It is characterized by being the unique gradient of a convex function φ s.t $T_{\#}\mu = \nabla\varphi_{\#}\mu = \nu$.*

Proof. Admitted. See [PC19b] or [Tho] Theorem 4.2; this is quite a long proof. Requires study of the dual formulation; all the necessary tools are introduced later though (e.g this theorem was the initial motivation for c-transforms). □

Example B.3. *(OT on 1D Gaussian) Assume $\mu = \mathcal{N}(m_\mu, s_\mu)$, $\nu = \mathcal{N}(m_\nu, s_\nu)$. One can verify that*

$$\varphi : x \mapsto \frac{\Sigma_\nu}{2\Sigma_\mu}(x - m_\mu)^2 + m_\nu x,$$

is convex and that with $T = \nabla\varphi$ we have $T_{\#}\mu = \nu$. Thus, Brenier theorem shows that for the Euclidean cost, T is the unique optimal transport map, and the associated Monge distance is

$$\mathbb{M}(\mu, \nu) = (m_\mu - m_\nu)^2 + (s_\mu - s_\nu)^2.$$

One could not hope for a better distance formula between two Gaussians! And the formula still holds for Dirac measures ($s_\mu = s_\nu = 0$). This should be contrasted with the KL geometry, where as we said $KL(\delta_x \parallel \delta_y) = +\infty$ whenever $x \neq y$, which is very undesirable for a geometry between measures (does not metrize convergence in distribution).

B.1.2 Kantorovitch Relaxation

Monge formulation is pretty limited. First, for discrete measures, the number of support points in the destination measure must necessarily be smaller than the initial one. Moreover, a transport map might not exist, as when sending a single Dirac to two ones. Each time, the deterministic nature of the transportation makes the method fall short. This motivates the following relaxation.

Definition 5. *(Kantorovitch Formulation for Optimal Transport) Given $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the Kantorovitch formulation for optimal transport consists in determining the following:*

$$K(\mu, \nu) = \inf_{\gamma \in \mathcal{M}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\gamma(x, y). \tag{18}$$

This is much more general, allowing for mass splitting for instance. The resulting infimum is at most Monge distance. But as the Brenier theorem stated, the usual continuous setting results in a degeneracy where they actually agree.
On a computational viewpoint, remark that when the cost function is convex, the problem is convex (constraint is convex too) and becomes

$$\inf_{P\mathbf{1}=M, P^T\mathbf{1}=N} \sum_{1\leq i,j\leq n} C_{ij}P_{ij},$$

which is solved by linear programming. But the reference network simplex algorithm typically scales like $O(n^3 \log n)$ ([BLO]).

Brenier theorem is interesting, but does not account for many contexts, such as discrete distributions. The following theorem assures us the Kantorovitch formulation is always well-defined.

Theorem B.4. *(Existence of Transport Plan, [Tho] Proposition 1.5) Let μ, ν be Radon measures on spaces $\mathcal{X}, \mathcal{Y}$. Assume $c : \mathcal{X} \times \mathcal{Y}$ is lower semicontinuous. Then there exists γ^* such that $K(\mu, \nu)$ is reached.*

Proof. We will first prove that μ, ν is compact, and then extract a minimizing sequence; hypothesis on c will let us prove that limit in μ, ν indeed reaches the infimum. First, $\mu \otimes \nu \in \mathcal{M}(\mu, \nu) \neq \emptyset$. μ, ν being Radon measures, they are inner regular. Take K, L compact sets such that $\mu(K), \nu(L) < \varepsilon/2$; then for any $\gamma \in \mathcal{M}(\mu, \nu)$ we have

$$\gamma(\mathcal{X} \times \mathcal{Y} \setminus K \times L) \leq \gamma((\mathcal{X} \setminus K) \times \mathcal{Y}) + \gamma(\mathcal{X} \times (\mathcal{Y} \setminus L)) = \mu(K) + \nu(L) \leq \varepsilon,$$

which proves that μ, ν is tight; Prokhorov's theorem ([Sch]) shows that the μ, ν is thus relatively compact in the topology of convergence in distribution (its closure is compact). We need to prove that $\mathcal{M}(\mu, \nu)$ is closed. Take a converging sequence $\gamma_n \in \mathcal{M}(\mu, \nu)$ s.t it converges to γ. Take any bounded continuous function f, and define $\hat{f} : (x, y) \mapsto f(x)$; as $\mathbb{E}_{\gamma_n}\hat{f}(X, Y) \to \mathbb{E}_{\gamma}\hat{f}(X, Y)$, we deduce

$$\int_{\mathcal{X}} f(x)d\mu(x) = \int_{\mathcal{X}\times\mathcal{Y}} f(x)d\gamma(x, y) = \int_{\mathcal{X}} f(x)d\Pi_X\gamma(x).$$

It follows that $\Pi_X\gamma = \mu$. Likewise, $\Pi_Y\gamma = \nu$ and $\mathcal{M}(\mu, \nu)$ is indeed compact. Take γ_n a minimizing sequence, by compacity just suppose it converges to γ. Then by the Portmanteau theorem (Theorem 2.1 [Bil68]):

$$K(\mu, \nu) = \lim_{n\to\infty} \mathbb{E}_{\gamma_n} c(X, Y) \geq \mathbb{E}_{\gamma} c(X, Y) = K(\mu, \nu),$$

which ends the proof. □

B.1.3 Metric Properties

OT defines a distance between measures. This is one of the main fact of the study here and one of the main reasons people are interested in the theory, as it introduces a convenient and natural geometry between measures. This property relies on a gluing lemma we now prove.

Lemma B.5. *(Gluing Lemma, [PC19b]) Let μ, ν, η be probability measures on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. Given $\gamma \in \mathcal{M}(\mu, \nu)$ and $\sigma \in \mathcal{M}(\nu, \eta)$, there exists at least a tensor coupling measure ξ such that:*

$$\Pi_{X,Y\#}\xi = \gamma, \quad \Pi_{Y,Z\#}\xi = \sigma.$$

Proof. Essentially disintegration of measures □

We can now construct the Wasserstein distance in a general setting of arbitrary distributions. Recall the tree axioms a distance d on some space $\mathcal{X}$ must verify:

- (Symmetry) $d(x, y) = d(y, x)$
- (Positive definite) $d(x, y) \geq 0$ with equality iff $x = y$
- (Triangle inequality) $d(x, z) \leq d(x, y) + d(y, z)$

Finally, we state the theorem.

Theorem B.6. *(Wasserstein Distance, [PC19b]) Assume $\mathcal{X} = \mathcal{Y}$ and that the cost function can be written $c(x, y) = d(x, y)^p$ for some integer p and distance d on $\mathcal{X}$. Then define $W(\mu, \nu)_{c,p}^p$ the p-Wasserstein distance as*

$$W(\mu, \nu)_{c,p}^p = K(\mu, \nu) = \inf_{\gamma \in \mathcal{M}(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}\ c(X, Y). \tag{19}$$

Remark that $W(\mu, \nu)_{c,p}^p$ depends on the cost function and on p. When $c(x, y) = \|x - y\|^p$, we write $W_{c,p}$ as W_p. As was implied, the p-Wasserstein distance is indeed a distance on $\mathcal{P}(\mathcal{X})$.

Proof. Clearly symmetric and positive. If $W(\mu, \nu)_{c,p} = 0$, since $d(x, x) = 0$ we can load a minimizing coupling on the diagonal $\Delta = (x, x)$ and by positivity of the quantities construct a a minimizing measure γ supported on Δ, denote by $\lambda(x)$ the associated measure on Δ. Then for any bounded continuous function f,

$$\int f(x, y)d\gamma(x, y) = \int f(x, x)d\lambda(x),$$

so since $\gamma \in \mathcal{M}(\mu, \nu)$, actually $\mu = \lambda = \nu$.

For the triangle inequality; take optimal couplings $\gamma \in \mathcal{M}(\mu, \rho), \sigma \in \mathcal{M}(\rho, \nu)$, and glue them to obtain ξ. Define $\rho = \Pi_{X,Z\#}\xi$. Write $\|f(X)\|_{\alpha,p} = (\int f(x)^p d\alpha(x))^{1/p}$. Apply Minkowski inequality to end the proof:

$$\begin{aligned} W_p(\mu, \nu) &\leq \|d(X, Z)\|_{\rho,p} = \|d(X, Z)\|_{\xi,p} \leq \|d(X, Y) + d(Y, Z)\|_{\xi,p} \\ &\leq \|d(X, Y)\|_{\xi,p} + \|d(Y, Z)\|_{\xi,p} \leq \|d(X, Y)\|_{\gamma,p} + \|d(Y, Z)\|_{\sigma,p} \\ &\leq W_p(\mu, \rho) + W_p(\rho, \nu). \end{aligned}$$

□

Theorem B.7. *(Wasserstein distance metrizes convergence in distribution) If $\mathcal{X}$ is compact, $\mu_n \rightharpoonup \mu$ iff $W_{c,p}(\mu_n, \mu) \to 0$. On a non-compact space, the distributions μ_n, μ must also verify convergence of moments up to order p.*

Proof. Admitted, needs duality. □

So the distance constructed possesses the most natural properties we would think of.

Immediate parametric estimation method Minimize $\theta \mapsto W(\alpha_\theta, \beta)$. However, fast computation methods still lack. As we will see, the entropy regularized setting holds much more promising properties.

B.2 Entropic Optimal Transport

B.2.1 Entropic Regularization

Definition 6. *Entropic optimal transport*
With cost function c between two measures μ, ν, add entropy term:

$$K_{c,\sigma^2}(\mu,\nu) = \inf_{\gamma \in \mathcal{M}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y) + \boldsymbol{\sigma^2 KL(\gamma \parallel \mu \otimes \nu)}. \tag{20}$$

Firstly, entropy smooths the optimal transport plan γ_{σ^2}:

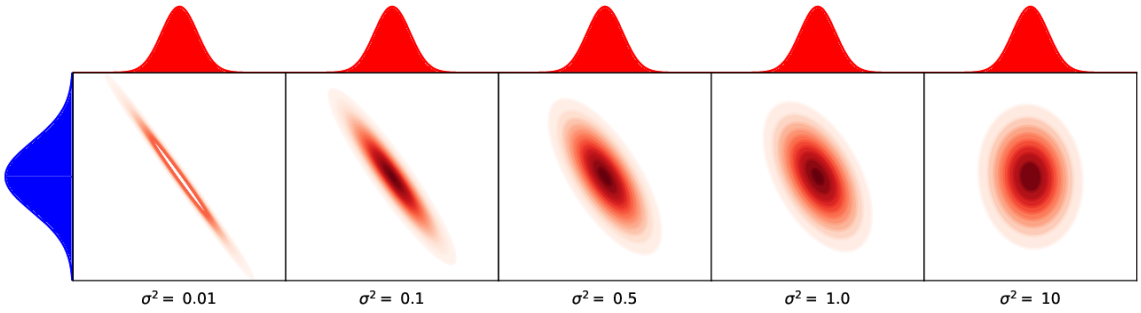

Figure 10: EOT coupling between two Gaussians [Jan+20]

Remark how $\gamma_{\sigma^2} \longrightarrow \mu \otimes \nu$.

Theorem B.8. *(Convergence with ε, [PC19a]) The unique solution γ_ε of 20 converges to the solution with maximal entropy within the set of optimal solutions of the Kantorovitch problem:*

$$\gamma_\varepsilon \xrightarrow{\varepsilon\to 0} \arg\min_\gamma \{KL(\gamma \parallel \mu\otimes\nu) \mid \gamma \in \mathcal{M}(\mu,\nu),\ \mathbb{E}_\gamma c(X,Y) = K(\mu,\nu)\}.$$

In particular

$$K_{c,\varepsilon}(\mu,\nu) \xrightarrow{\varepsilon\to 0} K_c(\mu,\nu).$$

Moreover

$$\gamma_\varepsilon \xrightarrow{\varepsilon\to+\infty} \mu\otimes\nu.$$

Proof. Proposition 12 in [PC19b]. □

Remark This is not a distance now, and $\arg\min_\beta W_\varepsilon(\alpha,\beta) \neq \alpha$. But the problem is also of its own interest; i.e., without letting $\varepsilon \to 0$. Numerous benefits result from regularization. First, smoothing enables optimization methods (**Sinkhorn**). Moreover, as we will see, it is entropic regularization that provides a bridge with MLE, and thus good statistical properties, or robustness. Regularization can be seen as an advantage rather than an approximation error.

B.2.2 Sinkhorn Algorithm

We are in the setting of the regularized problem. The minimized objective is equal to the projection on a Gibbs distribution. Computing that projection while keeping the bistochasticity constraint on P (alternating between $\Pi_x P = \alpha$ and $\Pi_x P = \beta$) constraints gives us a satisfying convergent algorithm, which is Sinkhorn Algorithm. This is noted in [PC19a]. First notice that:

$$\mathbb{E}_P c(X,Y) + \varepsilon KL(P \parallel \alpha \otimes \beta) = \varepsilon KL(P \parallel K), \tag{21}$$

where $dK(x,y) = e^{-c(x,y)/\varepsilon} d\alpha d\beta$ (Gibbs kernel). Write

$$P_\varepsilon = \arg\min_{P \in \alpha,\beta} KL(P \parallel K) := \Pi^{KL}_{\mathcal{M}(\alpha,\beta)}(K).$$

Then by defining

$$\mathcal{C}_\alpha = \{P \mid \Pi_X P = \alpha\} \quad \text{and} \quad \mathcal{C}_\beta = \{P \mid \Pi_Y P = \beta\},$$

one can use Bregman iterative projections to approximate a solution:

$$P^{l+1} = \Pi^{KL}_{\mathcal{C}_\alpha}(P^l), \quad P^{l+2} = \Pi^{KL}_{\mathcal{C}_\beta}(P^{l+1})$$

For finite distributions we have

$$\Pi^{KL}_{\mathcal{C}_\alpha}(P) = \mathrm{diag}(\frac{\alpha}{P\mathbf{1}_m})P \quad \text{and} \quad \Pi^{KL}_{\mathcal{C}_\beta}(P) = P\mathrm{diag}(\frac{\beta}{P^T\mathbf{1}_n}).$$

These iterate are equivalent to

$$u^{l+1} = \frac{\alpha}{Kv^l} \quad v^{l+1} = \frac{\beta}{K^T u^{l+1}}$$

Initialized with an arbitrary positive vector, say $v = \mathbf{1}$. This is proven to converge and we have some useful bounds on the quantities involved.

Theorem B.9. *(Theorem 3, 4, [PC19b]) This algorithm converges for the Hilbert metric at linear rate.*

In this context, measuring the error on the marginal constraints can be an effective stopping criterion ([PC19b]).

C Dual Formulation

First, the dual formulation to the optimization procedure of optimal transport is essential to many proofs of its properties. We will provide what we feel is important for a better understanding of the theory, and for the motivation behind its tools. More importantly for us, this further leads us to the formulation of the **semi-dual**, which admits exciting interpretations and eventually our path into robustness analysis.

C.1 Duality

C.1.1 General Setting for Duality

Consider the usual Lagragian duality setting. We wish to compute the following problem:

$$\min f_0(x) \text{ s.t } f_i(x) \leq 0,\ f_j(x) = 0.$$

Define

$$J(x) = \begin{cases} f_0(x) & \text{when } f_i(x) \leq 0, f_j(x) = 0 \\ +\infty & \text{else} \end{cases}$$

Then define:

$$L(x, \lambda_i, \lambda_j) = f_0(x) + \langle \lambda_i, f_i(x) \rangle + \langle \lambda_j, f_j(x) \rangle, \quad \lambda_i \succeq 0.$$

Thus $L(x, \lambda) \leq J(x)$ and $J(x) = \max_\lambda L(x, \lambda)$. We have the following primal and dual problems:

$$p^* = \min_x J(x) = \min_x \max_\lambda L(x, \lambda) \tag{Primal}$$

$$d^* = \max_\lambda \min_x L(x, \lambda) \tag{Dual}$$

and we always have $d^* \leq p^*$. When there is equality, we say we have strong duality.

C.1.2 Kantorovich Dual

Theorem C.1. *Designate by $D(\alpha, \beta)$ the dual to the Kantorovitch problem. Then*

$$K(\alpha, \beta) \geq D(\alpha, \beta) = \max_{f,g \in L^1, c - f \oplus g \geq 0} \int_{\mathcal{X}} f(x) d\alpha + \int_{\mathcal{Y}} g(y) d\beta. \tag{Kantorovich Dual}$$

Proof. Primal is

$$W^c(\alpha, \beta) = \inf_{P \in \mathcal{M}(\alpha, \beta)} \mathbb{E}_P(c(X, Y)).$$

Define

$$L(P, f, g) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) dP(x, y) + \int_{\mathcal{X}} f(x) d(\alpha - \Pi_X P) + \int_{\mathcal{Y}} g(y) d(\beta - \Pi_Y P).$$

Then the dual problem is:

$$\begin{aligned} d^* &= \max_{f,g \in L^1} \min_{P \geq 0} L(P, f, g) \\ &= \max_{f,g \in L^1} \min_{P \geq 0} \int_{\mathcal{X}} f(x) d\alpha + \int_{\mathcal{Y}} g(y) d\beta + \int_{\mathcal{X} \times \mathcal{Y}} c(x,y) - f(x) - g(y) dP(x,y) \end{aligned}$$

But remark that

$$\min_{P \geq 0} \int_{\mathcal{X} \times \mathcal{Y}} c(x,y) - f(x) - g(y) dP(x,y) = \begin{cases} 0 & \text{if } c - f \oplus g \geq 0 \\ -\infty & \text{else} \end{cases}$$

This leads to the conclusion. □

Theorem C.2. *(Kantorovitch Strong Duality) Assume the cost function c is lower semi-continuous. Then*

$$K(\mu, \nu) = D(\mu, \nu), \tag{22}$$

and the maximum is reached in the Dual problem.

Proof. Lemma 3.2,3.3, theorem 3.4, lemma 3.6 of [Tho]. □

C.2 Semi-Dual

Remember the dual formulation is

$$D_c(\alpha, \beta) = \sup_{c - f \oplus g \geq 0} \int f d\alpha + \int g d\beta.$$

As we said, showing the existence of solutions to the dual problem is non-trivial (Brenier Theorem B.2), and necessitates introducing another tool: the c-transform. Even though we are not planning on proving the latter theorem, we need this concept to get to the semi-dual formulation.

Keeping one function fixed, one can try to minimize with respect to the other. Define the c-transform as:

$$\forall y \in \mathcal{Y},\ f^c(y) := \inf_{x \in \mathcal{X}} c(x,y) - f(x). \tag{23}$$

Thus for a fixed f, $g = f^c$ is a solution to the dual problem. As a side note a useful property there is that c Lipschitz implies f^c, g^c Lipschitz, and this is important to show Kantorovitch Strong Duality. However, one could have thought about designing a maximization strategy alternating between $(f, g) \mapsto (f, f^c) \mapsto (f^{cc}, f^c) \mapsto \cdots$. However, $f^{ccc} = f^c$. This failure in the classical problem is one of the points where the entropic regularization shows its magic; unlike the original Kantorovitch dual formulation, the regularized problem is smooth and strictly convex. Not only all the theorems about existence and strong duality still hold, but we recover even more; among other things, the previous optimization strategy actually works, and is equivalent to Sinkhorn algorithm.

Consider the dual to the regularized optimal transport problem:

Theorem C.3. *(Dual of EOT) The dual to the regularized version of the optimal transport problem admits the following representation:*

$$D_\varepsilon(\alpha,\beta) = \sup_{f,g} \int f d\alpha + \int g d\beta - \varepsilon \int e^{-(c-f\oplus g)/\varepsilon} d\alpha d\beta. \tag{24}$$

Proof.

$$D_\varepsilon(\alpha,\beta) = \sup_{f,g} \int f d\alpha + \int g d\beta + \inf_{P\geq 0} \int c - f \oplus g + \varepsilon \log(\frac{dP}{d\alpha d\beta})\ dP,$$

then just see that the minimum over P is reached at $dP = e^{-(c(x,y)-f(x)-g(y))/\varepsilon} d\alpha d\beta$ by studying the function $\varphi(r) \mapsto (c - f - g + \varepsilon \log(r))r$. □

As proved in [PC19b], existence of functions reaching the maximum and strong duality ($K_\varepsilon(\alpha,\beta) = D_\varepsilon(\alpha,\beta)$) still holds. This time, the c-transforms are as follow. Fix f, and see that on a point in the support of β, we look for the following supremum:

$$\sup_{g(y)} g(y) - \varepsilon e^{g(y)/\varepsilon} \int_{\mathcal{X}} e^{(f(x)-c(x,y))/\varepsilon} d\alpha(x).$$

Thus,

$$f^c(y) = -\varepsilon \log \int e^{(f(x)-c(x,y))/\varepsilon} d\alpha(x).$$

Equivalently

$$g^c(y) = -\varepsilon \log \int e^{(g(y)-c(x,y))/\varepsilon} d\beta(x).$$

As it happens, alternating c-transforms in the discrete case gives a dual Sinkhorn algorithm, that can be adapted to recover its primal version, see [PC19b] Section 8.1, or [CP18] after Corollary 1. Now we finally get to the most important result of this section.

Theorem C.4. *(Semi-Dual Formulation for EOT [Men]) Let $\varepsilon \geq 0$. Then one has:*

$$K_\varepsilon(\alpha,\beta) = D_\varepsilon(\alpha,\beta) = \max_f \int f d\alpha - \varepsilon \int \log \int e^{(f(x)-c(x,y))/\varepsilon} d\alpha d\beta - \varepsilon. \tag{25}$$

The maximum is reached.

Proof. The proof is contained in [CP18] Proposition 2. and [CP15]. Once the existence of a maximum is proved we can just plug-in the appropriate c-transform:

$$\begin{aligned} D_\varepsilon(\alpha,\beta) &= \max_f \int f d\alpha + \int f^c d\beta - \varepsilon \int e^{(f(x)+f^c(y)-c(x,y))/\varepsilon} d\alpha d\beta \\ &= \max_f \int f d\alpha - \varepsilon \int \log \int e^{(f(x)-c(x,y))/\varepsilon} d\alpha d\beta - \varepsilon. \end{aligned}$$

□

Let us continue to build on this semi-dual representation. Write

$$G(f, Y) = \int f d\alpha - \varepsilon \log \int e^{(f(x)-c(x,y))/\varepsilon} d\alpha,$$

such that

$$D_{c,\varepsilon}(\alpha, \beta) = \max_f \mathbb{E}_{Y\sim\beta} G(f, Y).$$

We now get to the final representation we are interested in. Remember our way into parameter estimation is by solving the following projection type problem:

$$\arg\min_\theta D_{c,\varepsilon}(\alpha_\theta, \beta).$$

Theorem C.5. *(Parametric EOT Semi-Dual) Model the initial distribution α with some parametrization $\theta \in \Theta$ s.t $d\alpha_\theta = \alpha_\theta d\alpha$, denote by $d\alpha_{\theta,f} = \frac{e^f}{\mathbb{E}_{\alpha_\theta}(e^f)} d\alpha_\theta$ its f-tilting, let $c(x, y) = c(x - y)$, and take L_ε such that its density is $dL_\varepsilon = e^{-c(x)/\varepsilon} d\alpha$. Then the EOTE θ_{EOT} solves*

$$\theta_{EOT} = \arg\max_\theta \min_f KL(\alpha_\theta \parallel \alpha_{\theta,f/\varepsilon}) + \mathbb{E}_{Y\sim\beta}(\log L_\varepsilon * \alpha_{\theta,f/\varepsilon}(Y)). \tag{26}$$

Proof. See we can write

$$\begin{aligned} G(\theta, f, Y) &= \int f d\alpha_\theta - \varepsilon \log \int e^{(f(x)-c(y-x))/\varepsilon} d\alpha_\theta \\ &= -\varepsilon \int \log \frac{d\alpha_\theta}{d\alpha_{\theta,f/\varepsilon}} d\alpha_\theta - \varepsilon \log L_\varepsilon * \alpha_{\theta,f/\varepsilon}(Y) + cst \\ &= -\varepsilon \left(KL(\alpha_\theta \parallel \alpha_{\theta,f/\varepsilon}) + \log L_\varepsilon * \alpha_{\theta,f/\varepsilon}(Y) \right), \end{aligned}$$

and θ_{EOT} solves $\arg\min_\theta D_{c,\varepsilon}(\alpha_\theta, \beta)$. (Could redo all the calculations with no ε if c is replaced by c/ε). □

The Theorem C.5 uncovers an interesting (adversarial) estimator; the model tries to make the data look implausible while tilting not too far away from a distribution of the parametric family. As a final remark, as pointed out by [Men], we can do the same calculations as in Theorem C.4 and C.5 with g^c to obtain:

$$\theta_{EOT} = \arg\max_\theta \min_g D(\beta \parallel \beta_{g/\varepsilon}) + \mathbb{E}_{X\sim\alpha_\theta}(\log L_\varepsilon * \beta_{g/\varepsilon}(X)). \tag{27}$$

However, the usual setting is to model the observations $Y \sim \beta$, and thus the term $\mathbb{E}_{X\sim\alpha_\theta}(\log L_\varepsilon * \beta_{g/\varepsilon}(X))$ is not the right framework, as usually we do not know/have no proxy to the distribution of X. This approach cannot be very helpful.

D Computing the Estimators: EM Algorithms

Numerically maximizing the objective function of an estimator can be very hard, or impossible, for some statistical models. This is usually the case when the model can be formulated in a simpler fashion by assuming the existence of latent variables (unobserved variables); for instance, in a mixture model, they can indicate the specific mixture component a data point belongs. Expectation-Maximization (EM) type algorithms are used to address this problem.

D.1 Classical EM (MLE)

Assume we are given a parametric statistical model which generates variables X_i, Z_i, where X_i is the observation and Z_i the latent variable. Then the log-likelihood is

$$l(\theta) = \sum_i \log p_\theta(x_i) = \sum_i \log \int_{\mathcal{Z}} p_\theta(x_i|z) p_\theta(z) dz.$$

In the case of a finite mixture model, we have $Z \sim \sum_{j=1}^k \pi_j \delta_j$, so the log-likelihood becomes

$$l(\theta) = \sum_{i=1}^n \log \sum_{j=1}^k \pi_j p_\theta(x_i|z_j).$$

Notice how hard it is to solve, even numerically, as the log-likelihood is not convex. The optimization problem would be easier to solve if, first, the latent variables had been used implicitly to classify the observed data, and if second, the distribution of the data conditionally to the latent variable allows easy optimization. Indeed, given the latent variables, we have

$$l(\theta, z) = \sum \log \pi_{z_i} + \sum \log p_\theta(x_i|z_i),$$

and the subset of parameters corresponding to each mixture components can be estimated separately.
Let us come back to the general setting of non-necessarily mixture models. As we would like to make use of the potential convenient computations had we known the latent variable, the EM algorithm comes into play. Start with a guess θ_0;

- (Expectation) Determine the distribution of the latent variables knowing θ_t, X. To this end remark that

  $$p_{\theta_t}(Z|X) \propto p_{\theta_t}(X|Z) p_{\theta_t}(Z),$$

  and the normalization constant lets us recover these values entirely. Then compute

  $$\begin{aligned} Q(\theta|\theta_t) &= \mathbb{E}_{Z \sim p_{\theta_t}(\cdot|X)}(\log p_\theta(X, Z)) \\ &= \int_{\mathcal{Z}} l(\theta_t, z) p_{\theta_t}(z|x) dz \\ &= \sum_{i=1}^n \int_{\mathcal{Z}} l(\theta_t, z, x_i) p_{\theta_t}(z|x_i) dz. \end{aligned}$$

- (Maximization) Find the parameter that maximizes this quantity

$$\theta_{t+1} = \arg\max_{\theta} F(\theta|\theta_t).$$

Let us give the simple example of a Gaussian Mixture Model.

Example D.1. *(Gaussian Mixture Model) Consider $X|Z = z_j$ to be distributed as $\mathcal{N}(\mu_j, \Sigma_j)$. Start the EM algorithm with a guess θ_0.*

- *(Expectation) Focus first on the conditional probability $p_\theta(Z|X)$. See how $p_\theta(Z|X) = \prod_{i=1}^n p_\theta(Z_i|X_i) \propto \prod_{i=1}^n \pi_{z_i} p(X_i|Z_i)$. Denote by p_{ij} the probability that sample i comes from mixture j. Then*

$$p_{ij} \propto \pi_j p(X_i|Z_i = j).$$

This lets us finally compute the expectation step:

$$Q(\theta|\theta_t) = \sum_{i=1}^n \sum_{j=1}^k l(\theta_t, z_i = j, x_i) p_{\theta_t}(z_i = j|X) = \sum_{i=1}^n \sum_{j=1}^k p_{ij}(\log \pi_j + \log p(x_i|z_i = j)).$$

- *(Maximization) Define $n_j = \sum_{i=1}^n p_{ij}$ the expected number of observations belonging to mixture j. $Q(\theta|\theta_t)$ is maximal for the following empirical distribution, mean and covariance matrices:*

$$\hat{\pi}_j = \frac{n_j}{n}, \quad \hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^n p_{ij} x_i, \quad \hat{\Sigma}_j = \frac{1}{n_j} \sum_{i=1}^n p_{ij}(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T. \tag{28}$$

The total complexity of each iteration is $O(nk)$.

Another example, that we use in our subsequent computations, for Sinkhorn EM.

Example D.2. *(Symmetric Gaussian Mixture Model) Consider $X|Z = z_j$ to be distributed as $\mathcal{N}((-1)^j \theta, 1)$, $j = 0, 1$.*

- *(Expectation) Compute*

$$Q(\theta|\theta_t) = \sum_{i=1}^n p_{i0}(\log \pi_0 - \|x_i - \theta\|^2) + p_{i1}(\log \pi_1 - \|x_i + \theta\|^2).$$

- *(Maximization) $Q(\theta|\theta_t)$ is maximal for the following empirical distribution and mean:*

$$\hat{\pi}_j = \frac{\sum_i p_{ij}}{n}, \quad \hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n (2x_i - 1) p_{ij}. \tag{29}$$

(Proof; simply use Cauchy-Schwarz and minimize a quadratic polynomial).

Thus, starting from initial parameter θ_0, one can iterate to obtain a sequence expected to converge to a good approximation of the optimal parameter θ^*. We will now prove that this algorithm produces a non-decreasing sequence of log-likelihoods, which guarantees that the procedure converges to a local maximum.

Theorem D.3. *Each step of the EM algorithm increases the log-likelihood:*

$$l(\theta_t) \leq l(\theta_{t+1}),$$

and strictly so whenever θ_t is not a local optima.

Proof. In [Bon], Section 4. □

However, it is well-known that the EM algorithm does not converge to the true maximum in general. One also has to be careful to eliminate degenerate solutions, for instance where the algorithm might classify one and only point to a specific mixture component. In general, the strategy only consists in re-running the algorithm multiple times, from different seed points. One could design smart parameter initialization, alike what is done in k-means++ for instance. For Gaussian Mixture Models, the choice of initial value of the covariance matrix is critical too. Indeed, σ^2 the typical variance should be chosen with knowledge of the following equality:

$$\frac{1}{n}\sum_{i=1}^{n} \|x_i - \overline{x}\|^2 = \frac{1}{n^2}\sum_{i,j} \|x_i - x_j\|^2$$

If the variance is too big compared to the typical square distance between points, then the empirical distributions will tend to be uniform and the cluster centers will converge to the center of mass $\overline{x}$.

D.2 Maximization-Maximization Approach

Introduce the F-functional:

$$F(\mu, \theta) = \mathbb{E}_{Z\sim\mu} l_\theta(Y \mid X) - KL(\mu \parallel \mu_X), \tag{30}$$

where μ_X is the true mixing distribution of X. Computations in 2.1 show that standard EM can be reformulated as:

- E-step: Let $P^{t+1} = \arg\max_P \mathbb{E}_{Y\sim\mu_Y} F_Y(P(\cdot|Y), \theta^t)$.
- M-step: Let $\theta^{t+1} = \arg\max_\theta \mathbb{E}_{Y\sim\mu_Y} F_Y(P^{t+1}(\cdot|Y), \theta)$.

D.3 Sinkhorn EM (EOTE)

With the F-Functional, the Sinkhorn EM appears as a very straightforward readaptation of the EM algorithm. It suffices to slightly change the EM step;

- E-step: Let $P^{t+1} = \arg\max_{P\in\mu_\mathcal{X},\mu_\mathcal{Y}} \mathbb{E}_{Y\sim\mu_Y} F_Y(P(\cdot|Y), \theta^t)$.

- M-step: Let $\theta^{t+1} = \arg\max_\theta \mathbb{E}_{Y\sim\mu_Y} F_Y(P^{t+1}(\cdot|Y), \theta)$.

Indeed as seen in 2.1 (mainly using Lemma 2.1), it is equivalent to:

- E-step: Let $P^{t+1} = \arg\min_{P\in\mu_\mathcal{X},\mu_\mathcal{Y}} \mathbb{E}_P g_{\theta^t}(X,Y) + D(P \parallel \mu_x \otimes \mu_Y)$ (Sinkhorn).
- M-step: Let $\theta^{t+1} = \arg\min_\theta \mathbb{E}_{P^{t+1}} g_{\theta^t}(X,Y)$.

And the sequence $L(\theta^t)$ is decreasing, strictly so whenever θ^t is not a stationary point of L ([Men+20]).