Research statement

Paul Freulon

June 1, 2023

1 Introduction

Since the beginning of my PhD, I worked on optimal transport and its applications to statistics. I started working on these fields through a biological application: the analysis of flow cytometry data. In this document, the focus is on optimal transport as a tool for statistical inference. However, the research I will present has been motivated by applied biological questions.

2 Optimal transport

Informally stated, the optimal transport problem is to find the most efficient way to move a pile of sand from an area $\mathcal{X}$ toward an area $\mathcal{Y}$. This problem dates back to G.Monge in his *Mémoire sur la Théorie des Déblais et des Remblais* in 1781. In this document, we will work with a more modern formulation that L.Kantorovich introduced in [11].

For μ and ν two probability distributions on $\mathbb{R}^d$, we denote by $\Pi(\mu,\nu)$ the set of probability distributions with marginals μ and ν. With these notations, the optimal transport problem of Kantorovich reads

$$\mathcal{T}_0(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|^2 \mathrm{d}\pi(x,y). \tag{2.1}$$

We refer to the quantity $\mathcal{T}_0(\mu,\nu)$ as the optimal transport cost. Due to the applied context of my research, I often work with discrete measures such as $\mu = \sum_{i=1}^n a_i\delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j\delta_{y_j}$. Here, a and b denote probability vectors of $\mathbb{R}^n$ and $\mathbb{R}^m$, while $\{x_1,\dots,x_n\}$ and $\{y_1,\dots,y_m\}$ are the respective supports of μ and ν. In this case, the optimal transport problem (2.1) reads

$$\mathcal{T}_0(\mu,\nu) = \inf_{\pi\in\Pi(a,b)} \sum_{\substack{1\le i\le n,\\ 1\le j\le m}} \|x_i-y_j\|^2\pi_{i,j}. \tag{2.2}$$

In this discrete framework, the set of constraints $\Pi(\mu,\nu)$ of Problem (2.1) turns into the set of coupling matrices between a and b, that we denote by $\Pi(a,b) = \{\pi\in\mathbb{R}_+^{n\times m} \mid \pi 1_m = a \text{ and } \pi^T 1_n = b\}$.

From a theoretical point of view, a natural question is the well-posedness of problem (2.1). On the applied side, we are searching some efficient algorithms to solve problem (2.2).

These two questions were already well understood when I started working on these subjects. For the well-posedness of problem (2.1), its structure ensures the existence of a solution π^* to this minimization problem. The uniqueness of this solution π^* is a more involved question and requires additional assumptions. For instance, Y.Brenier proved in [2], that if μ is absolutely continuous with respect to the Lebesgue measure, then the solution π^* to the Kantorovich problem is unique. However, when both distributions are discrete, we can propose some elementary counter-examples to the uniqueness of a solution to the transport problem (2.2).

On the applied side, problem (2.2) is a linear optimization problem. Therefore, when both distributions are discrete with cardinal n, there is no algorithm to solve problem (2.2) with less than $\mathcal{O}(n^3 \log(n))$ operations [13, 12, 5].

To mitigate the computational cost of optimal transport, a recent technique introduced in [5] relies of the addition of a regularizing term to the optimal transport problem. In [5], M.Cuturi regularized the optimal transport thanks to a Kullback-Leibler penalization. We denote by KL this divergence and remind that for $\pi \in \Pi(\mu,\nu)$, it is defined by $\mathrm{KL}(\pi|\mu\otimes\nu) = \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu\otimes\nu}(x,y)\right)\mathrm{d}\pi(x,y)$ if π is absolutely continuous with respect to $\mu\otimes\nu$, and $\mathrm{KL}(\pi|\mu\otimes\nu) = +\infty$ otherwise. Thus, for μ and ν two probability distributions with compact supports, the entropic optimal transport problem with regularization parameter $\lambda \geq 0$ is

$$\mathcal{T}_\lambda(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|^2 \mathrm{d}\pi(x,y) + \lambda\,\mathrm{KL}(\pi|\mu\otimes\nu). \tag{2.3}$$

We refer to $\mathcal{T}_\lambda(\mu,\nu)$ as the regularized optimal transport cost. The regularizing term provides a computational advantage with respect to the unregularized optimal transport problem. In the case μ and ν have discrete supports of size ν, Sinkhorn algorithm [15] allows to compute $\mathcal{T}_\lambda(\mu,\nu)$ in $\mathcal{O}(n^2\log(n))$ operations. Moreover, the strong convexity of the Kullback-Leibler divergence ensures the uniqueness of a solution π^*_λ to the regularized transport problem (2.3).

Regularized [4] or not [14], the optimal transport problem admits a dual formulation. That is, substituting the constraint set $\Pi(\mu,\nu)$ by dual variables $(\varphi,\psi) \in L^\infty(\mu)\times L^\infty(\nu)$ allows to rewrite the optimal transport problem as a maximization problem. Thus, for $\lambda \geq 0$, the optimal transport cost between μ and ν reads

$$\mathcal{T}_\lambda(\mu,\nu) = \max_{\substack{\varphi\in L^\infty(\mu),\\ \psi\in L^\infty(\nu)}} \int_{\mathcal{X}} \varphi(x)\mathrm{d}x + \int_{\mathcal{Y}} \psi(y)\mathrm{d}y - \int_{\mathcal{X}\times\mathcal{Y}} m_\lambda(\varphi(x)+\psi(y)-\|x-y\|^2)\mathrm{d}\mu(x)\mathrm{d}\nu(y), \tag{2.4}$$

where m_λ is a constraint function defined as follows.

$$\text{For every } t\in\mathbb{R},\ m_\lambda(t) = \begin{cases} +\infty 1_{t>0} & \text{if } \lambda = 0 \\ \lambda(e^{\frac{t}{\lambda}} - 1) & \text{otherwise.} \end{cases}$$

Numerical schemes, such as the Sinkhorn algorithm, are based on this dual problem. Moreover, statistical properties of the transport cost are established thanks to this dual formulation.

3 A statistical point of view

In this section, I present two questions I studied during my PhD.

3.1 Estimation of an optimal transport cost

In statistics, a probability distributions μ on $\mathbb{R}^d$ is only accessible through its samples $X_1, \dots, X_n \sim \mu$. Then, the purpose is to derive some properties of μ based on the available samples. In optimal transport, two probability distributions are compared. We will thus assume to have access to a second series of observations $Y_1, \dots Y_n$, distributed with respect to an other probability measure ν. I studied the estimation of the optimal transport cost $\mathcal{T}_0(\mu, \nu)$ thanks to the observations from μ and ν. An existing, and natural strategy to estimate $\mathcal{T}_0(\mu, \nu)$ is to substitute μ and ν by their empirical versions, respectively defined by $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}$. Thus, the estimator of the optimal transport cost between μ and ν is defined by $\mathcal{T}_0(\hat{\mu}_n, \hat{\nu}_n)$.

In my previous work, I studied an other class of estimators defined thanks to regularized costs. For instance, I analyzed the estimation error of the family of estimators $(\mathcal{T}_\lambda(\hat{\mu}_n, \hat{\nu}_n))_{\lambda \geq 0}$, with $\mathcal{T}_\lambda$ defined by problem (2.3). The purpose was to derive rates of convergence of $\mathcal{T}_\lambda(\hat{\mu}_n, \hat{\nu}_n)$ toward $\mathcal{T}_0(\mu, \nu)$. To this end, I work under the assumption that both measures μ and ν have compact supports. I showed that an appropriate choice of the regularization parameter allows to reach the following non-asymptotic result. If n i.i.d. observations are available both for μ and ν, then,

$$\mathbb{E}\left[|\mathcal{T}_{\lambda_n}(\hat{\mu}_n, \hat{\nu}_n) - \mathcal{T}_0(\mu, \nu)|\right] \leq C n^{-2/d} \log(n). \tag{3.1}$$

This result derives from a decomposition of the error into an approximation term $|\mathcal{T}_\lambda(\mu, \nu) - \mathcal{T}_0(\mu, \nu)|$, and an estimation term $|\mathcal{T}_\lambda(\mu, \nu) - \mathcal{T}_\lambda(\hat{\mu}_n, \hat{\nu}_n)|$. Regarding the approximation term, I relied on error bounds established in [6]. In this article, Genevay et al. proved that $|\mathcal{T}_\lambda(\mu, \nu) - \mathcal{T}_0(\mu, \nu)|$ decreases with λ at the pace $\mathcal{O}(\lambda \log(\lambda))$. For the estimation term, I showed that it could be controlled by the supremum of an empirical process already studied in [4]. My study resulted on the following inequality,

$$\mathbb{E}\left[|\mathcal{T}_\lambda(\mu, \nu) - \mathcal{T}_\lambda(\hat{\mu}_n, \hat{\nu}_n)|\right] \leq C n^{-2/d}, \tag{3.2}$$

with a constant C independent of λ. This upper bound (3.2) allows for practical choice of regularizing parameter λ_n that depends on the number of observations n and their dimension d. From this choice of regularizing parameter, I could derive inequality (3.1).

3.2 Fitting a model with an optimal transport criterion

I came to the estimation of an optimal transport cost while fitting a model with an optimal transport criterion. In my PhD, I focused on mixture models parameterized by weights vector. However, under appropriate assumptions, these results extend to other statistical models. Let us assume to have a probability model $\{\mu_\theta \mid \theta \in \Theta\}$ where Θ denotes the parameter space. Then, given an unknown probability distribution ν, the purpose is to find a good approximation of ν inside $\{\mu_\theta \mid \theta \in \Theta\}$. More precisely, we denote by μ_{θ^*} the closest distribution, of the model, to ν and define it as follows.

$$\theta^* := \operatorname*{arg\,min}_{\theta \in \Theta} \mathcal{T}_0(\mu_\theta, \nu). \tag{3.3}$$

We notice that if ν belongs to the model $\{\mu_\theta \mid \theta \in \Theta\}$, and if this model is identifiable, then $\mu_\theta^* = \nu$. Let me also point out that problem (3.3) requires the measure ν to be known.

In the statistical framework where I work, the probability measure ν is only accessible through samples $Y_1, \dots, Y_n \sim \nu$. Therefore we substitute ν by its empirical version $\hat{\nu}_n$. Then, two arguments

motivate the use of a regularized cost $\mathcal{T}_\lambda$ instead of $\mathcal{T}_0$ in Problem (3.3). First, the regularized transport cost features an algorithmic improvement over the standard optimal transport problem. Second, the well-posedness of the function $\theta \mapsto \mathcal{T}_0(\mu_\theta, \hat{\nu}_n)$ is not ensured if the measures μ_θ are not absolutely continuous with respect to the Lebesgue measure. Therefore, I preferred to rely on the entropic optimal transport cost $\mathcal{T}_\lambda$ to compare $\hat{\nu}_n$ to the measures μ_θ. Hence, the family of estimators $(\hat{\theta}_\lambda)_{\lambda>0}$ I studied was defined thanks to the regularized optimal transport cost as follows.

$$\text{For } \lambda > 0,\ \hat{\theta}_\lambda := \operatorname*{arg\,min}_{\theta\in\Theta} \mathcal{T}_\lambda(\mu_\theta, \hat{\nu}_n). \tag{3.4}$$

Building on similar techniques as when estimating the optimal transport cost, I could derive non-asymptotic rates of convergence of $\hat{\theta}_\lambda$ toward θ^*. Under the assumption that there exists a bounded set $\mathcal{X}$ such that every μ_θ has its support in $\mathcal{X}$, it holds true that

$$R(\hat{\theta}_{\lambda_n}) := \mathbb{E}\left[\mathcal{T}_0(\mu_{\hat{\theta}_{\lambda_n}}, \nu) - \mathcal{T}_0(\mu_{\theta^*}, \nu)\right] \leq C n^{-2/d} \log(n). \tag{3.5}$$

As θ^* is defined as a minimizer of $\theta \mapsto \mathcal{T}_0(\mu_\theta, \nu)$, the quantity $R(\hat{\theta}_\lambda)$ is always positive or zeros. Again, the rate of convergence established in equation (3.5) is due to a specific choice of the regularizing parameter. Computing the minimizer of an optimal transport criterion, such as in equation (3.4), is also a question I addressed for the specific case of discrete mixture models.

4 Perspectives

My research could pursue on several directions. While I focused on *non*-asymptotic results in my previous works, I would be interested to study asymptotic properties of regularized transport cost. For instance, central limit theorems have been proven in [1] for regularized optimal transport costs on *discrete* spaces. Exploiting recent techniques applied in [10], it might be possible to establish central limit theorems for regularized optimal transport costs in a more general setting. In the same line of research, I would be motivated to study the asymptotic normality of estimators defined as minimum of regularized optimal transport criterion, such as in equation (3.4).

An other possibility of research could be to develop multivariate goodness-of-fit tests based on entropic optimal transport. Such tests have already been studied in [9], when exploiting unregularized transport costs. However, I do not know similar results when substituting the classic transport cost by its entropic regularized version.

I could also start working on the recent topic of distributional regression based on optimal transport. In this area of research, the regression function maps a probability distribution to an other probability distribution. In works such as [7], the estimation of the regression map is based on the Fréchet mean of probability distributions. I would be interested to apply entropic optimal transport barycenter techniques, recently proposed in [3], to the distributional regression problem. Indeed, the regularized strategy of [3] ensures the well posedness of the Fréchet mean problem, and allows the application of efficient algorithms. In the same direction, it would be possible to exploit these regularized transport barycenters to extend autoregression models already studied in [8].

Finally, I am ready to work on other questions related to the statistical properties of entropic optimal transport. I would also be interested to work on other topics in statistic, or optimal transport to improve my knowledge of these fields.

References

[1] J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. 2019.

[2] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[3] L. Chizat. Doubly regularized entropic wasserstein barycenters. *arXiv preprint arXiv:2303.11844*, 2023.

[4] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

[5] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[6] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR, 2019.

[7] L. Ghodrati and V. M. Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974, 2022.

[8] L. Ghodrati and V. M. Panaretos. On distributional autoregression and iterated transportation. *arXiv preprint arXiv:2303.09469*, 2023.

[9] M. Hallin, G. Mordant, and J. Segers. Multivariate goodness-of-fit tests based on wasserstein distance. 2021.

[10] S. Hundrieser, M. Klatt, T. Staudt, and A. Munk. A unifying approach to distributional limits for empirical optimal transport. *arXiv preprint arXiv:2202.12790*, 2022.

[11] L. V. Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382, 2006.

[12] O. Pele and M. Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

[13] 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.

[14] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[15] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.