




Domain : Data Science

Scholar Year : 2023-2024

Optimal Transport : Entropic Maps

Author : Reynaud Nils **Promotion :** 2024

Teachers : Marco Cuturi

1 Entropic Maps

1.1 General view of the paper

Optimal transport theory has been used in machine learning to study and characterize maps that can push-forward efficiently a probability measure onto another. Recent works have drawn inspiration from Brenier's theorem, which states that when the ground cost is the squared-Euclidean distance, the best map to morph a continuous measure in $\mathcal{P}()$ into another must be the gradient of a convex function.
The goal of optimal transport is to find a map between two probability distributions that minimizes the squared Euclidean transportation cost. This formulation leads to what is known as the Monge problem.

$$\min_{T \in \mathcal{T}(P,Q)} \int \frac{1}{2} \|x - T(x)\|_2^2 \, \mathrm{d}P(x) \tag{1}$$

where P and Q are two probability measures, and T (P, Q) is the set of admissible solutions. A solution to the Monge problem is guaranteed to exist if P and Q have finite second moments and P is absolutely continuous. In many samples we do not know P and Q but rather have samples X1, . . . , Xn P and Y1, . . . , Yn Q, and would like to estimate the optimal plan.
Previously, Hütter and Rigollet investigated this question and came up with an estimator. However, its computational cost scales exponentially in the dimension.
This paper proposes a new estimator, based on recent advances in entropic regularization. The new minimization problem to solve is the following :

$$\inf_{\pi \in \Pi(P,Q)} \int \int \frac{1}{2} \|x - y\|^2 \, d\pi(x, y) + \epsilon D_{KL}(\pi \| P \otimes Q),$$

where (P, Q) denotes the set of couplings between P and Q and DKL is the Kullback–Leibler divergence.
While OT problems can be solved as linear programs, adding an entropic smoothing term is known to result in solvers that are faster and more robust to outliers, differentiable and easier to parallelize. Entropic regularization is thus widely used as it leads to a problem that can be solved via the Sinkhorn algorithm.
The Sinkhorn fixed point algorithm is the cornerstone of this approach. Multiple attempts have been made to shorten its runtime using, for instance, annealing, momentum or acceleration. A better initialization of this algorithm could also prove useful and result in dramatic speed-ups.
This paper tries to develop estimator of optimal transport maps with convergence guarantees. It offers the first finite-sample convergence guarantees.
The estimator introduced is defined as the barycentric projection of the entropically optimal coupling between the empirical measures arising from the samples. The barycentric projection of the entropically optimal coupling is the gradient of the function which solves the dual problem.

1.2 Brenier's theorem

Central in Optimal transport is Brenier's theorem.
Let $P \in \mathcal{P}_{ac}(\Omega)$ and $Q \in \mathcal{P}(\Omega)$. Then

1. there exists a solution T_0 to Eq. (1), with $T_0 = \nabla \varphi_0$, for a convex function φ_0 solving

$$\inf_{\varphi \in L^1(P)} \int \varphi \mathrm{d}P + \int \varphi^* \, \mathrm{d}Q \tag{4}$$

where φ^* is the convex conjugate to φ.

2. If in addition $Q \in \mathcal{P}_{ac}(\Omega)$, then $\nabla\varphi_0^*$ is the optimal transport map from Q to P.

If P does not have a density, a convex relaxation gives us a new problem.

$$\frac{1}{2}W_2^2(P,Q) := \min_{\pi\in\Pi(P,Q)} \int \frac{1}{2}\|x-y\|_2^2 \, \mathrm{d}\pi(x,y) \tag{5}$$

where

$$\Pi(P,Q) := \{\pi \in \mathcal{P}(\Omega\times\Omega) \mid \pi(A\times\Omega) = P(A), \pi(\Omega\times A) = Q(A)\}$$

Unlike the first problem, this new one always admits a solution, as long as P and Q have finite moments. This solution is the optimal plan π_0
The solution-pair of its dual are the optimal potentials.
As stated previously, adding an entropic regularization term transfers us to entropic optimal transport.
$S_\varepsilon(P,Q) := \inf_{\pi\in\Pi(P,Q)} \iint \frac{1}{2}\|x-y\|_2^2 \, \mathrm{d}\pi(x,y) + \varepsilon D_{\mathrm{KL}}(\pi\|P\otimes Q)$
The solution to this new problem is the optimal entropic plan, π_ϵ, while the solutions to the dual are the optimal entropic potentials.
Several recent works have bridged the regularized and unregularized optimal transport regimes, with particular interest in the setting where $\to 0$.

1.3 Estimator

Given the optimal entropic plan π_ε between P and Q, its barycentric projection can be defined by

$$T_\varepsilon(x) := \int y \, \mathrm{d}\pi_\varepsilon^x(y) = \mathbb{E}_{\pi_\varepsilon}[Y \mid X = x] \tag{12}$$

An extension to R^d is then :

$$T_\varepsilon(x) := \frac{\int y e^{\frac{1}{\varepsilon}\left(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2\right)} \mathrm{d}Q(y)}{\int e^{\frac{1}{\varepsilon}\left(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2\right)} \mathrm{d}Q(y)} \tag{13}$$

The authors call this new element the entropic map.
Proposition 2. Let $(f_\varepsilon, g_\varepsilon)$ be optimal entropic potentials satisfying Eq. (9), and let T_ε be the entropic map. Then $T_\varepsilon = (Id - \nabla f_\varepsilon)$.
The authors are interested in the finite samples case.
In that case, we can write $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ for the empirical distributions corresponding to the samples from P and Q. The proposed estimator is $T_{\varepsilon,(n,n)}$, the entropic map between P_n and Q_n, which can be written

$$T_{\varepsilon,(n,n)}(x) = \frac{\frac{1}{n}\sum_{i=1}^n Y_i e^{\frac{1}{\varepsilon}\left(g_{\varepsilon,(n,n)}\right)(Y_i) - \frac{1}{2}\|x-Y_i\|^2\right)}}{\frac{1}{n}\sum_{i=1}^n e^{\frac{1}{\varepsilon}\left(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2\right)}} \tag{14}$$

where $g_{\varepsilon,(n,n)}$ is the entropic potential corresponding to Q_n in the optimal entropic plan between P_n and Q_n, which can be obtained as part of the output of Sinkhorn's algorithm : it is one of the dual potentials.

The authors then prove quantitative rates of convergence for their estimator, with the following regularity assumptions on P and Q.

(A1) $P, Q \in \mathcal{P}_{\text{ac}}(\Omega)$ for a compact set Ω, with densities satisfying $p(x), q(x) \leq M$ and $q(x) \geq m > 0$ for all $x \in \Omega$,
(A2) $\varphi_0 \in \mathcal{C}^2(\Omega)$ and $\varphi_0^* \in \mathcal{C}^{\alpha+1}(\Omega)$ for $\alpha > 1$,
(A3) $T_0 = \nabla\varphi_0$, with $\mu I \preceq \nabla^2\varphi_0(x) \preceq LI$ for $\mu, L > 0$ for all $x \in \Omega$.

The two following results help to prove the main one.

Theorem 4. Under assumptions (A1) to (A3) there exists a constant $\varepsilon_0 > 0$ such that for $\varepsilon \leq \varepsilon_0$, the entropic map $T_{\varepsilon,n}$ between P and Q_n satisfies

$$\mathbb{E}\,\|T_{\varepsilon,n} - T_0\|^2_{L^2(P)} \lesssim \varepsilon^{1-d'/2}\log(n)n^{-1/2} + \varepsilon^{(\bar{\alpha}+1)/2} + \varepsilon^2 I_0(P,Q) \tag{15}$$

with $d' = 2\lceil d/2\rceil$ and $\bar{\alpha} = \alpha \wedge 3$. Choosing $\varepsilon \asymp n^{-\frac{1}{d'+\bar{\alpha}-1}}$, we get the one-sample estimation rate

$$\mathbb{E}\,\|T_{\varepsilon,n} - T_0\|^2_{L^2(P)} \lesssim (1 + I_0(P,Q))\, n^{-\frac{\bar{\alpha}+1}{2(d'+\bar{\alpha}-1)}} \tag{16}$$

Theorem 5. Let $T_{\varepsilon,(n,n)}$ be the entropic map from P_n to Q_n, and let $T_{\varepsilon,n}$ be as in Theorem 4, Under assumptions ($\boldsymbol{A}$1) to (A3), for $\varepsilon \leq 1, T_{\varepsilon,(n,n)}$ satisfies

$$\mathbb{E}\left\|T_{\varepsilon,(n,n)} - T_{\varepsilon,n}\right\|^2_{L^2(P)} \lesssim \varepsilon^{-d'/2}\log(n)n^{-1/2}$$

where $d' = 2\lceil d/2\rceil$.

These two theorems combined give the main result, the convergence guarantee for the new estimator.

Theorem 3. Under assumptions (A1) to (A3), the entropic map $\hat{T} = T_{\varepsilon,(n,n)}$ from P_n to Q_n with regularization parameter $\varepsilon \asymp n^{-\frac{1}{d+\alpha+1}}$ satisfies

$$\mathbb{E}\left\|\hat{T} - T_0\right\|^2_{L^2(P)} \lesssim (1 + I_0(P,Q))\, n^{-\frac{(\bar{\alpha}+1)}{2(d'+\bar{\alpha}+1)}}\log n$$

where $d' = 2\lceil d/2\rceil$ and $\bar{\alpha} = \alpha \wedge 3$.

The authors prove successfully theorem 4, the one-sample estimates case, and 5, two-sample estimates, with duality arguments.
We show below the transported points obtained via Entropic Maps for some synthetic data.

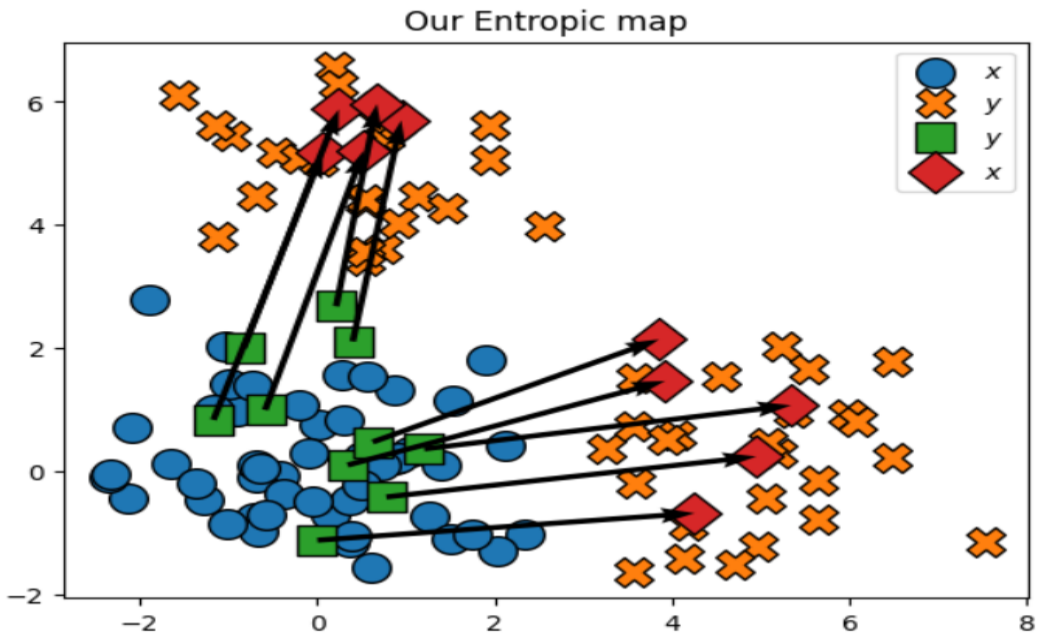

FIGURE 1 – Example of transportation via Entropic Maps

1.4 Computational aspects

The entropic map as an optimal transport estimator has strong computational benefits. The authors compare it to other estimators such as the wavelet-based estimator proposed by Hütter and Rigollet (2021) or the "1-Nearest Neighbor" by Manole et al.
It has the closed-form representation :

$$\hat{T}_{\varepsilon,(n,n)}(x) = \frac{\sum_{i=1}^{n} Y_i e^{\frac{1}{\varepsilon}\left(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2\right)}}{\sum_{i=1}^{n} e^{\frac{1}{\varepsilon}\left(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2\right)}} \tag{25}$$

Computing this estimator requires computing the optimal entropic potentials. The most popular way to do this is to use Sinkhorn's algorithm, an alternating maximization algorithm that computes approximations of the entropic potentials by iteratively updating f and g so that they satisfy one of the two dual optimality conditions. Starting with f^0, it does until termination :

$$g^{(k)}(y) = -\varepsilon \log \frac{1}{n} \sum_{i=1}^{n} e^{\frac{1}{\varepsilon}\left(f^{(k)}(X_i) - \frac{1}{2}\|X_i-y\|^2\right)}$$

$$f^{(k+1)}(y) = -\varepsilon \log \frac{1}{n} \sum_{j=1}^{n} e^{\frac{1}{\varepsilon}\left(g^{(k)}(Y_j) - \frac{1}{2}\|x-Y_j\|^2\right)}$$

To analyze the running time of this estimator, we consider the entropic map estimator obtained after k iterates of Sinkhorn's algorithm :

$$T^{(k)}(x) = \frac{\sum_{i=1}^{n} Y_i e^{\frac{1}{\varepsilon}\left(g^{(k)}(Y_i) - \frac{1}{2}\|x - Y_i\|^2\right)}}{\sum_{i=1}^{n} e^{\frac{1}{\varepsilon}\left(g^{(k)}(Y_i) - \frac{1}{2}\|x - Y_i\|^2\right)}} \tag{26}$$

The authors prove the following theorem, that states that T (k) is an acceptable estimator if k is sufficiently large.
Theorem 6. Suppose assumptions (A1) to (A3) hold, and we choose ε as in Theorem 3 . Then for any $k \gtrsim n^{7/(d'+\bar{\alpha}+1)} \log n$,

$$\mathbb{E}\left\|T^{(k)} - T_0\right\|^2_{L^2(P)} \lesssim (1 + I_0(P,Q))\, n^{-\frac{(\bar{\alpha}+1)}{2(d'+\bar{\alpha}+1)}} \log n$$

where $d' = 2\lceil d/2 \rceil$ and $\bar{\alpha} = 3 \wedge \alpha$. In particular, an estimator achieving the same rate as the estimator in Theorem 3 can be computed in $\tilde{O}\left(n^{2+7/(d'+\bar{\alpha}+1)}\right) = n^{2+o_d(1)}$ time.
The authors then test the computational efficiency of their algorithm. Reproducing their experiments gives us the following for their estimator, on synthetic data.

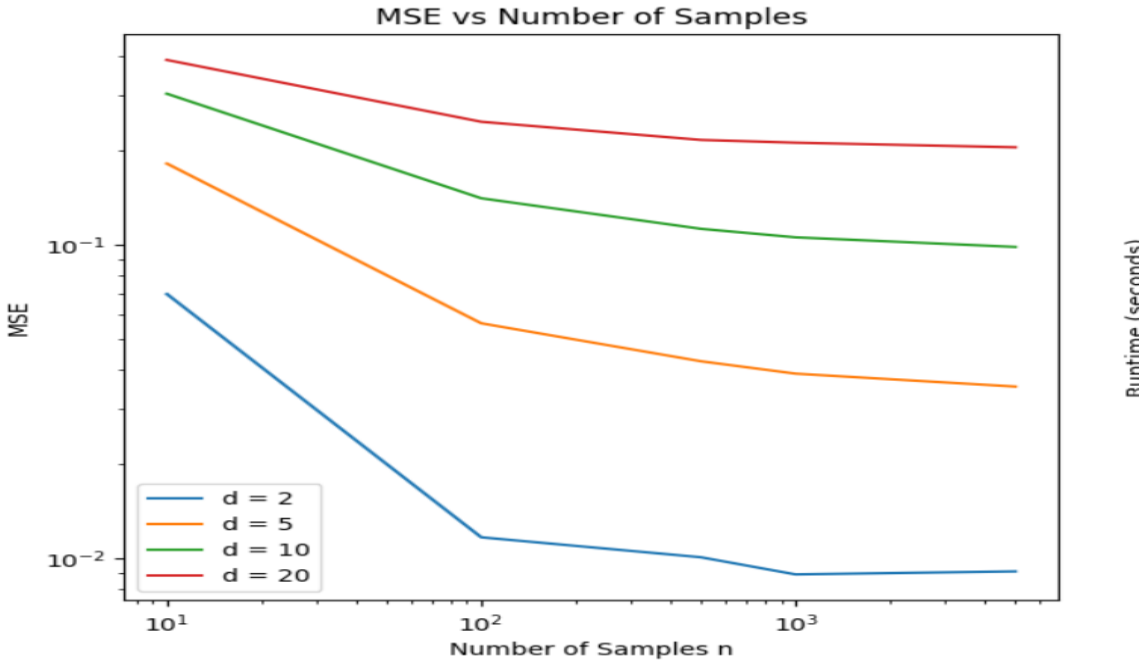

Figure 2 – Using CPU, MSE vs Number of Samples

As you can see, when the number of samples in the data reaches 10^4, the runtimes goes through the roof when using a CPU. However, as this estimator relies on Sinkhorn's algorithm, which can be parallelized, it runs efficiently on GPU's compared to its competitors, which solves the problem.

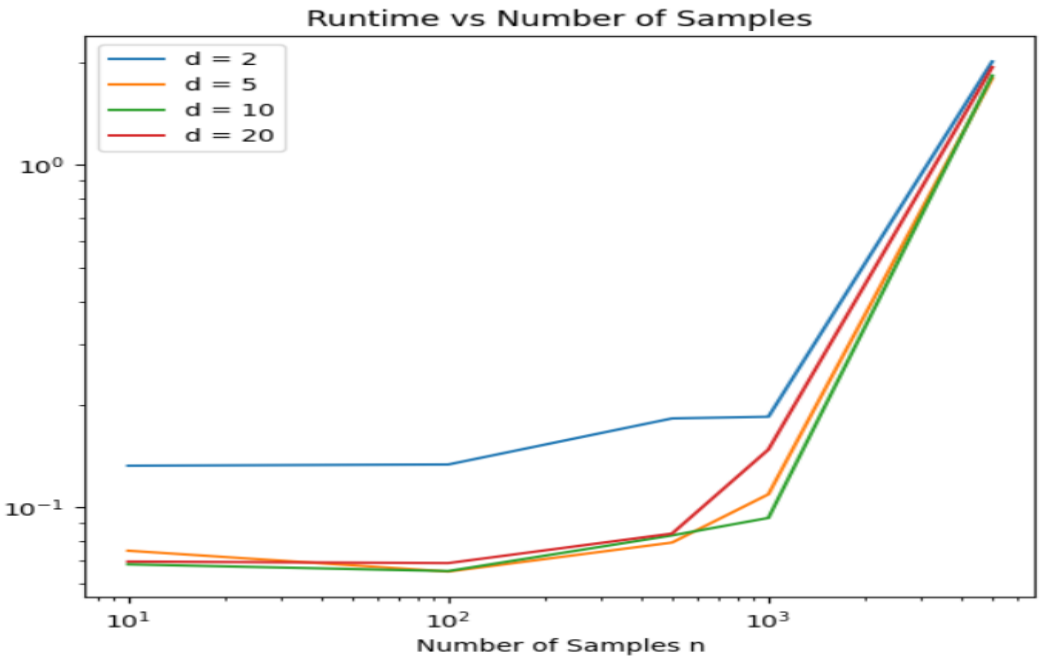

FIGURE 3 – Using CPU, Runtime vs Number of Samples

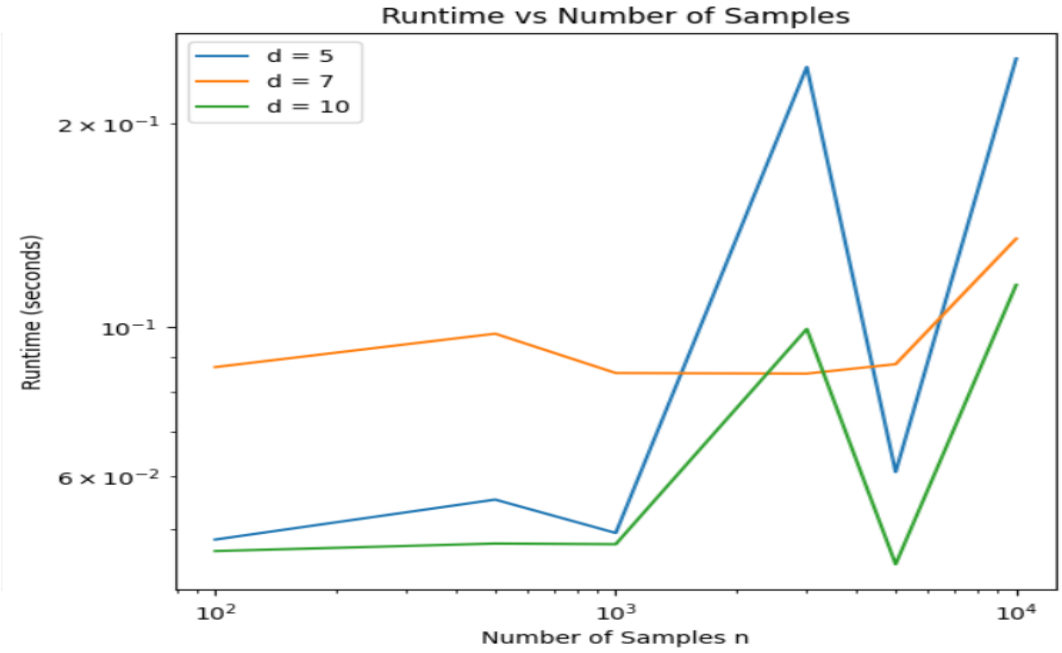

FIGURE 4 – Using GPU, Runtime vs Number of Samples

1.5 Conclusion and Future Work

Although we could not analyze the paper on its full extent, we were able to look into and recap its main results. We also developed a brief tutorial on Entropic maps, mainly on synthetic data, using the DualPotential and EntropicPotential package of OTT- JAX. The logical next step would consist in studying a large real dataset. Looking ahead, future work will focus on other ideas to find the optimal map T*. Using Input convex neural networks to model optimal transport maps could be ideal.
It could also be interesting to study the impact of the Monge Gap depending on the nature of the cost.