Sample Complexity of Sinkhorn Divergences

MVA Computational Optimal Transport Project

Simon Queric simon.queric@telecom-paris.fr

January 18, 2024

Abstract

In this project, we investigate properties of Sample Complexity of Sinkhorn divergences both from a theoretical and a practical point of view. Optimal Transport (OT) is a powerful tool for comparing probability distribution and has been widely adopted by Machine Learning and Data Science community. In particular, the regularized version of OT is widely used because of its simplicity from a computational point of view. However, the lack of closed forms in OT as well as in regularized OT (outside the Gaussian case) can be challenging to establish theoretical and numerical guarentees.

1 Introduction

Sample complexity in statistics

In statistical learning, Sample complexity refers to the number of training examples or data points required for a learning algorithm to achieve a certain level of performance or accuracy. Sample complexities of estimators are important to quantify and get bounds on approximation errors. In the case of parametric statistics, they've been widely studied and classical results such as Cramer Rao bound and asymptotical normality of the Maximum Likelihood Estimator are well established. It turns out that the classical sample complexity is $O\left(\frac{1}{n}\right)$ in parametric settings. However, in non-parametric statistics, sample complexity suffer from the curse of dimensionality, meaning its sample complexity is $O\left(\frac{1}{n^d}\right)$ where d is the dimension of the space considered.

If $\widehat{\theta}_n$ is an estimator of θ^*, the associated *sample complexity* is defined as the speed of convergence of $\|\widehat{\theta}_n - \theta^*\|$. The goal of studying sample complexity is to control the error of aproximating θ^* by $\widehat{\theta}_n$ and to obtain theoretical guarentees.

Distances between probability distribution

Several distances or divergences between probablity distribution have been studied. An interesting aspect of Optimal Transport is that it handles distributions with supports with no inclusion relationship.

Optimal Transport

Let $\alpha \in \mathcal{M}_1^+(\mathcal{X}), \beta \in \mathcal{M}_1^+(\mathcal{Y})$. Optimal Transport distance between α and β qunatify the minimal cost to move all the mass of α to the mass of β. It is also called *earth mover's distance*. It's defined as :

$$W(\alpha, \beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\pi(dx, dy)$$

where $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the cost function. Its interpretation is the following : $c(x, y)$ represents the cost of moving one unit of mass from the point x to the point y.

Maximum Mean Decrepency

Let k be a semi-definite positive kernel on $\mathcal{X}$. We denote $\mathcal{H}$ the associated RKHS. Then, the Maximum Mean Discrepancy (MMD) between α and β is defined by :

$$\text{MMD}(\alpha, \beta) = \mathbb{E}_{X,X' \sim \alpha \otimes \alpha}[k(X, X')] + \mathbb{E}_{Y,Y' \sim \beta \otimes \beta}[k(Y, Y')] - 2\mathbb{E}_{X,Y \sim \alpha \otimes \beta}[k(X, Y)]$$

An Integral Probability metric with respect to a class of function $\mathcal{F}$ is a semi-distance between probability distribution :

$$D_{\mathcal{F}}(\alpha, \beta) = \sup_{f \in \mathcal{F}} \mathbb{E}_{X \sim \alpha}[f(X)] - \mathbb{E}_{Y \sim \beta}[f(Y)]$$

It turns out that MMD is as specific case of Integral Probability metric, when the class of function is the unit ball, so that it satisfies :

$$\text{MMD}(\alpha, \beta) = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leqslant 1} \mathbb{E}_{X \sim \alpha}[f(X)] - \mathbb{E}_{Y \sim \beta}[f(Y)]$$

For other choices of $\mathcal{F}$, on recovers other well known statistical distances.

- Bounded continuous $\to$ Dudley's metric.
- Bounded variations $\to$ Kolmogorov metric.
- Bounded Lipschitz $\to$ 1-Wasserstein distance.

Bridging the gap between OT and MMD : Regularized Optimal Transport

We can compute Optimal Transport from samples by using classical Linear Programming. However, for a large number of samples, it becomes infeasible to solve in a reasonable time. This is why entropic regularization was introduced. It's an approximation of Optimal Transport that can be solved efficiently with Sinkhorn's algorithm. It concists to add a regularization term which is the Kullback-Leibler divergence between the transportation plan (also known as coupling) and the product measure between α and β.

$$W_\varepsilon(\alpha, \beta) = \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y)\pi(dx, dy) + \varepsilon KL(\pi \mid \alpha \otimes \beta)$$

Regularized OT can be interpreted as an interpolation between OT and MMD. Indeed, when $\varepsilon \to 0$, we recover the classical OT, whereas $\varepsilon \to +\infty$, we recover MMD for kenerl $k = -c/2$.

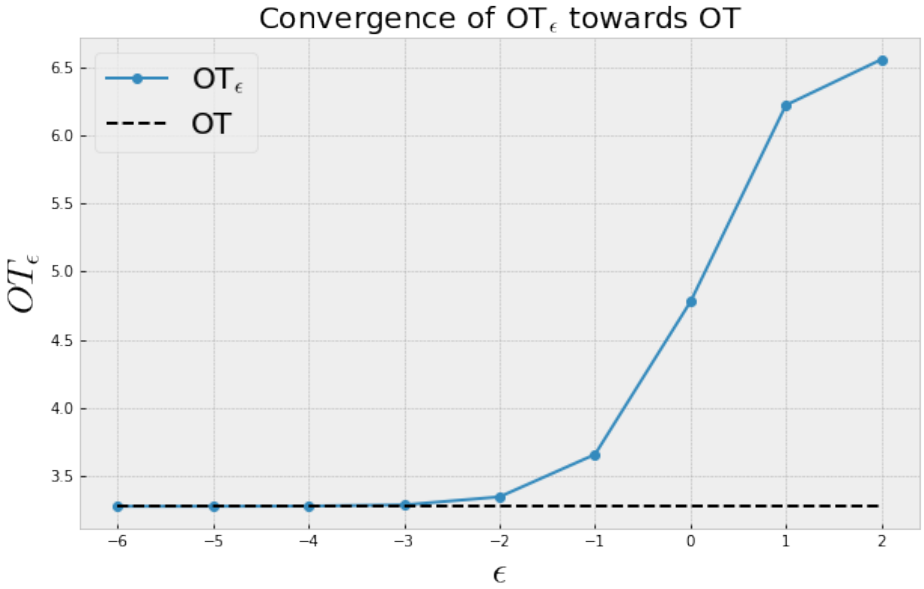

Figure 1: Convergence of OT_ε towards OT, ε in log scale. Gaussian case, $d = 3$.

We will also deal with a normalized distance to bypass the fact that W_ε is not a true distance. We define $\overline{W}_\varepsilon(\alpha, \beta) = W_\varepsilon(\alpha, \beta) - \frac{1}{2}(W_\varepsilon(\alpha, \alpha) + W_\varepsilon(\beta, \beta))$.

Computing OT, Regularized OT and MMD in practice

The three distances OT, regularized OT and MMD can be computed from samples. Algorithm time complexity have been grouped together in the table below.

OT	MMD	Regularized OT
$O(n^3(\log n)^2)$	n^2	$O(n^2)$

Table 1: Algorithms time complexity.

Let $X_{1:n}, Y_{1:n}$ be two samples drawn from α and β. We denote $\alpha_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ the empirical distributions. Computing MMD is straightforward as it's just the Monte-Carlo estimate :

$$\widehat{\mathrm{MMD}}(\alpha_n, \beta_n) = \frac{1}{n(n-1)}\sum_{i\neq j} k(X_i, X_j) + \frac{1}{n(n-1)}\sum_{i\neq j} k(Y_i, Y_j) - 2\frac{1}{n^2}\sum_{i,j} k(X_i, Y_j)$$

for kernel $k = -c/2$.

Computing OT is more time-consuming since it requires Linear Programming that has a time complexity in $O(n^3(\log n)^2)$. It consists to solve the constraint optimization problem :

$$P^* = \min_{P1=a, P^T=b} \langle P, C\rangle$$

Finally, regularized OT can be efficiently computed with an alternate minimization scheme called Sinkhorn's algorithm which has been a revolution in the field of OT especially with development of GPUs and parallel computing.

Sinkhorn's algorithm

Algorithm 1 Sinkhorn's algorithm

Input : Samples $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ of size n, Cost matrix $C_{i,j}$, Regularizer ε, Histograms a, b
$v_0 \leftarrow (1, \cdots, 1)$
$k = 1$
$K_{i,j} = \exp(-C_{i,j}/\varepsilon)$
Until convergence :
$u_k \leftarrow a/Kv_{k-1}$
$v_k \leftarrow b/K^T u_k$
$k \leftarrow k + 1$
Output : u, v

Sinkhorn divergences and sample complexity

We can know introduce the *Sinkhorn divergence* which is defined as the approximation of the true regularized distance from samples, denoted as $\widehat{W}_\varepsilon(\alpha_n, \beta_n)$. Its *sample complexity* is the distance of Sinkhorn divergence from the true distance, denoted as : $|W_\varepsilon(\alpha, \beta) - \widehat{W}_\varepsilon(\alpha_n, \beta_n)|$. The study of sample complexity aims at finding bounds of error approximation.

2 Theoretical results

Previous work by Dudley

A well-known result has been established by Dudley for OT sample complexity : Optimal Transport suffer from curse of dimensionality that is : $\mathbb{E}\left[|W(\alpha, \beta) - W(\alpha_n, \beta_n)|\right] = O(n^{-d})$ where α and β are probability measures on $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$.

For Maximum Mean Discrepency, it turns out that sample complexity is independent from the dimension. We get $\mathbb{E}\left[\mathrm{MMD}(\alpha, \beta) - \mathrm{MMD}(\alpha_n, \beta_n)\right] = O(1/\sqrt{n})$

Results of the article Sample Complexity of Sinkhorn divergences by Genevay et al.

In the article [2], Genevay et al. have established the following results for Sample Complexity for Sinkhorn divergences :

$$\mathbb{E}\left[|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\alpha_n, \beta_n)|\right] = O\left(\frac{e^{\kappa/\varepsilon}}{\sqrt{n}}\left(1 + \frac{1}{\varepsilon^{d/2}}\right)\right)$$

Two behaviors os the regularizer hyperparameter ε are of interests. First, when $\varepsilon \to +\infty$, one recovers the sample complexity of MMD :

$\mathbb{E}\left[|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\alpha_n, \beta_n)|\right] = O\left(\frac{e^{\kappa/\varepsilon}}{\varepsilon^{d/2}\sqrt{n}}\right)$ when $\varepsilon \to +\infty$

On the other side of the spectrum, when $\varepsilon \to 0$, we recover the curse of dimensionality of OT :

$\mathbb{E}\left[|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\alpha_n, \beta_n)|\right] = O(1/\sqrt{n})$ when $\varepsilon \to +\infty$.

This can also be interpreted as the fact that Sinkhorn Divergence is an interpolation between OT and MMD.

OT closed forms, Janati et al.

As we said in the introduction, OT lacks of closed forms which makes theoretical analysis of error estimations quite challenging. However, the Gaussian case is remarkable because there are closed forms in the regularized case and even in unbalanced OT, although we won't discuss about it. This is

OT between Gaussians

Let $\alpha = \mathcal{N}(a, A)$ and $\beta = \mathcal{N}(b, B)$, then :

$$W(\alpha, \beta) = \|a - b\|_2^2 + \mathcal{B}^2(A, B)$$

where $\mathcal{B}^2$ is the so-called Bures distance defined in $\mathcal{B}^2(A, B) = \mathrm{Tr}(A + B - 2(A^{1/2}BA^{1/2})^{1/2})$

Regularized OT between Gaussian measures

Let $\alpha = \mathcal{N}(a, A)$ and $\beta = \mathcal{N}(b, B)$. In the case of entropic regularization, closed form have been established by [3].

$$W_\varepsilon(\alpha, \beta) = \|a - b\|_2^2 + \mathcal{B}_\varepsilon^2(A, B)$$

where $\mathcal{B}_\varepsilon^2$ can be seen as en extension fo the Bures distance, defined by :

$\mathcal{B}_\varepsilon^2(A, B) = \mathrm{Tr}(A+B-D_\varepsilon)+d\varepsilon/2(1-\log\varepsilon)+\varepsilon/2\log\det(D_\varepsilon+\varepsilon/2I_d)$ where $D_\varepsilon = (4A^{1/2}BA^{1/2}+\varepsilon^2/4I_d)^{1/2}$.

3 Numerical experiments

Numerical experiments can corroborate theory and give some intuition on what's going on. We'll plot sample complexities of regularized distances fo various values of ε.

For the numerical experiments, we draw samples $X_{1:n}, Y_{1:n}$ from distributions α and β. Several choices are possible for α and β. Gaussians are conveniants because we have closed forms. I also chose to study uniform and beta distribution because they have bounded support.

We work with normalized and regularized distance : $\overline{W}_\varepsilon$ and with the regularized distance W_ε .

Estimation of regularized distance :

$$\widehat{W}_\varepsilon(\alpha_n, \beta_n) = \langle P_{\varepsilon,n}, C \rangle$$

where $P_{\varepsilon,n}$ is the optimal coupling for the $\varepsilon-$regularized OT problem.

When working in log-domain, there is another estimate of these distance which ca be computed from the sample. It is the Monte-Carlo estimation of regularized distance :

$$\widehat{W}_\varepsilon(\alpha_n, \beta_n) = \frac{1}{n}\sum_{i=1}^{n} u(X_i) + \frac{1}{n}\sum_{j=1}^{n} v(Y_j) - \frac{\varepsilon}{n^2}\sum_{i,j} \exp\left(\frac{-C_{i,j} + u(X_i) + v(Y_j)}{\varepsilon}\right) + \varepsilon$$

Here, $u(X_i), v(Y_j)$ are the dual potentials and can be computed with to Sinkhorn algorithm in log-domain.

In practice we just compute : $\widehat{W}_\varepsilon(\alpha_n, \beta_n) = \frac{1}{n}\sum_{i=1}^{n} u(X_i) + \frac{1}{n}\sum_{j=1}^{n} v(Y_j)$. Why is it possible ? Why could we get rid of the last two terms in the equation above ? Well, working in log-domain means that we deal with the dual of Kantorovitch's problem. The functions $u \in \mathcal{C}(\mathcal{X})$ and $v \in \mathcal{C}(\mathcal{Y})$ are the dual potentials and satisfy $(P^*_\varepsilon)_{i,j} = \exp\left(\frac{u_i + v_j - C_{i,j}}{\varepsilon}\right)$, where $P^*_\varepsilon = \underset{P1=a, P^T1=b}{\arg\min} \langle P, C\rangle + \varepsilon H(P \mid \alpha \otimes \beta)$ so that the sum of all entries of P^*_ε sum to one *i.e* $\frac{1}{n^2}\sum_{i,j} \exp\left(\frac{u_i + v_j - C_{i,j}}{\varepsilon}\right) = 1$ and the last two terms cancel each other. It couldn't be done if we where working in the unbalanced setting of OT, when $\sum a_i \neq \sum b_j$.

Sample complexity of Sinkhorn divergences for Gaussians

Experiments have been conducted for OT between gaussians, essentially because closed form are available for the regularized case ([3]). Then, it's possibe to plot the empirical convergence of $OT_\varepsilon(\alpha_n, \beta_n)$ towards $OT_\varepsilon(\alpha, \beta)$.

Implementation

For small ε, Sinkhorn's algorithm lead to numerical instabilities. A log-domain Sinkhorn's algorithm was introduced in [1] to overcome this difficulty. It takes advantage of the dual formulation of OT.

Dual of Kantorovitch

The lagrangian dual of Kantorovitch formulation is the following :

$$\sup_{f\in\mathcal{C}(\mathcal{X}), g\in\mathcal{C}(\mathcal{Y}), f\oplus g\leqslant c} \int_{\mathcal{X}} f(x)\alpha(dx) + \int_{\mathcal{Y}} g(y)\beta(dy) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{(f(x)+g(y)-c(x,y))/\varepsilon}\alpha(dx)\beta(dy) + \varepsilon$$

Sinkhorn's algorithm in log-domain

In order to compute dual potentials $u(X_i), v(Y_j)$ we Sinkhorn's algorithm in log domain. Sinkhorn's algorithm in log-domain returns the dual potentials u and v. It is also an altenate minimization algorithm to update u and v, and uses the log-sum-exp trick to ensure numerical stability.

Algorithm 2 Sinkhorn's algorithm in log domain

Input : Samples $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ of size n, Cost matrix $C_{i,j}$, Regularizer ε
$K_{i,j} = \exp\left(-C_{i,j}/\varepsilon\right)$
$u_0 \leftarrow (0, \cdots, 0)$
$k = 1$
Until convergence :
$u_k \leftarrow v_{k-1}^{c,\varepsilon}$
$v_k \leftarrow u_k^{c,\varepsilon}$
$k \leftarrow k + 1$
Output : u, v

where $u_i^{c,\varepsilon} = -\varepsilon \log \sum_j \exp((c_{i,j} - v_j)/\varepsilon) b_j$ and $v_j^{c,\varepsilon} = -\varepsilon \log \sum_i \exp((c_{i,j} - u_i)/\varepsilon) a_i$

Experiments

Theory suggests :

1. Faster convergence for bigger ε
2. Slower convergence for bigger dimension d (curse of dimensionality).

Gaussian measures

For two different gaussian measures

For α, β gaussian measures such that $m_\alpha, m_\beta \sim \mathcal{U}([-1,1]^d)$, $\Sigma_\alpha^{1/2}, \Sigma_\beta^{1/2} \sim \mathcal{N}(0, 0.5)$.

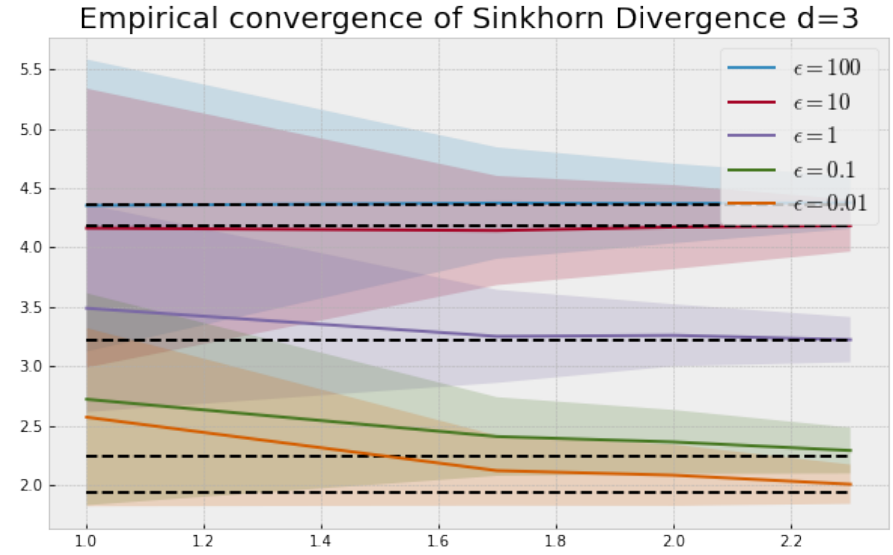

Figure 2: $|W_\varepsilon(\alpha_n, \beta_n) - W(\alpha, \beta)|$ as a function of n in semi-log. Experiment for gaussian measures in dimension 3 with $c(x, y) = \|x - y\|_2^2$.

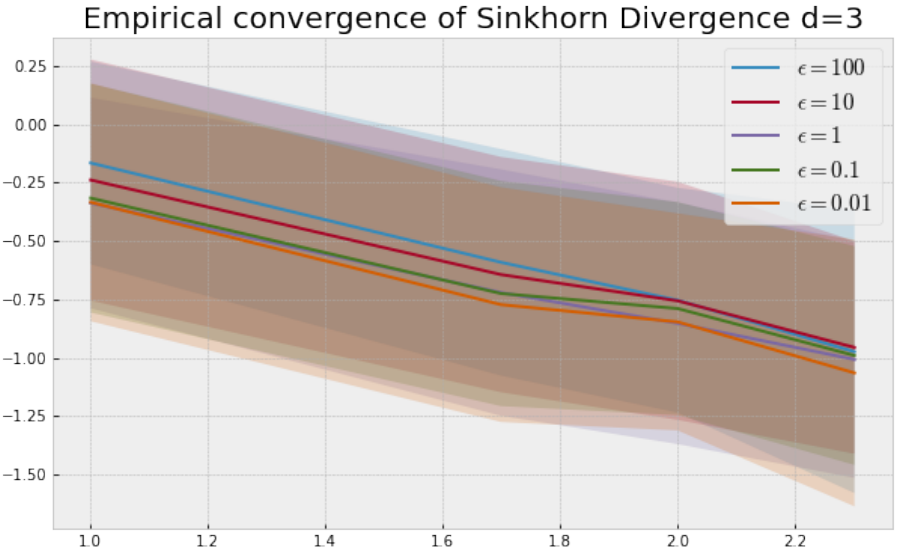

Figure 3: $|W_\varepsilon(\alpha_n, \beta_n) - W(\alpha, \beta)|$ as a function of n in the log-log space. Experiment for gaussian measures in dimension 3 with $c(x, y) = \|x - y\|_2^2$.

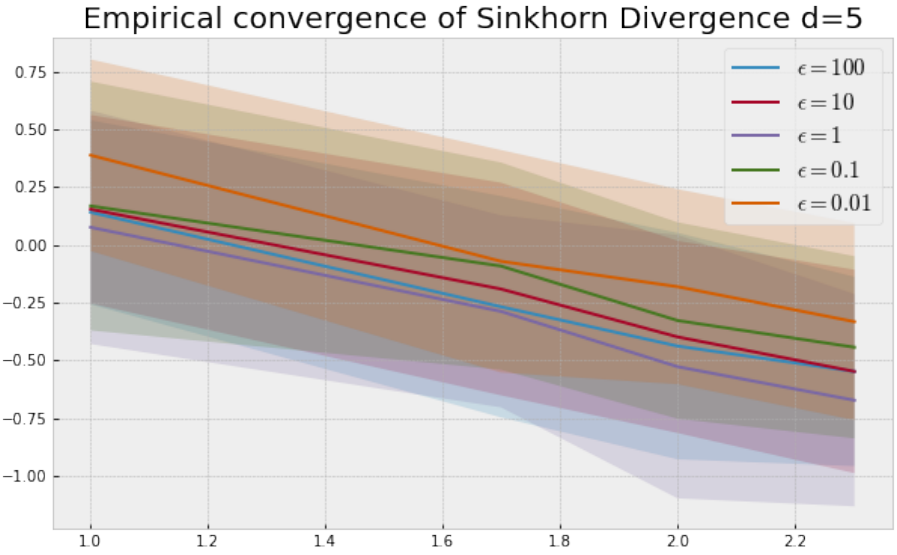

Figure 4: $|W_\varepsilon(\alpha_n, \beta_n) - W(\alpha, \beta)|$ as a function of n in the log-log space. Experiment for gaussian measures in dimension 5 with $c(x, y) = \|x - y\|_2^2$.

Standard normal distributions for $d = 5$

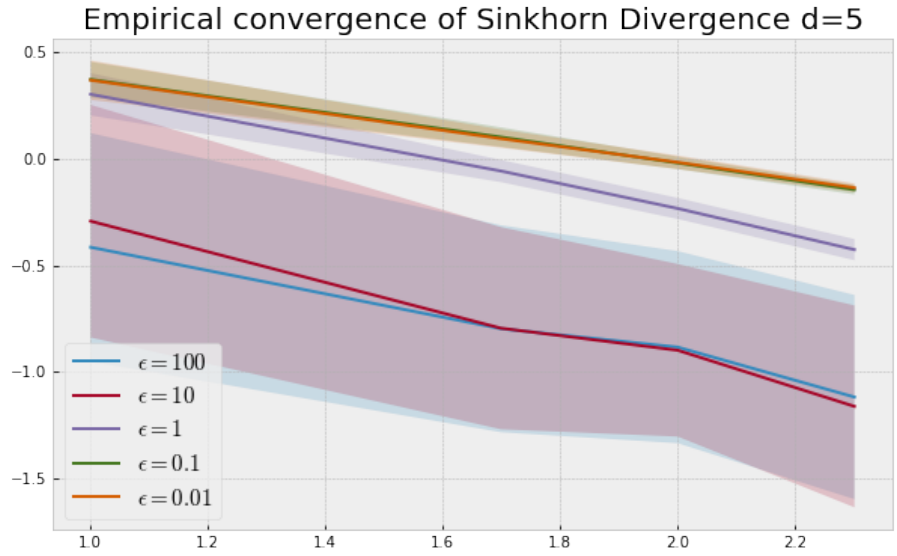

Figure 5: $\overline{W}_\varepsilon(\alpha_n, \beta_n)$ as a function of n in the log-log space. Experiment for standard normal distribution in dimension 5 with $c(x, y) = \|x - y\|_2^2$.

Uniform measures

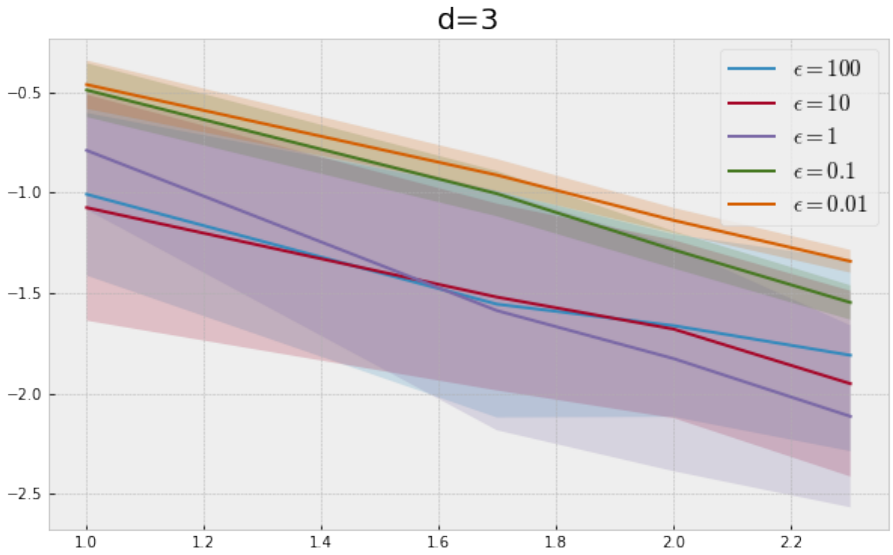

Figure 6: $\overline{W}_\varepsilon(\alpha_n, \beta_n)$ as a function of n in the log-log space. Experiment for uniform measures in $[-1, 1]^3$ with $c(x, y) = \|x - y\|_2^2$.

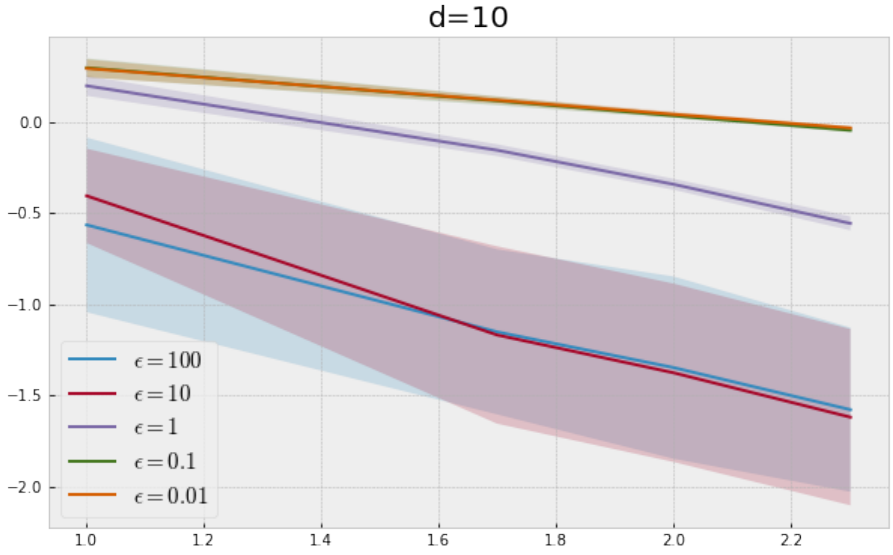

Figure 7: $\overline{W}_\varepsilon(\alpha_n, \beta_n)$ as a function of n in the log-log space. Experiment for uniform measures in $[-1, 1]^{10}$ with $c(x, y) = \|x - y\|_2^2$.

Beta(2, 5) distribution, dimension 2

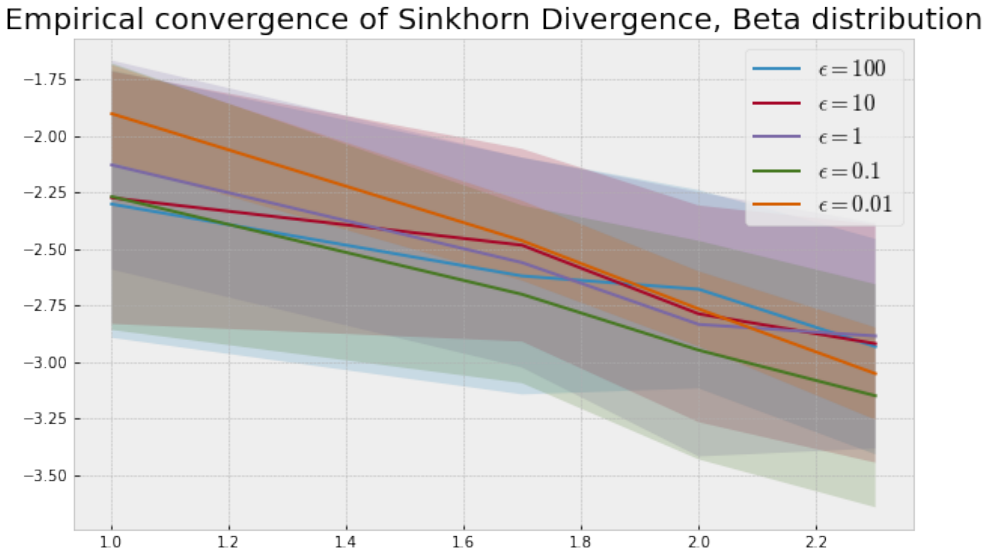

Figure 8: $\overline{W}_\varepsilon(\alpha_n, \beta_n)$ as a function of n in the log-log space. Experiment for Beta distributions of parameters $(2, 5)$ in dimension $d = 2$, with $c(x, y) = \|x - y\|_2^2$.

Analysis of the results

What's interesting is that we can see the linear convergence of Sinkhorn's algorithm when we plot in log-log scale.

For two different gaussian measures (Figure 2 & 3), the dependance of sample complexity in ε is not clear at all, it's clearer in dimension 5 (Figure 4).

In the case of dimension 5 for Gaussian (Figure 5) and 10 for Uniform distribution (Figure 6), we can clearly see flater slope for small regularizer ε which corroborates theory. In particular, for uniform distribution we can see the difference between low dimension $d = 3$ (Figure 6) and higher dimension (Figure 7).

In the case of Beta distribution, the dependance of sample complexity in ε is not clear probably because we are in low dimension $d = 2$ and with a bounded support.

4 Conclusion and perspective

In this project, we where stuck in the world of Gaussian measures and of probability measures with bounded support.

Furthermore, I've investigated Sinkhorn divergences i.e how our approximation of reguarized distance behave when the number of samples increases. A possible extension would be to quantify the distance between approximation of regularized OT and OT itself which seems quite challenging.

5 Connexion with the course

In the course, we constructed Sinkhorn's algorithm that estimate regularized Wasserstein distance. The article [2] established theoretical results for sample complexity of Sinkhorn divergences helping to understand the behavior of our estimates. I also used closed forms established by [3] to compute approximation errors for gaussian measures.

References

[1] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

[2] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences, 2019.

[3] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced gaussian measures has a closed form, 2020.

[4] Gabriel Peyré's Numerical tours on Optimal Transport

[5] Gabriel Peyré and Marco Cuturi's textbook on Computational Optimal Transport.