Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem

Gonzalo E. Mena[1], Jonathan Weed[2], 1. Harvard University 2. Massachussets institute of technology,

Introduction

Optimal transport (OT) has become a popular analysis tool for large datasets in high dimension, and **entropic regularization** has shown to provide computationally efficient approximations (Cuturi, 2013).

However, it also appears to have useful **statistical** properties. For instance Genevay et al. (2019) established that even though standard OT suffers from the **curse of dimensionality**, entropic OT always converges at the **parametric** $1/\sqrt{n}$ for compactly supported probability measures.

Definitions

Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ be probability measures and P_n, Q_n be their empirical versions. Their squared Wasserstein distance is:

$$W_2^2(P,Q) := \inf_{\pi \in \Pi(P,Q)} \left[\int_{\mathcal{X}\times\mathcal{Y}} \frac{1}{2}\|x-y\|^2 \, \mathrm{d}\pi(x,y) \right],$$

where $\Pi(P,Q)$ is the set of joints with marginals P and Q. We focus on an entropy regularized version of the above:

$$S(P,Q) := \inf_{\pi \in \Pi(P,Q)} \left[\int_{\mathcal{X}\times\mathcal{Y}} \frac{1}{2}\|x-y\|^2 \, \mathrm{d}\pi(x,y) + H(\pi | P \otimes Q) \right],$$

with $H(\alpha|\beta)$ the relative entropy between α and β. $S(P,Q)$ has a dual formulation (Csiszar 1975):

$$S(P,Q) = \sup_{f \in L_1(P), g \in L_1(Q)} \int f(x)\,\mathrm{d}P(x) + \int g(y)\,\mathrm{d}Q(y) - \int e^{f(x)+g(y)-\frac{1}{2}\|x-y\|^2} \,\mathrm{d}P(x)\mathrm{d}Q(y) + 1.$$

We say that a distribution $P \in \mathcal{P}(\mathbb{R}^d)$ is σ^2-subgaussian if $E_P e^{\frac{\|X\|^2}{2d\sigma^2}} \leq 2$.

Main Results

- **Theorem 1**: New **sample complexity** bounds, extending the results of Genevay et al. (2019) to the subgaussian case.
- **Theorem 2**. **Central Limit Theorem** for the fluctiations of the empirical version of entropic optimal transport around its expected value, extending the results of Del Barrio and Loubes (2019) and Bigot et al. (2018).
- **Theorem 3**. As an application, we show how entropic OT can be used to **estimate the entropy** of random variables corrupted by subgaussian noise.

Sample Complexity

Theorem 1 Let P and Q be σ^2-subgaussian, then

$$E_{P,Q}|S(P,Q) - S(P_n,Q_n)| \leq K_d(1+\sigma^{\lceil 5d/2\rceil + 6})\frac{1}{\sqrt{n}}.$$

Remark: the Wasserstein distance is cursed by dimensionality (Dudley, 1969),

$$E_{P,Q}|W_2(P,Q) - W_2(P_n,Q_n)| \leq O(n^{-1/d}).$$

Application: entropy estimation

If Q has a density q, its differential entropy is defined as $h(Q) := -\int q(x)\log q(x)dx$. The main result is the following

Theorem 3 Let P be subgaussian, $G \sim \mathcal{N}(0, \sigma_g^2 I_d)$ and $Q = P * G$ with density q. Define the plug-in $\hat{h}(Q) = S(P_n, Q_m) + \frac{d}{2}\log\left(2\pi\sigma_g^2\right)$ where P_n and Q_m are independent. Then,

(a) If $m = n$,

$$\sup_P E_P|\hat{h}(Q) - h(Q)| \leq O\left(\frac{1}{\sqrt{n}}\right).$$

(b)

$$\sqrt{\frac{mn}{m+n}}\left(\hat{h}(Q) - E(\hat{h}(Q))\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \lambda \operatorname{Var}_Q(\log q(Y)))$$

Theorem 3 is a simple application of Theorems 1 and 2 based on the observation that

$$h(P*G) = S(P, P*G) + \frac{d}{2}\log\left(2\pi\sigma_g^2\right).$$

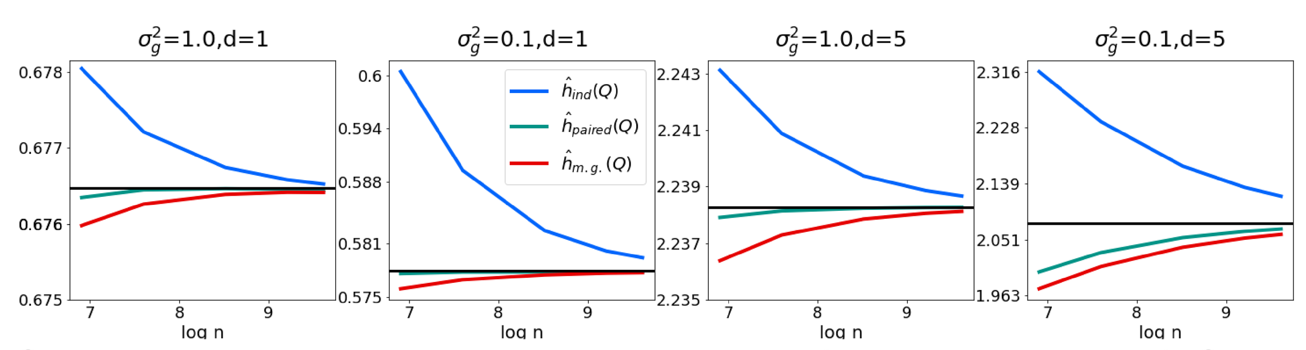

Figure 2: Comparison between three estimators for the entropy of Q, (same as Figure 1). $h_{m.g}$ is a naive estimator based on a mixture of gaussians (Goldfeld et al, 2019), h_{ind} is the one from Theorem 1 and h_{paired} is the same, but Q-samples are not independent of P: they are taken by adding gaussian noise to each P sample.

Central limit theorem

Theorem 2. Let $X_1, \ldots, X_n \sim P$ and $Y_1, \sim Y_m \sim Q$ are two i.i.d. sequences independent of each other. Assume P and Q are both subgaussian. Denote $\lambda := \lim_{m,n\to\infty} \frac{n}{m+n} \in (0,1)$. Then

$$\sqrt{\frac{mn}{m+n}}\left(S(P_n,Q_m) - E(S(P_n,Q_m))\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Sigma)$$
$$\text{with } \Sigma = (1-\lambda)\operatorname{Var}_P(f(X_1)) + \lambda \operatorname{Var}_Q(g(Y_1)).$$

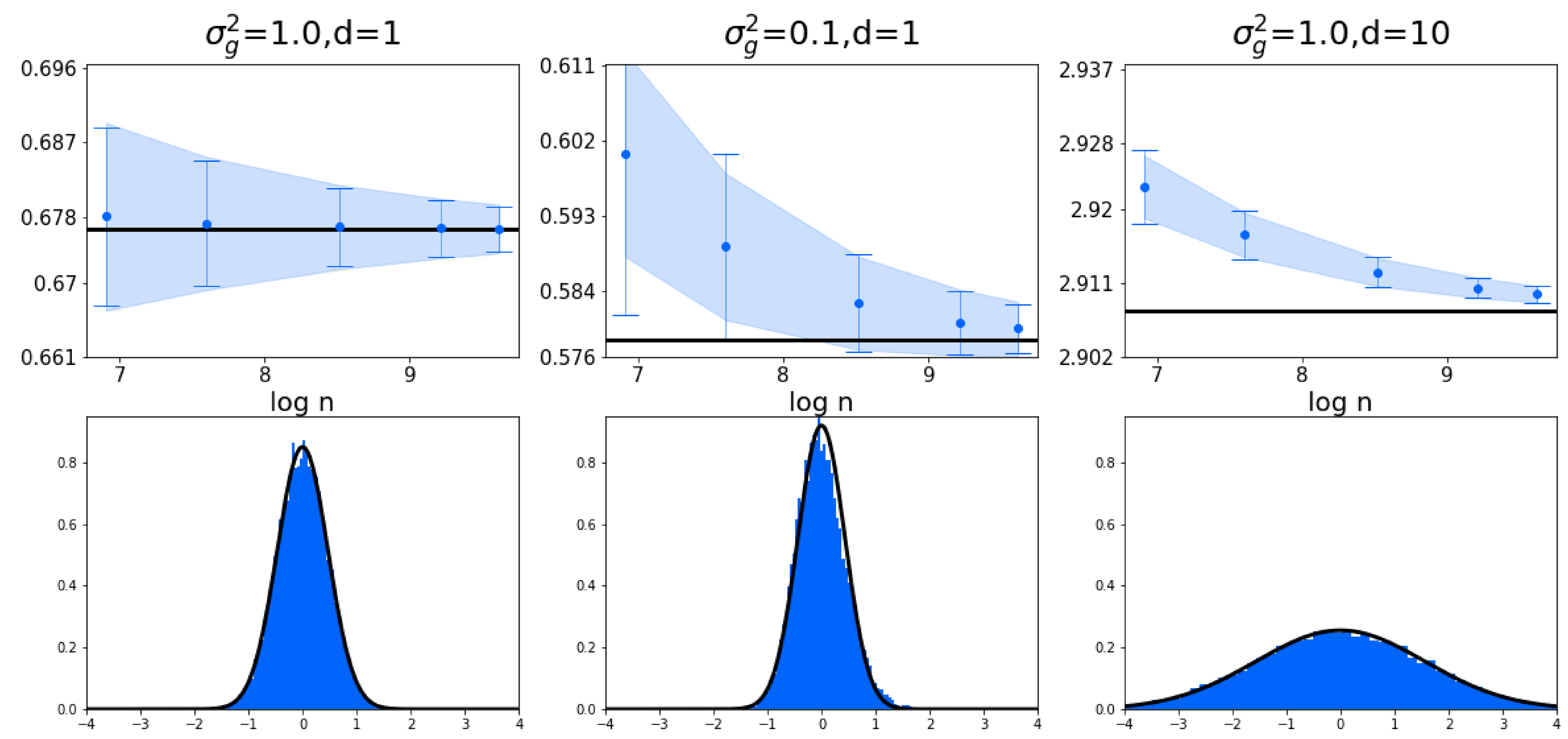

Figure 1: Example on $P = \frac{1}{2}(\delta_{-1} + \delta_1)$, $Q = P * \mathcal{N}(0,\sigma^2)$. Top: $ES(P_n,Q_n)$ as a function of n. The shading corresponds to one standard deviation of $S(P_n,Q_n) - ES(P_n,Q_n)$, given by Theorem 2. Error bars are sample standard deviations. Bottom: histograms of $\sqrt{\frac{nn}{n+n}}(S(P_n,Q_n) - ES(P_n,Q_n)))$ when $n = 15000$. Ground truth is indicated with solid lines

References

- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. NIPS
- Csiszar, I. (1975), I-divergence geometry of probability distributions and minimization problems. The Annals of Probability,
- Del Barrio, E. and Loubes, J.M (2019). Central limit theorems for empirical transportation cost in general dimension. The Annals of Probability.
- Bigot, J., Cazelles, E. and Papadakis, N. (2018). Central limit theorems for Sinkhorn divergence between probability distributions on finite spaces and statistical applications. arXiv.
- Dudley, R.M. (1969). The Speed of Mean Glivenko-Cantelli Convergence. The Annals of Mathematical Statistics.
- Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample Complexity of Sinkhorn divergences. AISTATS.
- Goldfeld, Z., Greenewald, K., Polyanskiy, Y. and Weed, J., (2019). Convergence of Smoothed Empirical Measures with Applications to Entropy Estimation. arXiv.

Email: gomena@fas.harvard.edu