Regularity properties of Entropic Optimal Transport in applications to machine learning

Giulia Luise
University College London

MAGA Days, 20/11/2019

Table of contents

Notation

$\mathcal{X} \subset \mathbb{R}^d$ compact

$\mathcal{M}(\mathcal{X})$ space of finite measures over $\mathcal{X}$

$\mathcal{P}(\mathcal{X})$ probability measures over $\mathcal{X}$

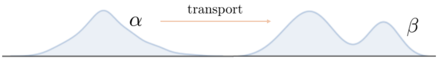
$\mathsf{c} : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ continuous, symmetric cost function ($\mathsf{c}(\cdot,\cdot) = \|\cdot - \cdot\|^p, p \in [1, \infty)$).

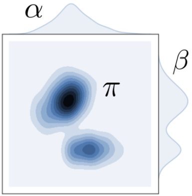
Given $\alpha, \beta \in \mathcal{P}(\mathcal{X})$, the Optimal Transport (OT) problem is

Optimal Transport Problem

$$W(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} \mathsf{c}(x, y)\, d\pi(x, y) \tag{1}$$

where $\Pi(\alpha, \beta) = \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) \quad \text{s.t.} \;\; \mathrm{Proj}^1_{\#}\pi = \alpha, \mathrm{Proj}^2_{\#}\pi = \beta\}$.

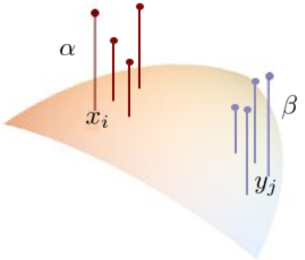




Given $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ finite discrete probability measures

$$\alpha = \sum_{i=1}^{n} \mathsf{a}_i \delta_{x_i}, \qquad \beta = \sum_{j=1}^{m} \mathsf{b}_j \delta_{y_j},$$

with $x_1, \ldots, x_n \in \mathcal{X}$ and $y_1, \ldots y_m \in \mathcal{X}$ and $\mathsf{a} := (\mathsf{a}_1, \ldots, \mathsf{a}_n) \in \Delta_n$, $\mathsf{b} := (\mathsf{b}_1, \ldots, \mathsf{b}_n) \in \Delta_n$ (Δ_n is the simplex).

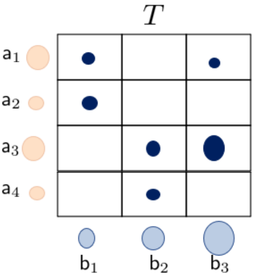


Discrete Setting

Optimal Transport Problem: Set $C \in \mathbb{R}^{n\times m}$ with $C_{ij} = \mathsf{c}(x_i, y_j)$,

$$W(\alpha, \beta) = \min_{T\in\Pi(\mathsf{a},\mathsf{b})} \langle T, C\rangle \tag{2}$$

where $\Pi(\mathsf{a}, \mathsf{b}) = \{T \in \mathbb{R}_+^{n\times m} \;\; \text{s.t.} \;\; T\mathbf{1} = \mathsf{a},\, T^\top\mathbf{1} = \mathsf{b}\}$ is the transport polytope.



Given $\alpha, \beta \in \mathcal{P}(\mathcal{X})$, the Entropic Optimal Transport (OT) problem is

$$\mathrm{OT}_{\varepsilon}(\alpha, \beta) := \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} \mathsf{c}(x, y)\, d\pi(x, y) + \varepsilon \mathrm{KL}(\alpha \otimes \beta, \pi)$$

Discrete setting:

$$\mathrm{OT}_{\varepsilon}(\alpha, \beta) = \min_{T \in \Pi(\mathsf{a},\mathsf{b})} \langle T, C\rangle + \varepsilon \sum_{i,j} T_{ij}(\log\left(\frac{T_{ij}}{\mathsf{a}_i \mathsf{b}_j}\right) - 1).$$

Sinkhorn divergence

To remove the bias ($\mathrm{OT}_\varepsilon(\alpha, \alpha) \neq 0$) introduced by the KL, one could consider the unbiased Sinkhon divergence [Feydy et al., 2019]:

$$\mathsf{S}_\varepsilon(\alpha, \beta) = \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\mathrm{OT}_\varepsilon(\alpha, \alpha) - \frac{1}{2}\mathrm{OT}_\varepsilon(\beta, \beta).$$

Figure: S_ε

Computational cost

Optimal transport:

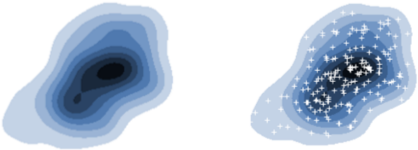
Hungarian algorithm / others

$\tilde{O}(n^3)$ [Pele, et al., 2009]

Entropic OT:

Sinkhorn algorithm/ variants

$\tilde{O}(n^2/\varepsilon^2)$ [Cuturi, 2013, Altschuler et al., 2018]



Sample complexity

Optimal Transport

$\mathbb{E}W(\alpha, \hat{\alpha}_n)) \asymp n^{-\frac{1}{d}}$ (on $\mathbb{R}^d$)

curse of dimensionality

[Dudley, 1969]

Entropic OT

$\mathbb{E}|\mathrm{OT}_\varepsilon(\alpha, \hat{\alpha}_n)| \leq C(\varepsilon) n^{-\frac{1}{2}}$

no curse!

[Genevay et al., 2019]

Part II: Advantages of Entropic OT in terms of regularity

Entropic regularization provides advantages in terms of regularity itself.

Regularity in which sense?

This regularity enables to show theoretical guarantees of different nature, namely from statistical and optimization point of view.

Advantages in terms of smoothness

Entropic regularization $\xrightarrow{\text{[L. et al., 2019]}}$ Lipschitzness of the gradient of Sinkhorn divergence

Entropic regularization $\xrightarrow{\text{[Genevay. et al., 2019]}}$ High order regularity of Sinkhorn potentials $\xrightarrow{\text{[L. et al., 2019]}}$ sample complexity of Sinkhorn gradients

Entropic regularization $\xrightarrow{\text{[L. et al., 2018]}}$ high order differentiability of Sinkhorn divergence (in a restricted setting, simplex)

$$\mathrm{OT}_\varepsilon(\alpha,\beta) = \max_{u,v\in\mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} u\, d\alpha + \int_{\mathcal{X}} v\, d\beta - \varepsilon \int_{\mathcal{X}\times\mathcal{X}} e^{\frac{u\oplus v - \mathsf{c}}{\varepsilon}}\, d\alpha\, d\beta, \tag{3}$$

First order optimality conditions read as

$$e^{-\frac{u(x)}{\epsilon}} = \int_{\mathcal{X}} e^{\frac{v(y)-\mathsf{c}(x,y)}{\epsilon}}\, d\beta(y) \text{ for } x \in \operatorname{supp}(\alpha),$$
$$e^{\frac{-v(y)}{\epsilon}} = \int_{\mathcal{X}} e^{\frac{u(x)-\mathsf{c}(x,y)}{\epsilon}}\, d\alpha(x) \text{ for } y \in \operatorname{supp}(\beta).$$

Formulas above provide a canonical extension of u, v on the whole domain $\mathcal{X}$.

Result on Lipschitzness

Gradient of Entropic OT is given by the optimal potentials extended on the whole domain $\mathcal{X}$. We write

$$\nabla \mathrm{OT}_\varepsilon(\alpha, \beta) = (u, v).$$

Result on Lipschitzness

Gradient of Entropic OT is given by the optimal potentials extended on the whole domain $\mathcal{X}$. We write

$$\nabla \mathrm{OT}_\varepsilon(\alpha, \beta) = (u, v).$$

Theorem. The gradient $\nabla \mathrm{OT}_\varepsilon$ is Lipschitz continuous:

for every $\alpha, \alpha', \beta, \beta' \in \mathcal{P}(\mathcal{X})$, let $(u, v) = \nabla \mathrm{OT}_\varepsilon(\alpha, \beta)$ and

$(u', v') = \nabla \mathrm{OT}_\varepsilon(\alpha', \beta')$. Then,

$$\left\|u - u'\right\|_\infty + \left\|v - v'\right\|_\infty \quad \leq C_\varepsilon(\left\|\alpha - \alpha'\right\|_{TV} + \left\|\beta - \beta'\right\|_{TV}),$$

Moreover, $\nabla \mathsf{S}_\varepsilon$ is Lipschitz continuous.

The proof relies on:

- Hilbert metric and its relation with $\|\cdot\|_\infty$
- Contraction properties under Hilbert metric
- Estimates of

$$ {}_{\mathcal{C}(\mathcal{X})}\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}, \beta - \beta' \rangle_{\mathcal{M}(\mathcal{X})}, \qquad {}_{\mathcal{C}(\mathcal{X})}\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{u(\cdot)}{\varepsilon}}, \alpha - \alpha' \rangle_{\mathcal{M}(\mathcal{X})}. $$

Entropic reg $\xrightarrow{\text{[L. et al., 2019]}}$ Lipschitzness of the gradient of Sinkhorn divergence ✓

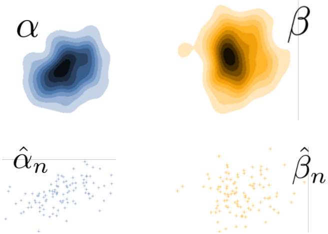
Entropic reg $\xrightarrow{\text{[Genevay. et al., 2019]}}$ High order regularity of Sinkhorn potentials $\xrightarrow{\text{[L. et al., 2019]}}$ Sample complexity of Sinkhorn gradients

Entropic reg $\xrightarrow{\text{[L. et al., 2018]}}$ high order differentiability of Sinkhorn divergence (in a restricted setting, simplex)

Sample complexity of Sinkhorn potential

$$(\alpha, \beta) \xrightarrow{\text{gradient}} (u, v)$$

$$(\hat{\alpha}_n, \hat{\beta}_n) \xrightarrow{\text{gradient}} (u_n, v_n)$$



We know that $|\mathrm{OT}_\varepsilon(\alpha, \beta) - \mathrm{OT}_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| \leq C_\varepsilon n^{-\frac{1}{2}}$ with high probability.

What can we say on $\|u - u_n\|_\infty$?

We have

$$\|u - u_n\|_\infty \lesssim \|\alpha - \hat{\alpha}_n\|_{TV} + \|\beta - \hat{\beta}_n\|_{TV}.$$

We have

$$\|u - u_n\|_\infty \lesssim \|\alpha - \hat{\alpha}_n\|_{TV} + \|\beta - \hat{\beta}_n\|_{TV}.$$

If $\hat{\alpha}_n$ and $\hat{\beta}_n$ converged to α, β in TV norm with some given rate, we could deduce a sample complexity result.

We have

$$\|u - u_n\|_\infty \lesssim \|\alpha - \hat{\alpha}_n\|_{TV} + \|\beta - \hat{\beta}_n\|_{TV}.$$

If $\hat{\alpha}_n$ and $\hat{\beta}_n$ converged to α, β in TV norm with some given rate, we could deduce a sample complexity result.

But it is not the case…

We have

$$\|u - u_n\|_\infty \lesssim \|\alpha - \hat{\alpha}_n\|_{TV} + \|\beta - \hat{\beta}_n\|_{TV}.$$

If $\hat{\alpha}_n$ and $\hat{\beta}_n$ converged to α, β in TV norm with some given rate, we could deduce a sample complexity result.

But it is not the case…

Exploiting the fact that the potentials belong not only to $\mathcal{C}(\mathcal{X})$ but also to $W^{s,2}(\mathcal{X})$ for s big enough [Genevay et al., 2019], we can get a similar bound with a weaker norm (MMD) on the r.h.s.

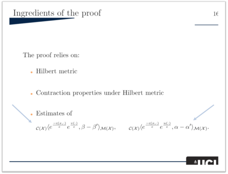
Sample complexity of Sinkhorn potentials

Ingredients of the proof

The proof relies on:

- Hilbert metric
- Contraction properties under Hilbert metric
- Estimates of

$$\mathcal{C}(\mathcal{X})\langle e^{\frac{-c(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}, \beta - \beta' \rangle_{\mathcal{M}(\mathcal{X})}, \qquad \mathcal{C}(\mathcal{X})\langle e^{\frac{-c(x,\cdot)}{\varepsilon}} e^{\frac{u(\cdot)}{\varepsilon}}, \alpha - \alpha' \rangle_{\mathcal{M}(\mathcal{X})}.$$



$\beta \in \mathcal{P}(\mathcal{X})$ can be represented as an element μ_β (mean embedding) in a suitable Hilbert space $\mathcal{H}$ (with $\mathcal{H} \subset \mathcal{C}(\mathcal{X})$) [Gretton et al., 2013].

If $f \in \mathcal{H}$, it holds that

$$_{\mathcal{C}(\mathcal{X})}\langle f, \beta \rangle_{\mathcal{M}(\mathcal{X})} = {}_{\mathcal{H}}\langle f, \mu_\beta \rangle_{\mathcal{H}}.$$

$\beta \in \mathcal{P}(\mathcal{X})$ can be represented as an element $\mu_\beta \in \mathcal{H}$ (mean embedding).

If $f \in \mathcal{H} \longrightarrow \ {}_{\mathcal{C}(\mathcal{X})}\langle f, \beta \rangle_{\mathcal{M}(\mathcal{X})} = \ {}_{\mathcal{H}}\langle f, \mu_\beta \rangle_{\mathcal{H}}.$

$\beta \in \mathcal{P}(\mathcal{X})$ can be represented as an element $\mu_\beta \in \mathcal{H}$ (mean embedding).

If $f \in \mathcal{H} \longrightarrow \; {}_{\mathcal{C}(\mathcal{X})}\langle f, \beta \rangle_{\mathcal{M}(\mathcal{X})} = {}_{\mathcal{H}}\langle f, \mu_\beta \rangle_{\mathcal{H}}$.

Now, $e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}$ belongs to a ball with some fixed radius in $\mathcal{H} = W^{s,2}(\mathcal{X})$[Genevay et al, 2019]. Hence,

$\beta \in \mathcal{P}(\mathcal{X})$ can be represented as an element $\mu_\beta \in \mathcal{H}$ (mean embedding).

If $f \in \mathcal{H} \longrightarrow {}_{\mathcal{C}(\mathcal{X})}\langle f, \beta \rangle_{\mathcal{M}(\mathcal{X})} = {}_{\mathcal{H}}\langle f, \mu_\beta \rangle_{\mathcal{H}}$.

Now, $e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}$ belongs to a ball with some fixed radius in $\mathcal{H} = W^{s,2}(\mathcal{X})$[Genevay et al, 2019]. Hence,

$$ {}_{\mathcal{C}(\mathcal{X})}\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}, \beta - \beta' \rangle_{\mathcal{M}(\mathcal{X})} = {}_{\mathcal{H}}\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}, \mu_\beta - \mu'_\beta \rangle_{\mathcal{H}} $$

$\beta \in \mathcal{P}(\mathcal{X})$ can be represented as an element $\mu_\beta \in \mathcal{H}$ (mean embedding).

If $f \in \mathcal{H} \longrightarrow {}_{\mathcal{C}(\mathcal{X})}\langle f, \beta \rangle_{\mathcal{M}(\mathcal{X})} = {}_{\mathcal{H}}\langle f, \mu_\beta \rangle_{\mathcal{H}}$.

Now, $e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}$ belongs to a ball with some fixed radius in $\mathcal{H} = W^{s,2}(\mathcal{X})$[Genevay et al., 2019]. Hence,

$$
\begin{aligned}
{}_{\mathcal{C}(\mathcal{X})}\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}, \beta - \beta' \rangle_{\mathcal{M}(\mathcal{X})} &= {}_{\mathcal{H}}\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}, \mu_\beta - \mu_{\beta'} \rangle_{\mathcal{H}} \\
&\leq \mathsf{r}\|\mu_\beta - \mu_{\beta'}\|_{\mathcal{H}} =: \mathsf{MMD}(\beta, \beta').
\end{aligned}
$$

$\beta \in \mathcal{P}(\mathcal{X})$ can be represented as an element $\mu_\beta \in \mathcal{H}$ (mean embedding).

If $f \in \mathcal{H} \longrightarrow {}_{\mathcal{C}(\mathcal{X})}\langle f, \beta \rangle_{\mathcal{M}(\mathcal{X})} = {}_{\mathcal{H}}\langle f, \mu_\beta \rangle_{\mathcal{H}}$.

Now, $e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}$ belongs to a ball with some fixed radius in $\mathcal{H} = W^{s,2}(\mathcal{X})$[Genevay et al., 2019]. Hence,

$$
\begin{aligned}
{}_{\mathcal{C}(\mathcal{X})}\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}, \beta - \beta' \rangle_{\mathcal{M}(\mathcal{X})} &= {}_{\mathcal{H}}\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} e^{\frac{v(\cdot)}{\varepsilon}}, \mu_\beta - \mu_{\beta'} \rangle_{\mathcal{H}} \\
&\leq \mathsf{r} \|\mu_\beta - \mu_{\beta'}\|_{\mathcal{H}} =: \mathsf{MMD}(\beta, \beta').
\end{aligned}
$$

Note: $\mathsf{MMD}(\beta, \hat{\beta}_n) \leq C n^{-\frac{1}{2}}$ in high probability.

Theorem (Sample Complexity of Sinkhorn Potentials)

Suppose that $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$ *with* $s > d/2$*. Then, there exists a constant* $\bar{\mathsf{r}} = \bar{\mathsf{r}}(\mathcal{X}, \mathsf{c}, d)$ *such that for any* $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ *and any empirical measure* $\hat{\beta}$ *of a set of* n *points independently sampled from* β*, we have, for every* $\tau \in (0, 1]$

$$\|u - u_n\|_\infty = \|\nabla_1 \mathrm{OT}_\varepsilon(\alpha, \beta) - \nabla_1 \mathrm{OT}_\varepsilon(\alpha, \hat{\beta})\|_\infty \leq \frac{C_\varepsilon \log \frac{3}{\tau}}{\sqrt{n}} \quad (4)$$

with probability at least $1 - \tau$*.*

Smoothness properties

Entropic reg $\xrightarrow{\text{[L. et al., 2019]}}$ Lipschitzness of the gradient of Sinkhorn divergence ✓

Entropic reg $\xrightarrow{\text{[Genevay. et al., 2019]}}$ High order regularity of Sinkhorn potentials $\xrightarrow{\text{[L. et al., 2019]}}$ Sample complexity of Sinkhorn gradients ✓

Entropic reg $\xrightarrow{\text{[L. et al., 2018]}}$ High order differentiability of Sinkhorn divergence (in a restricted setting, simplex)

Let's consider the setting: $\mathsf{a}, \mathsf{b} \in \Delta_n$, Δ_n is the simplex.

Theorem

$\mathrm{OT}_\varepsilon : \Delta_n \times \Delta_n \to \mathbb{R}$ *is* C^∞ *differentiable in the interior of the domain.*

The proof is an application of the implicit function theorem.

Entropic regularization provides advantages in terms of regularity itself.

Regularity in which sense? ✓

This regularity enables to show theoretical guarantees of different nature, namely from statistical and optimization point of view.

Part III, Applications:

1. Theoretical guarantees for Sinkhorn barycenters

2. Statistical guarantees for supervised learning with Sinkhorn loss

Barycenter problem

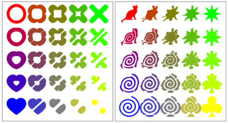


Figure: 2D-Sinkhorn barycenters, taken from [Cuturi and Peyré, Computational OT]

Given $\beta_1, \ldots, \beta_m \in \mathcal{P}(\mathcal{X})$, the barycenter with respect to Sinkhron divergence is

$$\alpha^* = \operatorname*{argmin}_{\alpha \in \mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha), \qquad \mathsf{B}_\varepsilon(\alpha) = \sum_{j=1}^{m} w_j \mathsf{S}_\varepsilon(\alpha, \beta_j)$$

with $w_j \geq 0$, $\sum_j w_j = 1$.

Fixed support and free support

Fixed support methods: fix $\{x_i\}_{i=1}^N$ and set $\alpha^* = \sum_{i=1}^N \mathsf{a}_i \delta_{x_i}$. Optimize B_ε on $\mathsf{a} = (\mathsf{a}_1, \ldots, \mathsf{a}_N)$. E.g. Iterative Bregman projections. Well understood theoretical guarantees.

[Benamou et al., 2015, Dvurechensky et al., 2018]

Free support methods: usually alternate minimization to optimize weights a and support points locations x_i, $i = 1, \ldots, N$ [Cuturi et al., 2014]. Other approaches? Theoretical guarantees of convergence?

FW approach

We propose an approach based on Frank-Wolfe algorithm. The features of this method are the following:

- There is no alternation in optimizing w.r.t points and w.r.t weights
- The barycenter is populated via an iterative procedure
- There is no parameter tuning

Frank-Wolfe algorithm

$\mathcal{W}$ is a real Banach space, with $\mathcal{W}^*$ topological dual

$\mathcal{D} \subset \mathcal{W}^*$ nonempty, convex, closed, bounded set

$\mathsf{G} : \mathcal{D} \to \mathbb{R}$ convex function with $\nabla \mathsf{G} : \mathcal{D} \to \mathcal{W}$ Lipschitz

Algorithm 1 Frank-Wolfe

input: initial $w_0 \in \mathcal{D}$, threshold Δ_k s.t. $\Delta_k(k+2)$ is nondecreasing

For $k = 1, 2, \ldots$

take z_{k+1} s.t. $\langle \nabla \mathsf{G}(w_k), z_{k+1} - w_k \rangle \leq \min_{z \in \mathcal{D}} \langle \nabla \mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_k}{2}$

$w_{k+1} = w_k + \frac{2}{k+2}(z_{k+1} - w_k)$

Convergence rate $O(1/k)$

[Jaggi, 2013]

Recall: $\mathsf{B}_\varepsilon(\alpha) = \sum_{j=1}^m w_j \mathsf{S}_\varepsilon(\alpha, \beta_j)$.

$$
\begin{array}{lcl}
\mathcal{W}^* & \longrightarrow & \mathcal{M}(\mathcal{X}) \\
\mathcal{D} \subset \mathcal{W}^* & \longrightarrow & \mathcal{P}(\mathcal{X}) \subset \mathcal{M}(\mathcal{X}) \\
\mathcal{W} & \longrightarrow & \mathcal{C}(\mathcal{X}) \\
\mathsf{G} : \mathcal{D} \longrightarrow \mathbb{R} & \longrightarrow & \mathsf{B}_\varepsilon : \mathcal{P}(\mathcal{X}) \longrightarrow \mathbb{R}
\end{array}
$$

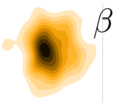
Note that since $\nabla \mathsf{S}_\varepsilon$ is Lipschitz, $\nabla \mathsf{B}_\varepsilon$ is Lipschitz.

Theorem

Suppose that $\beta_1, \ldots \beta_m \in \mathcal{P}(\mathcal{X})$ have finite support and let α_k be the k-th iterate of Alg1 applied to B_ε. Then,

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{C_\varepsilon}{k+2}. \tag{5}$$

Convergence guarantees for this free support method.



What if β_j are not finite and we only have access to samples?

Frank-Wolfe algorithm allows to use approximations of the gradient rather than the real gradient.

Algorithm 1 Frank-Wolfe

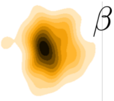
---

input: initial $w_0 \in \mathcal{D}$, treshold Δ_k s.t. $\Delta_k(k+2)$ is nondecreasing

For $k = 1, 2, \ldots$

take z_{k+1} s.t. $\langle \nabla \mathsf{G}(w_k), z_{k+1} - w_k \rangle \leq \min_{z \in \mathcal{D}} \langle \nabla \mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_k}{2}$

$w_{k+1} = w_k + \frac{2}{k+2}(z_{k+1} - w_k)$



What if β_j are not finite and we only have access to samples?

Frank-Wolfe algorithm allows to use approximations of the gradient rather than the real gradient.

We need to control the approximation $\nabla \mathsf{B}_\varepsilon(\cdot, \hat{\beta})$ of $\nabla \mathsf{B}_\varepsilon(\cdot, \beta)$ $\longrightarrow$ this is doable because we have a result on the sample complexity.

Algorithm 1 Frank-Wolfe

input: initial $w_0 \in \mathcal{D}$, treshold Δ_k s.t. $\Delta_k(k+2)$ is nondecreasing

For $k = 1, 2, \ldots$

take z_{k+1} s.t. $\langle \nabla \mathsf{G}(w_k), z_{k+1} - w_k \rangle \leq \min_{z \in \mathcal{D}} \langle \nabla \mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_k}{2}$

$w_{k+1} = w_k + \frac{2}{k+2}(z_{k+1} - w_k)$

Setting:

- $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$ with $s > d/2$
- $\hat{\beta}_1, \ldots, \hat{\beta}_m$ be empirical distributions with $n \in \mathbb{N}$ support points, each independently sampled from $\beta_1, \ldots, \beta_m$.

Let α_k be the k-th iterate of FW applied to $\hat{\beta}_1, \ldots, \hat{\beta}_m$. Then for any $\tau \in (0, 1]$,

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{C_\varepsilon \log \frac{3m}{\tau}}{\min(k, \sqrt{n})}.$$

with probability larger than $1 - \tau$.

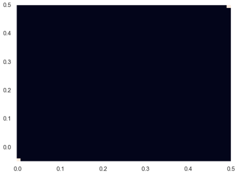
Convergence rate in general setting

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha\in\mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \le \frac{C_\varepsilon \log\frac{3m}{\tau}}{\min(k,\sqrt{n})} \qquad \text{w.h.p.}$$

If $\hat{\beta}_j$, $j = 1, \ldots m$, are sampled with $n = k^2$ points at iteration k:
$\longrightarrow$ rate of convergence: $O(\frac{1}{k})$

If $\hat{\beta}_j$, $j = 1, \ldots m$, are sampled with $n = k$ points at iteration k:
$\longrightarrow$ rate of convergence: $O(\frac{1}{\sqrt{k}})$.

Barycenter of 30 randomly generated nested ellipses on a 50×50 grid [Cuturi et al., 2014]. Each image is interpreted as a probability distribution in 2D.



Learning problem:

- input space $\mathcal{X}$
- output space $\mathcal{Y}$
- unknown probability measure ρ on $\mathcal{X} \times \mathcal{Y}$, accessed through $\{(x_i, y_i)\}_{i=1}^N$ sampled iid from ρ
- loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- expected risk of a function $f : \mathcal{X} \to \mathcal{Y}$

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y) d\rho$$

Goal: find a good approximation $\hat{f}_N$ of the minimizer f^* of $\mathcal{E}$ using $\{(x_i, y_i)\}_{i=1}^N$.

Desirable property: Intuitively we would want that as the number of points increases, so “we get to know ρ better”, then the error that we expect to make using $\hat{f}_N$ rather than f^* should get smaller

$$\mathcal{E}(\hat{f}_N) \xrightarrow{N \to +\infty} \mathcal{E}(f^*) \qquad \text{with high probability}$$

The property above is called consistency.

$\mathcal{X}$	$\longrightarrow$	$\mathcal{X}$
$\mathcal{Y}$	$\longrightarrow$	$\mathcal{P}(\{1, 2, \dots, k\}) = \Delta_k$
unknown ρ	$\longrightarrow$	unknwon ρ
ℓ	$\longrightarrow$	Sinkhorn divergence S_ε

C. Frogner et al. 2015: ‘Learning with Wasserstein loss’:

(a) **Flickr user tags**: street, parade, dragon; **our proposals**: people, protest, parade; **baseline proposals**: music, car, band.

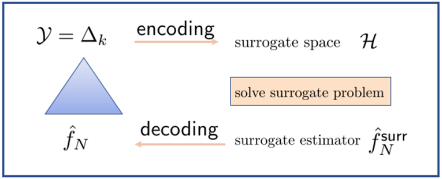
(b) **Flickr user tags**: water, boat, reflection, sunshine; **our proposals**: water, river, lake, summer; **baseline proposals**: river, water, club, nature.

Application: tag prediction, i.e. predicting probability over tags of an image.

The estimator that they proposed was not shown to be consistent and this is what motivated our work [L., et al, 2018].

We interpret the problem of learning with Sinkhorn loss with simplex Δ_k as output space as a *structured prediction* problem which is to be solved using a surrogate framework.

Intuition behind *surrogate framework*:



Where do the regularity properties of Entropic OT come to play?

High order smoothness of S_ε in the interior of Δ_k

$$\downarrow$$

encoding+surrogate+decoding is a valid procedure

$$\downarrow$$

consistent estimator for learning with Sinkhorn loss.

Conclusion

We showed that entropic regularization leads to a range of smoothness properties

- lipschitzness of the gradient
- sample complexity of the potentials
- high order differentiability on the simplex

We used the smoothness properties to show theoretical guarantees in:

- Sinkhorn barycenter problem with free support
- supervised learning with Sinkhorn loss function

Thank you for the attention!

References

Genevay, A. et al., *Sample complexity of Sinkhorn divergences*, AISTATS2019

Luise, G. et al., *Differential Properties of Sinkhorn approximation for Learning with Wasserstein distance*, NeurIPS2018

Luise, G. et al., *Free Support Sinkhorn Barycenter via Frank Wolfe algorithm*, NeurIPS2019

Cuturi, M., Doucet A, *Fast computation of Wasserstein barycenters*, ICML2014

Feydy, J. et al., *Interpolating between Optimal Transport and MMD using Sinkhorn Divergences* , AISTATS2019

Benamou, J.D. et al., *Iterative Bregman Projections for Regularized Transportation Problems*, SIAMJ.Sci.Comput., 37(2)2015

Jaggi, M., *Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization* , ICML2013

Dudley, R.M., *The Speed of Mean Glivenko-Cantelli Convergence*, Ann. Math. Statist. 40, 1969

Pele, O., et al. *Fast and robust earth mover's distances*, ICCV, 2009

Cuturi, M., *Sinkhorn distances: lightspeed computation of optimal trasportation distances*, NIPS, 2013

Proof Lipschitzness

Set $\mathsf{D} := \sup_{y,y\in\mathcal{X}} \mathsf{c}(x,y)$, the diameter of $\mathcal{X}$

Denote by L the operator $\mathsf{L}_\alpha \colon \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is defined as

$$(\forall f \in \mathcal{C}(\mathcal{X})) \qquad \mathsf{L}_\alpha f \colon x \mapsto \int e^{\frac{-\mathsf{c}(x,z)}{\varepsilon}} f(z)\, d\alpha(z); \tag{6}$$

Set $\mathsf{D} := \sup_{y,y\in\mathcal{X}} \mathsf{c}(x,y)$, the diameter of $\mathcal{X}$

Denote by L the operator $\mathsf{L}_\alpha \colon \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is defined as

$$(\forall f \in \mathcal{C}(\mathcal{X})) \qquad \mathsf{L}_\alpha f \colon x \mapsto \int e^{\frac{-\mathsf{c}(x,z)}{\varepsilon}} f(z)\, d\alpha(z); \tag{6}$$

Theorem (Birkhoff-Hopf Theorem)

Let $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$ and $\alpha \in \mathcal{P}(\mathcal{X})$. Then, for every $f, f' \in \mathcal{C}_+(\mathcal{X})$ such that $f \sim f'$, we have

$$d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \leq \lambda\, d_H(f, f'). \tag{7}$$

Proof Lipschitzness II

Let $\alpha \in \mathcal{P}(\mathcal{X})$. We define the map $\mathsf{A}_\alpha \colon \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$, such that

$$(\forall\, f \in \mathcal{C}_{++}(\mathcal{X})) \qquad \mathsf{A}_\alpha(f) = 1/(\mathsf{L}_\alpha f), \tag{8}$$

Set $f := e^{\frac{u}{\varepsilon}}$, $g := e^{\frac{v}{\varepsilon}}$. Recall that

$$\begin{cases} e^{-\frac{u(x)}{\varepsilon}} = \displaystyle\int_{\mathcal{X}} e^{\frac{v(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d\beta(y) \quad (\forall\, x \in \operatorname{supp}(\alpha)) \\ e^{-\frac{v(y)}{\varepsilon}} = \displaystyle\int_{\mathcal{X}} e^{\frac{u(x)-\mathsf{c}(x,y)}{\varepsilon}}\, d\alpha(x) \quad (\forall\, y \in \operatorname{supp}(\beta)), \end{cases}$$

Then it holds

$$f = \mathsf{A}_\beta(g) \qquad \text{and} \qquad g = \mathsf{A}_\alpha(f), \tag{9}$$

or equivalently, by setting $\mathsf{A}_{\beta\alpha} = \mathsf{A}_\beta \circ \mathsf{A}_\alpha$ and $\mathsf{A}_{\alpha\beta} = \mathsf{A}_\alpha \circ \mathsf{A}_\beta$,

$$f = \mathsf{A}_{\beta\alpha}(f) \qquad \text{and} \qquad g = \mathsf{A}_{\alpha\beta}(g). \tag{10}$$

Theorem (Hilbert's metric contraction for $\mathsf{A}_{\beta\alpha}$)

The map $\mathsf{A}_{\beta\alpha} : \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ has a unique fixed point up to positive scalar multiples. Moreover, let $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$. Then, for every $f, f' \in \mathcal{C}_{++}(\mathcal{X})$,

$$d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta\alpha}(f')) \leq \lambda^2 \; d_H(f, f'). \tag{11}$$

Relation between Hilbert distance and infinity norm:

$$\frac{\varepsilon}{2} d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}) \leq \left\| u - u' \right\|_\infty \leq \varepsilon \; d_H(e^{u/\varepsilon}, e^{u'/\varepsilon})$$

Proof Lipschitzness V

Putting everything together:

$$d_H(f, f') \le \frac{1}{1-\lambda^2}\, d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)).$$

Using triangle inequality and some computations on $d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f))$, we arrive at a point where we only need to estimate:

$$\begin{aligned}[(\mathsf{L}_{\beta'} - \mathsf{L}_{\beta})g](x) &= \int e^{\frac{-\mathsf{c}(x,z)}{\varepsilon}} g(z)\, d(\beta - \beta')(z) \\ &= \left\langle e^{\frac{-\mathsf{c}(x,\cdot)}{\varepsilon}} g, \beta - \beta' \right\rangle \le \|g\|_\infty \left\|\beta - \beta'\right\|_{TV}.\end{aligned}$$

Lipschitz constant