Uniform Quantitative Stability of Sinkhorn

George Deligiannidis

Joint with V. de Bortoli and A. Doucet

Department of Statistics, Oxford University

Athens Probability Colloquium, March 2023

Optimal Transport

- Coupling: A coupling of two probability measures μ, ν, on spaces $X, Y \subset \mathbb{R}^d$ respectively, a coupling is a joint distribution π, with μ, ν as its marginals,

$$\int f(x)\pi(dx, dy) = \int f(x)\mu(dx), \qquad \int g(y)\pi(dx, dy) = \int g(y)\nu(dy).$$

- We write $\mathcal{C}(\mu, \nu)$ for the collection of couplings of μ, ν.
- Optimal Transport: the basic problem of Optimal Transport is to find a coupling $\pi \in \mathcal{C}(\mu, \nu)$ that minimizes

$$\inf_{\pi \in \mathcal{C}(\mu,\nu)} \int \pi(dx, dy)c(x, y); \qquad \mathrm{OT}(\mu, \nu)$$

where $c : X \times Y \to \mathbb{R}$ is any cost function.

Optimal Transport

$$\inf_{\pi \in \mathcal{C}(\mu,\nu)} \int \pi(\mathrm{d}x, \mathrm{d}y) c(x,y);$$

- this formulation goes back to Kantorovich;
- the earlier formulation by Monge considers mappings $y = T(x)$, that is $\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)\delta_{T(x)}(\mathrm{d}y)$ and results in non-convex optimisation.
- Brenier 1987: Assuming $c(x,y) = |x-y|^2$, μ, ν have finite second moments and μ gives measure 0 to all sets of Hausdorff dimension $\leq d-1$, there exists a convex mapping $\varphi : \mathcal{X} \mapsto \mathcal{Y}$ such that optimal coupling takes the form of an optimal transport map $\mu(\mathrm{d}x)\delta_{\nabla\varphi(x)}(\mathrm{d}y)$
- When μ is discrete, π is a proper coupling.
- For discrete measures with N atoms, worst case computational cost is $O(N^{5/2})$ for assignment problem, using Hungarian (auction) algorithm, see e.g. Mérigot and Oudet 2016.
- This can be improved by considering the entropy regularised version.

Entropy Regularised Optimal Transport

- Cuturi 2013 realised that the entropy regularised problem

$$\Pi_\epsilon^{\mu,\nu} = \underset{\pi \in \mathcal{C}(\mu,\nu)}{\arg\min} \int \pi(\mathrm{d}x, \mathrm{d}y)\|x-y\|^2 + \epsilon\, \mathrm{KL}\left(\pi | \mu \otimes \nu\right) \qquad \mathrm{OT}_\epsilon(\mu,\nu)$$

  can be solved efficiently using the Iterative Proportional Fitting Procedure (IPFP), aka Sinkhorn's algorithm.

- the computational cost is roughly $O(N^2)$, see Altschuler, Weed, and Rigollet 2017, ignoring dimensionality.

- Equivalent to (static) Schrödinger bridge problem

$$\inf_{\pi \in \mathcal{C}(\mu,\nu)} \mathrm{KL}\left(\pi | \Gamma_\epsilon\right), \quad \Gamma_\epsilon(\mathrm{d}x, \mathrm{d}y) = \exp[-c(x,y)/\epsilon]\mu(\mathrm{d}x)\nu(\mathrm{d}y).$$

- Solution always in the form

$$\frac{\mathrm{d}\Pi_\epsilon^{\mu,\nu}}{\mathrm{d}\mu \otimes \nu} = \exp[\varphi_\epsilon(x) + \psi_\epsilon(y) - c(x,y)/\epsilon],$$

  where $(\varphi_\epsilon, \psi_\epsilon)$ are known as the potentials.

Schrödinger bridge problem

- Question: what is the most likely evolution of a stochastic process $\{X_t : t \in [0,1]\}$, $X_0 \sim \mu$ and $X_1 \sim \nu$?
- one formulation is

$$\operatorname*{argmin}_{\mathbb{P}} \mathrm{KL}\left(\mathbb{P} | \mathbb{W}_\epsilon\right) \tag{1}$$

- minimization over $\mathcal{P}\left(\mathcal{C}([0,1]; \mathbb{R}^n)\right)$ with marginals μ and ν at times 0 and 1 respectively;
- $\mathbb{W}_\epsilon$ is the law of Brownian motion multiplied by $\sqrt{\epsilon/2}$.
- The solution to (1) can then be expressed in terms of the solution to $\mathrm{OT}_\epsilon(\mu, \nu)$ and the bridges of the measure $\mathbb{W}_\epsilon$.
- The quadratic cost arises from the log-transition density of Brownian motion—different reference measures can give rise to different costs.

Recent applications of Schrödinger bridges

How is the static Schrödinger bridge useful in practice?

- Idea is to use solution π_ϵ^* of OT_ϵ and treat it as an approximation to the solution π^* of OT.
- A lot of recent progress quantifying $\pi_\epsilon^* \to \pi^*$.
- In recent applications, the Schrödinger bridge is used for its own benefits rather than as a computationally feasible approximation to the standard optimal transport problem.

Differentiable Particle Filtering

- Corenflos et al. 2021 used the solution to Schrödinger bridge to build a particle filtering scheme.
- Suppose $\widehat{\pi}_t = N^{-1} \sum_{j=1}^{N} \delta_{X_j^t}$ is particle approximation of π_t;
- Let

$$\widetilde{X}_j^{t+1} \sim q_{t+1}(\cdot|X_j^t), j \in [N]; \qquad \omega_j^{t+1} = \frac{f_\theta(\widetilde{X}_j^{t+1}|X_j^t) g_\theta(Y^{t+1}|\widetilde{X}_j^{t+1})}{q_{t+1}(\widetilde{X}_j^{t+1}|X_j^t)}$$

$$\mathbb{P}(I_j^{t+1} = k) = \frac{\omega_k^{t+1}}{\sum_{j=1}^{N} \omega_j^{t+1}} \qquad X_j^{t+1} \overset{\text{iid}}{\sim} \sum_{k=1}^{N} \frac{\omega_k^{t+1}}{\sum_{j=1}^{N} \omega_j^{t+1}} \delta_{\widetilde{X}_k^{t+1}}$$

- Resampling: estimators not differentiable wrt θ.
- There are approximate ways to bypass this.

Differentiable Particle Filtering

- Corenflos et al. 2021 solves the ϵ-entropy regularised OT problem between

$$\alpha_N^{(t)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widetilde{X}_t^i}, \quad \beta_N^{(t)} = \sum_i w_t^i \delta_{\widetilde{X}_t^i}$$

  to obtain a matrix $\boldsymbol{P}_\epsilon = \boldsymbol{P}_\epsilon(\theta)$;
- resampling replaced by ensemble transform $\boldsymbol{X}_t = \boldsymbol{P}_\epsilon \tilde{\boldsymbol{X}}_t$
- for $\epsilon > 0$ $\theta \mapsto \boldsymbol{P}_\epsilon(\theta)$ differentiable $\implies$ end to end differentiable estimators.
- Proof of consistency: compare with the solution of the unregularised OT problem (needs $\epsilon = \epsilon_N \to 0$)
- OR we can compare the solution of $\mathrm{OT}_\epsilon(\alpha_N, \beta_N)$ with that of $\mathrm{OT}_\epsilon(\alpha_N, \beta_N)$ where $\alpha = \lim_{N\to\infty} \alpha_N, \beta = \lim_N \beta_N$.

Stability of OT

This brings us to the following question

- Question: suppose $\alpha_n \to \alpha$ and $\beta_n \to \beta$ (say weakly);
- *D*oes the solution of $\mathrm{OT}(\alpha_n, \beta_n)$ converges in some sense to that of $\mathrm{OT}(\alpha, \beta)$?
- In the realm of Brenier's theorem, we could talk about convergence of the transport maps.
- In classical Optimal Transport there is a classical, qualitative solution.

Stability of OT

Theorem (5.20 in Villani 2009)

Let $\mathcal{X}, \mathcal{Y}$ be Polish spaces and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a a continuous, lower bounded cont function. Let $(\mu_k), (\nu_k)$ be sequences of probability measures on $\mathcal{X}, \mathcal{Y}$ respectively, such that $\mu_k \to \mu$ and $\nu_k \to \nu$ weakly. For each k, let π_k be an optimal transference plan between μ_k and ν_k.
If for all $k \in \mathbb{N}$, $\int c d\pi_k < +\infty$, then up to extraction of a subsequence, π_k converges weakly to some c-cyclically monotone transference plan $\pi \in \mathcal{C}(\mu, \nu)$.
If moreover $\liminf_k \int c d\pi_k < \infty$, then the optimal transport cost between μ, ν is finite and π is an optimal transport plan.

Stability of OT

- If the optimal transport plan between μ and ν is unique then there is no need to extract a subsequence.
- If plan is given by map, maps converge in probability (Corollary 5.23 Villani 2009).
- Proof is based on compactness and characterization of optimal transport plans in terms of c-cyclical monotonicity: a set $\Gamma \subset X \times Y$ is called c-cyclically monotone if for all collections $(x_1, y_1), \ldots, (x_N, y_N) \subset \Gamma$ we have

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}), \quad y_{N+1} := y_1 .$$

a transport plan is called c-cyclically monotone if its support is.

Quantitative Stability of OT

Li and Nochetto 2021

- $\mu_h \approx \mu$, $\nu_h \approx \nu$, γ_h solves $\mathrm{OT}(\mu_h, \nu_h)$, and $T = \nabla\varphi$ is the unique λ-Lipschitz optimal transport map between μ, ν. Then

$$\left(\int_{X\times Y} |T(x)-y|^2 \mathrm{d}\gamma_h(x,y)\right)^{1/2} \leq 2\lambda^{1/2}\Delta_h^{1/2}\left[\mathbf{W}_2(\mu,\nu)+\Delta_h\right]^{1/2} \quad + (\lambda \vee 1)\Delta_h, \tag{2}$$

$$\Delta_h := \mathbf{W}_2(\mu,\mu_h) + \mathbf{W}_2(\nu,\nu_h) \tag{3}$$

Mérigot, Delalande, and Chazal 2020

- ρ, μ, ν p.m.s, $T_\mu^\# \rho = \mu$, $T_\nu^\# \rho = \nu$ Brenier maps.
- Then

$$\mathbf{W}_2(\mu,\nu) \leq \|T_\mu - T_\nu\|_{L^2(\rho)} \leq C\mathbf{W}_p(\mu,\nu)^{2/15},$$

  for any $p \geq 1$, where C depends on the dimension and the sets X, Y.
- If T_μ is K-Lipschitz then the exponent can be improved to $1/2$.
- They also show that the Monge embedding $\mu \mapsto T_\mu$ is in general not better than $1/2$-Holder.

Stability of Regularised OT

- Until 2-3 years ago much less was known on stability of regularised OT.
- Analogue of cyclical monotonicity recently done in Ghosal, Nutz, and Bernton 2021 who proved the qualitative stability of Schödinger bridges (no compactness required).
- For compact state spaces Luise et al. 2019 prove that

$$\|\varphi_\epsilon^{\mu,\nu} - \varphi_\epsilon^{\mu',\nu'}\|_\infty \leq C(d,\epsilon,X,Y)\left\{\|\mu-\mu'\|_{TV} + \|\nu-\nu'\|_{TV}\right\}.$$

  NOTE: TV is too strong to capture convergence of empirical measures.

- For smooth costs, $c \in C^{s+1}$, $s > d/2$, and ν_n an empirical version of ν, using MMD metrics, w.prob $> 1-\tau$

$$\|\varphi_\epsilon^{\mu,\nu} - \varphi_\epsilon^{\mu,\nu_n}\|_\infty \leq C(d,\epsilon,X,Y)\log(3/\tau)n^{-1/2},$$

Sample complexity of cost

Stability is related to sample complexity.

- Suppose we want to estimate $\mathbf{W}_p(\mu, \nu)$ using samples from μ, ν.
- How many samples do we need to be within ϵ from the truth?
- Let $S_\epsilon(\mu, \nu)$ be the cost of the OT_ϵ, write $S(\mu, \nu)$ for the unregularised.
- If μ_n, ν_n empirical approximations to μ, ν in general $S(\mu, \nu) = S(\mu_n, \nu_n) + O(n^{-1/d})$;
- BUT miraculously for any $\epsilon > 0$, $S_\epsilon(\mu, \nu) = S_\epsilon(\mu_n, \nu_n) + O(n^{-1/2})$, see Genevay et al. 2019; Mena and Niles-Weed 2019.

Sample complexity of Coupling

- Interested in full coupling $\Pi_\epsilon^{\mu,\nu}$: when measured say in wasserstein distance, sample complexity depends on dimension;
- Say μ_n, ν_n are empirical versions of μ, ν. Then trivially we have that

$$\mathbf{W}_1\left(\Pi_\epsilon^{\mu_n,\nu_n}, \Pi_\epsilon^{\mu,\nu}\right) \geq \mathbf{W}_1\left(\mu_n, \mu\right) = \Omega(n^{-1/d}),$$

  see e.g. Fournier and Guillin 2015.
- But $\mathbf{W}$ is a uniform metric. What if we only care about expectations of just a few functionals?
- Rigollet and Stromme 2022 obtain $O(n^{-1/2})$ rates in non-uniform metrics, e.g. measuring expectation of a single function.

Iterative Proportional Fitting (aka Sinkhorn) I

- Given two probability measures μ, ν and $\epsilon > 0$ let $\Pi^{(0)} = e^{-c}\mu \otimes \nu$;
- the IPFP iteratively solves the following one-sided Bridge problems

$$\Pi^{(2j+1)} = \arg\ \min_{\Pi:\Pi_0=\mu} \mathrm{KL}\left(|() \Pi|\Pi^{(2j)}\right)$$
$$\Pi^{(2j+2)} = \arg\ \min_{\Pi:\Pi_1=\nu} \mathrm{KL}\left(|() \Pi|\Pi^{(2j+1)}\right).$$

- it alternates between projecting on measures with first marginal μ and measures with second marginal ν.
- It's easy to see that we can write

$$\frac{d\Pi^{(j)}}{d\Pi^{j-1}}(x,y) := \frac{d\nu}{d\Sigma_2^{\#}\Pi^{j-1}}(y), \quad j \text{ odd},$$
$$\frac{d\Pi^{(j)}}{d\Pi^{j-1}}(x,y) := \frac{d\mu}{d\Sigma_1^{\#}\Pi^{j-1}}(y), \quad j \text{ even},$$

where $\Sigma_1(x,y) = x,\ \Sigma_2(x,y) = y$.

Iterative Proportional Fitting (aka Sinkhorn) II

- The dual formulation in terms of potentials is

$$\varphi^{(t+1)}(x) := -\log \int \exp\{\psi^{(t)}(y) - c(x,y)/\epsilon\}\nu(\mathrm{d}y)$$
$$\psi^{(t+1)}(y) := -\log \int \exp\{\varphi^{(t+1)}(x) - c(x,y)/\epsilon\}\mu(\mathrm{d}x)$$

- For any $c \in \mathbb{R}$, $(\varphi^{(t)} - c, \psi^{(t)} + c)$ defines the same measure; we fix this choice following Carlier 2021, so that $\mu[\varphi^{(t+1)}] = 0$.

Main results

Theorem 1

Suppose that X, Y *are compact metric spaces and* $c \in \mathrm{Lip}(X \times Y)$. *For any* $\pi_0, \widehat{\pi}_0 \in \mathscr{P}(X)$, $\pi_1, \widehat{\pi}_1 \in \mathscr{P}(Y)$ *let* $(\mathbb{P}^n)_{n\in\mathbb{N}}$ *and* $(\widehat{\mathbb{P}}^n)_{n\in\mathbb{N}}$ *the IPFP sequence with marginals* (π_0, π_1) *respectively* $(\widehat{\pi}_0, \widehat{\pi}_1)$. *Then any* $n \in \mathbb{N}$ *we have*

$$\mathbf{W}_1(\mathbb{P}^n, \widehat{\mathbb{P}}^n) \leq C\left\{\mathbf{W}_1(\pi_0, \widehat{\pi}_0) + \mathbf{W}_1(\pi_1, \widehat{\pi}_1)\right\}, \tag{4}$$

with

$$C = \mathrm{e}^{10\|c\|_\infty}\{1 + (2\mathrm{Lip}(c) + 10)(\mathrm{diam}(X) + \mathrm{diam}(Y))\}. \tag{5}$$

- Good: The result is uniform in the iterations of Sinkhorn.
- Good: Sample complexity is sharp.
- The bad: Constant is terrible.

Main results (ctd)

As an immediate consequence of Theorem 1 and the fact that the IPFP sequence converges, we obtain the quantitative stability of Schrödinger bridge.

Corollary 2

For any $\pi_0, \widehat{\pi}_0 \in \mathscr{P}(\mathsf{X})$, $\pi_1, \widehat{\pi}_1 \in \mathscr{P}(\mathsf{Y})$ *let* $\mathbb{P}^*$, *respectively* $\widehat{\mathbb{P}}^*$, *be the Schrödinger bridge with marginals* (π_0, π_1), *respectively* $(\widehat{\pi}_0, \widehat{\pi}_1)$. *Then we have*

$$\mathbf{W}_1(\mathbb{P}^*, \widehat{\mathbb{P}}^*) \leq C\{\mathbf{W}_1(\pi_0, \widehat{\pi}_0) + \mathbf{W}_1(\pi_1, \widehat{\pi}_1)\}, \tag{6}$$

with C *as in Theorem 1.*

Background on Hilbert projective metric I

Will now sketch the main idea for the Schrödinger bridge, rather than IPFP as it is a little clearer.

- In compact spaces we can employ the machinery of the Birkhoff-Hopf contraction theorem;
- contraction w.r.t. the Hilbert metric.
- It is a projective metric in the sense that it measures distances between rays $\{\lambda x : \lambda \geq 0\}$ rather than points.
- For function spaces: oscilation in the log-scale, ie

$$d_H(f,g) = \|\log(f/g)\|_{\mathrm{osc}} := \sup_x \log \frac{f(x)}{g(x)} - \inf_y \log \frac{f(y)}{g(y)}.$$

We are now ready to state the Birkhoff contraction theorem.

- $(V, \|\cdot\|), (V', \|\cdot\|)$ normed real vector spaces;
- $K \subset V, K' \subset V'$ cones;
- $C \subset K, C' \subset K'$ convex parts;

Background on Hilbert projective metric II

- d_H, d'_H the Hilbert metric on C, C' respectvely;
- $T : V \to V'$ is a linear mapping such that $T(C) \subset C'$.

Theorem (Birhoff Contraction Theorem)

Then

$$\kappa(T) := \sup_{x,y \in C} \frac{d'_H(T(x), T(y))}{d_H(x,y)} \leq \tanh(\Delta(T)/4), \tag{7}$$

where the projective diameter $\Delta(T)$ of T is defined by

$$\Delta(T) := \sup \{ d'_H(T(x), T(y)) : x, y \in C, \|x\| = \|y\| = 1 \}.$$

Since $\Delta(T)$ is finite $\kappa(T) < 1$.

Sketch of proof I

- In our context
$$C = \mathscr{C}(X; (0, \infty)), \quad C' = \mathscr{C}(Y; (0, \infty)).$$
- Letting $\mu \in \mathscr{P}(X), \nu \in \mathscr{P}(Y)$ define two linear maps of interest through
$$(\mathcal{E}_\mu f)(y) := \int f(x) K(x,y) \mu(dx); \quad \mathcal{E}_\mu : C \to C' \tag{8}$$
$$(\mathcal{E}_\nu g)(x) := \int g(y) K(x,y) \nu(dy); \quad \mathcal{E}_\nu : C' \to C. \tag{9}$$
- Let also $\mathscr{I} : f \mapsto f^{-1}$, where we overload the operator to act on both C and C'.
- As pointed out in Chen, Georgiou, and Pavon 2016 it is easy to show that $\mathscr{I}$ is an isometry w.r.t. the Hilbert metric;
- Also we can bound the projective diameter of $\mathcal{E}_\mu, \mathcal{E}_\nu$.

Sketch of proof II

- In this notation the Sinkhorn iteration that takes $e^{\varphi^{(t)}} \to e^{\varphi^{(t+1)}}$ can be written as

$$\mathscr{S}_{\mu,\nu} \exp[\varphi^{(t)}] := \left[\mathscr{I} \circ \mathcal{E}_\nu \circ \mathscr{I} \circ \mathcal{E}_\mu\right] [\exp \varphi^{(t)}], \tag{10}$$

$$\mathscr{S}^\dagger_{\mu,\nu} \exp[\psi^{(t)}] := \left[\mathscr{I} \circ \mathcal{E}_\mu \circ \mathscr{I} \circ \mathcal{E}_\nu\right] \exp[\psi^{(t)}]. \tag{11}$$

- The Birkhoff contraction theorem show that both of these maps are contractions in the Hilbert metric.
- The potentials $(\varphi_\epsilon, \psi_\epsilon), (\widehat{\varphi}_\epsilon, \widehat{\psi}_\epsilon)$ corresponding to $(\mu, \nu), (\widehat{\mu}, \widehat{\nu})$ are fixed points of $(\mathscr{S}_{\mu,\nu}, \mathscr{S}^\dagger_{\mu,\nu})$ and $(\mathscr{S}_{\widehat{\mu},\widehat{\nu}}, \mathscr{S}^\dagger_{\widehat{\mu},\widehat{\nu}})$ resp.

Sketch of proof:main ideas I

For $h \in \mathrm{Lip}(X \times Y)$, writing

$$F_\epsilon(x,y) = \exp\{\varphi_\epsilon(x) + \psi_\epsilon(y)\}, \quad \widehat{F}_\epsilon(x,y) = \exp\{\widehat{\varphi}_\epsilon(x) + \widehat{\psi}_\epsilon(y)\}$$

we want to write

$$\begin{aligned}
&\int h(x,y)\left[\Pi_\epsilon^{\mu,\nu}(\mathrm{d}x,\mathrm{d}y) - \Pi_\epsilon^{\widehat{\mu},\widehat{\nu}}(\mathrm{d}x,\mathrm{d}y)\right] \\
&= \int h(x,y)F_\epsilon(x,y)K_\epsilon(x,y)\mu(\mathrm{d}x)\nu(\mathrm{d}y)s - \int h(x,y)\widehat{F}_\epsilon(x,y)K_\epsilon(x,y)\widehat{\mu}(\mathrm{d}x)\widehat{\nu}(\mathrm{d}y) \\
&= \int h\left[F_\epsilon - \widehat{F}_\epsilon\right]K_\epsilon(x,y)\mu(\mathrm{d}x)\nu(\mathrm{d}y) + \int h\widehat{F}_\epsilon K_\epsilon(x,y)\left[\mu\otimes\nu - \widehat{\mu}\otimes\widehat{\nu}\right].
\end{aligned}$$

- The second term looks like it can be controlled by $\mathbf{W}_1(\mu\otimes\nu, \widehat{\mu}\otimes\widehat{\nu})$;
- But the potentials $\widehat{F}_\epsilon$ are not even necessarily smooth, or even well-defined everywhere….

Sketch of proof:main ideas II

- Canonical extensions: although $\widehat{\varphi}_\epsilon, \widehat{\psi}_\epsilon$ are only defined on the supports of $\widehat{\mu}, \widehat{\nu}$ they are fixed points of the Sinkhorn iteration, that is

$$\widehat{\varphi}_\epsilon(y) = -\log \int \exp\{\widehat{\psi}_\epsilon(y) - c(x,y)/\epsilon\}\widehat{\nu}(\mathrm{d}y) \tag{12}$$

$$\widehat{\psi}_\epsilon(y) = -\log \int \exp\{\widehat{\varphi}_\epsilon(x) - c(x,y)/\epsilon\}\widehat{\mu}(\mathrm{d}x). \tag{13}$$

- We can use this to extend $\widehat{\varphi}_\epsilon, \widehat{\psi}_\epsilon$ X, Y;
- they inherit smoothness from cost, see also Luise et al. 2019,
- So indeed

$$\int h(x,y)\left[\Pi_\epsilon^{\mu,\nu}(\mathrm{d}x,\mathrm{d}y) - \Pi_\epsilon^{\widehat{\mu},\widehat{\nu}}(\mathrm{d}x,\mathrm{d}y)\right] \leq \int h\left[F_\epsilon - \widehat{F}_\epsilon\right] K_\epsilon(x,y)\mu(\mathrm{d}x)\nu(\mathrm{d}y) + \mathbf{W}_1(\mu\otimes\nu, \widehat{\mu}\otimes\widehat{\nu}).$$

- If we can control $\|F_\epsilon - \widehat{F}_\epsilon\|_\infty$ we should be done.

Sketch of proof:main ideas I

- Next we use the fact that $F_\epsilon, \widehat{F}_\epsilon$ are fixed points of Sinkhorn iterations to control $F_\epsilon - \widehat{F}_\epsilon$.
- Idea here is

$$
\begin{aligned}
d_H(F_\epsilon, \widehat{F}_\epsilon) &= d_H\left(\mathscr{S}_{\mu,\nu} F_\epsilon, \mathscr{S}_{\widehat{\mu},\widehat{\nu}} \widehat{F}_\epsilon\right) \\
&= \underbrace{d_H\left(\mathscr{S}_{\mu,\nu} F_\epsilon, \mathscr{S}_{\mu,\nu} \widehat{F}_\epsilon\right)}_{\text{contraction of Sinkhorn}} + \underbrace{d_H\left(\mathscr{S}_{\mu,\nu} \widehat{F}_\epsilon, \mathscr{S}_{\widehat{\mu},\widehat{\nu}} \widehat{F}_\epsilon\right)}_{\leq \mathbf{W}_1(\mu\otimes\nu, \widehat{\mu}\otimes\widehat{\nu})} \\
&= \kappa d_H\left(F_\epsilon, \widehat{F}_\epsilon\right) + C\mathbf{W}_1(\mu\otimes\nu, \widehat{\mu}\otimes\widehat{\nu}) \\
d_H(F_\epsilon, \widehat{F}_\epsilon) &\leq \frac{C}{1-\kappa}\mathbf{W}_1(\mu\otimes\nu, \widehat{\mu}\otimes\widehat{\nu}).
\end{aligned}
$$

- Final issue is that $d_H(F_\epsilon, \widehat{F}_\epsilon)$ only controls the oscillations $\|\log F_\epsilon - \log \widehat{F}_\epsilon\|_{\mathrm{osc}}$ rather than the supremum;

Sketch of proof:main ideas II

- To bypass this issue notice that $\widehat{F}_\epsilon$ defines a probability density w.r.t. reference measure $K_\epsilon \mathrm{d}\widehat{\mu} \otimes \widehat{\nu}$.

$$\int \frac{\widehat{F}_\epsilon}{F_\epsilon} F_\epsilon K_\epsilon \mathrm{d}\mu \otimes \nu = \int \widehat{F}_\epsilon K_\epsilon \mathrm{d}\widehat{\mu} \otimes \widehat{\nu} + C\mathbf{W}_1(\mu \otimes \nu, \widehat{\mu} \otimes \widehat{\nu}) \tag{14}$$

$$= \int \widehat{F}_\epsilon K_\epsilon \widehat{\mu} \otimes \widehat{\nu} + C\mathbf{W}_1(\mu \otimes \nu, \widehat{\mu} \otimes \widehat{\nu}) \tag{15}$$

$$= 1 + C\mathbf{W}_1(\mu \otimes \nu, \widehat{\mu} \otimes \widehat{\nu}) \tag{16}$$

$$\tag{17}$$

- Recall $F_\epsilon K_\epsilon \mu \otimes \nu$ is a probability measure;
- thus the random variable $\widehat{F}_\epsilon / F_\epsilon(X, Y)$ with $(X, Y) \sim F_\epsilon K_\epsilon \mu \otimes \nu$, must either be a.s. equal to $1 + C\mathbf{W}_1(\mu \otimes \nu, \widehat{\mu} \otimes \widehat{\nu})$, or must take values both above and below $1 + C\mathbf{W}_1(\mu \otimes \nu, \widehat{\mu} \otimes \widehat{\nu})$.
- But recall that we have controlled the oscillation $\| \log F_\epsilon - \log \widehat{F}_\epsilon \|_{\mathrm{osc}}$;

Sketch of proof:main ideas III

- therefore we can anchor $\log F_\epsilon / \widehat{F}_\epsilon$ near 1 and therefore

$$\log \frac{F_\epsilon}{\widehat{F}_\epsilon} \leq C\mathbf{W}_1(\mu \otimes \nu, \widehat{\mu} \otimes \widehat{\nu}) + \| \log \widehat{F}_\epsilon(x,y) - \log F_\epsilon(x,y) \|_{\text{osc}}.$$

- We can similarly control $\widehat{F}_\epsilon / F_\epsilon$.
- A little more work obtains stability uniform in the iterations of Sinkhorn. □

Recent progress (an incomplete summary)

Eckstein and Nutz 2021

- non-compact space; consider the Schrödinger bridge, ie the limit of Sinkhorn and prove Hölder continuity, rather than Lipschitz, in the marginals; rate for Sinkhorn convergence.

Rigollet and Stromme 2022

- Dimension free rates for barycentric approximation, density in L_2 and expectations of single functionals.

Carlier, Chizat, and Laborde 2022

- Mapping sending marginals to potentials Lipschitz continuous wrt $(W_2(\mu,\mu')^2 + W_2(\nu,\nu')^2)^{1/2}$.

Chiarini et al. 2022

- gradient estimates for potentials and stability in terms of KL and a negative order Sobolev norm.

Bayraktar, Eckstein, and Zhang 2022

- Quantitative stability estimates for general f-divergences.

Thanks for your attention

An incomplete list of references I

[1] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems*. 2017.

[2] Erhan Bayraktar, Stephan Eckstein, and Xin Zhang. "Stability and Sample Complexity of Divergence Regularized Optimal Transport". In: *arXiv preprint arXiv:2212.00367* (2022).

[3] Yann Brenier. "Polar decomposition and increasing rearrangement of vector-fields". In: *COMPTES RENDUS DE L ACADEMIE DES SCIENCES SERIE I-MATHEMATIQUE* 305.19 (1987), pp. 805–808.

[4] Guillaume Carlier. "On the linear convergence of the multi-marginal Sinkhorn algorithm". In: *HAL preprint hal-03176512* (2021).

[5] Guillaume Carlier, Lénaïc Chizat, and Maxime Laborde. "Lipschitz Continuity of the Schr\" odinger Map in Entropic Optimal Transport". In: *arXiv preprint arXiv:2210.00225* (2022).

An incomplete list of references II

[6] Yongxin Chen, Tryphon Georgiou, and Michele Pavon. "Entropic and displacement interpolation: a computational approach using the Hilbert metric". In: *SIAM Journal on Applied Mathematics* 76.6 (2016), pp. 2375–2396.

[7] Alberto Chiarini, Giovanni Conforti, Giacomo Greco, and Luca Tamanini. "Gradient estimates for the Schr\" odinger potentials: convergence to the Brenier map and quantitative stability". In: *arXiv preprint arXiv:2207.14262* (2022).

[8] Adrien Corenflos, James Thornton, George Deligiannidis, and Arnaud Doucet. "Differentiable particle filtering via entropy-regularized optimal transport". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2100–2111.

[9] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*. 2013.

[10] Stephan Eckstein and Marcel Nutz. "Quantitative stability of regularized optimal transport and convergence of sinkhorns algorithm". In: *arXiv preprint arXiv:2110.06798* (2021).

An incomplete list of references III

[11] Nicolas Fournier and Arnaud Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability Theory and Related Fields* 162.3 (2015), pp. 707–738.

[12] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. "Sample complexity of sinkhorn divergences". In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 1574–1583.

[13] Promit Ghosal, Marcel Nutz, and Espen Bernton. "Stability of Entropic Optimal Transport and Schrödinger Bridges". In: *arXiv preprint arXiv:2106.03670* (2021).

[14] Wenbo Li and Ricardo H Nochetto. "Quantitative stability and error estimates for optimal transport plans". In: *IMA Journal of Numerical Analysis* 41 (3 2021), pp. 1941–1965.

An incomplete list of references IV

[15] Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. "Sinkhorn Barycenters with Free Support via Frank-Wolfe Algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[16] Gonzalo Mena and Jonathan Niles-Weed. "Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem". In: *Advances in Neural Information Processing Systems* 32 (2019).

[17] Quentin Mérigot, Alex Delalande, and Frédéric Chazal. "Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3186–3196.

[18] Quentin Mérigot and Edouard Oudet. "Discrete optimal transport: complexity, geometry and applications". In: *Discrete & Computational Geometry* 55.2 (2016), pp. 263–283.

An incomplete list of references V

[19] Philippe Rigollet and Austin J Stromme. "On the sample complexity of entropic optimal transport". In: *arXiv preprint arXiv:2206.13472* (2022).

[20] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer, 2009.