Entropic Regularization of Optimal Transport

Tianyu Xie
2022/5/15

1. Entropic Regularization

Entropic Regularization of Optimal Transport

Definition

- The **discrete entropy of a coupling matrix** is defined as

$$\mathbf{H}(\mathbf{P}) \stackrel{\text{def.}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1),$$

- This definition is a little different to the common definition. The whole expression is just $\sum_{i,j} \int_0^{P_{i,j}} \ln pdp$, which gives us an alternative motivation for Shannon entropy.

- Remarks:
  - $H(P)$ is defined to be $-\infty$ if one of the entries is 0 or negative.
  - The function H is 1-strongly concave, because its Hessain matrix is $\partial^2 H(P) = -\operatorname{diag}(1/P_{i,j})$, and $P_{i,j} \leq 1$.

Entropic Regularization of Optimal Transport

Regularization term

- Using $-H$ as a regularizing function,we obtain approximate solutions to the original transport problem:

$$\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C}\rangle - \varepsilon \mathbf{H}(\mathbf{P}).$$

- This objective function is ε-strongly convex and thus has a **unique optimal solution.**

- The effect of the entropy to regularize a linear program over the simplex:

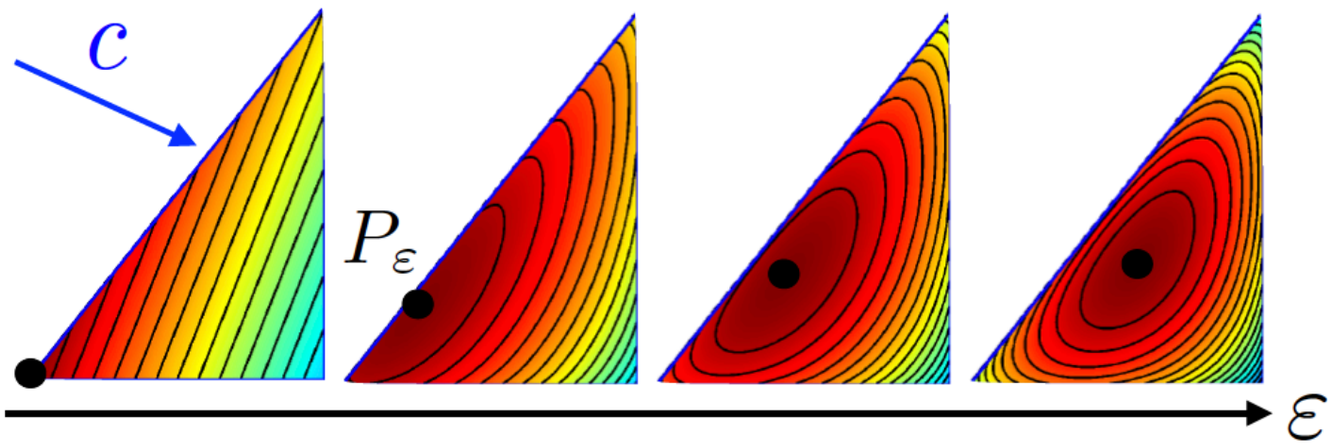


Figure 4.1: Impact of ε on the optimization of a linear function on the simplex, solving $\mathbf{P}_\varepsilon = \mathrm{argmin}_{\mathbf{P}\in\Sigma_3}\langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon\mathbf{H}(\mathbf{P})$ for a varying ε.

Entropic Regularization of Optimal Transport

Convergence with ε

- The unique solution P_ε of the entropic regularized objective function converges to the optimal solution with maximal entropy within the set of all optimal solutions of Kantorovich problem, namely

$$\mathbf{P}_\varepsilon \overset{\varepsilon \to 0}{\longrightarrow} \operatorname*{argmin}_{\mathbf{P}} \{-\mathbf{H}(\mathbf{P}) \ : \ \mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b}), \langle \mathbf{P}, \mathbf{C} \rangle = \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}), \}$$

- In particular,

$$\mathrm{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\varepsilon \to 0}{\longrightarrow} \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}).$$

- One also has

$$\mathbf{P}_\varepsilon \overset{\varepsilon \to \infty}{\longrightarrow} \mathbf{a} \otimes \mathbf{b} = \mathbf{a}\mathbf{b}^{\mathrm{T}} = (\mathbf{a}_i \mathbf{b}_j)_{i,j}.$$

Entropic Regularization of Optimal Transport

Convergence with ε

- (Proof) take $\varepsilon_l \to 0$, we can extract e subsequence of the resulting P_l such that $P_l \to P^* \in U(a,b)$ (since $U(a,b)$ is compact and closed).

- For any P such that $\langle C, P\rangle = L_c(a,b)$, by the optimality one has

$$0 \leq \langle \mathbf{C}, \mathbf{P}_\ell \rangle - \langle \mathbf{C}, \mathbf{P} \rangle \leq \varepsilon_\ell (\mathbf{H}(\mathbf{P}_\ell) - \mathbf{H}(\mathbf{P})).$$

- By the boundness of H, letting $l \to \infty$ yields $\langle C, P\rangle = \langle C, P^*\rangle$. Thus $\langle C, P^*\rangle = L_c(a,b)$. By the continuity of H, $H(P) \leq H(P^*)$ for any P. Moreover, the convexity of H implies the uniqueness of P^*.

- For the $\varepsilon_l \to \infty$ case, we only have to note that
$H(a \otimes b) - H(P) = KL(P \,|\, a \otimes b) \geq 0$
for any $P \in U(a,b)$

Entropic Regularization of Optimal Transport

Convergence with ε

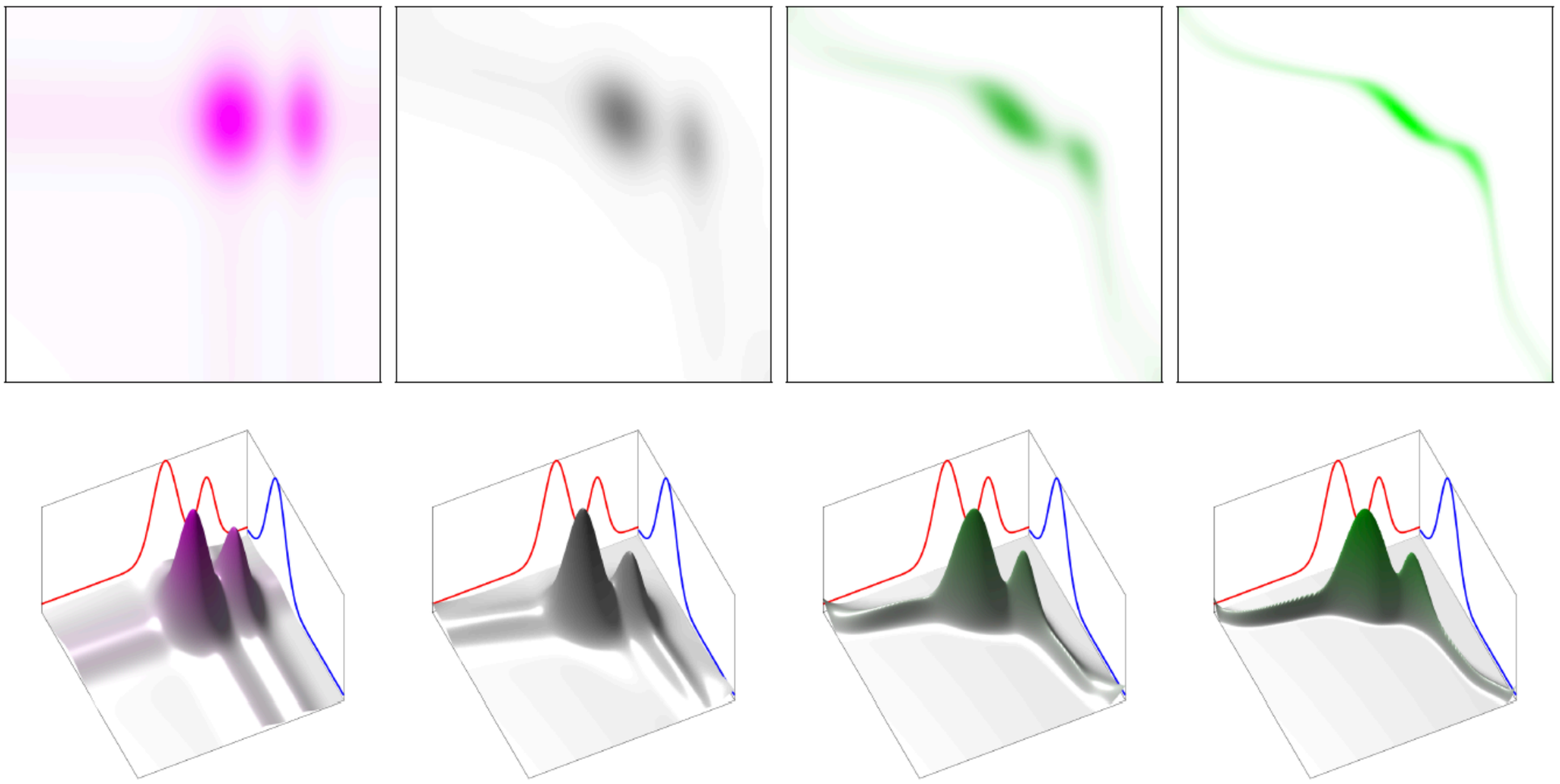


Figure 4.2: Impact of ε on the couplings between two 1-D densities, illustrating Proposition 4.1. Top row: between two 1-D densities. Bottom row: between two 2-D discrete empirical densities with the same number $n = m$ of points (only entries of the optimal $(\mathbf{P}_{i,j})_{i,j}$ above a small threshold are displayed as segments between x_i and y_j).

Entropic Regularization of Optimal Transport

Convergence with ε

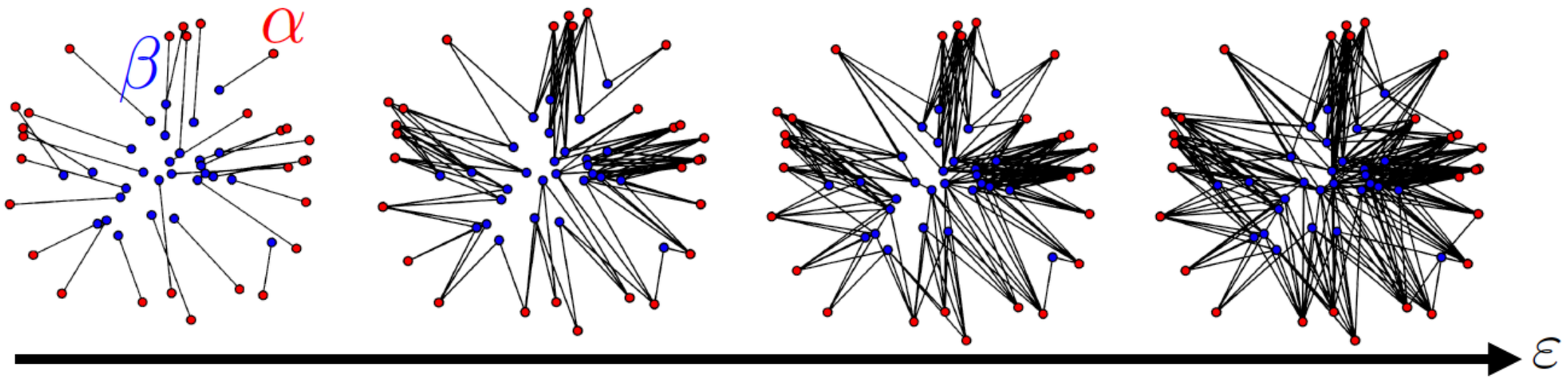


Figure 4.3: Impact of ε on coupling between two 2-D discrete empirical densities with the same number $n = m$ of points (only entries of the optimal $(\mathbf{P}_{i,j})_{i,j}$ above a small threshold are displayed as segments between x_i and y_j).

Entropic Regularization of Optimal Transport
KL divergence regularization

- Defining the KL divergence between couplings as

$$\mathbf{KL}(\mathbf{P}|\mathbf{K}) \overset{\text{def.}}{=} \sum_{i,j} \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{K}_{i,j}}\right) - \mathbf{P}_{i,j} + \mathbf{K}_{i,j},$$

- Given the cost matrix C, define

$$\mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}.$$

- The entropic regularization problem can be transfomed into

$$\mathbf{P}_\varepsilon = \text{Proj}^{\mathbf{KL}}_{\mathbf{U(a,b)}}(\mathbf{K}) \overset{\text{def.}}{=} \underset{\mathbf{P} \in \mathbf{U(a,b)}}{\text{argmin}} \ \mathbf{KL}(\mathbf{P}|\mathbf{K}).$$

- Intuitively, the unique solution P_ε is a projection of K onto $U(a, b)$.

Entropic Regularization of Optimal Transport
General formulation

- For arbitrary measures, we can defined a regularized counterpart using

$$\mathcal{L}_c^{\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta),$$

  where the $KL(\pi|\alpha\otimes\beta)$ can also be considered as relative entropy, defined as

$$\begin{aligned}\mathrm{KL}(\pi|\xi) \overset{\text{def.}}{=} &\int_{\mathcal{X}\times\mathcal{Y}} \log\Big(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\Big)\mathrm{d}\pi(x,y)\\ &+\int_{\mathcal{X}\times\mathcal{Y}}(\mathrm{d}\xi(x,y)-\mathrm{d}\pi(x,y)),\end{aligned}$$

- To avoid the case $KL=\infty$, we choose the reference measure $\alpha\otimes\beta$. Indeed, we can also choose other measure since

$$\mathrm{KL}(\pi|\alpha\otimes\beta) = \mathrm{KL}(\pi|\alpha'\otimes\beta') - \mathrm{KL}(\alpha\otimes\beta|\alpha'\otimes\beta').$$

  as long as $\alpha'\otimes\beta'$ has the same zero measure sets as $\alpha\otimes\beta$.

2. Sinkhorn's Algorithm and Its Convergence

Sinkhorn's Algorithm and Its Convergence

Parameterization of P_ε

Proposition 4.3. The solution to (4.2) is unique and has the form

$$\forall\,(i,j)\in[\![n]\!]\times[\![m]\!],\quad \mathbf{P}_{i,j}=\mathbf{u}_i\mathbf{K}_{i,j}\mathbf{v}_j \tag{4.12}$$

for two (unknown) scaling variable $(\mathbf{u},\mathbf{v})\in\mathbb{R}_+^n\times\mathbb{R}_+^m$.

Proof. Introducing two dual variables $\mathbf{f}\in\mathbb{R}^n,\mathbf{g}\in\mathbb{R}^m$ for each marginal constraint, the Lagrangian of (4.2) reads

$$\mathcal{E}(\mathbf{P},\mathbf{f},\mathbf{g})=\langle\mathbf{P},\,\mathbf{C}\rangle-\varepsilon\mathbf{H}(\mathbf{P})-\langle\mathbf{f},\,\mathbf{P}\mathbb{1}_m-\mathbf{a}\rangle-\langle\mathbf{g},\,\mathbf{P}^{\mathrm{T}}\mathbb{1}_n-\mathbf{b}\rangle.$$

First order conditions then yield

$$\frac{\partial\mathcal{E}(\mathbf{P},\mathbf{f},\mathbf{g})}{\partial\mathbf{P}_{i,j}}=\mathbf{C}_{i,j}+\varepsilon\log(\mathbf{P}_{i,j})-\mathbf{f}_i-\mathbf{g}_j=0,$$

which result, for an optimal $\mathbf{P}$ coupling to the regularized problem, in the expression $\mathbf{P}_{i,j}=e^{\mathbf{f}_i/\varepsilon}e^{-\mathbf{C}_{i,j}/\varepsilon}e^{\mathbf{g}_j/\varepsilon}$, which can be rewritten in the form provided above using nonnegative vectors $\mathbf{u}$ and $\mathbf{v}$. □

Sinkhorn's Algorithm and Its Convergence

Sinkhorn's algorithm

- Write the optimal coupling P as $P = \mathrm{diag}(u)\, K\, \mathrm{diag}(v)$. The parameter (u, v) must satisfy the following nonlinear equations:

$$\mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})\mathbb{1}_m = \mathbf{a}, \quad \text{and} \quad \mathrm{diag}(\mathbf{v})\mathbf{K}^\top\,\mathrm{diag}(\mathbf{u})\mathbb{1}_n = \mathbf{b}.$$

- These two equations can be further simplified as

$$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a} \quad \text{and} \quad \mathbf{v} \odot (\mathbf{K}^{\mathrm{T}}\mathbf{u}) = \mathbf{b},$$

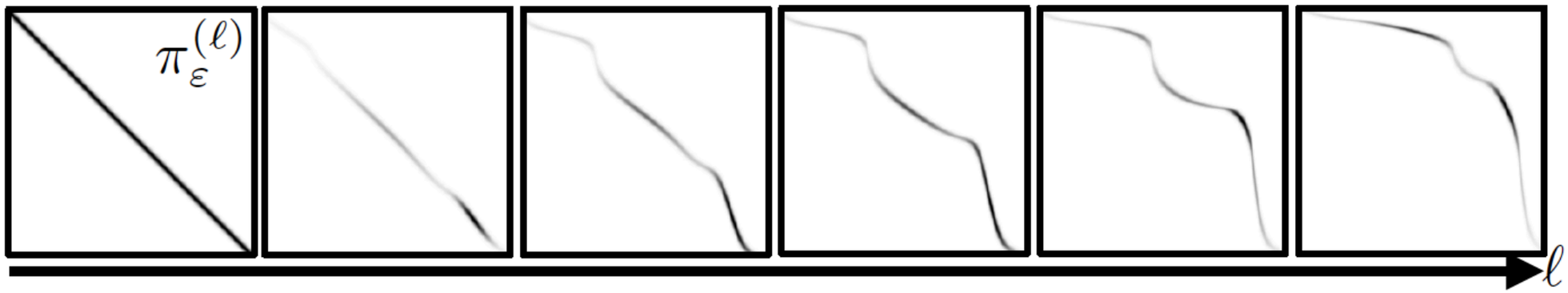
- An intuitive way to solve these two nonlinear equations is iteration, that is

$$\mathbf{u}^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}} \quad \text{and} \quad \mathbf{v}^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{b}}{\mathbf{K}^{\mathrm{T}}\mathbf{u}^{(\ell+1)}},$$

initialized with $v^{(0)} = 1_m$. This update scheme is called **Sinkhorn's algorithm.**

Sinkhorn's Algorithm and Its Convergence

Experiments



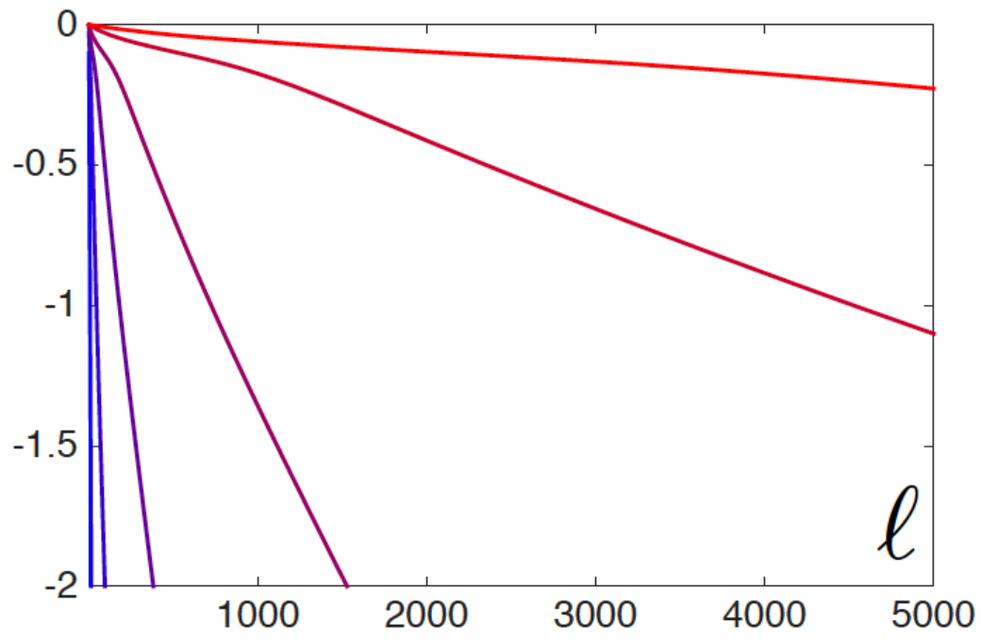


Figure 4.5: Top: evolution of the coupling $\pi_\varepsilon^{(\ell)} = \mathrm{diag}(\mathbf{u}^{(\ell)})\mathbf{K}\,\mathrm{diag}(\mathbf{v}^{(\ell)})$ computed at iteration ℓ of Sinkhorn's iterations, for 1-D densities on $\mathcal{X} = [0,1]$, $c(x,y) = |x-y|^2$, and $\varepsilon = 0.1$. Bottom: impact of ε the convergence rate of Sinkhorn, as measured in term of marginal constraint violation $\log(\|\pi_\varepsilon^{(\ell)}\mathbb{1}_m - \mathbf{b}\|_1)$.

Sinkhorn's Algorithm and Its Convergence

Convergence analysis

- Altschuler et al. [2017] showed that by setting $\varepsilon = \dfrac{4\log n}{\tau}$, $O(||C||_\infty^3 \log(n)\tau^{-3})$ Sinkhorn iterations are enough to ensure that

$$\langle \hat{\mathbf{P}}, \mathbf{C} \rangle \leq \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) + \tau.$$

- Therefore, Sinkhorn computes a τ-approximate solution of the unregularized OT problem in $O(n^2 \log(n)\tau^{-3})$ operations. (One iteration needs $O(n^2)$ operations.

- A serious problem with Sinkhorn's analysis is that, the convergence of Sinkhorn's algorithm requires $\varepsilon \to 0$. However, too small ε will make the kennel K too large to be stored in memory.

Sinkhorn's Algorithm and Its Convergence

An alternative formulation

- Denoting

$$\mathcal{C}_{\mathbf{a}}^1 \stackrel{\text{def.}}{=} \{\mathbf{P} \,:\, \mathbf{P}\mathbb{1}_m = \mathbf{a}\} \quad \text{and} \quad \mathcal{C}_{\mathbf{b}}^2 \stackrel{\text{def.}}{=} \left\{\mathbf{P} \,:\, \mathbf{P}^{\mathrm{T}}\mathbb{1}_m = \mathbf{b}\right\}$$

- One can use Bregman's iterative projections to approximate the solution

$$\mathbf{P}^{(\ell+1)} \stackrel{\text{def.}}{=} \mathrm{Proj}_{\mathcal{C}_{\mathbf{a}}^1}^{\mathbf{KL}}(\mathbf{P}^{(\ell)}) \quad \text{and} \quad \mathbf{P}^{(\ell+2)} \stackrel{\text{def.}}{=} \mathrm{Proj}_{\mathcal{C}_{\mathbf{b}}^2}^{\mathbf{KL}}(\mathbf{P}^{(\ell+1)}).$$

- These iterates are equivalent to Sinkhorn's iterations if we define

$$\mathbf{P}^{(2\ell)} \stackrel{\text{def.}}{=} \mathrm{diag}(\mathbf{u}^{(\ell)})\mathbf{K}\,\mathrm{diag}(\mathbf{v}^{(\ell)}),$$

$$\mathbf{P}^{(2\ell+1)} \stackrel{\text{def.}}{=} \mathrm{diag}(\mathbf{u}^{(\ell+1)})\mathbf{K}\,\mathrm{diag}(\mathbf{v}^{(\ell)})$$
$$\mathbf{P}^{(2\ell+2)} \stackrel{\text{def.}}{=} \mathrm{diag}(\mathbf{u}^{(\ell+1)})\mathbf{K}\,\mathrm{diag}(\mathbf{v}^{(\ell+1)}).$$

Sinkhorn's Algorithm and Its Convergence

Convergence analysis

- Assume for simpicity $P^{(0)} = 1_n 1_m^T$, the Sinkhorn iterations has the form

$$\begin{aligned} P^{(\ell+1)} &= \operatorname{diag}(\mathbf{u}^{(\ell)})(e^{-\frac{\mathbf{C}}{\varepsilon}} \odot \mathbf{P}^{(\ell)}) \operatorname{diag}(\mathbf{v}^{(\ell)}) \\ &= \operatorname{diag}(\mathbf{u}^{(\ell)} \odot \cdots \odot \mathbf{u}^{(0)}) e^{-\frac{(\ell+1)\mathbf{C}}{\varepsilon}} \odot \mathbf{P}^{(\ell)}) \operatorname{diag}(\mathbf{v}^{(\ell)} \odot \cdots \odot \mathbf{v}^{(0)}). \end{aligned}$$

  to calculate the coupling matrix.

- The regularization parameter ε/ℓ should decay.

- The decaying schedule of ε/ℓ should be carefully chosen. See, for instance, [Kosowsky and Yuille, 1994], [Schmitzer, 2016b].

Sinkhorn's Algorithm and Its Convergence

Convergence under Hilbert metric

- The Hilbert projective metric on the set of positive vectors is defined as

$$\forall (\mathbf{u}, \mathbf{u}') \in (\mathbb{R}^n_{+,*})^2, \quad d_{\mathcal{H}}(\mathbf{u}, \mathbf{u}') \overset{\text{def.}}{=} \log \max_{i,j} \frac{\mathbf{u}_i \mathbf{u}'_j}{\mathbf{u}_j \mathbf{u}'_i}.$$

- This definition is a distnace on the projective cone $\mathbb{R}^n_{+,*}/\sim$, where $u \sim u'$ means that there exists a positive scalar.

- The Hilbert projective metric can be equivalently defined as

$$d_{\mathcal{H}}(\mathbf{u}, \mathbf{u}') = \left\| \log(\mathbf{u}) - \log(\mathbf{u}') \right\|_{\text{var}}$$

$$\text{where} \quad \|\mathbf{f}\|_{\text{var}} \overset{\text{def.}}{=} (\max_i \mathbf{f}_i) - (\min_i \mathbf{f}_i).$$

- One always has $||f||_{\text{var}} \leq 2||f||_\infty$. If $f_i = 0$ for some fixed i, then a converse inequality also holds since $||f||_\infty \leq ||f||_{\text{var}}$.

Sinkhorn's Algorithm and Its Convergence
Convergence under Hilbert metric

- [Birkhoff, 1957] proved the floowing fundamental theorem.

Theorem 4.1. Let $\mathbf{K} \in \mathbb{R}_{+,*}^{n \times m}$; then for $(\mathbf{v}, \mathbf{v}') \in (\mathbb{R}_{+,*}^{m})^2$

$$d_{\mathcal{H}}(\mathbf{K}\mathbf{v}, \mathbf{K}\mathbf{v}') \leq \lambda(\mathbf{K}) d_{\mathcal{H}}(\mathbf{v}, \mathbf{v}'), \text{ where } \begin{cases} \lambda(\mathbf{K}) \overset{\text{def.}}{=} \frac{\sqrt{\eta(\mathbf{K})}-1}{\sqrt{\eta(\mathbf{K})}+1} < 1, \\ \eta(\mathbf{K}) \overset{\text{def.}}{=} \max_{i,j,k,\ell} \frac{\mathbf{K}_{i,k}\mathbf{K}_{j,\ell}}{\mathbf{K}_{j,k}\mathbf{K}_{i,\ell}}. \end{cases}$$

- This theorem has following illustration:

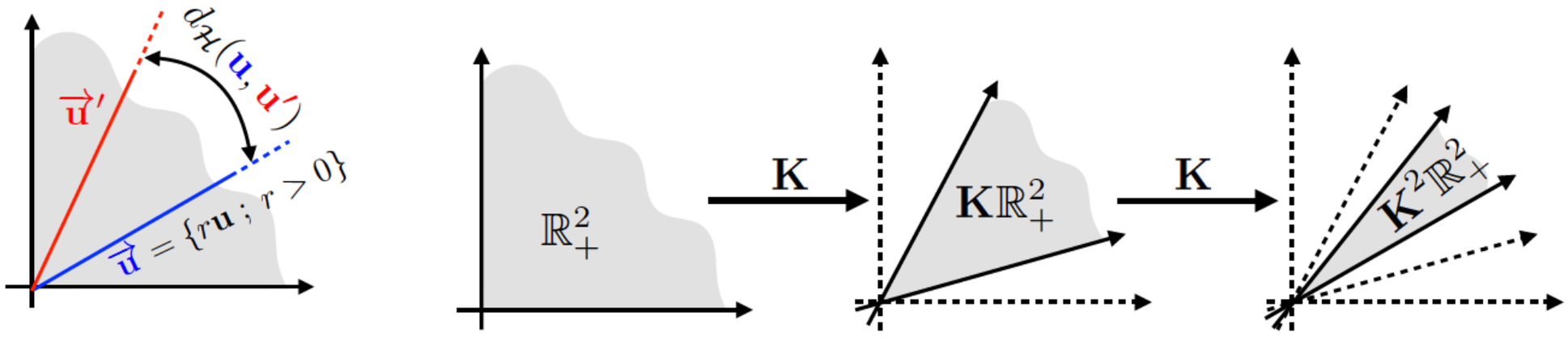


Figure 4.7: Left: the Hilbert metric $d_{\mathcal{H}}$ is a distance over rays in cones (here positive vectors). Right: visualization of the contraction induced by the iteration of a positive matrix $\mathbf{K}$.

Sinkhorn's Algorithm and Its Convergence

Convergence under Hilbert metric

- The following theorem show the linear convergence of Sinkhorn's iterations.

Theorem 4.2. One has $(\mathbf{u}^{(\ell)}, \mathbf{v}^{(\ell)}) \to (\mathbf{u}^\star, \mathbf{v}^\star)$ and

$$d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) = O(\lambda(\mathbf{K})^{2\ell}), \quad d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star) = O(\lambda(\mathbf{K})^{2\ell}). \tag{4.22}$$

One also has

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) &\leq \frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell)}\mathbb{1}_m, \mathbf{a})}{1-\lambda(\mathbf{K})^2}, \\ d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star) &\leq \frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell),\top}\mathbb{1}_n, \mathbf{b})}{1-\lambda(\mathbf{K})^2}, \end{aligned} \tag{4.23}$$

where we denoted $\mathbf{P}^{(\ell)} \overset{\text{def.}}{=} \operatorname{diag}(\mathbf{u}^{(\ell)})\mathbf{K}\operatorname{diag}(\mathbf{v}^{(\ell)})$. Last, one has

$$\|\log(\mathbf{P}^{(\ell)}) - \log(\mathbf{P}^\star)\|_\infty \leq d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) + d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star), \tag{4.24}$$

where $\mathbf{P}^\star$ is the unique solution of (4.2).

Sinkhorn's Algorithm and Its Convergence

Convergence under Hilbert metric

- To prove the first conclusion, note that for any (v, v'), one has

$$d_{\mathcal{H}}(\mathbf{v}, \mathbf{v}') = d_{\mathcal{H}}(\mathbf{v}/\mathbf{v}', \mathbb{1}_m) = d_{\mathcal{H}}(\mathbb{1}_m/\mathbf{v}, \mathbb{1}_m/\mathbf{v}').$$

  This shows that

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(\ell+1)}, \mathbf{u}^\star) &= d_{\mathcal{H}}\left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}}, \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^\star}\right) \\ &= d_{\mathcal{H}}(\mathbf{K}\mathbf{v}^{(\ell)}, \mathbf{K}\mathbf{v}^\star) \leq \lambda(\mathbf{K}) d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star), \end{aligned}$$

- To prove the second conclusion, use the triangular inequality

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) &\leq d_{\mathcal{H}}(\mathbf{u}^{(\ell+1)}, \mathbf{u}^{(\ell)}) + d_{\mathcal{H}}(\mathbf{u}^{(\ell+1)}, \mathbf{u}^\star) \\ &\leq d_{\mathcal{H}}\left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}}, \mathbf{u}^{(\ell)}\right) + \lambda(\mathbf{K})^2 d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) \\ &= d_{\mathcal{H}}\left(\mathbf{a}, \mathbf{u}^{(\ell)} \odot (\mathbf{K}\mathbf{v}^{(\ell)})\right) + \lambda(\mathbf{K})^2 d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star), \end{aligned}$$

- The second conclusion shows that, marginal constraints violation, i.e. $||P1_m - a||$ or $||P^T 1_m - b||$ are useful stopping criteria.

Sinkhorn's Algorithm and Its Convergence

Other regularization

- It is possible to replace the entropic term $-H(P)$ by any other strictly convex penalty $R(P)$. For instance, a typical example is the squared ℓ^2 norm

$$R(\mathbf{P}) = \sum_{i,j} \mathbf{P}_{i,j}^2 + \iota_{\mathbb{R}_+}(\mathbf{P}_{i,j});$$

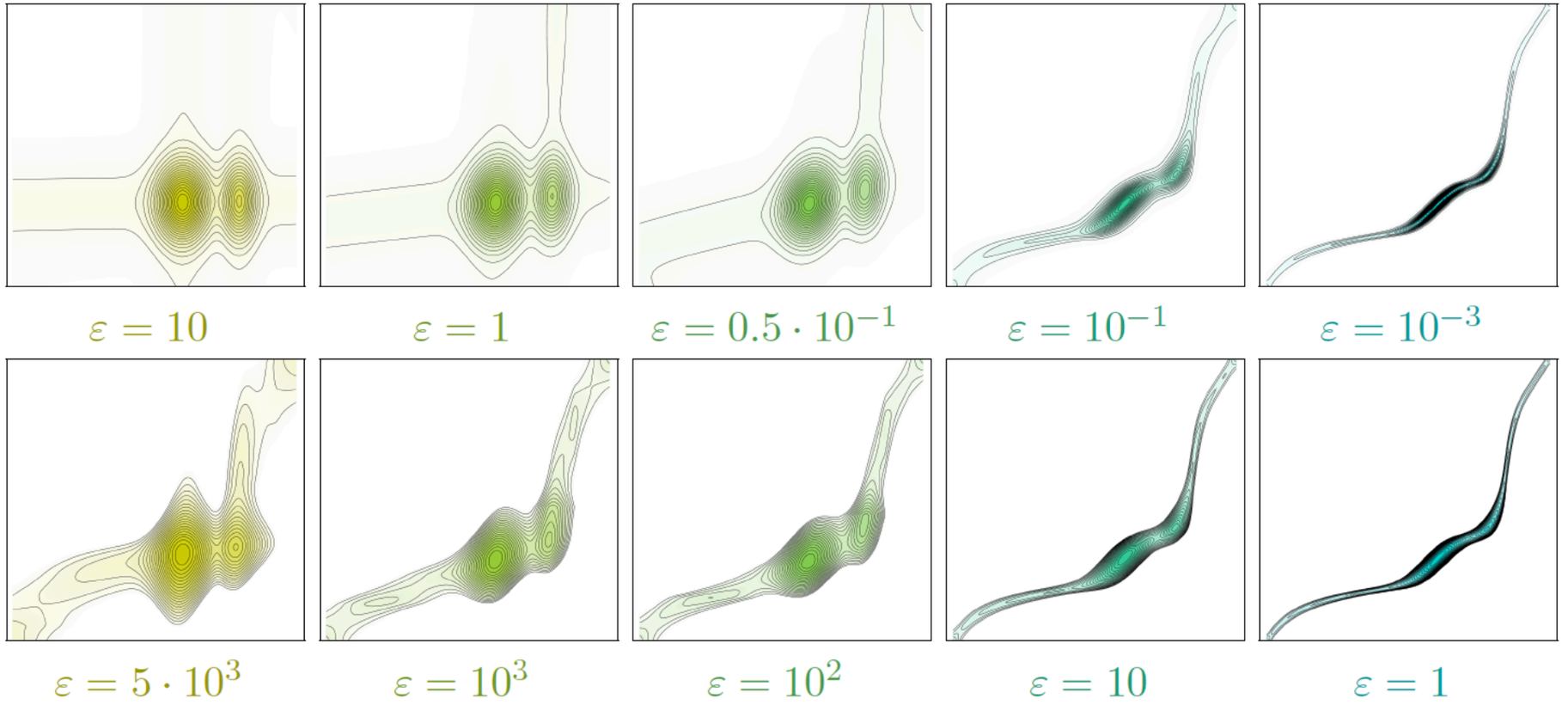


Figure 4.6: Comparison of entropic regularization $R = -\mathbf{H}$ (top row) and quadratic regularization $R = \|\cdot\|^2 + \iota_{\mathbb{R}_+}$ (bottom row). The (α, β) marginals are the same as for Figure 4.4.

Sinkhorn's Algorithm and Its Convergence

Barycentric projection

- Under some conditions, Monge problem is equivalent to Kantorovich problem (see Section 2).

- For finite case, if the Monge map is a permutation matrix and is unique, the barycentric projection map

$$: x_i \in \mathcal{X} \longmapsto \frac{1}{\mathbf{a}_i} \sum_j \mathbf{P}_{i,j} y_j \in \mathcal{Y},$$

  will converge to the Monge map.

- For arbitrary case, if the solution π to the Kantorovich problem is supported on the graph of the Monge map, then the map

$$x \in \mathcal{X} \longmapsto \int_{\mathcal{Y}} y \frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)} \mathrm{d}\beta(y).$$

  will converge to the Monge map.

3. Speeding Up Sinkhorn's Iterations

Speeding Up Sinkhorn's Iterations
Computational complexity

- The main computational bottleneck of Sinkhorn's iterations is the vector-matrix multiplication against kernels K and K^T.

- The time complexity of vector-matrix multiplication is $O(mn)$ if implemented naively

- In many situations, such as solving more than one coupling matrix, or the high dimension case (curse of dimension), mn can be very large.

Speeding Up Sinkhorn's Iterations
Parallelization

- Assume we are to solve the OT problem for pairs $(a_1, b_1), \ldots, (a_N, b_N)$ (with a common cost matrix C) simultaneously.

- Let $A = [a_1, a_2, \ldots, a_N], B = [b_1, b_2 \ldots, b_N]$ be $n \times N$ and $m \times N$ matices storing all measures. All Sinkhorn iterations for these N pairs can be carried out in parallell, i.e.

$$\mathbf{U}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{A}}{\mathbf{K}\mathbf{V}^{(\ell)}} \quad \text{and} \quad \mathbf{V}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{B}}{\mathbf{K}^{\mathrm{T}}\mathbf{U}^{(\ell+1)}},$$

initialized with $V^{(0)} = 1_{m \times N}$, where $\frac{\cdot}{\cdot}$ is elementwise division.

- The author said the vector of regularized distances is

$$\mathbb{1}_n{}^{\mathrm{T}}(\mathbf{U} \odot \log \mathbf{U} \odot ((\mathbf{K} \odot \mathbf{C})\mathbf{V}) + \mathbf{U} \odot ((\mathbf{K} \odot \mathbf{C})(\mathbf{V} \odot \log \mathbf{V}))) \in \mathbb{R}^N.$$

- He didnot define regularized distance. The i-th element correspond to

$$\sum_{all\ elements} C \odot P_i \odot \log(\frac{P_i}{K})$$

Speeding Up Sinkhorn's Iterations
Higher dimension

- In the *d*-dimensional case, the indices of a histogram becomes *d*-vector

$$i = (i_k)_{k=1}^d, j = (j_k)_{k=1}^d \in [\![n_1]\!] \times \cdots \times [\![n_d]\!].$$

  Thus $n = n_1 n_2 \cdots n_d$, and this problem becomes untractable as *d* gets larger.

- To alleviate the curse of dimension, we assume a model of additive cost matrix. That is, there exists *d* matrices $C^1, \ldots, C^d$ of size $n_1 \times n, \ldots, n_d \times n$, such that

$$\mathbf{C}_{ij} = \sum_{k=1}^{d} \mathbf{C}^k_{i_k, j_k},$$

  and thus the kernel matrix has a seperable multiplicative structure

$$\mathbf{K}_{i,j} = \prod_{k=1}^{d} \mathbf{K}^k_{i_k, j_k}.$$

  where *i* and *j* are both d-vector.

Speeding Up Sinkhorn's Iterations
Higher dimension: an example

- Consider the case $\mathcal{X} = \mathcal{Y} = [0,1]^d$, the ground cost ht the q-th power of the q-norm,

$$c(x,y) = \|x-y\|_q^q = \sum_{i=1}^{d} |x_i - y_i|^q,\ q > 0;$$

and the space is discretized using a regular grid containing only points $x_i = (i_1/n_1, \ldots, i_d/n_d)$ for $i = (i_1, \ldots, i_d) \in [n_1] \times \cdots \times [n_d]$.

- The kernel matrix K can be represented by the multiplication of

$$\mathbf{K}^k = \left[\exp(-\left|\tfrac{r-s}{n_k}\right|^q/\varepsilon)\right]_{1\le r,s\le n_k}$$

- For instance, if $d = 2$, the matrix-vector multiplication Ku where u is a vector of length $n_1 n_2$ and $K = K^1 \odot K^2$. If we define a matrix U of size $n_1 \times n_2$, then:

$$Ku = K^1 U K^2.$$

Speeding Up Sinkhorn's Iterations
Higher dimension: an example

- In this way, we recover the iteration with only $n_1^2 n_2 + n_1 n_2^2 = n(n_1 + n_2)$ operations instead of $(n_1 n_2)^2$ operations.

- For general $d \geq 2$, the kernel matrix can be decomposed as $K = K^1 \odot K^2 \odot \cdots \odot K^d$, and U is tensor of size $n_1 \times n_2 \times \cdots \times n_d$. In this way, the matrix-vector multiplication can be rewritten as

$$Ku = Dot(\cdots Dot(Dot(U, K^1, 1), K^2, 2), \cdots, K^d, d)$$

  where $Dot(U, K^i, i)$ refers to multiply U and K^i along the i-th dimension. (Recall the definition of torch.tensordot)

- For general $d \geq 2$, the total computation cost is

$$nn_1 + nn_2 + \cdots + nn_d \sim n^{1+1/d}$$

  intead of $O(n^2)$.

Speeding Up Sinkhorn's Iterations
Higher dimension: another approach

- In planar domains, a simplest but common case is translation invariant kernels $K_{i,j} = k_{i-j}$.

- It is typically the case of distance on $\mathbb{Z}^d$.

- In this case, the matrix-vector multiplication is a convolution, $Ku = k * u$.

- There are several algorithms to approximate the convolution in nearly linear time. For example, by Fourier transform $\mathscr{F}$, we have

$$\mathscr{F}(k * v) = \mathscr{F}(k) \odot \mathscr{F}(v).$$

- At last, Sinkhorn's iterations is a fixes point algorithm, one can use the standard **extrapolation schemes** to enhance the conditioning around the dixed point.

4. Stability and Log-Domain Computations

Stability and Log-Domain Computations

Alternative formulation

Proposition 4.4. One has

$$\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a},\mathbf{b}) = \max_{\mathbf{f}\in\mathbb{R}^n,\mathbf{g}\in\mathbb{R}^m} \langle \mathbf{f},\, \mathbf{a}\rangle + \langle \mathbf{g},\, \mathbf{b}\rangle - \varepsilon\langle e^{\mathbf{f}/\varepsilon},\, \mathbf{K}e^{\mathbf{g}/\varepsilon}\rangle. \tag{4.30}$$

The optimal $(\mathbf{f},\mathbf{g})$ are linked to scalings $(\mathbf{u},\mathbf{v})$ appearing in (4.12) through

$$(\mathbf{u},\mathbf{v}) = (e^{\mathbf{f}/\varepsilon}, e^{\mathbf{g}/\varepsilon}). \tag{4.31}$$

Proof. We start from the end of the proof of Proposition 4.3, which links the optimal primal solution $\mathbf{P}$ and dual multipliers $\mathbf{f}$ and $\mathbf{g}$ for the marginal constraints as

$$\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon} e^{-\mathbf{C}_{i,j}/\varepsilon} e^{\mathbf{g}_j/\varepsilon}.$$

Substituting in the Lagrangian $\mathcal{E}(\mathbf{P},\mathbf{f},\mathbf{g})$ of Equation (4.2) the optimal $\mathbf{P}$ as a function of $\mathbf{f}$ and $\mathbf{g}$, we obtain that the Lagrange dual function equals

$$\mathbf{f},\mathbf{g} \mapsto \langle e^{\mathbf{f}/\varepsilon},\, (\mathbf{K}\odot\mathbf{C})\, e^{\mathbf{g}/\varepsilon}\rangle - \varepsilon\mathbf{H}(\mathrm{diag}(e^{\mathbf{f}/\varepsilon})\mathbf{K}\,\mathrm{diag}(e^{\mathbf{g}/\varepsilon})). \tag{4.32}$$

The neg-entropy of $\mathbf{P}$ scaled by ε, namely $\varepsilon\langle \mathbf{P},\, \log\mathbf{P} - \mathbb{1}_{n\times m}\rangle$, can be stated explicitly as a function of $\mathbf{f},\mathbf{g},\mathbf{C}$,

$$\begin{aligned}&\langle \mathrm{diag}(e^{\mathbf{f}/\varepsilon})\mathbf{K}\,\mathrm{diag}(e^{\mathbf{g}/\varepsilon}),\, \mathbf{f}\mathbb{1}_m{}^{\mathrm{T}} + \mathbb{1}_n\mathbf{g}^{\mathrm{T}} - \mathbf{C} - \varepsilon\mathbb{1}_{n\times m}\rangle\\ &= -\langle e^{\mathbf{f}/\varepsilon},\, (\mathbf{K}\odot\mathbf{C})\, e^{\mathbf{g}/\varepsilon}\rangle + \langle \mathbf{f},\, \mathbf{a}\rangle + \langle \mathbf{g},\, \mathbf{b}\rangle - \varepsilon\langle e^{\mathbf{f}/\varepsilon},\, \mathbf{K}e^{\mathbf{g}/\varepsilon}\rangle;\end{aligned}$$

therefore, the first term in (4.32) cancels out with the first term in the entropy above. The remaining terms are those appearing in (4.30). □

Stability and Log-Domain Computations
Alternative formulation

- Using this formulation, one can calculate the gradients of the objective function $Q(f, g)$ w.r.t. f and g, and then use gradient based method.

- The gradients of $Q(f, g)$ w.r.t. f and g are

$$\nabla|_{\mathbf{f}} Q(\mathbf{f}, \mathbf{g}) = \mathbf{a} - e^{\mathbf{f}/\varepsilon} \odot \left(\mathbf{K} e^{\mathbf{g}/\varepsilon} \right),$$
$$\nabla|_{\mathbf{g}} Q(\mathbf{f}, \mathbf{g}) = \mathbf{b} - e^{\mathbf{g}/\varepsilon} \odot \left(\mathbf{K}^{\mathrm{T}} e^{\mathbf{f}/\varepsilon} \right).$$

- To approximate the zero points of the gradients, coordinate ascent gives the following updates: (indeed, this is equivalent to Sinkhorn's updates.)

$$\mathbf{f}^{(\ell+1)} = \varepsilon \log \mathbf{a} - \varepsilon \log \left(\mathbf{K} e^{\mathbf{g}^{(\ell)}/\varepsilon} \right),$$
$$\mathbf{g}^{(\ell+1)} = \varepsilon \log \mathbf{b} - \varepsilon \log \left(\mathbf{K}^{\mathrm{T}} e^{\mathbf{f}^{(\ell+1)}/\varepsilon} \right).$$

Stability and Log-Domain Computations
Alternative formulation

- The iterations in the last slide can be given an alternative interpretation.

- Definition: Given a vector z of real numbers, we write $\min_\varepsilon z$ for the soft-minimum of its coordinates, namely,

$$\min_\varepsilon \mathbf{z} = -\varepsilon \log \sum_i e^{-\mathbf{z}_i/\varepsilon}.$$

- $\min_\varepsilon z \to \min z$ as $\varepsilon \to 0$. Indeed, $\min_\varepsilon z$ can be interpreted as a differentiable approximation of the min function.

- Using this notation, these two updates can be rewritten as

$$(\mathbf{f}^{(\ell+1)})_i = \min_\varepsilon{}' (\mathbf{C}_{ij} - \mathbf{g}_j^{(\ell)})_j + \varepsilon \log \mathbf{a}_i,$$
$$(\mathbf{g}^{(\ell+1)})_j = \min_\varepsilon{}' (\mathbf{C}_{ij} - \mathbf{f}_i^{(\ell)})_i + \varepsilon \log \mathbf{b}_j.$$

Stability and Log-Domain Computations
Alternative formulation

- To get a more compact form, we define

$$\mathrm{Min}_\varepsilon^{\mathrm{row}}(\mathbf{A}) \stackrel{\text{def.}}{=} \left(\min_\varepsilon (\mathbf{A}_{i,j})_j\right)_i \in \mathbb{R}^n,$$
$$\mathrm{Min}_\varepsilon^{\mathrm{col}}(\mathbf{A}) \stackrel{\text{def.}}{=} \left(\min_\varepsilon (\mathbf{A}_{i,j})_i\right)_j \in \mathbb{R}^m.$$

  for any matrix $A \in \mathbb{R}^{n,m}$.

- Using this notation, Sinkhorn's iterates read

$$\mathbf{f}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{row}}(\mathbf{C} - \mathbb{1}_n \mathbf{g}^{(\ell)\mathrm{T}}) + \varepsilon \log \mathbf{a},$$
$$\mathbf{g}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{col}}(\mathbf{C} - \mathbf{f}^{(\ell)} \mathbb{1}_m{}^{\mathrm{T}}) + \varepsilon \log \mathbf{b}.$$

Stability and Log-Domain Computations

Avoid overflow

- Recall we may encounter with overflow when calculating $e^{-z/\varepsilon}$. Define $\underline{z} = \min z$, the log-sum-exp stabilization trick suggests evaluating $\min_\varepsilon z$

$$\min_\varepsilon \mathbf{z} = \underline{\mathrm{z}} - \varepsilon \log \sum_i e^{-(\mathbf{z}_i - \underline{\mathrm{z}})/\varepsilon}.$$

- This leads to stablized iteration

$$\mathbf{f}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{row}} \left(\mathbf{S}(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)})\right) + \mathbf{f}^{(\ell)} + \varepsilon \log(\mathbf{a}),$$
$$\mathbf{g}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{col}} \left(\mathbf{S}(\mathbf{f}^{(\ell+1)}, \mathbf{g}^{(\ell)})\right) + \mathbf{g}^{(\ell)} + \varepsilon \log(\mathbf{b}),$$

where

$$\mathbf{S}(\mathbf{f}, \mathbf{g}) = \left(\mathbf{C}_{i,j} - \mathbf{f}_i - \mathbf{g}_j\right)_{i,j}.$$

- A proper decaying schedule of ε is still important.