ICML | 2020
Thirty-seventh International Conference on Machine Learning

Mind the entropy bias in regularized OT :

Debiased Sinkhorn barycenters

Hicham Janati — Marco Cuturi — Alexandre Gramfort

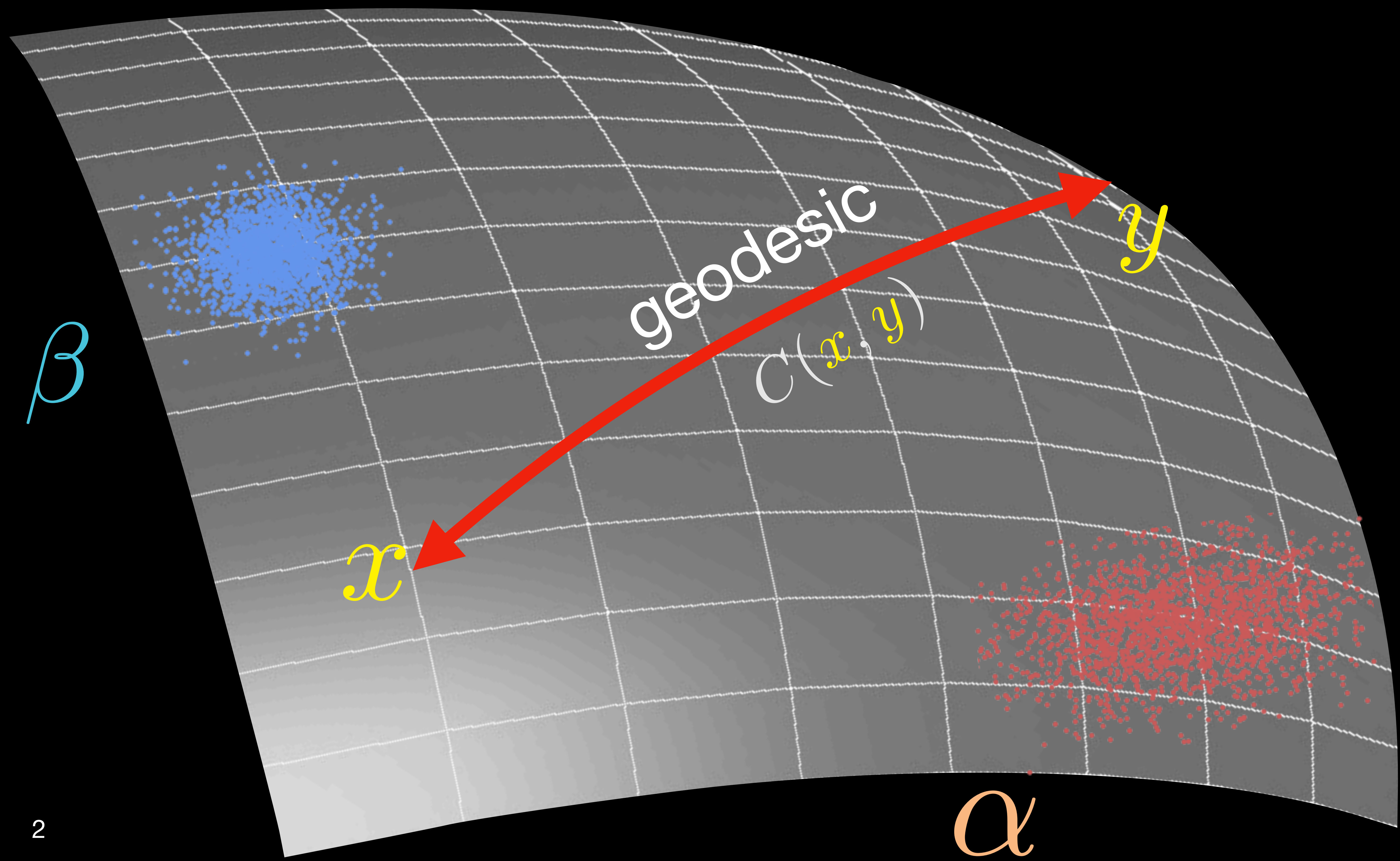
Inria Saclay / ENSAE   Google Brain / ENSAE   Inria Saclay

ENSAE IP PARIS   Inria   Google AI   erc European Research Council a metric space $(\mathcal{X}, C)$

$$C(x,y)\mathrm{d}\pi(x,y)$$

“unit transport cost”

$$\alpha, \beta \in \mathcal{P}(\mathcal{X})$$



a metric space $(\mathcal{X}, C)$

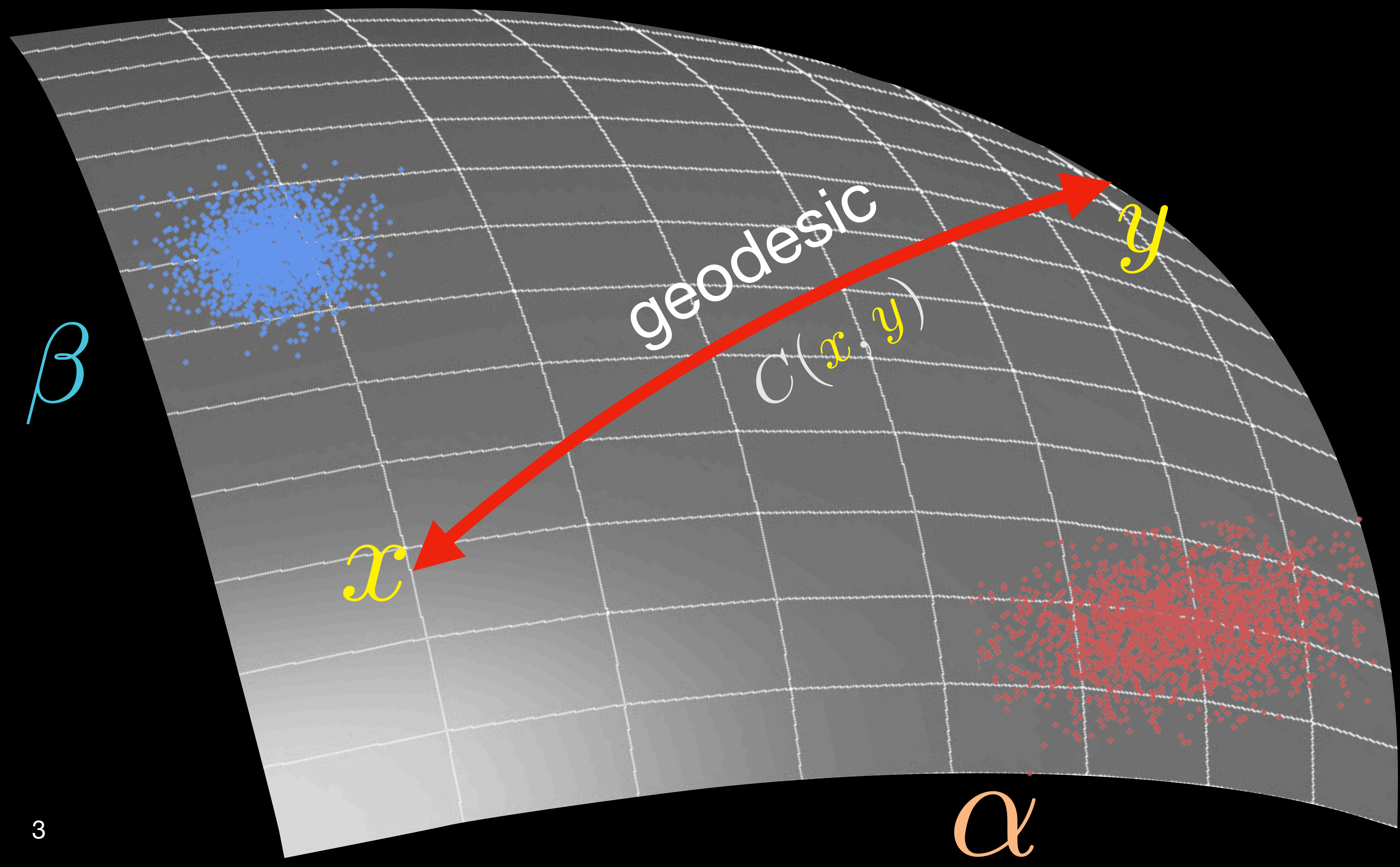
$$\mathrm{OT}(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \int_{\mathcal{X} \times \mathcal{X}} C(x, y) \mathrm{d}\pi(x, y)$$

"Wasserstein distance"

Linear program

$O(n^3)$

$$\alpha, \beta \in \mathcal{P}(\mathcal{X})$$



Entropy regularized OT

$$\mathrm{OT}_{\varepsilon}^{m_1, m_2}(\alpha, \beta) \overset{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \int_{\mathcal{X} \times \mathcal{X}} C(x, y) \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}(\pi | m_1 \otimes m_2)$$

regularizer: relative entropy

$m_1, m_2 \in$ reference measures in $\mathcal{P}(\mathcal{X})$

+ Computational cost: cubic → quadratic
+ Breaks the curse of dimension

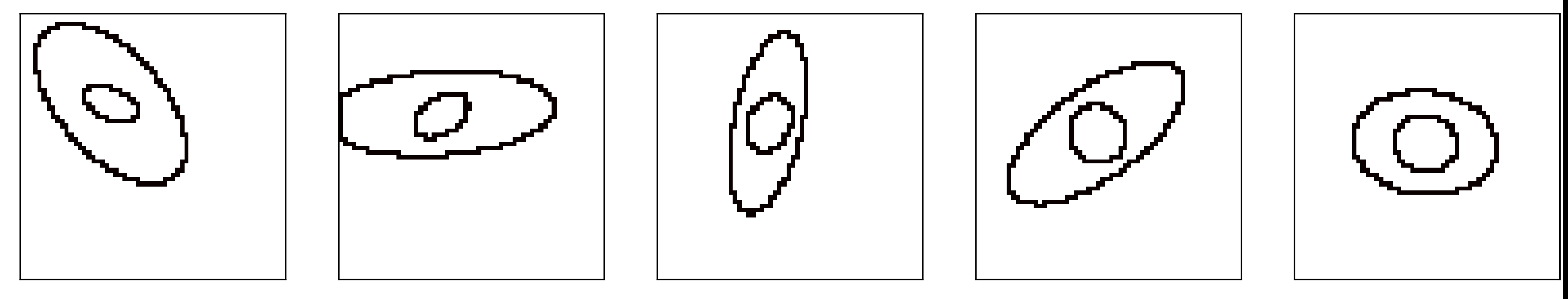
- Entropy bias
- Not a metric

$$\arg\min_{\alpha} \mathrm{OT}_{\varepsilon}(\alpha, \beta) \neq \beta$$

Entropy OT barycenters and bias

$\mathcal{X}$ finite $\qquad m_1, m_2 = \mathcal{U}(\mathcal{X})$

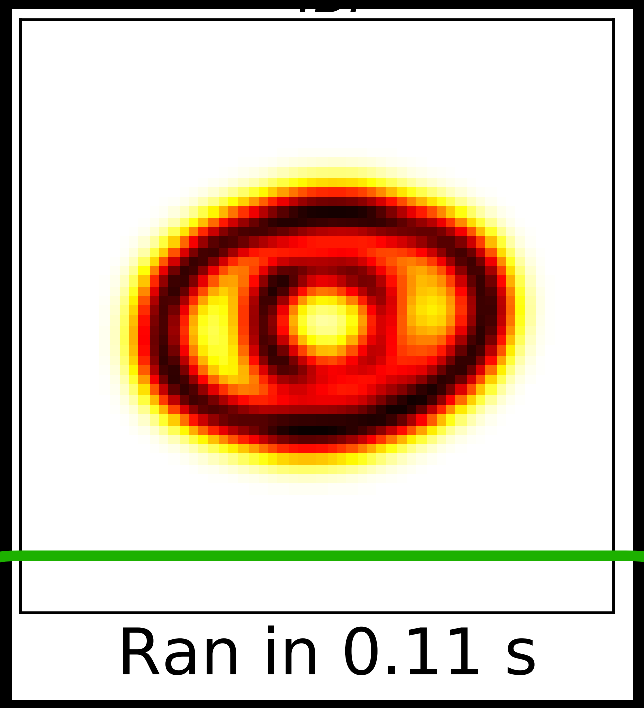
$\alpha_1, \quad \alpha_2 \quad \dots$



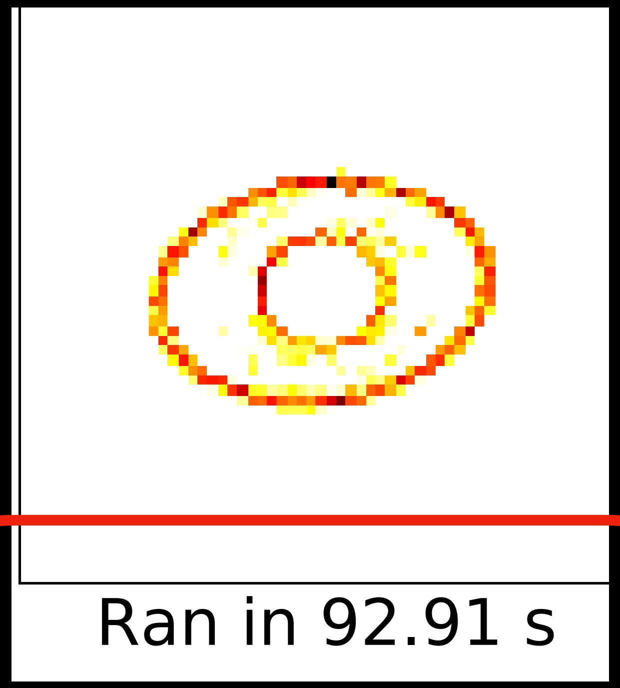
$$\arg\min_{\alpha} \sum_{k=1}^{K} w_k \, \mathrm{OT}_{\varepsilon}(\alpha_k, \alpha)$$

α

$\varepsilon = 0.5$



$\varepsilon = 0$



Using Sinkhorn's algorithm

(Benamou et al, 15')

Entropy regularized OT

$$\mathrm{OT}^{m_1,m_2}_{\varepsilon}(\alpha,\beta) \overset{\text{def}}{=} \min_{\substack{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X})\\ \pi_1=\alpha,\pi_2=\beta}} \int_{\mathcal{X}\times\mathcal{X}} C(x,y)\mathrm{d}\pi(x,y) + \varepsilon\mathrm{KL}(\pi|m_1\otimes m_2)$$

regularizer: relative entropy

$m_1, m_2 \in$ reference measures in $\mathcal{P}(\mathcal{X})$

+ Computational cost: cubic → quadratic
+ Breaks the curse of dimension

- Entropy bias
- Not a metric

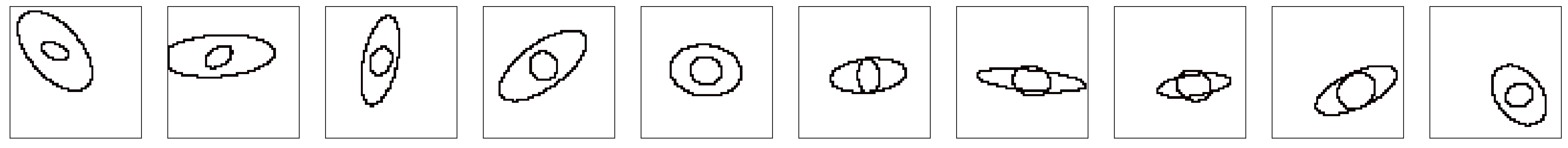
(Feydy et al, 19): $\mathcal{X}$ is compact $m_1 = \alpha \quad m_2 = \beta$

$$S_\varepsilon(\alpha,\beta) \overset{\text{def}}{=} \mathrm{OT}_\varepsilon(\alpha,\beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon(\alpha,\alpha) + \mathrm{OT}_\varepsilon(\beta,\beta))$$

- ~~Entropy bias~~

+ $S_\varepsilon(\alpha,\beta) \geq 0;$
+ $S_\varepsilon(\alpha,\beta) = 0 \Rightarrow \alpha = \beta$

Entropy regularized OT

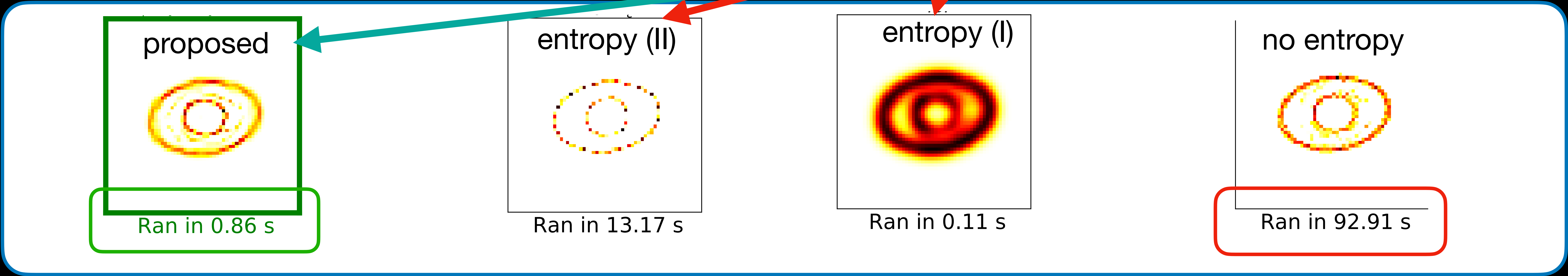


$\pi_1 = \alpha, \pi_2 = \beta$

regularizer: relative entropy

Contributions:

1) S_ε and OT_ε are convex and differentiable on non compact spaces ($\mathcal{X} = \mathbb{R}^d$)
2) Depending on m_1, m_2 the entropy bias can be blurring or shrinking
3) a Fast modified Sinkhorn algorithm to compute debiased barycenters



Entropy regularized OT

$$\mathrm{OT}_{\varepsilon}^{m_1, m_2}(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \int_{\mathcal{X} \times \mathcal{X}} C(x, y) \mathrm{d}\pi(x, y) + \underline{\varepsilon \mathrm{KL}(\pi | m_1 \otimes m_2)}$$

regularizer: relative entropy

$m_1, m_2 \in$ reference measures in $\mathcal{P}(\mathcal{X})$

$$\mathrm{KL}(\pi | r) \stackrel{\text{def}}{=} \int_{\mathcal{X} \times \mathcal{X}} \log \left(\frac{\mathrm{d}\pi}{\mathrm{d}r} \right) \mathrm{d}\pi \quad \text{if} \quad \pi \ll r \quad \text{else} \quad +\infty$$

(0) Discrete case $\mathcal{X} = \{x_1, \dots, x_n\}$

$$\alpha, \beta \in \Delta_n \overset{\text{def}}{=} \{x \in \mathbb{R}^d_+ | \sum_{i=1}^n x_i = 1\}$$

$$\mathrm{OT}^{\mathcal{U}}_{\varepsilon}(\alpha, \beta) = \min_{\substack{\pi \in \mathbb{R}^{n \times n}_+ \\ \pi \mathbb{1} = \alpha, \pi^\top \mathbb{1} = \beta}} \langle (C(x_i, x_j)_{ij}), \pi \rangle + \varepsilon \langle \pi, \log(\pi) - 1 \rangle$$

(Cuturi, Neurips 13')

(I) Lebesgue continuous case

$$\mathrm{OT}^{\mathcal{L}}_{\varepsilon}(\alpha, \beta) \overset{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \int_{\mathcal{X} \times \mathcal{X}} C(x, y) \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}(\pi | \mathcal{L} \otimes \mathcal{L})$$

(II) General case

$$\mathrm{OT}^{\otimes}_{\varepsilon}(\alpha, \beta) \overset{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \int_{\mathcal{X} \times \mathcal{X}} C(x, y) \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta)$$

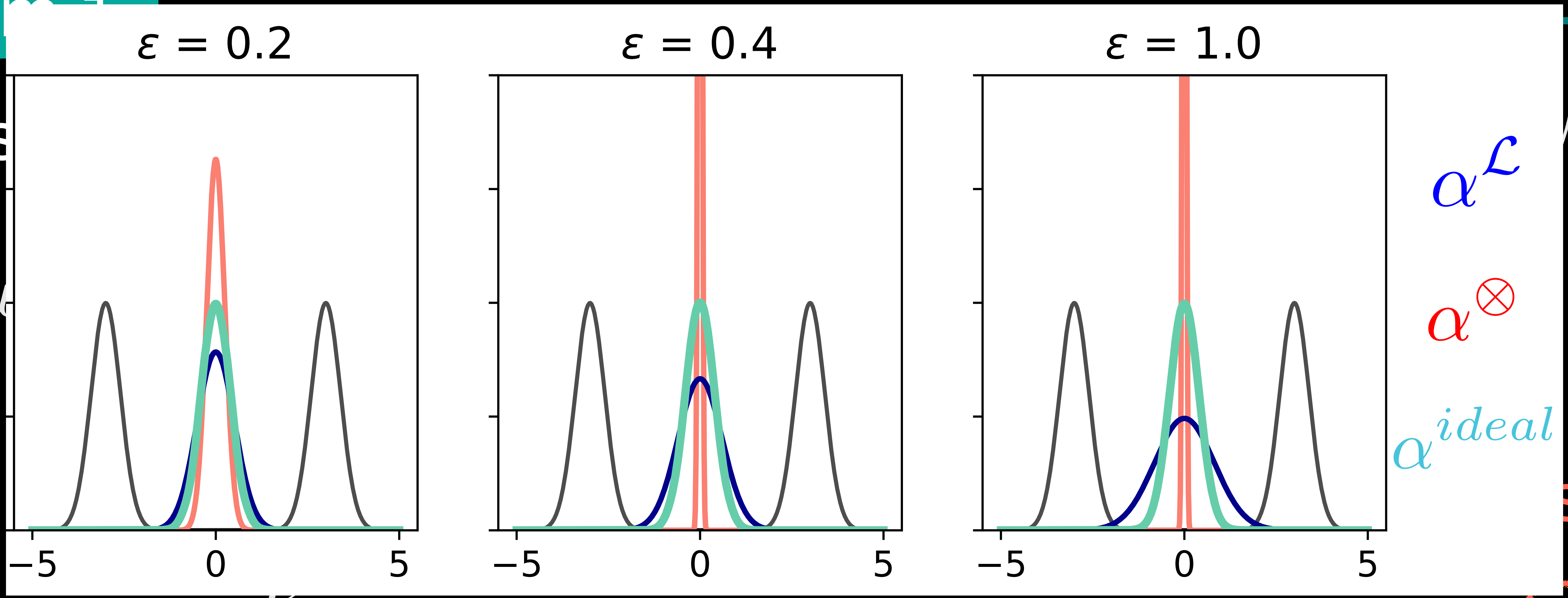
(Genevay 18', Feydy 19')

Quantifying the entropy bias

Theorem 1

Let $\mathcal{G}$ the s … $\mathcal{N}(\mu_k, \sigma^2)$

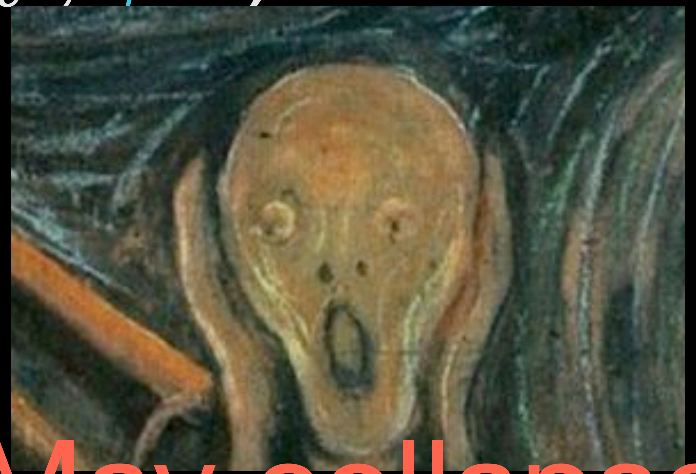
$\bar{\mu} = \sum_{k=1}^{K} w$ …



smoothing

$$\alpha^{\mathcal{L}} \overset{\text{def}}{=} \arg\min_{\beta \in \mathcal{G}} \sum_{k=1}^{K} w_k \, \mathrm{OT}_{\varepsilon}^{\mathcal{L}}(\alpha_k, \beta) \sim \mathcal{N}\left(\bar{\mu}, \sigma^2 + \frac{\varepsilon^2}{2}\right)$$

(II) General case



May collapse to a dirac

$$\alpha^{\otimes} \overset{\text{def}}{=} \arg\min_{\beta \in \mathcal{G}} \sum_{k=1}^{K} w_k \, \mathrm{OT}_{\varepsilon}^{\otimes}(\alpha_k, \beta) \sim \mathcal{N}\left(\bar{\mu}, \left(\sigma^2 - \frac{\varepsilon^2}{2}\right)_+\right)$$

Entropy shrinking

H. Janati - Inria Saclay

Quantifying the entropy bias

Theorem 1

Let $\mathcal{G}$ the set of sub-Gaussian measures in $\mathbb{R}$ and $\alpha_k \sim \mathcal{N}(\mu_k, \sigma^2)$

"Entropic OT is maximum likelihood deconvolution"

$$Y = X + \sigma^2 Z, \quad Z \sim \mathcal{N}(0, \mathrm{Id})$$

$$P_X = \arg\min_P \mathrm{OT}^{\otimes}_{\sigma^2}(P, \sum_{i=1}^{n} \frac{1}{n}\delta_{y_i})$$

(Rigollet & Weed, 18')

$$\alpha^{\mathcal{L}} \stackrel{\text{def}}{=} \arg\min_{\beta \in \mathcal{G}} \sum_{k=1} w_k \, \mathrm{OT}^{\mathcal{L}}_{\varepsilon}(\alpha_k, \beta) \sim \mathcal{N}\left(\bar{\mu}, \sigma^2 + \frac{\varepsilon^2}{2}\right)$$

(II) General case

$$\alpha^{\otimes} \stackrel{\text{def}}{=} \arg\min_{\beta \in \mathcal{G}} \sum_{k=1}^{K} w_k \, \mathrm{OT}^{\otimes}_{\varepsilon}(\alpha_k, \beta) \sim \mathcal{N}\left(\bar{\mu}, (\sigma^2 - \frac{\varepsilon^2}{2})_+\right)$$

Yet, useful

Entropy shrinking

Fixing the entropy bias

(Feydy et al, 19): $\mathcal{X}$ is compact

$$S_\varepsilon(\alpha, \beta) \stackrel{\text{def}}{=} \overset{\otimes}{\mathrm{OT}}_\varepsilon(\alpha, \beta) - \frac{1}{2}\left(\overset{\otimes}{\mathrm{OT}}_\varepsilon(\alpha, \alpha) + \overset{\otimes}{\mathrm{OT}}_\varepsilon(\beta, \beta)\right)$$

- ~~Entropy bias~~

+ $S_\varepsilon(\alpha, \beta) \geq 0;$

+ $S_\varepsilon(\alpha, \beta) = 0 \Rightarrow \alpha = \beta$

Theorem

Why ?

Let $\mathcal{G}$ the set of sub-Gaussian measures in $\mathbb{R}$ and $\alpha_k \sim \mathcal{N}(\mu_k, \sigma^2)$

$\bar{\mu} = \sum_{k=1}^{K} w_k \mu_k$ $\quad C(x, y) = (x - y)^2$ then:

$$\alpha^S \stackrel{\text{def}}{=} \underset{\beta \in \mathcal{G}}{\arg\min} \sum_{k=1}^{K} w_k S_\varepsilon(\alpha_k, \beta) \sim \mathcal{N}\left(\bar{\mu}, \sigma^2\right)$$

To prove the convexity and differentiability of S_ε and OT_ε and characterize the barycentric optimality

Barycentric Algorithms: two different views

Lagrangian (Free supports)

Distributions represented

by point clouds (positions, weights)

- \+ scalable for sparse distributions in high dimensions
- \- Optimization over weights and positions: requires several Sinkhorn loops

Done by (Luise, 19')

Eulerian (Fixed supports)

Distributions represented

by histograms on a fixed grid

- \+ Parallelizable on regular grids (e.g. images)
- \- Not scalable memory footprint

Sinkhorn algorithm and IBP

$$\alpha^{\mathcal{U}} \overset{\text{def}}{=} \underset{\beta \in \Delta_n}{\arg\min} \sum_{k=1}^{K} w_k \, \mathrm{OT}_{\varepsilon}^{\mathcal{U}}(\alpha_k, \beta)$$

can be written as a KL projection:

$$\min_{\substack{\pi_1, \ldots, \pi_K \in \mathbb{R}_+^{n \times n} \\ \pi_k \mathbb{1} = \alpha_k \\ \pi_1^\top \mathbb{1} = \cdots = \pi_K^\top \mathbb{1}}} \sum_{k=1}^{K} w_k \mathrm{KL}(\pi_k | e^{-\frac{(C_{ij})}{\varepsilon}})$$

Solved using

"Iterative Bregman Projections" (IBP)

a.k.a. generalized Sinkhorn

Not possible with the product measure as reference :(

$$\alpha^{\otimes} \overset{\text{def}}{=} \underset{\beta \in \Delta_n}{\arg\min} \sum_{k=1}^{K} w_k \, \mathrm{OT}_{\varepsilon}^{\otimes}(\alpha_k, \beta)$$

What about

S_ε ?

Debiased Sinkhorn algorithm and IBP

$\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}), \quad \pi_1 = \alpha$ and $\pi_2 = \beta$ then:

$$\mathrm{KL}(\pi | m_1 \otimes m_2) = \mathrm{KL}(\pi | \alpha \otimes \beta) + \mathrm{KL}(\alpha | m_1) + \mathrm{KL}(\beta | m_2)$$

(Di marino, 19')

Then S_ε is independent of the reference measure:

If discrete: $S_\varepsilon^{\mathcal{U}}(\alpha, \beta) = S_\varepsilon^{\otimes}(\alpha, \beta)$

If Lebesgue-continuous: $S_\varepsilon^{\mathcal{L}}(\alpha, \beta) = S_\varepsilon^{\otimes}(\alpha, \beta)$

(Special thanks to T. Séjourné & F-X Vialard)

Debiased Sinkhorn algorithm and IBP

Using $S_\varepsilon^{\mathcal{U}}(\alpha, \beta) = S_\varepsilon^{\otimes}(\alpha, \beta)$

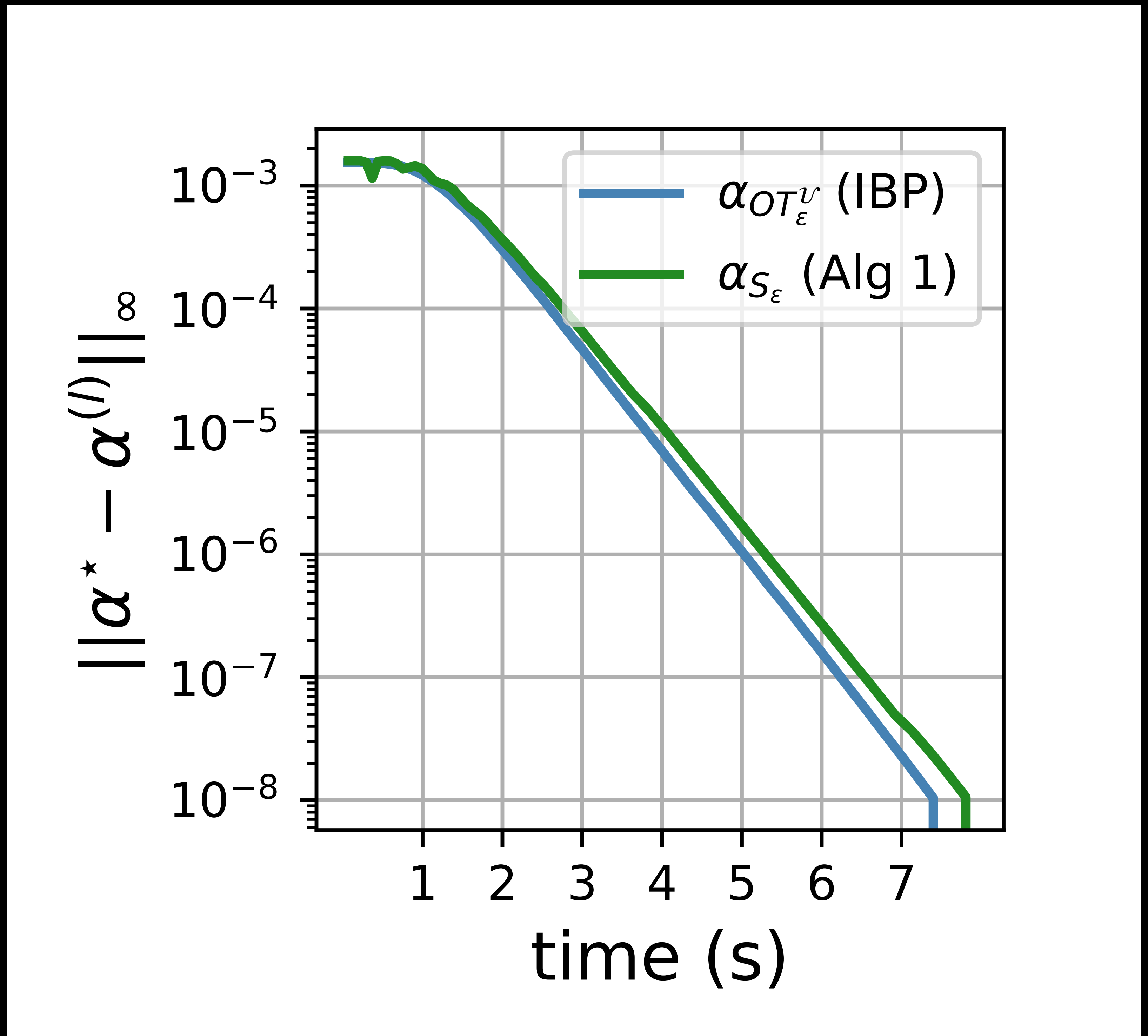
$$\alpha^S \stackrel{\text{def}}{=} \operatorname*{arg\,min}_{\beta \in \Delta_n} \sum_{k=1}^{K} w_k S_\varepsilon^{\mathcal{U}}(\alpha_k, \beta)$$ equivalent to:

$$\min_{\substack{\pi_1, \dots, \pi_K \in \mathbb{R}_+^{n \times n} \\ \pi_k \mathbb{1} = \alpha_k \\ \pi_1^\top \mathbb{1} = \dots = \pi_K^\top \mathbb{1} \\ d \in \mathbb{R}_+^n}} \sum_{k=1}^{K} w_k \mathrm{KL}(\pi_k | e^{-\frac{(C_{ij})}{\varepsilon}} d_j) + \frac{\varepsilon}{2} \langle d - \mathbb{1}, e^{-\frac{(C_{ij})}{\varepsilon}} (d - \mathbb{1}) \rangle$$

Alternating block minimization:

1) IBP loop
2) Symmetric Sinkhorn

Debiased Sinkhorn algorithm and IBP



Algorithm 1 Debiased Sinkhorn Barycenter

Input: $\alpha_1, \ldots, \alpha_K, \mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$

Output: $\alpha_{\mathrm{S}_\varepsilon}$

Initialize all scalings (b_k), d to $\mathbb{1}$,

repeat

  for $k = 1$ **to** K **do**

    $a_k \leftarrow \left(\frac{\alpha_k}{\mathbf{K} b_k}\right)$

  end for

  $\alpha \leftarrow d \odot \prod_{k=1}^{K} (\mathbf{K}^\top a_k)^{w_k}$

  for $k = 1$ **to** K **do**

    $b_k \leftarrow \left(\frac{\alpha}{\mathbf{K}^\top a_k}\right)$

  end for

  $d \leftarrow \sqrt{d \odot \left(\frac{\alpha}{\mathbf{K} d}\right)}$

until convergence

Take home message

- The entropic barycenter of univariate Gaussians is Gaussian
- The entropic bias can be smoothing or shrinking
- The debiased barycenter can be computed on a GPU-friendly modified Sinkhorn algorithm