Information Geometry of Entropy-Regularized Optimal Transport

Ryo Karakida (National Institute of AIST, Japan)

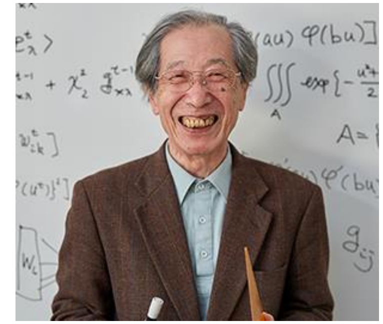
Shun-ichi Amari (RIKEN CBS, Japan),

Masafumi Oizumi (Univ. of Tokyo, Japan),

Marco Cuturi (Google Brain & ENSAE, France)

ICIAM2019 minisymposium
"Distance Metrics and Mass Transfer Between High Dimensional Point Clouds"
July 17th , 2019

"**Information geometry of Wasserstein divergence**"
Ryo Karakida, Shun-ichi Amari
Geometric Science of Information (GSI 2017)



“**Information Geometry Connecting Wasserstein Distance and Kullback-Leibler Divergence via the Entropy-Relaxed Transportation Problem**"
Shun-ichi Amari, Ryo Karakida, Masafumi Oizumi
Information Geometry (2018)

"**Iformation Geometry for Regularized Optimal Transport and Barycenters of Patterns**"
Shun-ichi Amari, Ryo Karakida, Masafumi Oizumi, Marco Cuturi
Neural Computation, (2019)

Outline

- **Background**
  - Entropy-regularized Optimal Transport (OT)

- **Information geometry of entropy-regularized OT**
  - Optimal transportation plan as exponential family
  - Dually flat structure

- **Alternative divergence to entropy-regularized cost**
  - An information geometric viewpoint
  - Barycenter of patterns

Optimal Transport Problem in S_{n-1}

(a.k.a. Hitchcock Problem)

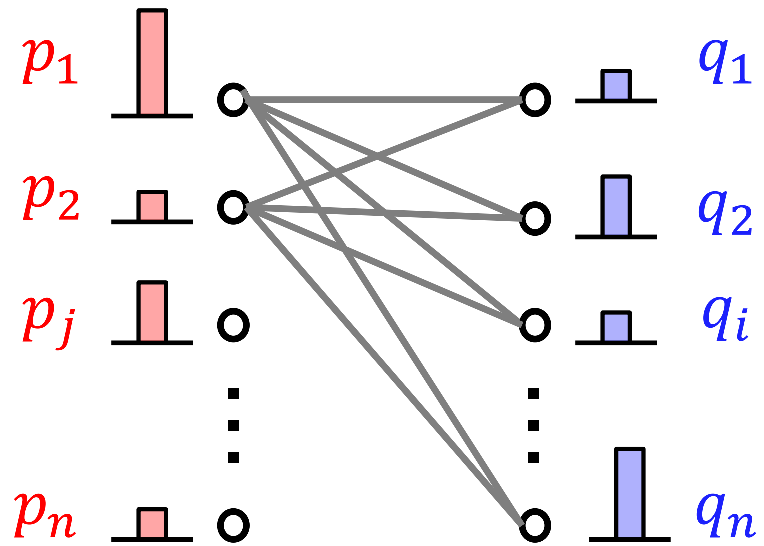
- **Probability simplex:**

$$S_{n-1} = \left\{ \boldsymbol{p} \in \mathbb{R}^n \mid \sum_i p_i = 1, \quad p_i \geq 0 \right\}$$

$\boldsymbol{p}, \boldsymbol{q} \in S_{n-1}$

$\Pi(\boldsymbol{p}, \boldsymbol{q})$: Set of joint distributions P

$$\sum_i P_{ij} = p_j \quad \sum_j P_{ij} = q_i$$



Metric: M_{ij}

ex. distance $|i-j|^2$

Wasserstein distance:

$$W(\boldsymbol{p}, \boldsymbol{q}) = \min_{P \in \Pi(\boldsymbol{p}, \boldsymbol{q})} \sum_{i,j} M_{ij} P_{ij}$$

Optimal P is solvable by linear programming

Optimal Transport Problem in S_{n-1}

Wasserstein distance:

$$W(\boldsymbol{p},\boldsymbol{q}) = \min_{P\in\Pi(\boldsymbol{p},\boldsymbol{q})} \sum_{i,j} M_{ij}P_{ij}$$

$$\boldsymbol{p},\boldsymbol{q} \in S_{n-1}$$

Difficulties

- Computationally demanding $O(n^3 \ln n)$
- Solution is not necessarily unique
- Undifferentiable with $\boldsymbol{p}$ and $\boldsymbol{q}$

Entropy-regularized OT [M. Cuturi, NIPS 2013]:

$$W(\boldsymbol{p},\boldsymbol{q}) = \min_{P\in\Pi(\boldsymbol{p},\boldsymbol{q})} \sum_{i,j} M_{ij}P_{ij} + \lambda \sum_{ij} P_{ij} \ln P_{ij} \qquad (\lambda > 0)$$

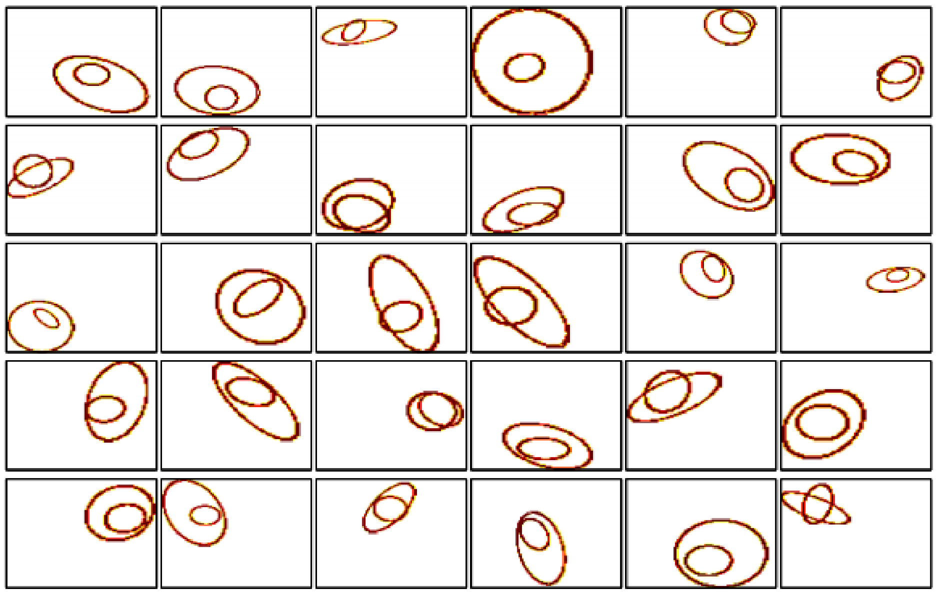
Advantages

- Fast O (n^2) and GPU-friendly
- Unique solution
- Differentiable

Entropy-R Optimal Transport in S_{n-1}

- **Barycenter** of images $\boldsymbol{q}_i$ $(i = 1, 2, \dots, N)$

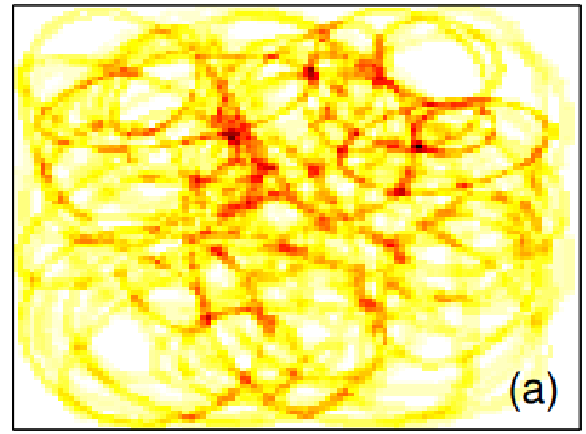
$$\min_{\boldsymbol{p}} \sum_{i=1}^{N} W(\boldsymbol{p}, \boldsymbol{q}_i)$$



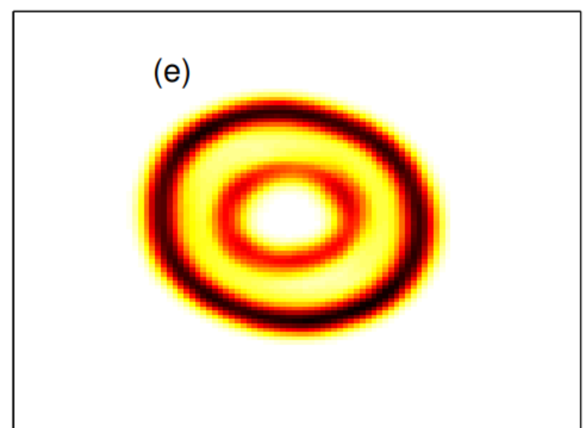
[Cuturi&Doucet, ICML 2014]

Spatially close pixels tend to take similar values

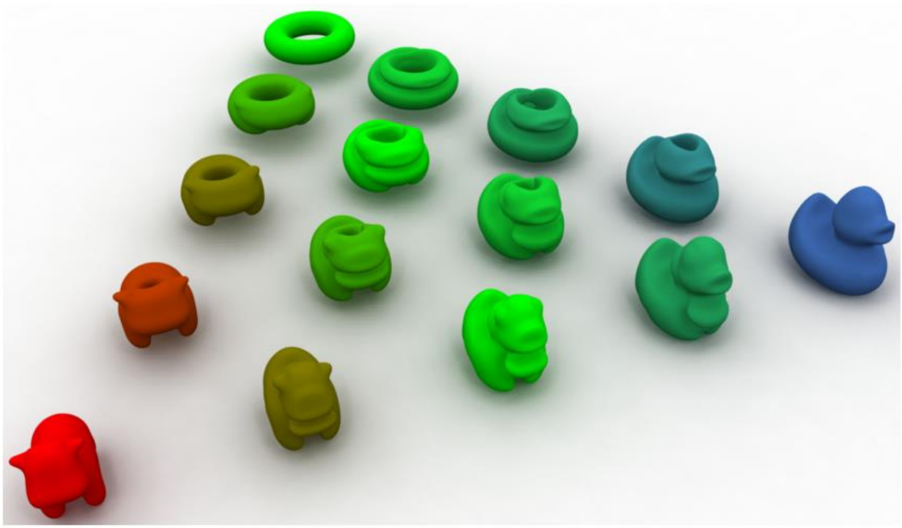
KL divergence



Wasserstein



Application

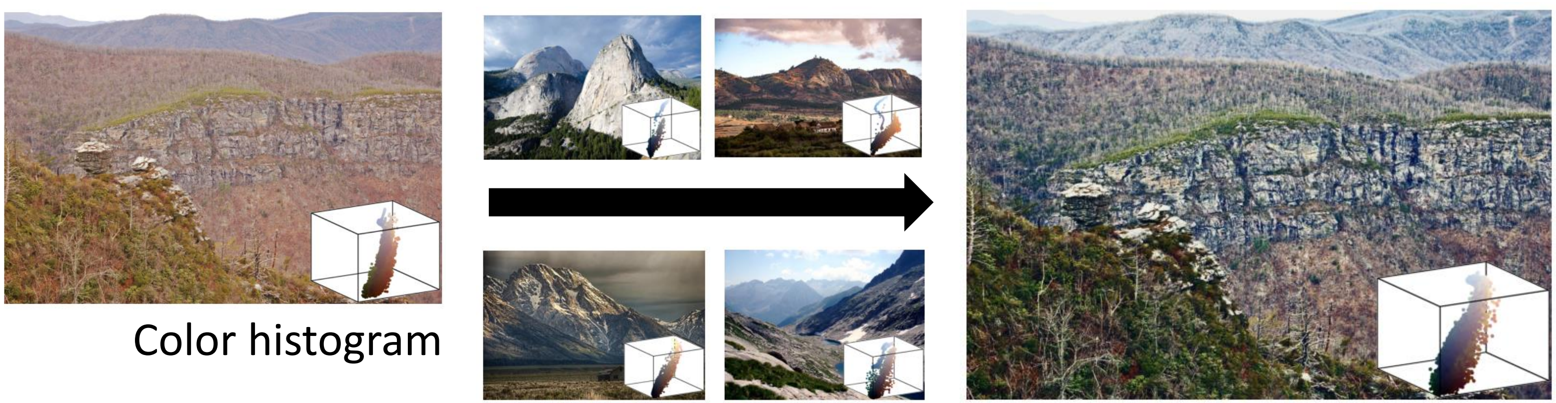


Imaging

[Solomon+, SIGGRAPH 2015]

Color Grading

[Bonneel+, SIGGRAPH 2016]



Color histogram

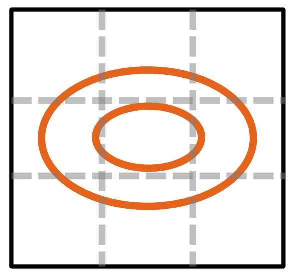
Dictionary learning [Schmitz+ 2018], generative models [Genevay+ 2017], ...

Geometrical structures of probabilistic distributions

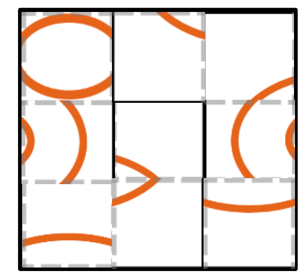
Kullback-Leibler (KL) divergence

Invariant under reversible transformations of random variables

$$\int dp(x) \ln \frac{p(x)}{q(x)}$$



Permutation



Discretization

Destruction of neighboring structure

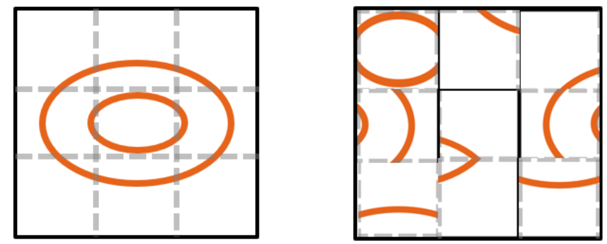
Information Geometry [Amari 1985]: Fisher information metric (Riemannian metric), dual affine connections

Geometry of probabilistic distributions

Kullback-Leibler (KL) divergence

Invariant under reversible transformations of random variables

$$\int dp(x) \ln \frac{p(x)}{q(x)}$$

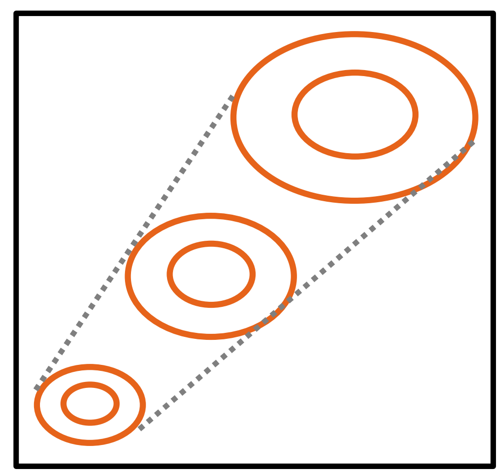


Information Geometry [Amari 1985]

Wasserstein distance

Reflects the ground metric
supporting probability measures

$$\left(\inf_{P \in \Pi(\mu,\nu)} \int d(x,y)^p dP(x,y) \right)^{1/p}$$



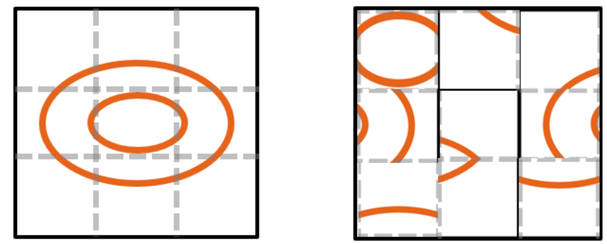
*Is **Information Geometry** not related?*

Geometry of probabilistic distributions

Kullback-Leibler (KL) divergence

Invariant under reversible transformations of random variables

$$\sum_i p_i \ln \frac{p_i}{q_i}$$

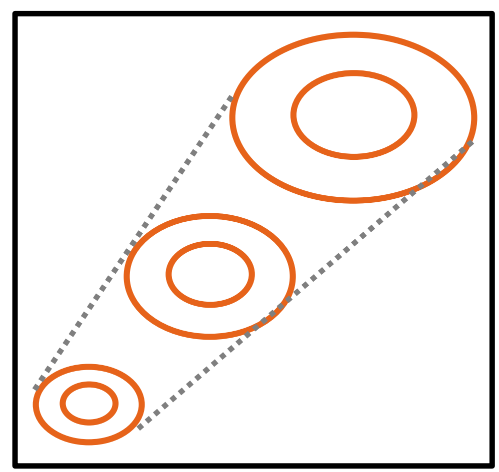


Information Geometry [Amari 1985]

Wasserstein distance

Reflects the ground metric
supporting probability measures

$$\min_{P \in \Pi(\boldsymbol{p},\boldsymbol{q})} \sum_{i,j} M_{ij} P_{ij} + \underbrace{\lambda \sum_{ij} P_{ij} \ln P_{ij}}_{\text{Entropy}}$$



We can introduce ***Information Geometry of OT plans***

Outline

- **Background**
  - Entropy-regularized Optimal Transport (OT)

- **Information geometry of entropy-regularized OT**
  - Optimal transportation plan as exponential family
  - Dually flat structure

- **Alternative divergence to entropy-regularized cost**
  - An information geometric viewpoint
  - Barycenter of patterns

Entropy-regularized OT in S_{n-1}

Probability simplex: $S_{n-1} = \left\{ \boldsymbol{p} \in \mathbb{R}^n \mid \sum_i p_i = 1, \quad p_i \geq 0 \right\}$

Marginal distributions: $\boldsymbol{p}, \boldsymbol{q} \in S_{n-1}$

Metric between p_i and q_j : M_{ij}

Entropy-regularized OT [M. Cuturi, NIPS 2013] **:**

$$\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \min_{\mathbf{P}} \frac{1}{1+\lambda} \langle \mathbf{M}, \mathbf{P} \rangle - \frac{\lambda}{1+\lambda} H(\mathbf{P}) \qquad (\lambda > 0)$$

$$\langle \mathbf{M}, \mathbf{P} \rangle = \sum_{ij} M_{ij} P_{ij}, \quad H(\mathbf{P}) = -\sum P_{ij} \log P_{ij}$$

- How to solve: **Method of Lagrange multiplier**

$$L_\lambda(\mathbf{P}) = \frac{1}{1+\lambda} \langle \mathbf{M}, \mathbf{P} \rangle - \frac{\lambda}{1+\lambda} H(\mathbf{P}) - \underline{\sum_{i,j} (\alpha_i + \beta_j) P_{ij}}$$

Multipliers: α_i, β_j

Entropy-regularized OT: Unique Solution

$$L_\lambda(\mathbf{P}) = \frac{1}{1+\lambda}\langle \mathbf{M}, \mathbf{P} \rangle - \frac{\lambda}{1+\lambda} H(\mathbf{P}) \underline{- \sum_{i,j} (\alpha_i + \beta_j) P_{ij}}$$

Multipliers: α_i, β_j

[M. Cuturi, NIPS 2013] revealed

Entropy-regularized OT has **a unique optimal solution** P_λ^*

$$P_{\lambda ij}^* = c a_i b_j K_{ij} \quad \text{with} \quad K_{ij} = \exp\left(-\frac{M_{ij}}{\lambda}\right)$$

$$a_i = \exp\left(\frac{1+\lambda}{\lambda}\alpha_i\right), \quad b_j = \exp\left(\frac{1+\lambda}{\lambda}\beta_j\right), \quad c = \frac{1}{\sum a_i b_j K_{ij}}$$

- Multipliers $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are determined by iterative computations

Sinkhorn algorithm: $\boldsymbol{a} \leftarrow \boldsymbol{p}./K^T\boldsymbol{b}, \quad \boldsymbol{b} \leftarrow \boldsymbol{q}./K\boldsymbol{a}$

Entropy-regularized OT: Convexity

- **Cost function φ_λ** $\quad \varphi_\lambda(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{1+\lambda}\langle \mathbf{M}, \mathbf{P}^*_\lambda \rangle - \frac{\lambda}{1+\lambda} H\left(\mathbf{P}^*_\lambda\right)$

> **Lemma.** Entropy-regularized cost $\varphi_\lambda(\boldsymbol{p},\boldsymbol{q})$ convex over $(\boldsymbol{p},\boldsymbol{q})$

Proof sketch : Wasserstein distance (linear in P) + Entropy (convex over P)

- Set of OT plans is an **Exponential Family**

$$P^*_\lambda(x;\boldsymbol{\theta}) = \exp\left\{\sum_{i,j}\theta^{ij}\delta_{ij}(x) - \psi_\lambda\right\}$$

$$\theta^{ij} = \frac{1+\lambda}{\lambda}(\alpha_i + \beta_j) - \frac{M_{ij}}{\lambda}$$

$$\delta_{ij}(x) = \begin{cases} 1 & \text{when } x = (i,j), \\ 0 & \text{otherwise} \end{cases}$$

- **Normalization factor ψ_λ is convex** over (α, β)
- $2(n-1)$-dimensional manifold; $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{n-1}$

Information geometry of entropy-regularized OT

$$\boldsymbol{\eta} = (\boldsymbol{p}, \boldsymbol{q})^T \qquad \boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^T$$

Theorem. Normalization factor ψ_λ and cost φ_λ are both convex and connected by the Legendre Transformation;

$$\psi_\lambda(\boldsymbol{\theta}) + \varphi_\lambda(\boldsymbol{\eta}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta} \ , \quad \boldsymbol{\theta} = \nabla_{\boldsymbol{\eta}} \varphi_\lambda(\boldsymbol{\eta}), \quad \boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}} \psi_\lambda(\boldsymbol{\theta})$$

Information geometry of entropy-regularized OT

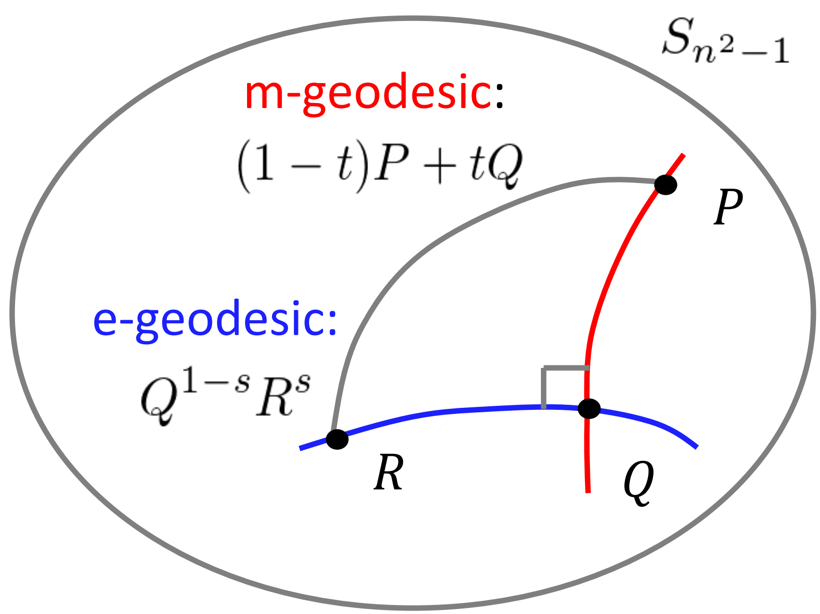
$$\psi_\lambda(\boldsymbol{\theta}) + \varphi_\lambda(\boldsymbol{\eta}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta} \qquad \boldsymbol{\theta} = \nabla_{\boldsymbol{\eta}} \varphi_\lambda(\boldsymbol{\eta}), \quad \boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}} \psi_\lambda(\boldsymbol{\theta})$$

- The set of P_λ^* is a $2(n-1)$-dimensional **dually flat manifold**

$$P_\lambda^*(x;\boldsymbol{\theta}) = \exp\left\{ \sum_{i,j} \theta^{ij} \delta_{ij}(x) - \psi_\lambda \right\}$$

$\boldsymbol{\eta} = (\boldsymbol{p}, \boldsymbol{q})^T$ **m-flat**

$\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^T$ **e-flat**



- **Generalized Pythagorean Theorem**

$$KL[P:R] = KL[P:Q] + KL[Q:R]$$

for $P, Q, R \in$ Optimal transportation plans

Information Geometry of Optimal Transportation Plans

Riemannian metric of dually flat manifold :

$$\mathbf{G}_\lambda = \nabla_{\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} \varphi_\lambda(\boldsymbol{\eta}), \quad \mathbf{G}_\lambda^{-1} = \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \psi_\lambda(\boldsymbol{\theta})$$

Fisher information matrix (in the θ-coordinates) is explicitly given by

$$\mathbf{G}_\lambda^{-1} = \left[\begin{array}{c|c} p_i\delta_{ij} - p_i p_j & P_{ij} - p_i q_j \\ \hline P_{ij} - p_i q_j & q_i\delta_{ij} - q_i q_j \end{array}\right]$$

- Useful in estimation; $\min_{\boldsymbol{q}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q})$

Dual affine connections : By using a cubic tensor $T_{ijk} = \partial_i \partial_j \partial_k \psi_\lambda$,

$$\Gamma_{ijk} = [ij;k] - \frac{1}{2}T_{ijk}, \quad \Gamma^*_{ijk} = [ij;k] + \frac{1}{2}T_{ijk}$$

$[ij;k]$: Levi-Civita connection

Sinkhorn algorithm = **iterations of e-projection**

[Amari, Karakida & Oizumi, Information Geometry (2018)]

Sinkhorn algorithm :
$$\boldsymbol{a} \leftarrow \boldsymbol{p}./K^T\boldsymbol{b}$$
$$\boldsymbol{b} \leftarrow \boldsymbol{q}./K\boldsymbol{a}$$

$$P^*_{\lambda ij} = c a_i b_j K_{ij}$$

$M(\boldsymbol{p},\cdot)$: Subspace conditioned by a fixed $\boldsymbol{p}$

$M(\cdot,\boldsymbol{q})$: Subspace conditioned by a fixed $\boldsymbol{q}$

- e-projection to $M(\boldsymbol{p},\cdot)$

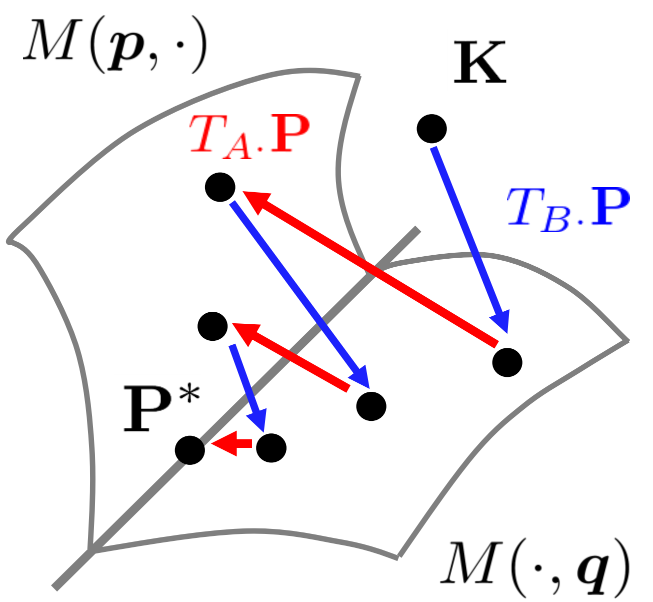
$$T_A.\mathbf{P} = (a_i P_{ij}) \ , \ a_i = \frac{p_i}{\sum_j P_{ij}}$$

$$KL[T_A.\mathbf{P} : \mathbf{P}] + KL[\mathbf{P}^* : T_A.\mathbf{P}] = KL[\mathbf{P}^* : \mathbf{P}]$$

$$KL[T_A.\mathbf{P} : \mathbf{P}^*] \leq KL[\mathbf{P} : \mathbf{P}^*]$$

- e-projection to $M(\cdot,\boldsymbol{q})$

$$T._B\mathbf{P} = (b_j P_{ij}) \ , \ b_j = \frac{q_j}{\sum_i P_{ij}}$$



Outline

- **Background**
  - Entropy-regularized Optimal Transport (OT)
- **Information geometry of entropy-regularized OT**
  - Optimal transportation plan as exponential family
  - Dually flat structure
- **Alternative divergence to entropy-regularized OT** [AKOC'19]
  - An information geometric viewpoint
  - Barycenter of patterns

Disadvantage of entropy-regularized cost

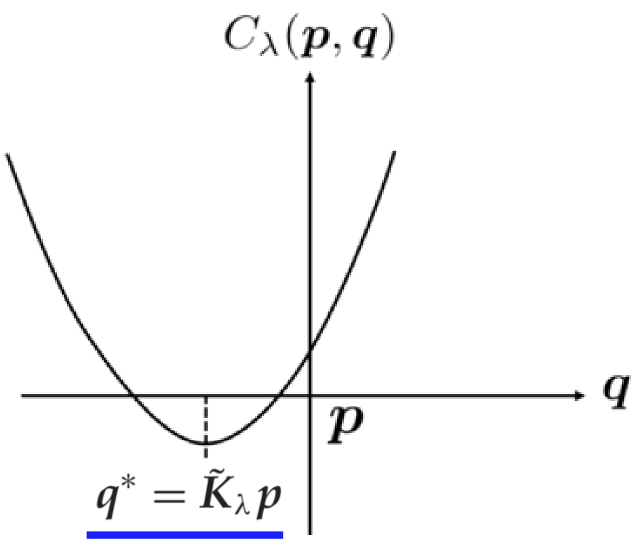
Entropy regularized cost:

$$C_\lambda(\boldsymbol{p},\boldsymbol{q}) = \min_{P\in\Pi(\boldsymbol{p},\boldsymbol{q})} \langle \mathbf{M},\mathbf{P}\rangle - \lambda H(\mathbf{P}) \qquad = (1+\lambda)\varphi_\lambda(\boldsymbol{p},\boldsymbol{q})$$

- It does **not** satisfy a criterion of divergence (distance);

$$\exists \boldsymbol{q} \quad C_\lambda(\boldsymbol{p},\boldsymbol{p}) > C_\lambda(\boldsymbol{p},\boldsymbol{q})$$

- Minimum is given by $\underline{\boldsymbol{q}^* = \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p}}$



Conditional metric

$$\boxed{\tilde{\boldsymbol{K}}_\lambda = \left[\frac{K_{ji,\lambda}}{K_{j\cdot}}\right]_{ij} \quad K_{ij} = \exp\left(-\frac{M_{ij}}{\lambda}\right) \quad K_{j\cdot} = \sum_i K_{ji}}$$

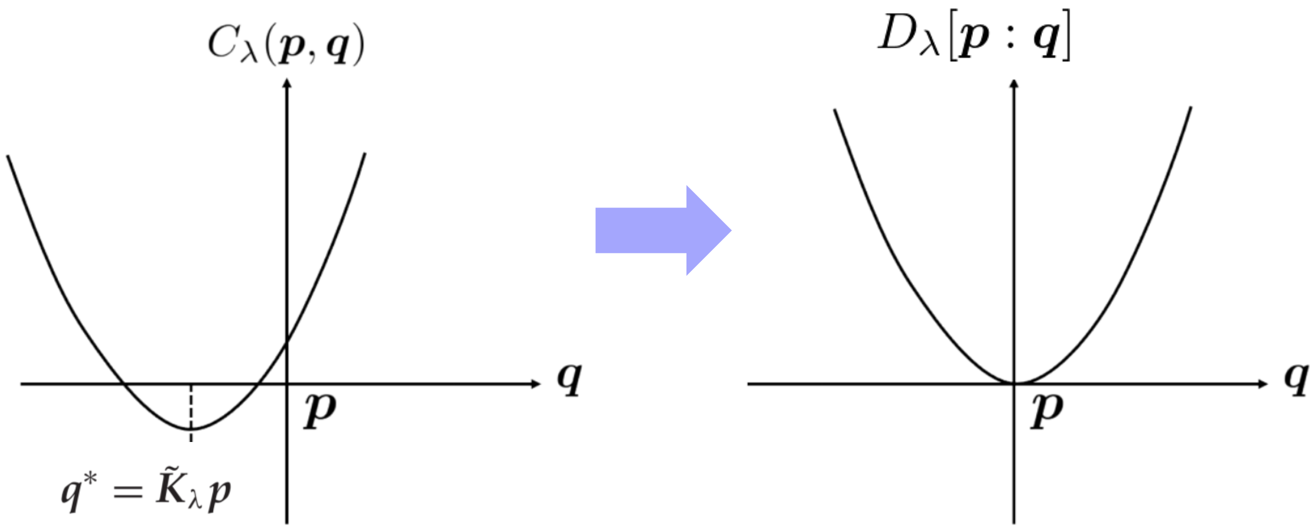
Divergence derived from entropy-regularized cost

- **A new divergence**

$$\boxed{D_\lambda[\boldsymbol{p}:\boldsymbol{q}] = (1+\lambda)\left(\mathrm{C}_\lambda(\boldsymbol{p}, \tilde{\mathbf{K}}_\lambda \boldsymbol{q}) - \mathrm{C}_\lambda(\boldsymbol{p}, \tilde{\mathbf{K}}_\lambda \boldsymbol{p})\right)}$$

- Convex with respect to $\boldsymbol{p}$ and $\boldsymbol{q}$

Proof based on the Riemannian metric $\mathbf{G}_\lambda = \nabla_{\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} \varphi_\lambda(\boldsymbol{\eta})$



Divergence derived from entropy-regularized cost

- **A new divergence**

$$\boxed{D_\lambda[\boldsymbol{p}:\boldsymbol{q}] = (1+\lambda)\left(\mathrm{C}_\lambda(\boldsymbol{p},\tilde{\mathrm{K}}_\lambda \boldsymbol{q}) - \mathrm{C}_\lambda(\boldsymbol{p},\tilde{\mathrm{K}}_\lambda \boldsymbol{p})\right)}$$

- Convergence to Wasserstein distance ($\lambda \to 0$)

- Convergence to a squared distance ($\lambda \to \infty$)

$$\lim_{\lambda\to\infty} D_\lambda[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{2}(\boldsymbol{q}-\boldsymbol{p})^T \tilde{\boldsymbol{M}}(\boldsymbol{q}-\boldsymbol{p})$$

$\tilde{\boldsymbol{M}}$: Symmetric matrix determined by $\boldsymbol{M}$

This overcomes disadvantage of C_λ

$$\lim_{\lambda\to\infty} C_\lambda(\boldsymbol{p},\boldsymbol{q}) = -\lambda(H(\boldsymbol{p})+H(\boldsymbol{q}))$$

which cannot measure the distance between $\boldsymbol{p}$ and $\boldsymbol{q}$

Barycenter of Patterns

$$D_\lambda[\boldsymbol{p}:\boldsymbol{q}] = (1+\lambda)\left(C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q}) - C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p})\right)$$

- Find the barycenter of input samples $\boldsymbol{p}_i$ $(i = 1, 2, \ldots, N)$

$$\boldsymbol{q}_C^* = \operatorname*{argmin}_{\boldsymbol{q}} \sum_i C_\lambda(\boldsymbol{p}_i, \boldsymbol{q}) \qquad \boldsymbol{q}_D^* = \operatorname*{argmin}_{\boldsymbol{q}} \sum_i D_\lambda[\boldsymbol{p}_i : \boldsymbol{q}]$$

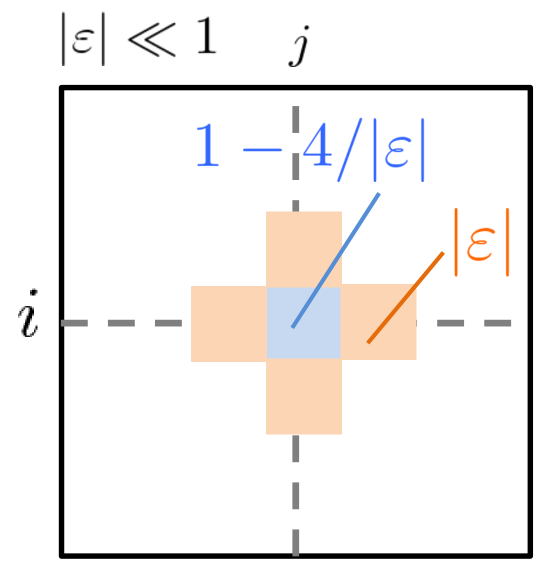
- **Sharpening**

$$\boldsymbol{q}_C^* = \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q}_D^*$$

Conditional metric

$$\boxed{\tilde{\boldsymbol{K}}_\lambda = \left[\frac{K_{ji,\lambda}}{K_{j\cdot}}\right]_{ij}} \underset{\lambda \to 0}{=}$$

Discrete diffusion operator



Under certain conditions,

$$\boldsymbol{q}_D^* = \tilde{\boldsymbol{K}}_\lambda^{-1} \boldsymbol{q}_C^*$$; sharper barycenter by "Inverse diffusion"

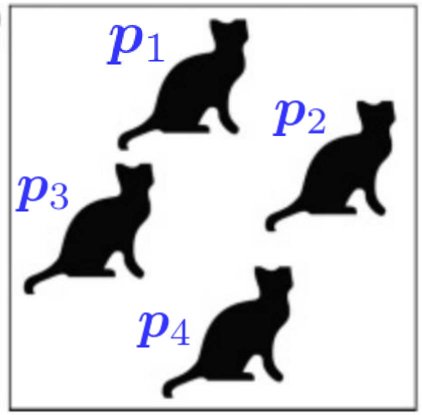
Barycenter of Patterns

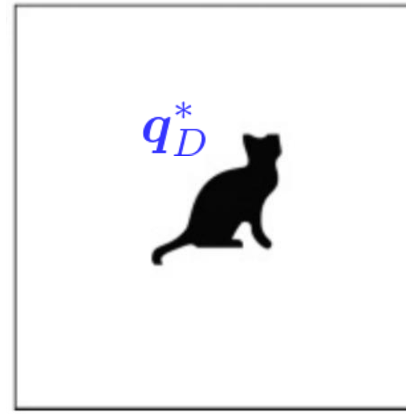
$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = (1+\lambda)\left(C_\lambda(\boldsymbol{p}, \tilde{K}_\lambda \boldsymbol{q}) - C_\lambda(\boldsymbol{p}, \tilde{K}_\lambda \boldsymbol{p})\right)$$

- **Shape-location separation**

  $$\boldsymbol{q}_D^* = \operatorname*{argmin}_{\boldsymbol{q}} \sum_i D_\lambda[\boldsymbol{p}_i : \boldsymbol{q}]$$

  is located at the barycenter of $\boldsymbol{p}_i$'s locations



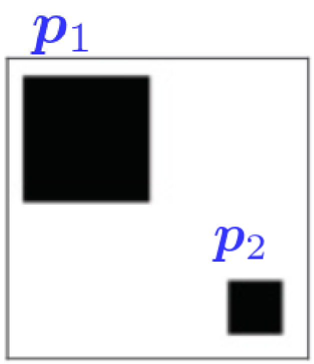


Remark: λ-Divergence [AKO'18]

$$\boxed{\widetilde{D}_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \varphi_\lambda(\boldsymbol{p}, \boldsymbol{p}) - \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) - \nabla_{\boldsymbol{q}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) \cdot (\boldsymbol{p} - \boldsymbol{q})}$$

- ***Canonical divergence*** from the Legendre duality
- It does **not** satisfy sharpening & shape-location separation

Barycenter of Patterns



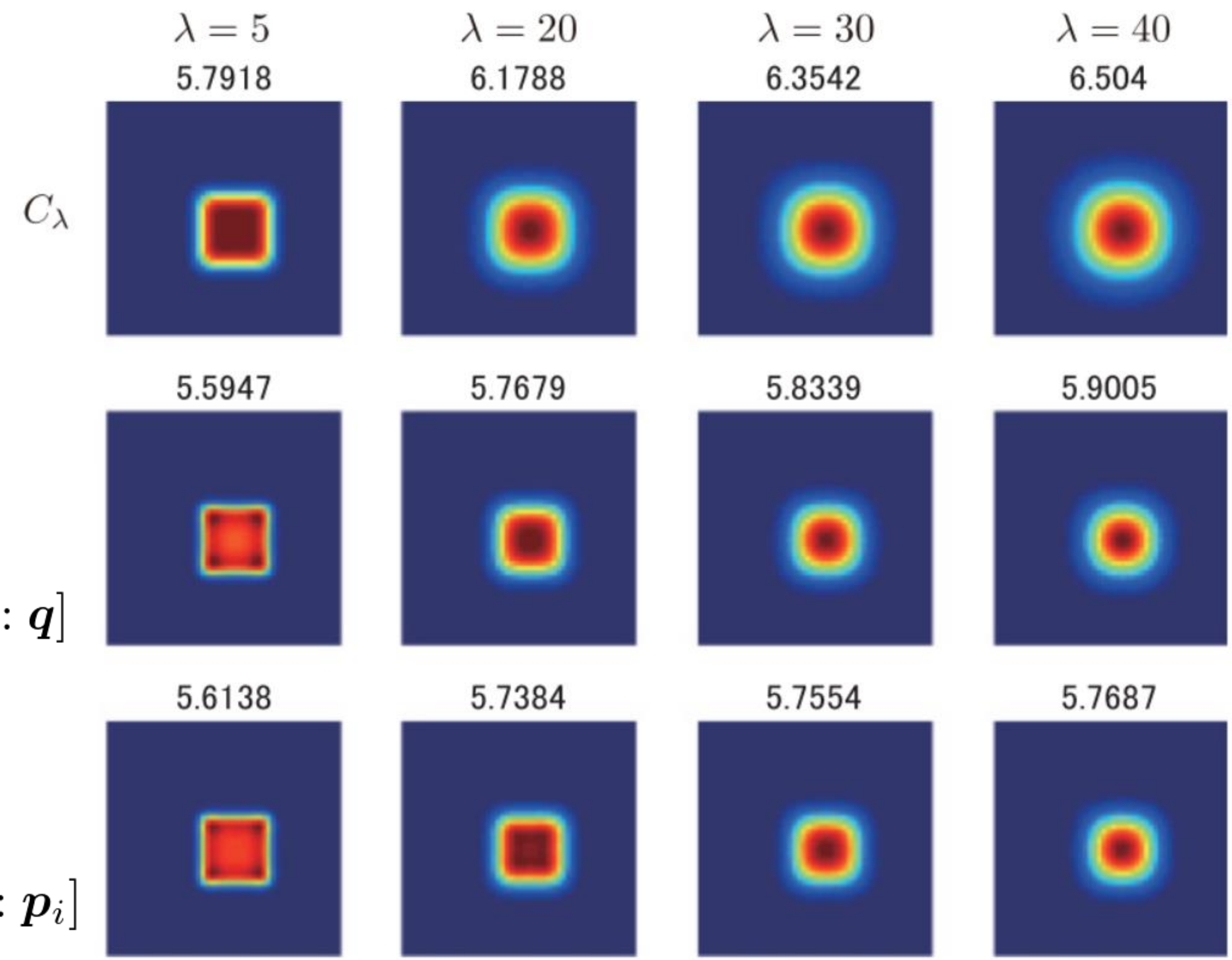
Two patterns

Right barycenter

$$\boldsymbol{q}_D^* = \underset{\boldsymbol{q}}{\operatorname{argmin}} \sum_i D_\lambda[\boldsymbol{p}_i : \boldsymbol{q}]$$

Left barycenter

$$\boldsymbol{q}_D^{*'} = \underset{\boldsymbol{q}}{\operatorname{argmin}} \sum_i D_\lambda[\boldsymbol{q} : \boldsymbol{p}_i]$$



D_λ-barycenters are obtained by a gradient method

Summary

- The set of entropy-regularized OT plans is an exponential family and naturally defines the **dually flat manifold**

  - We obtained its Riemannian metric & dual affine connections

- A new divergence D_λ

  - Inherits essential properties of Wasserstein geometry

    Sharpening, shape-location separation

  - Better than C_λ

    satisfying criterion of divergence, shaper barycenter

Future work

Information geometry of Entropy-regularized OT with continuous measures

- continuous $\boldsymbol{p}$ vs. continuous $\boldsymbol{q}$, discrete $\boldsymbol{p}$ vs. continuous $\boldsymbol{q}$
- Generative model