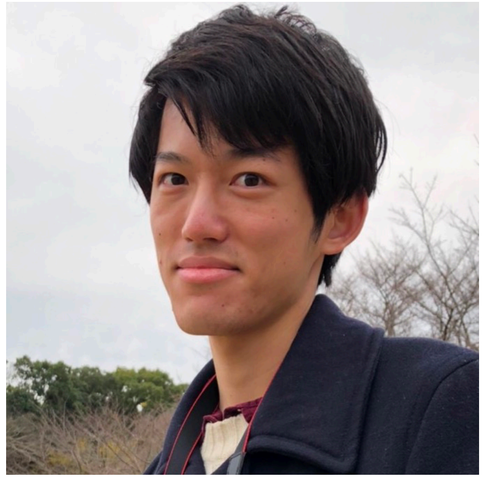
Sparse Regularized Optimal Transport with Deformed q-Entropy

March 15th, 2023

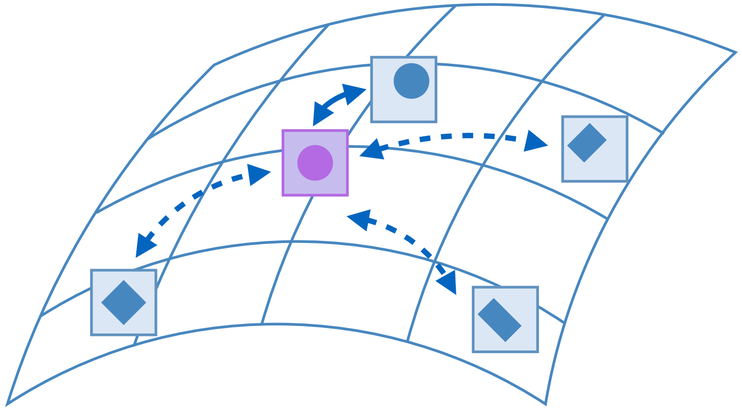
包 含 (Han Bao) / Kyoto University

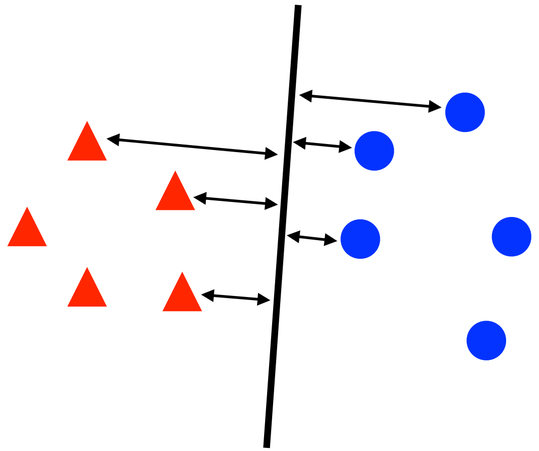
@Workshop OT

Short bio | Han Bao / 包　含（つつみ ふくむ）



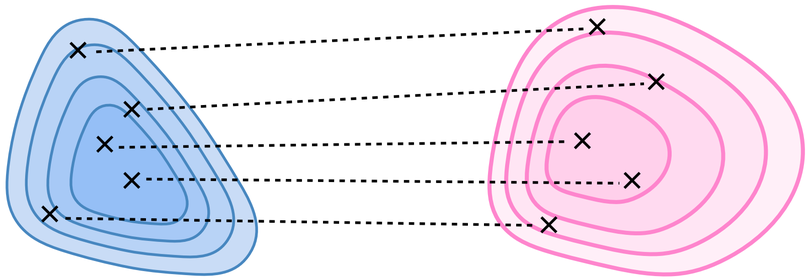
- 2017 April - 2022 March: Graduate student @ The University of Tokyo
  - My ex-office was in Science Bldg. #7 (very close from here!)
- 2022 April -: Project assistant professor @ Kyoto University (Hakubi center)
- Research field: computer science, machine learning
  - Design of robust loss functions
  - Theory of representation learning



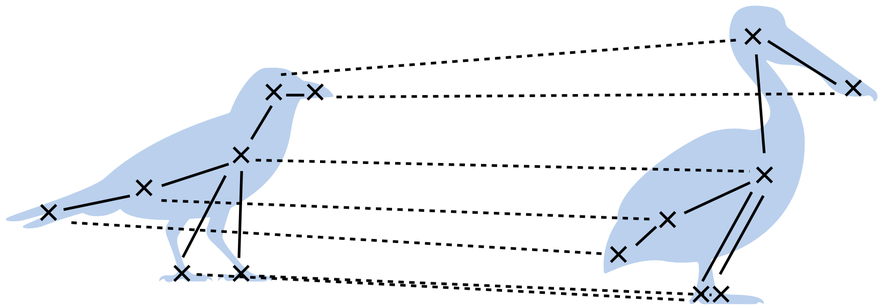


Optimal transport in machine learning

- Measure the discrepancy of probability distributions (statistical inference)



- Compute alignments between objects (for information retrieval)



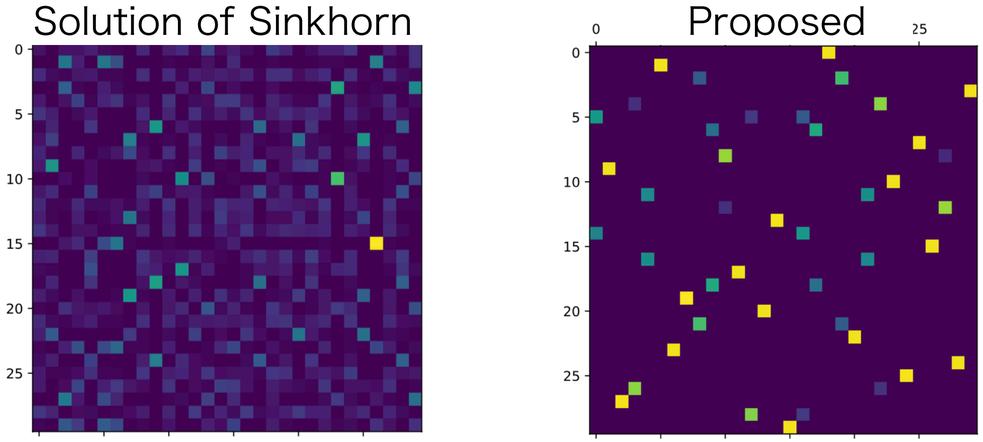
- **Q.** What kind of divergences are preferred?
- **Q.** Can we compute accurately and faster?
- **Q.** What can it be used for?

今日紹介する研究

B. & Sakaue.
Sparse Regularized Optimal Transport with Deformed q-Entropy. Entropy, 2022.

Summary in one page

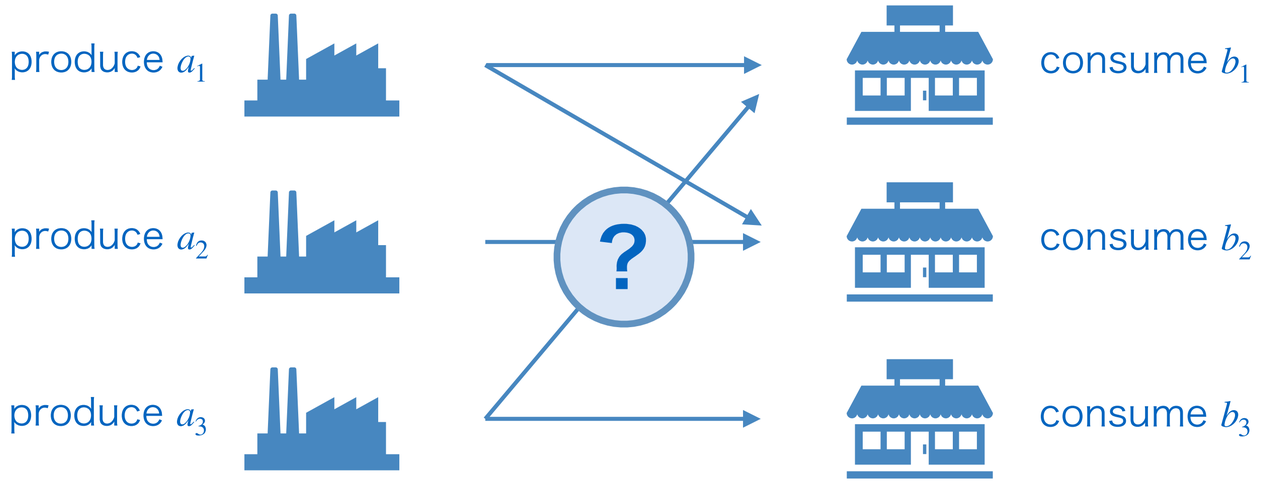
- What we do
  - Propose a new (entropic) regularizer to enhance sparsity of optimal transport solution



- What's interesting from practical viewpoint
  - Controllable trade-off between sparsity and computational time
- What's interesting from theoretical viewpoint
  - Connection to convex duality and statistical mechanics via Fenchel–Legendre transform

Introduction | Optimal transport

- Goal: to find an optimal way to transport goods from sources to targets



- LP formulation (Kantorovich's relaxation)

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle \quad \text{where} \quad U(\mu, \nu) = \left\{ \mathbf{\Pi} \in \mathbb{R}^{n \times n} \;\middle|\; \mathbf{\Pi 1} = \mathbf{a},\ \mathbf{\Pi}^\top \mathbf{1} = \mathbf{b} \right\}$$

$\mathbf{D}_{ij}$: cost between (i, j)

transport all goods from sources

satisfy demand constraints of targets

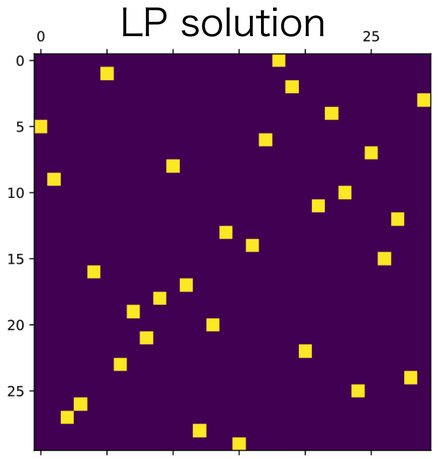
- Issue: $O(n^3 \log n)$ time complexity ☹

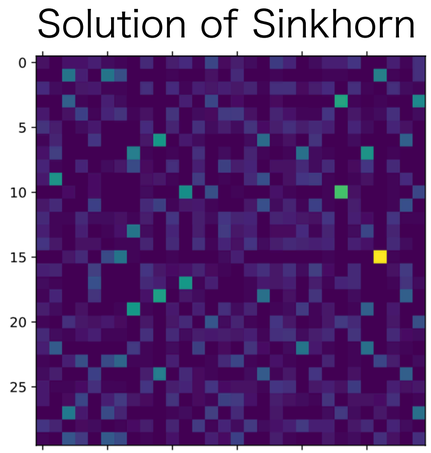
Introduction | Regularized optimal transport

- Motivation: to accelerate optimization (by using convex optimization)
- Primal formulation [Cuturi, 2013]

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \lambda \sum_{i,j} H(\mathbf{\Pi}_{ij}) \quad \text{where} \quad H(\pi) = \pi \log \pi - \pi \quad \text{(negative Shannon entropy)}$$

  - Sinkhorn algorithm solves the primal by block coordinate descent with $O(n^2)$ time 😊
- Issue: the regularized solution deviate from the original solution ☹
  - In particular: “sparsity” is lost

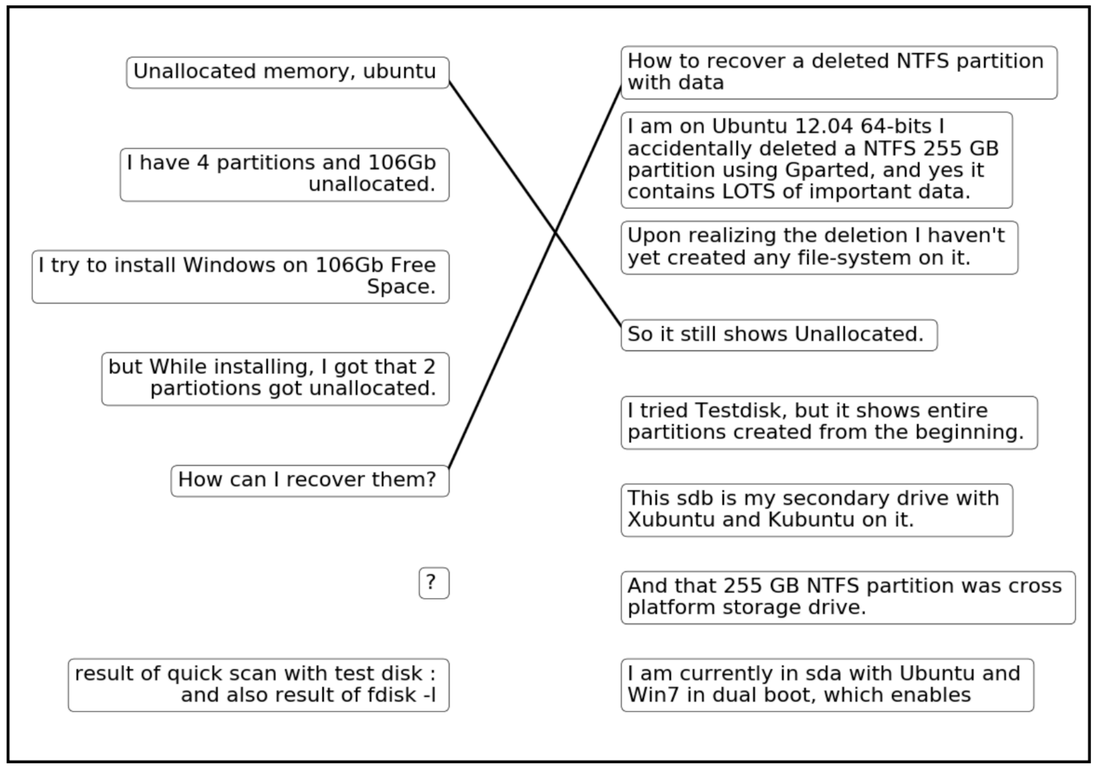




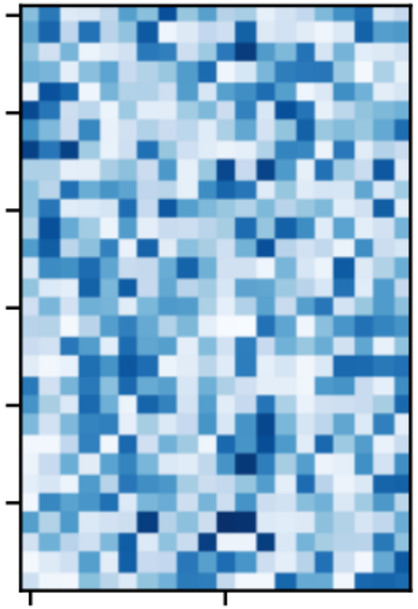
Cuturi, M. (NeurIPS2013). Sinkhorn distances: Lightspeed computation of optimal transport.

Introduction | Sparsity of transport matrix

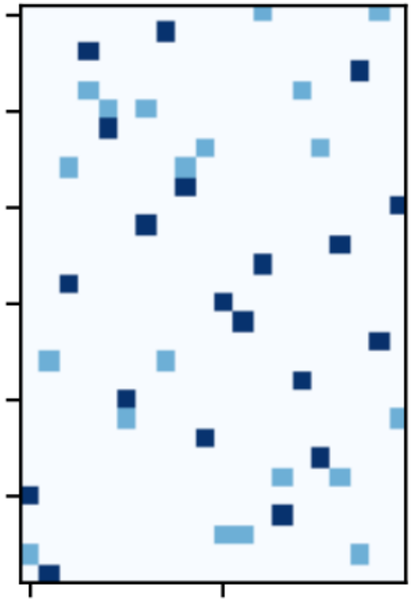
- Application of optimal transport: alignment
- Example: matching two different sentence pairs [Swanson *et al.*, 2020]



Dense alignment



Sparse alignment



Sparse alignment is more salient to extract meaningful matches

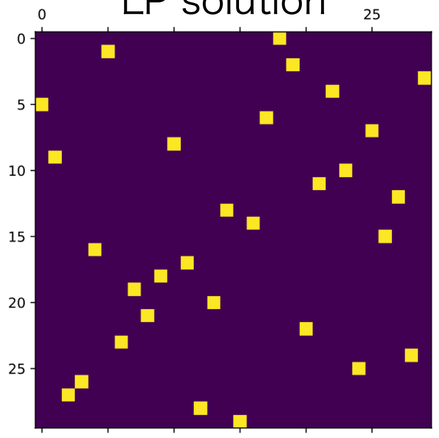
Motivation | New regularizer to balance sparsity/runtime

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle$$

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega(\mathbf{\Pi}_{ij})$$

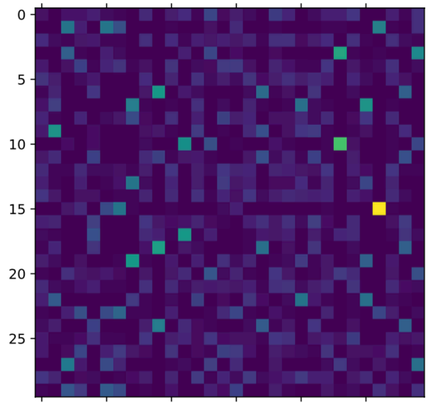
$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \lambda \sum_{i,j} H(\mathbf{\Pi}_{ij})$$

LP solution



?

Solution of Sinkhorn



- Sparse solution
- Cubic time

- Quadratic time
- Dense solution

Formulation | Regularized OT & dual

- Regularized OT (primal)

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega\big(\mathbf{\Pi}_{ij}\big) \quad \text{for a convex regularizer } \Omega : \mathbb{R} \to \mathbb{R}$$

- Regularized OT (dual)

$$\sup_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{R}^n} -\langle \mathbf{a}, \boldsymbol{\alpha} \rangle - \langle \mathbf{b}, \boldsymbol{\beta} \rangle - \sum_{i,j} \Omega^\star\big(-\mathbf{D}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j\big) \quad \text{where } \Omega^\star \text{ is Fenchel–Legendre transform}$$

  - By Lagrangian ($\boldsymbol{\alpha}, \boldsymbol{\beta}$: multipliers) & strong duality
  - Note: dual is unconstrained

- Primal-dual correspondence (inverse link)

$$\mathbf{\Pi}_{ij} = \nabla \Omega^\star\big(-\mathbf{D}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j\big) \quad \text{for all i, j}$$

⇒ Primal transport plan $\mathbf{\Pi}$ can be obtained from dual variables $\boldsymbol{\alpha}, \boldsymbol{\beta}$

Fenchel–Legendre transform

$$\Omega^\star(\eta) = \sup_{\pi \in \mathrm{dom}(\Omega)} \langle \pi, \eta \rangle - \Omega(\pi)$$

Examples

- (Negative) Shannon entropy
  - Regularizer $\Omega(\pi) = \lambda(\pi \log \pi - \pi)$
  - Inverse link $\nabla\Omega^\star(\eta) = \exp(\eta/\lambda)$
- 2-norm
  - Regularizer $\Omega(\pi) = \frac{\lambda}{2}\pi^2$
  - Inverse link $\nabla\Omega^\star(\eta) = \frac{1}{\lambda}\max\{0,\eta\}$

- Primal-dual correspondence

$$\mathbf{\Pi}_{ij} = \nabla\Omega^\star\left(-\mathbf{D}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j\right)$$

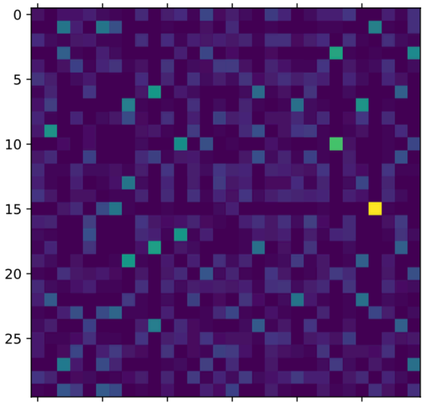
Gibbs kernel

Energy
(= cost + multipliers)

$$\mathbf{\Pi}_{ij} = \exp\left(-\frac{\mathbf{D}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j}{\lambda}\right)$$

Temperature
(= reg. strength)

Dense solution



Idea | Regularizer from alternative (inverse) link function

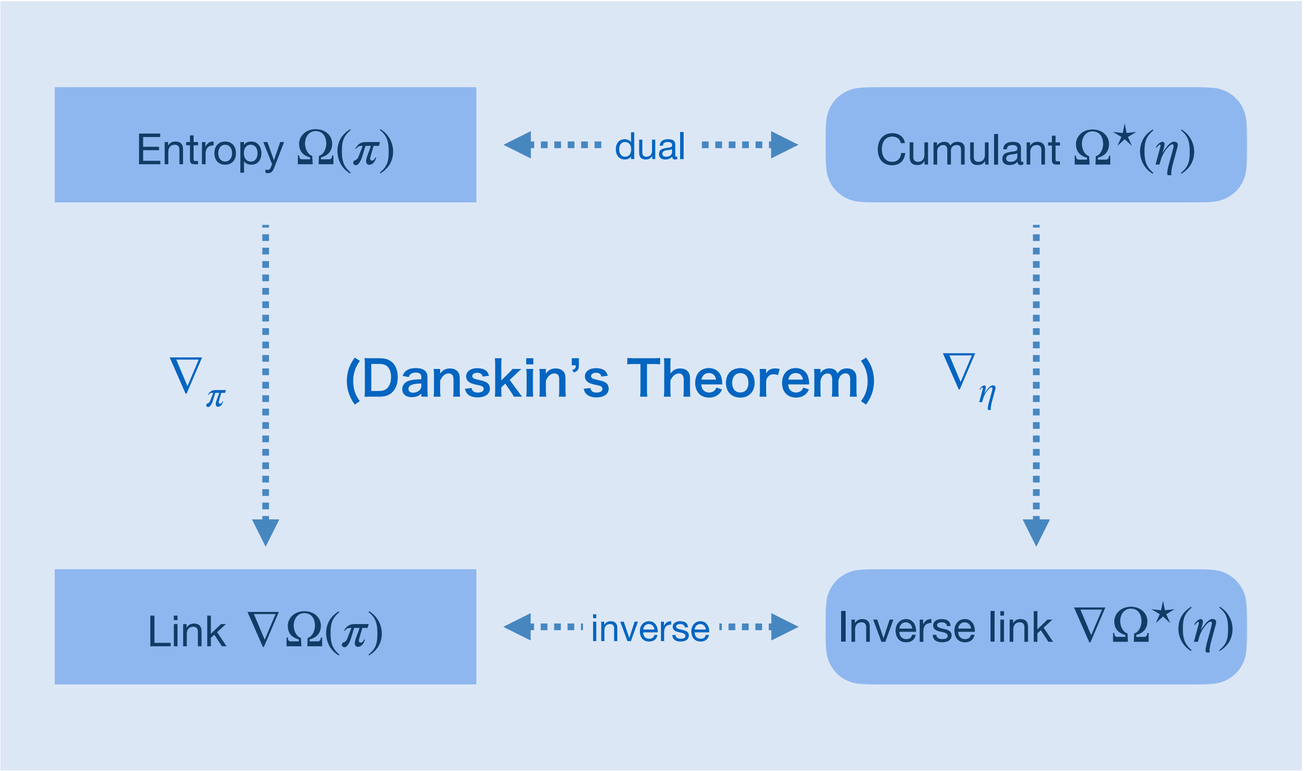
- Instead of a “dense” inverse link …

$$\mathbf{\Pi}_{ij} = \exp\left(-\frac{\mathbf{D}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j}{\lambda}\right)$$

- Design a “sparse” inverse link

$$\mathbf{\Pi}_{ij} = \boxed{?}\left(-\frac{\mathbf{D}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j}{\lambda}\right)$$

- Then, recover an entropy-like function
  - ❖ By convex duality (see right figure)



- **Q.** How to design a sparse inverse link?

Idea | q-exponential functions

- q-exponential

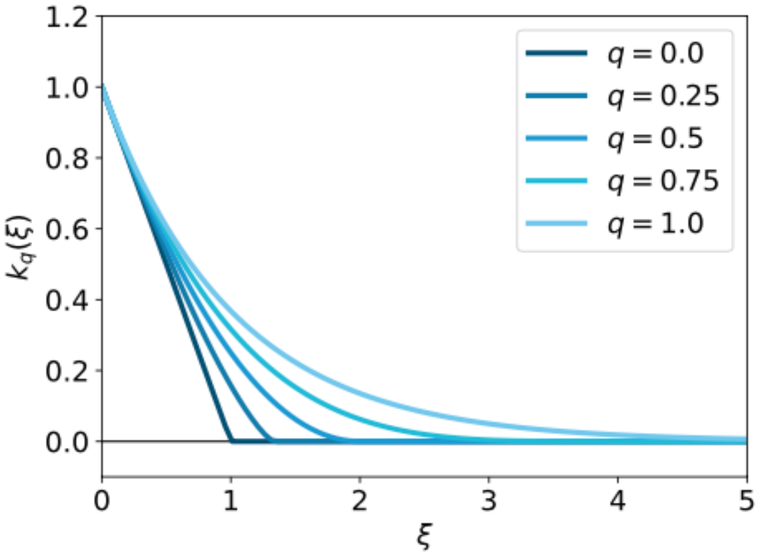
$$\exp_q(x) := \begin{cases} \max\{0, 1 + (1-q)x\}^{1/(1-q)} & \text{if } 0 \le q < 1 \\ \exp(x) & \text{if } q = 1 \end{cases}$$

  - Used in Tsallis statistics, robust Bayesian inference, etc.

- Proposed: replace Gibbs kernel with q-exponential distributions

Before: $\mathbf{\Pi}_{ij} = \exp\left(-\frac{\mathbf{D}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j}{\lambda}\right)$ (support is $\mathbb{R}_{\geq 0}$)

After: $\mathbf{\Pi}_{ij} = \boxed{\exp_q}\left(-\frac{\mathbf{D}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j}{\lambda}\right)$ (support is subset of $\mathbb{R}_{\geq 0}$)



Proposed | Regularizer based on q-exponential functions

- Derive primal regularizer from q-exp

$$\Omega(\pi) = \frac{\lambda}{2-q}(\pi \log_q(\pi) - \pi)$$

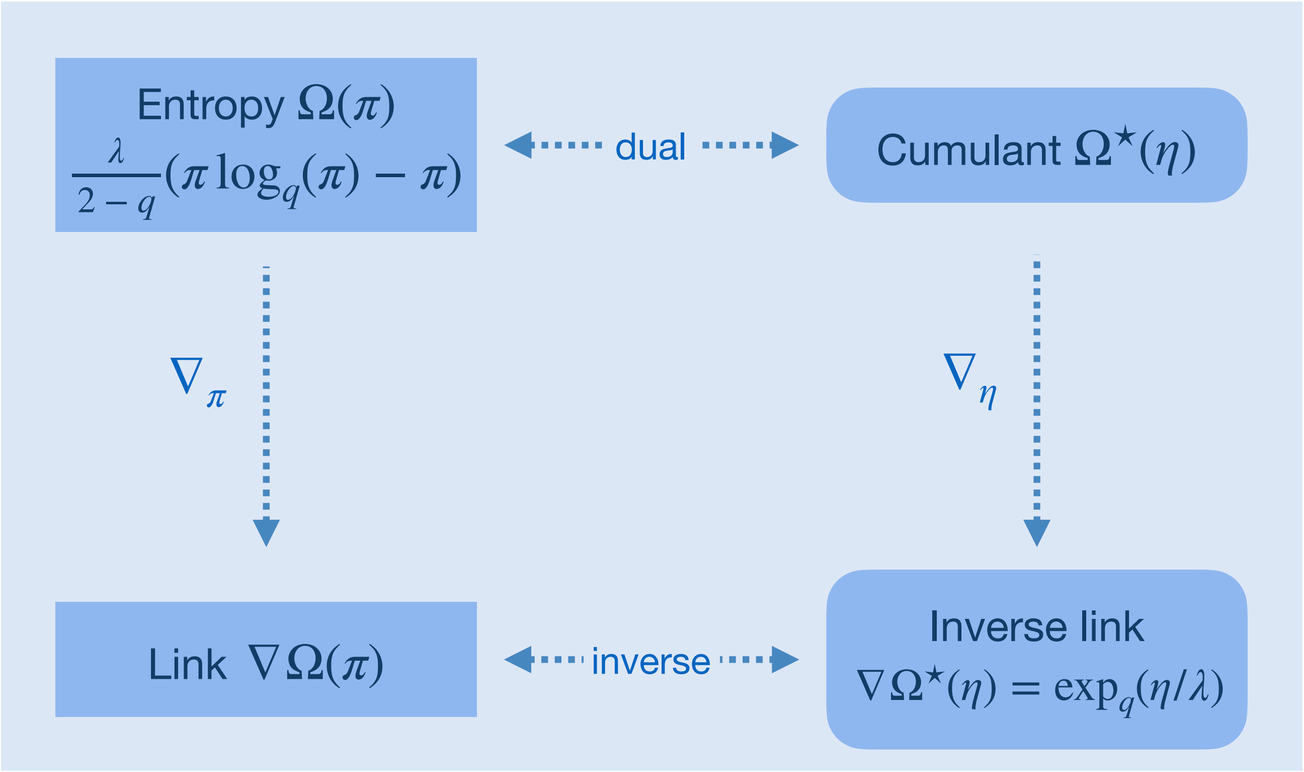
  - λ: regularizer strength
  - q: control support of kernel

- Solve dual optimization

$$\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^n} -\langle \mathbf{a}, \boldsymbol{\alpha}\rangle - \langle \mathbf{b}, \boldsymbol{\beta}\rangle - \sum_{i,j} \Omega^\star\big(-\mathbf{D}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j\big)$$

- Obtain primal solution via inverse link

$$\mathbf{\Pi}_{ij} = \exp_q\left(-\frac{\mathbf{D}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j}{\lambda}\right)$$



Convergence analysis

- Suppose dual optimization $\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^n} -\langle \mathbf{a},\boldsymbol{\alpha}\rangle - \langle \mathbf{b},\boldsymbol{\beta}\rangle - \sum_{i,j} \Omega^\star\big(-\mathbf{D}_{ij}-\boldsymbol{\alpha}_i-\boldsymbol{\beta}_j\big)$ via BFGS

Theorem. Under some conditions, the gradient at the point obtained by K-th iteration of BFGS is upper-bounded by

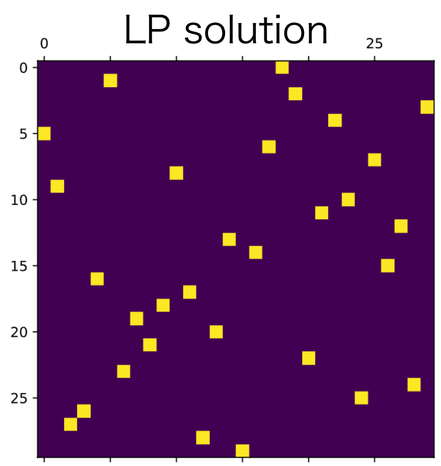
$$\sqrt{\frac{Cn\tau^q}{\lambda}}r^K,$$

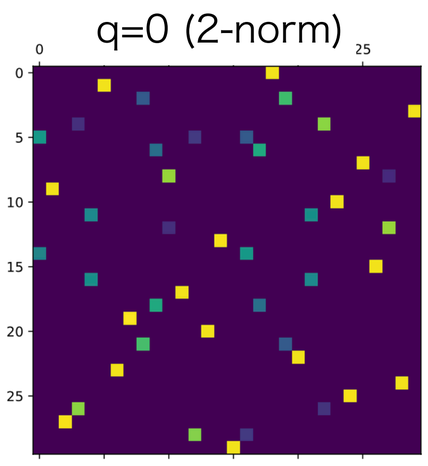
where the rate satisfies $0<r<1$ and τ is a constant in $(0,1)$.

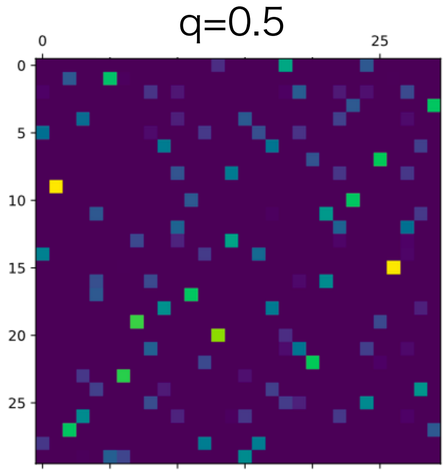
- Implication: larger q is beneficial in terms of convergence

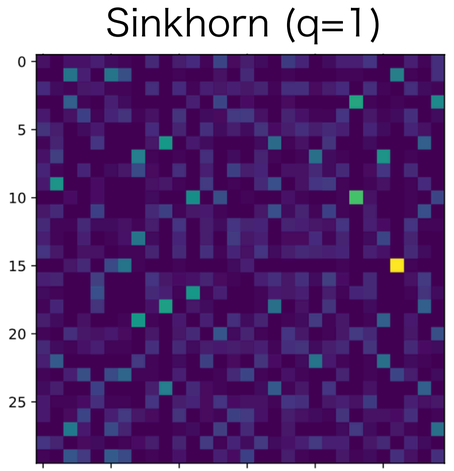
Experiments | Sparsity of transport matrices

- Solution becomes closer to LP solution as $q \to 0$







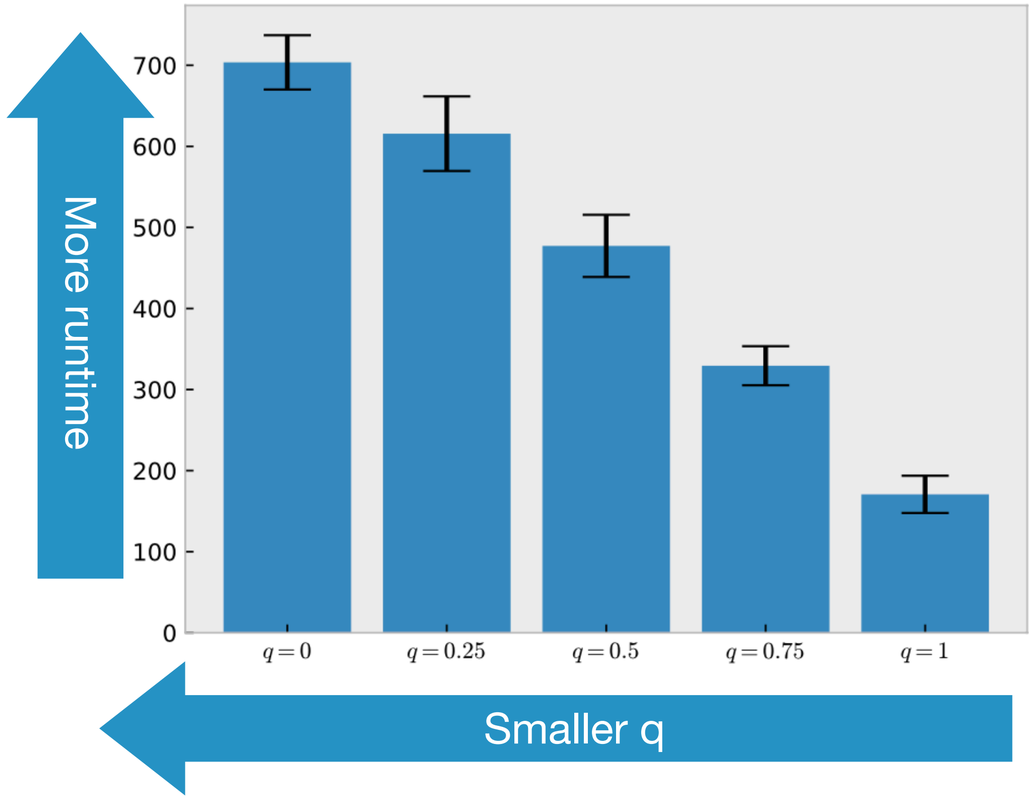


- Quantitative results
  - $q = 1$: fully dense (sparsity = 0)
  - $q < 1$: far more sparse than $q = 1$ (often sparsity > 0.6)

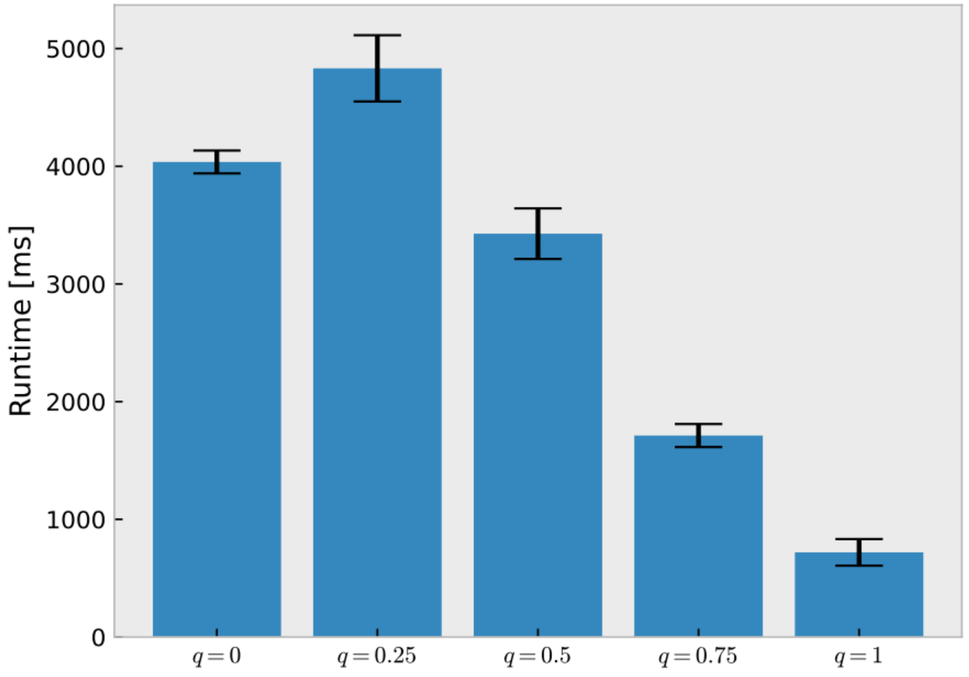
Experiments | Runtime comparison

- Smaller q often requires more runtime
  - Trade-off between sparsity and runtime

Dataset size = 100



Dataset size = 300



Comparison of q-entropy and Tsallis entropy

- Primal

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega\left(\mathbf{\Pi}_{ij}\right)$$

- q-entropy

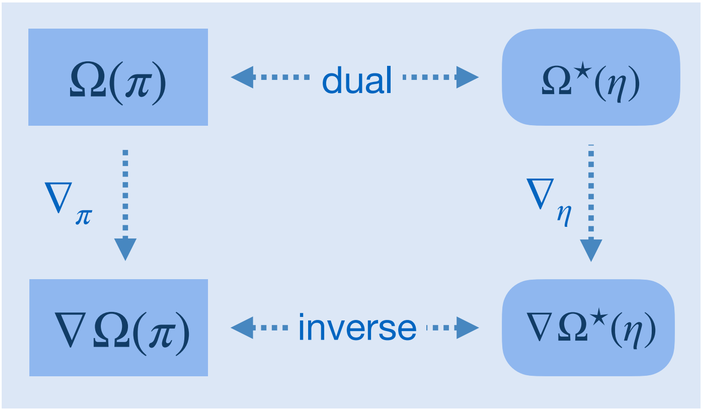
$$\Omega(\pi) = \frac{\lambda}{2-q}(\pi \log_q(\pi) - \pi) \quad \xrightarrow{\text{Inverse link}} \quad \nabla\Omega^\star(\eta) = \exp_q(\eta/\lambda)$$

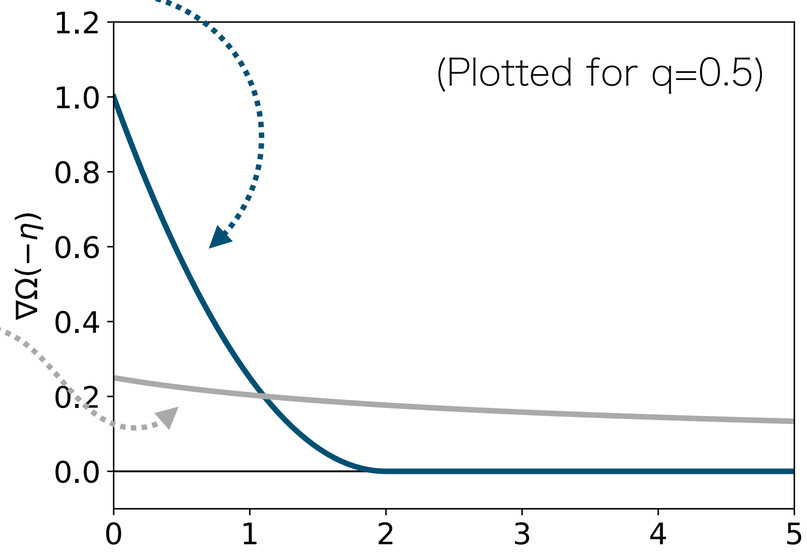
  - ❖ Inverse link is finitely supported for q < 1

- Tsallis entropy

$$T(\pi) = \lambda \pi^q \log_q(\pi) \quad \xrightarrow{\text{Inverse link}} \quad \nabla T^\star(\eta) = \frac{q^{1/(1-q)}}{\exp_q(-\eta/\lambda)}$$

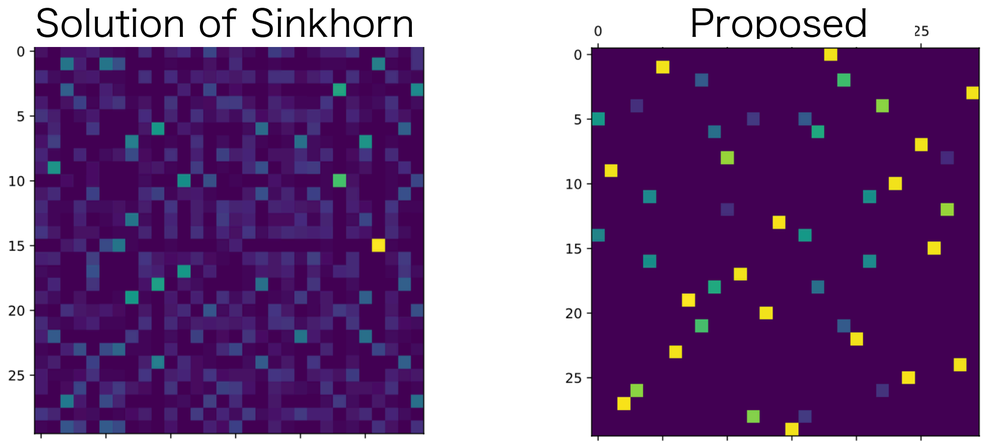
  - ❖ Inverse link is not finitely supported for any q





Summary

- What we do
  - Propose a new (entropic) regularizer to enhance sparsity of optimal transport solution



- Practically: trade-off between sparsity and runtime
- Theoretically: design of a good regularizer/entropy/divergence/etc.
  - Many machine learning problems have convex duality structures!