Revisiting Dimensionality Reduction with Optimal Transport: Within and Across Spaces

Hugues Van Assel

Cédric Vincent-Cuaz

Rémi Flamary

Nicolas Courty

Pascal Frossard

Titouan Vayer

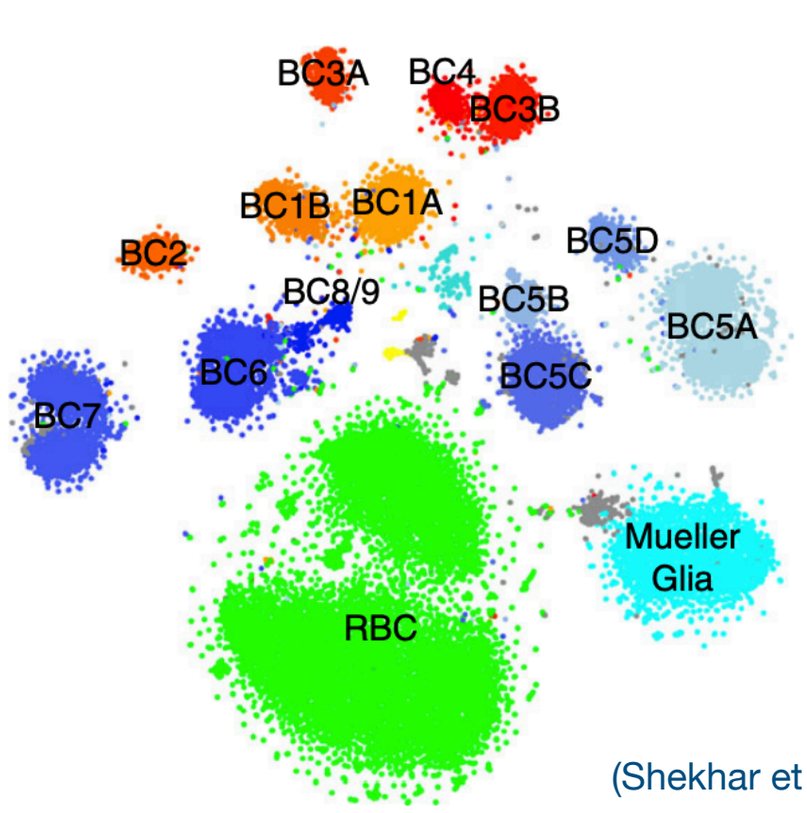
My talk

Overview of Dimensionality Reduction as Affinity Matching

Constructing Adaptive Symmetric Affinities

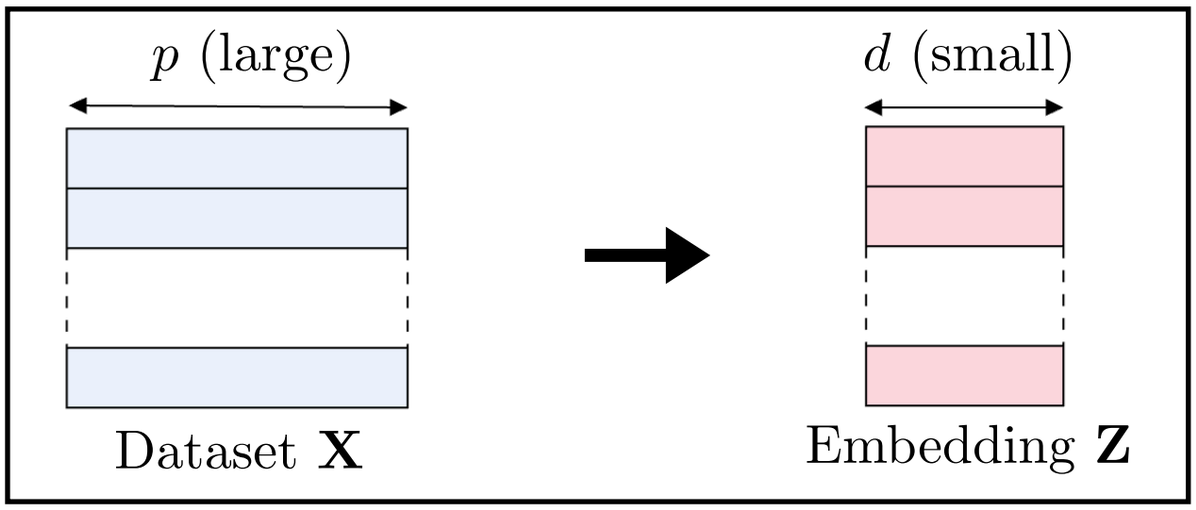
Distributional Reduction : a Framework to Embed Distributions

Dimension Reduction

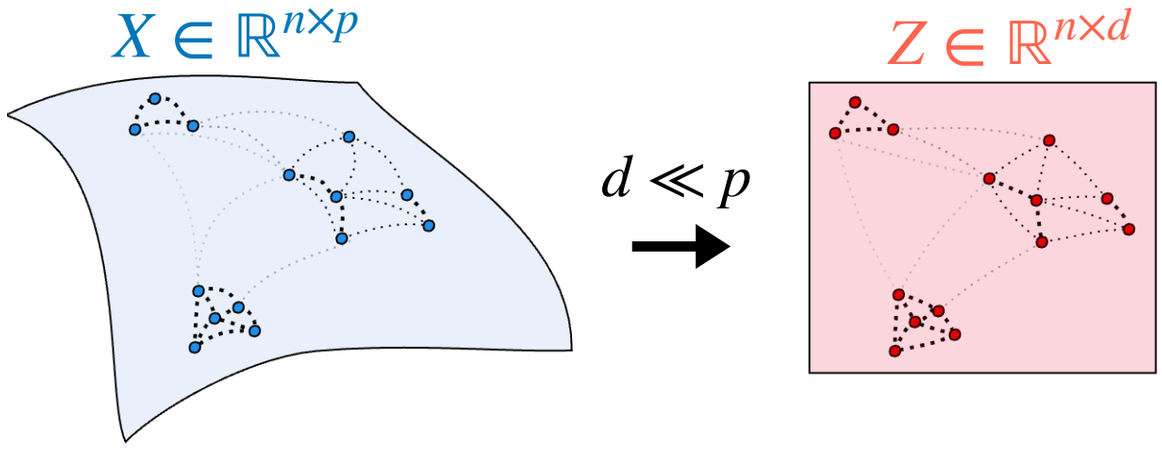


(Shekhar et al., 2016)

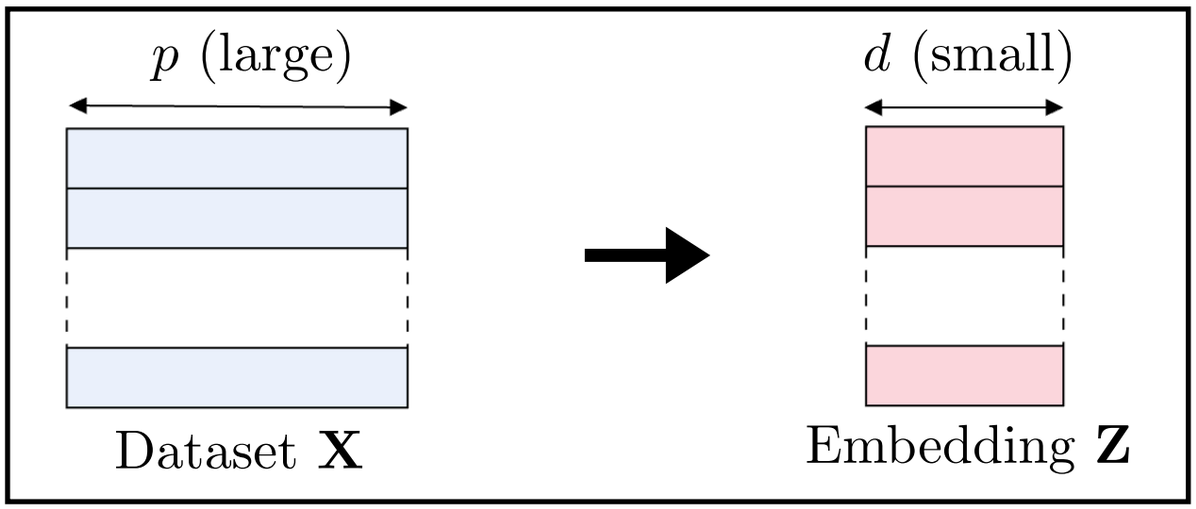
Dimension reduction



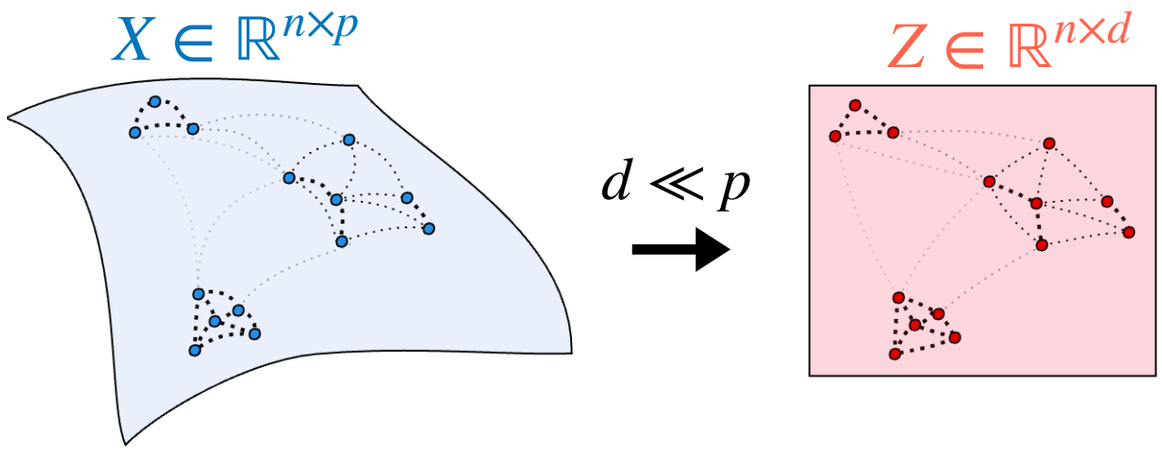
✦ **Preserving geometric structure**



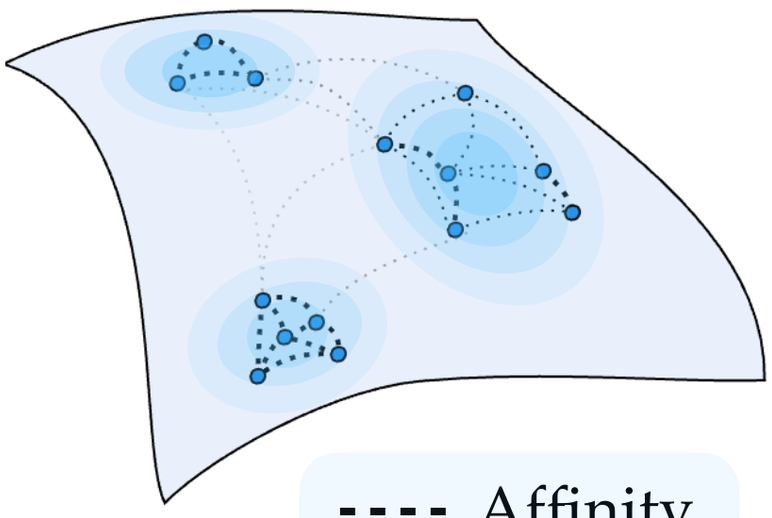
Dimension reduction

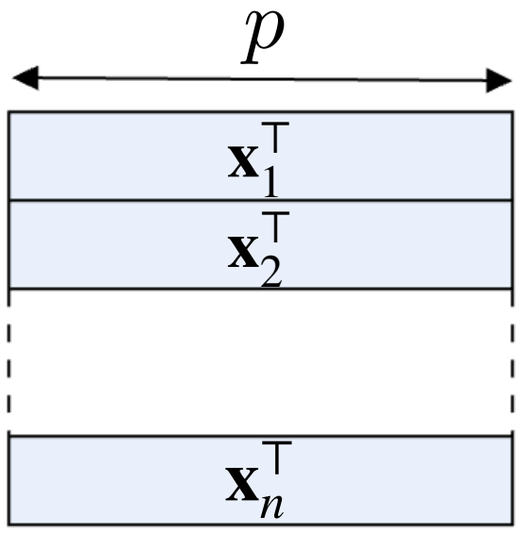


✦ Preserving geometric structure



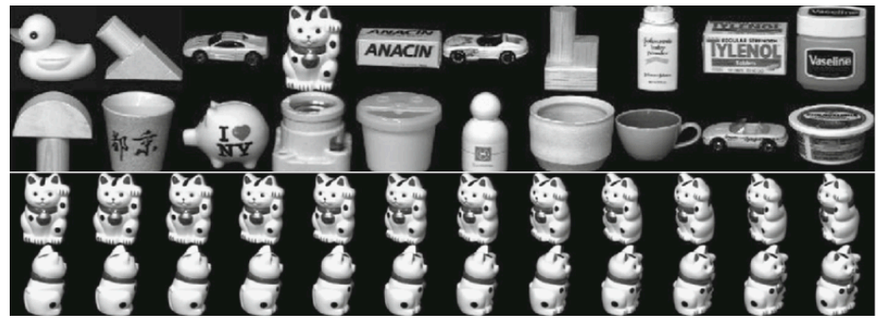
✦ Affinity Matrices



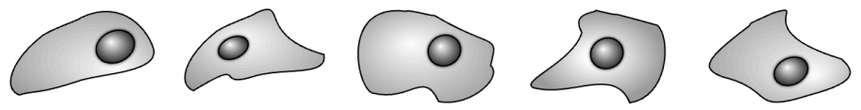


Images

COIL Dataset [Nene et al., 1996]

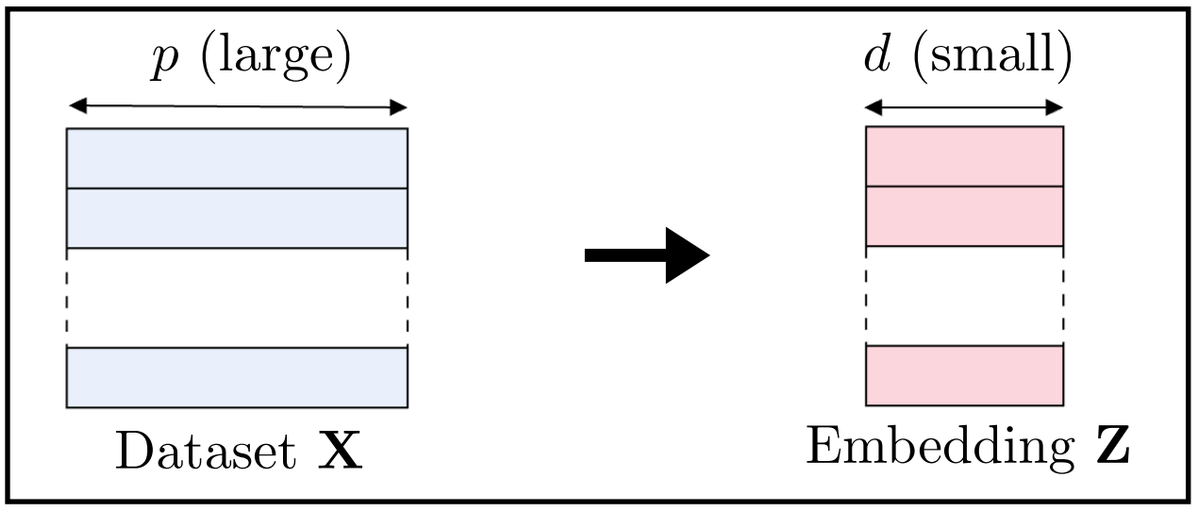


Cells

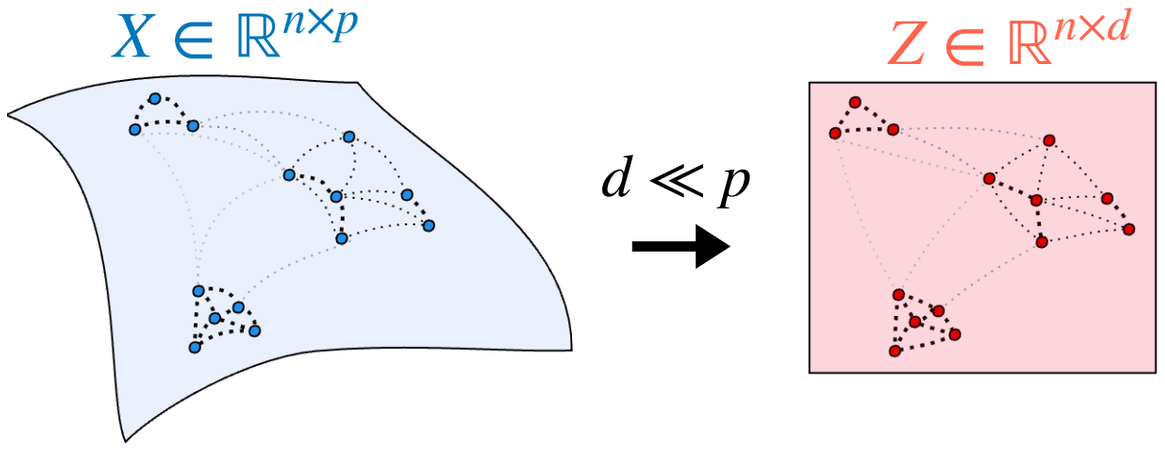


Coefficient (i, j) = similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

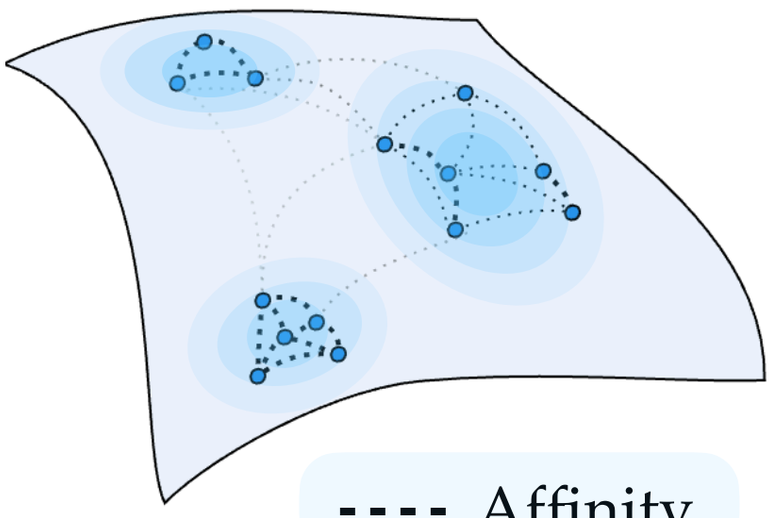
Dimension reduction

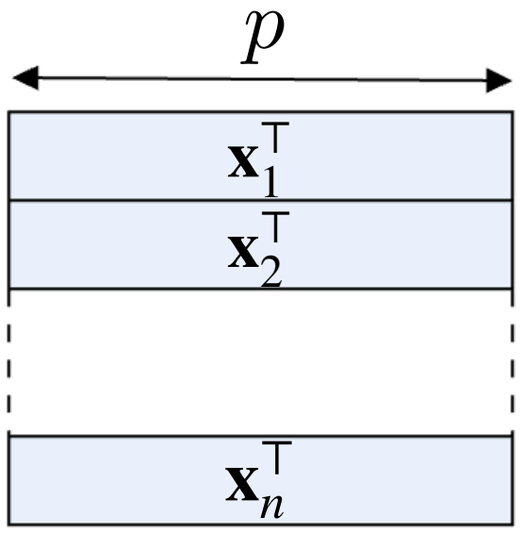


✦ **Preserving geometric structure**



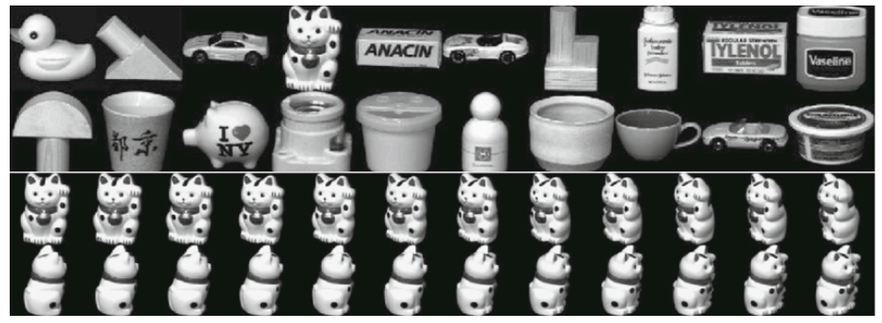
✦ **Affinity Matrices**



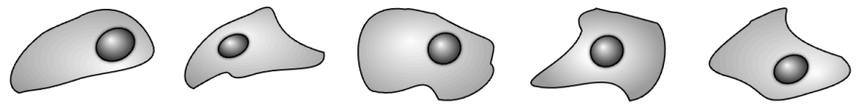


Images

COIL Dataset [Nene et al., 1996]

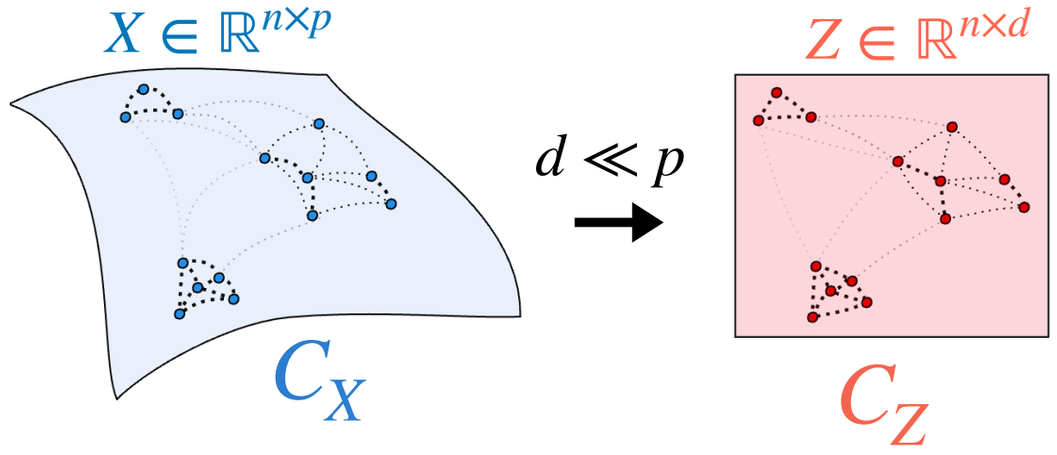


Cells



Requirement : a **meaningful metric** in input space !

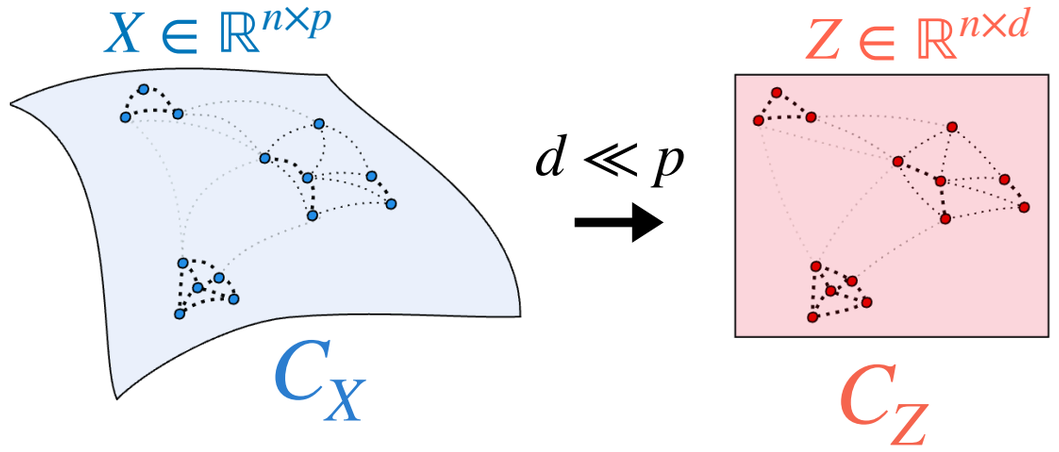
Dimension reduction



✦ **A general optimization problem**

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij} \, , \, [C_Z]_{ij} \right) \quad \text{for some loss} \quad L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

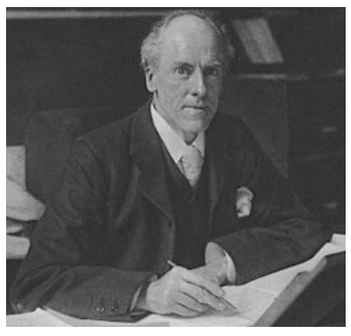
Dimension reduction



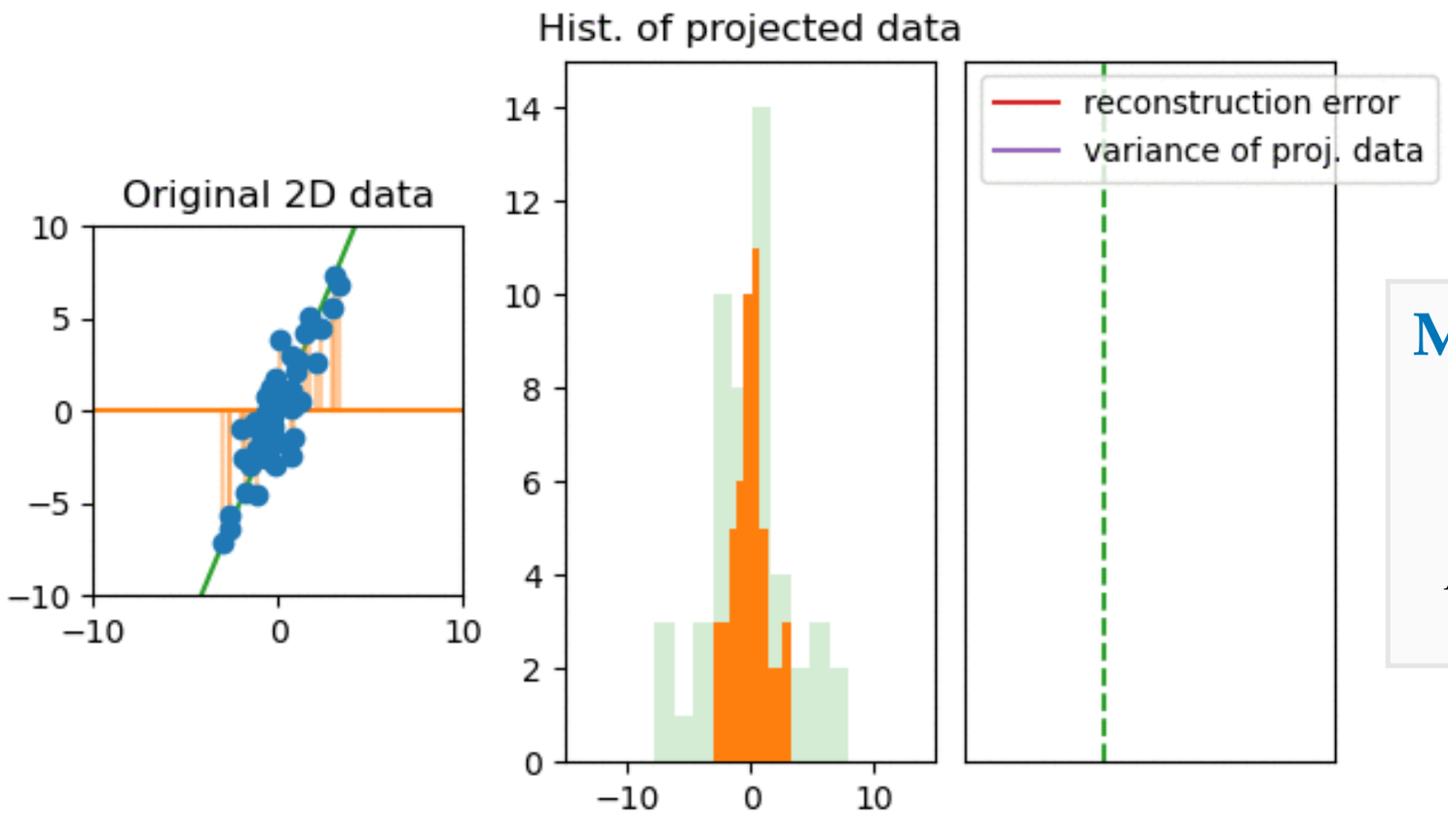
✦ A general optimization problem

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\,,\,[C_Z]_{ij}\right) \quad \text{for some loss} \quad L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

✦ Principal components analysis



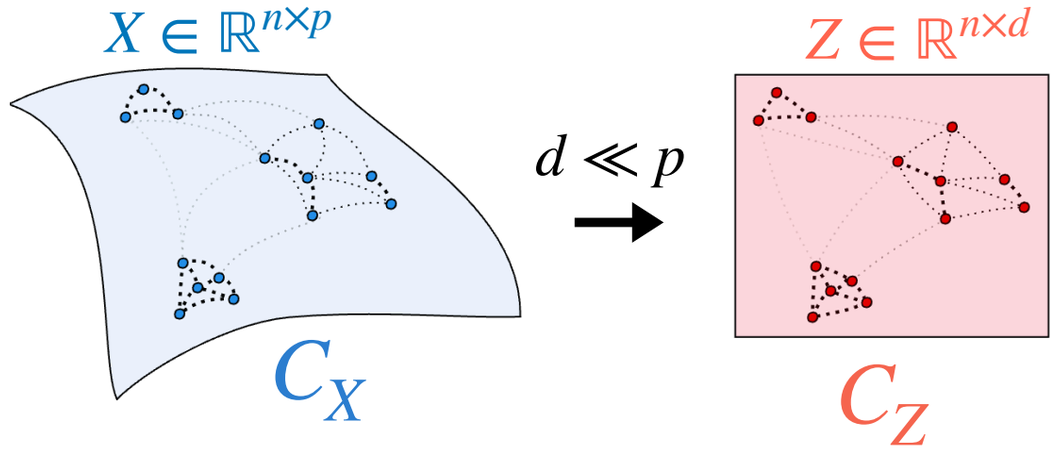
(Pearson, 1901)



Minimizing the reconstruction error

$$\min_{H:\dim(H)=d} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - P_H(\mathbf{x_i})\|_2^2$$

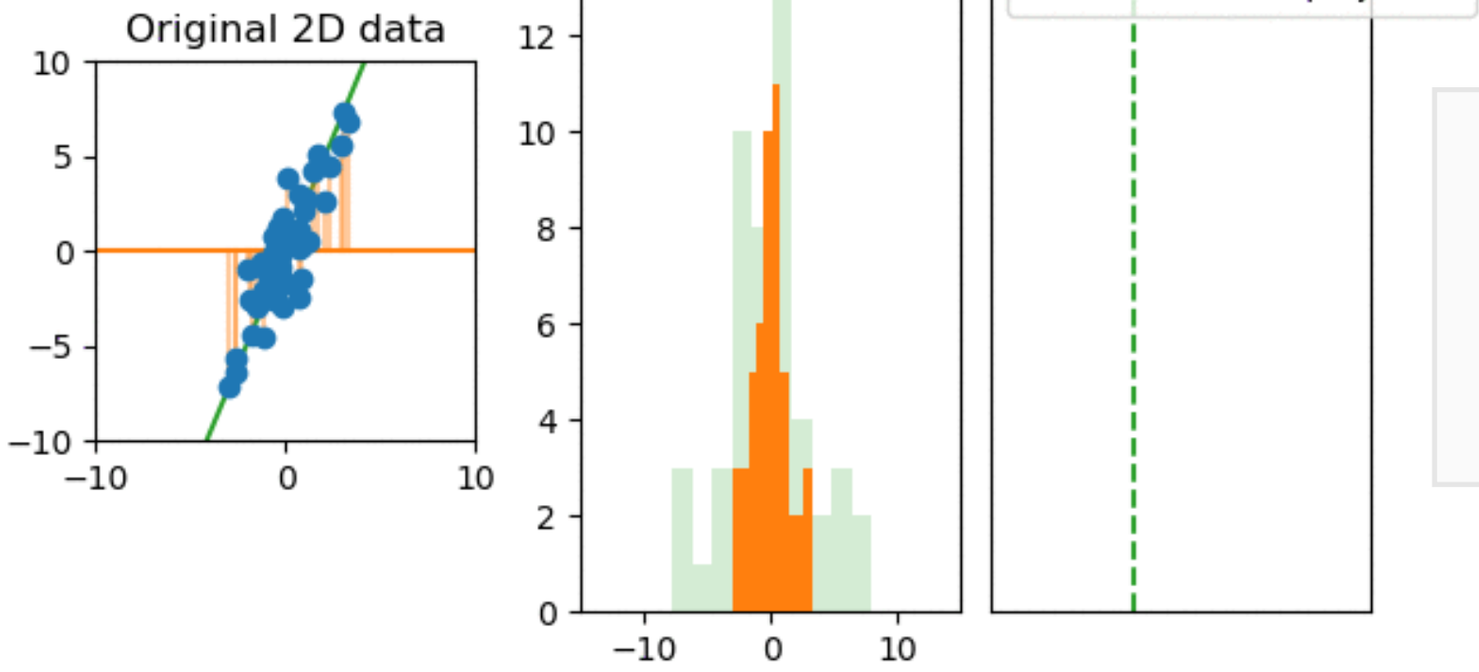
Dimension reduction

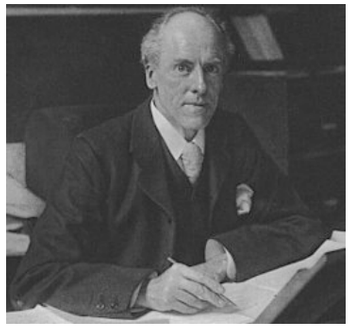


✦ A general optimization problem

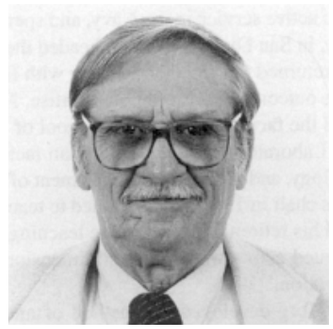
$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij} \, , \, [C_Z]_{ij} \right) \quad \text{for some loss} \quad L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

✦ Principal components analysis





(Pearson, 1901)



(Torgerson, 1958)

Preserving the inner products

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i=1}^{n} \left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle - \langle \mathbf{z}_i, \mathbf{z}_j \rangle \right)^2$$

$$Z \leftarrow \text{EVD}(\frac{1}{n} X X^\top)$$

Dimension reduction

✦ Spectral methods

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i=1}^{n} \left([C_X]_{ij} - \langle \mathbf{z}_i, \mathbf{z}_j \rangle \right)^2$$

Dimension reduction

✦ **Spectral methods**

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i=1}^{n} \left([C_X]_{ij} - \langle \mathbf{z}_i, \mathbf{z}_j \rangle \right)^2$$

$C_X \succeq 0$

solution

(Eckart & Young, 1936)

$$Z^\star = (\sqrt{\lambda_1}\mathbf{v}_1, \cdots, \sqrt{\lambda_d}\mathbf{v}_d)^\top$$

λ_i i-th **largest** eigenvalue of C_X with eigenvector $\mathbf{v}_i$

$$[C_X]_{ij} = \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$$

Dimension reduction

✦ **Spectral methods**

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i=1}^{n} \left([C_X]_{ij} - \langle \mathbf{z}_i, \mathbf{z}_j \rangle \right)^2$$

$C_X \succeq 0$

solution

(Eckart & Young, 1936)

$$Z^\star = (\sqrt{\lambda_1}\mathbf{v}_1, \cdots, \sqrt{\lambda_d}\mathbf{v}_d)^\top$$

λ_i i-th **largest** eigenvalue of C_X with eigenvector $\mathbf{v}_i$

✦ **Kernel PCA** $C_X \succeq 0$ $[C_X]_{ij} = \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$

(Schölkopf, 1997)

→ PCA: $C_X = XX^\top$ $(Z \leftarrow \mathrm{SVD}(X))$

Dimension reduction

✦ **Spectral methods**

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i=1}^{n} \left([C_X]_{ij} - \langle \mathbf{z}_i, \mathbf{z}_j \rangle \right)^2 \xrightarrow[\text{(Eckart \& Young, 1936)}]{C_X \succeq 0 \ \textbf{solution}} Z^\star = (\sqrt{\lambda_1}\mathbf{v}_1, \cdots, \sqrt{\lambda_d}\mathbf{v}_d)^\top$$

λ_i i-th **largest** eigenvalue of C_X with eigenvector $\mathbf{v}_i$

✦ **Kernel PCA** $C_X \succeq 0$ $\quad [C_X]_{ij} = \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$

(Schölkopf, 1997)

→ PCA: $C_X = XX^\top$ $(Z \leftarrow \text{SVD}(X))$

→ (classical) Multidimensional scaling: $C_X = -\frac{1}{2} H D_X H$

Dimension reduction

✦ **Spectral methods**

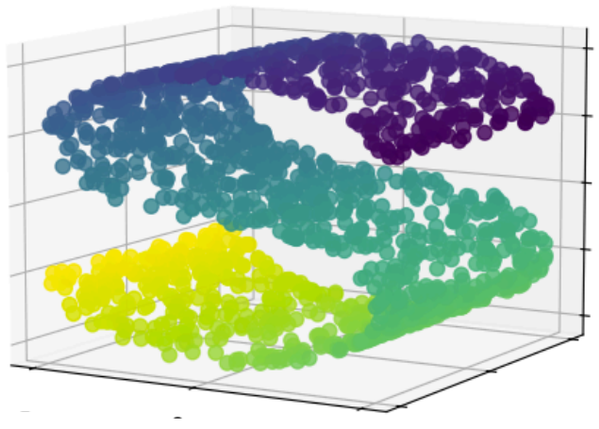
$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i=1}^{n} \left([C_X]_{ij} - \langle \mathbf{z}_i, \mathbf{z}_j \rangle \right)^2 \xrightarrow[\text{(Eckart \& Young, 1936)}]{C_X \succeq 0 \text{ solution}} Z^\star = (\sqrt{\lambda_1}\mathbf{v}_1, \cdots, \sqrt{\lambda_d}\mathbf{v}_d)^\top$$

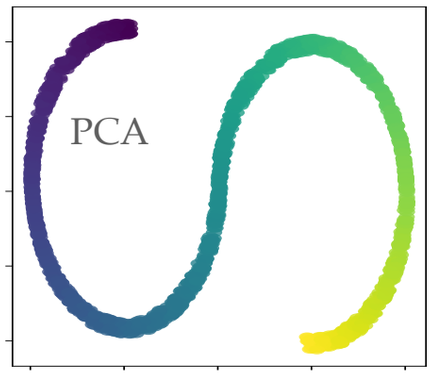
λ_i i-th **largest** eigenvalue of C_X with eigenvector $\mathbf{v}_i$

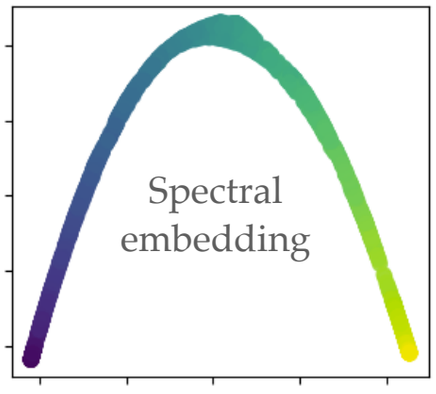
✦ **Kernel PCA** $C_X \succeq 0$ $\quad [C_X]_{ij} = \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$

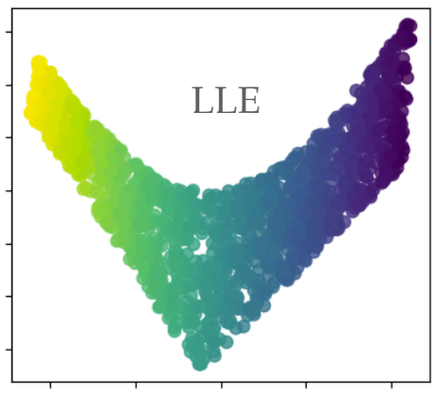
(Schölkopf, 1997)

- PCA: $C_X = XX^\top$ $(Z \leftarrow \text{SVD}(X))$
- (classical) Multidimensional scaling: $C_X = -\frac{1}{2} H D_X H$
- Laplacian Eigenmap (spectral embedding): $C_X = L_X^\dagger$ (Belkin & Niyogi, 2003)
- Locally Linear Embedding, Diffusion Map … (Roweis & Saul, 2000) (Coifman & Lafon, 2006)



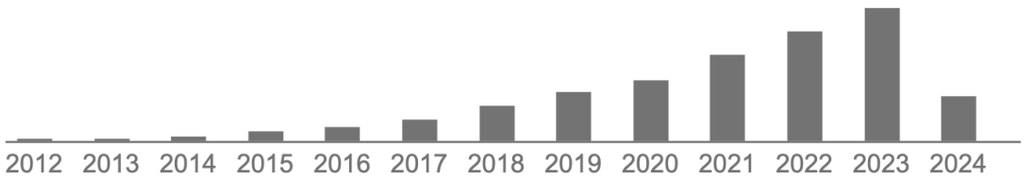


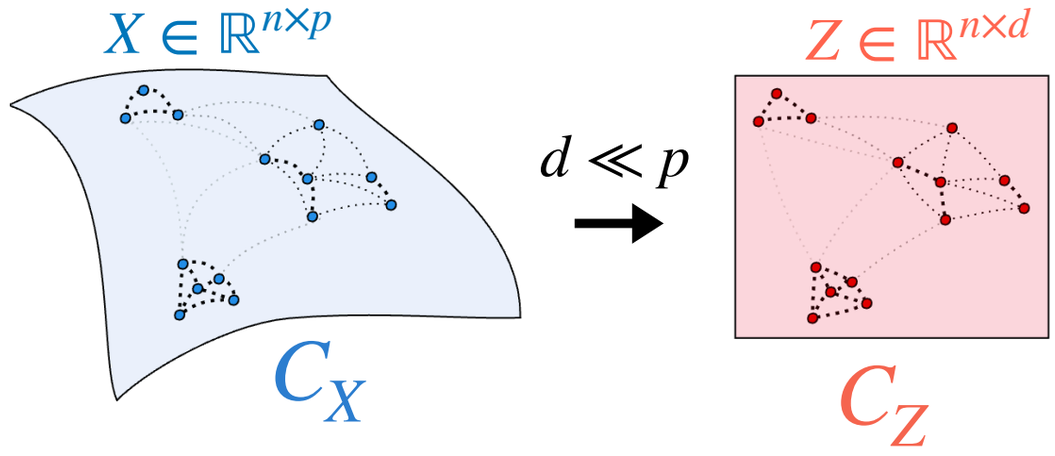




Dimension reduction

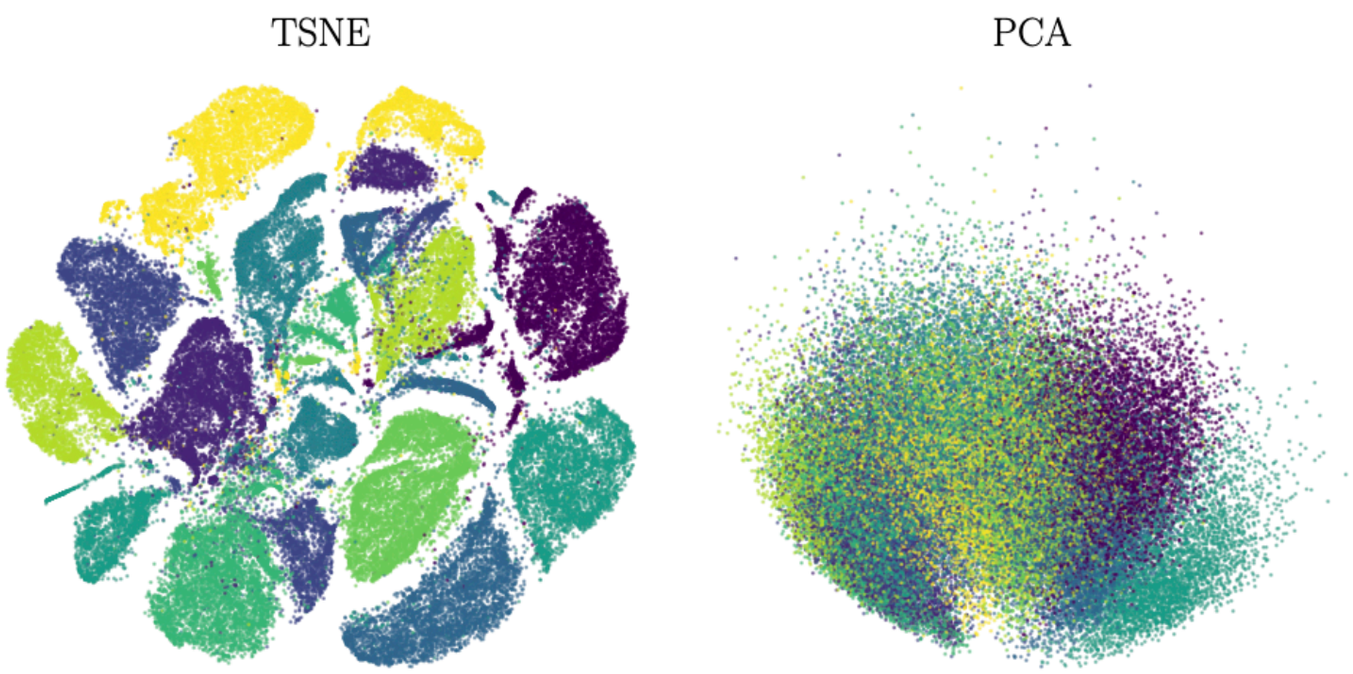
Total citations Cited by 36223





✦ Neighbor embedding methods

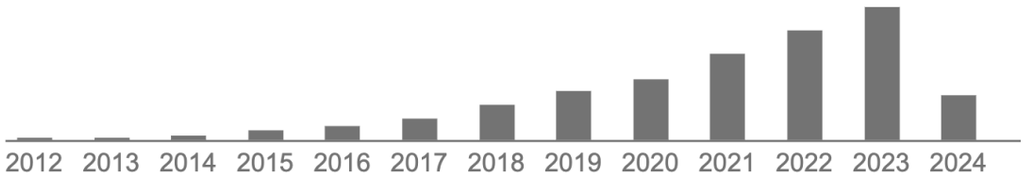
$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} \mathrm{KL}\left([C_X]_{ij} \, , \, [C_Z]_{ij} \right)$$

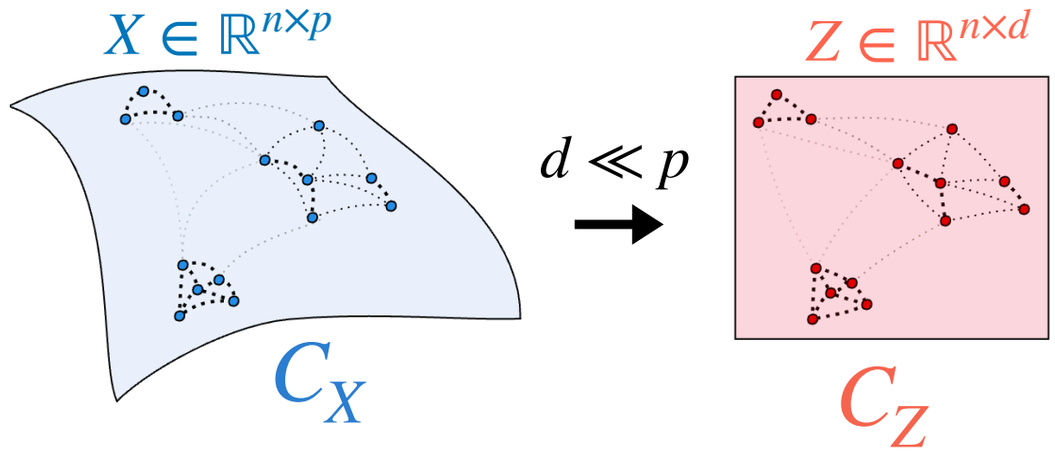


Kuzushiji - Japanese Letters

Dimension reduction

Total citations Cited by 36223



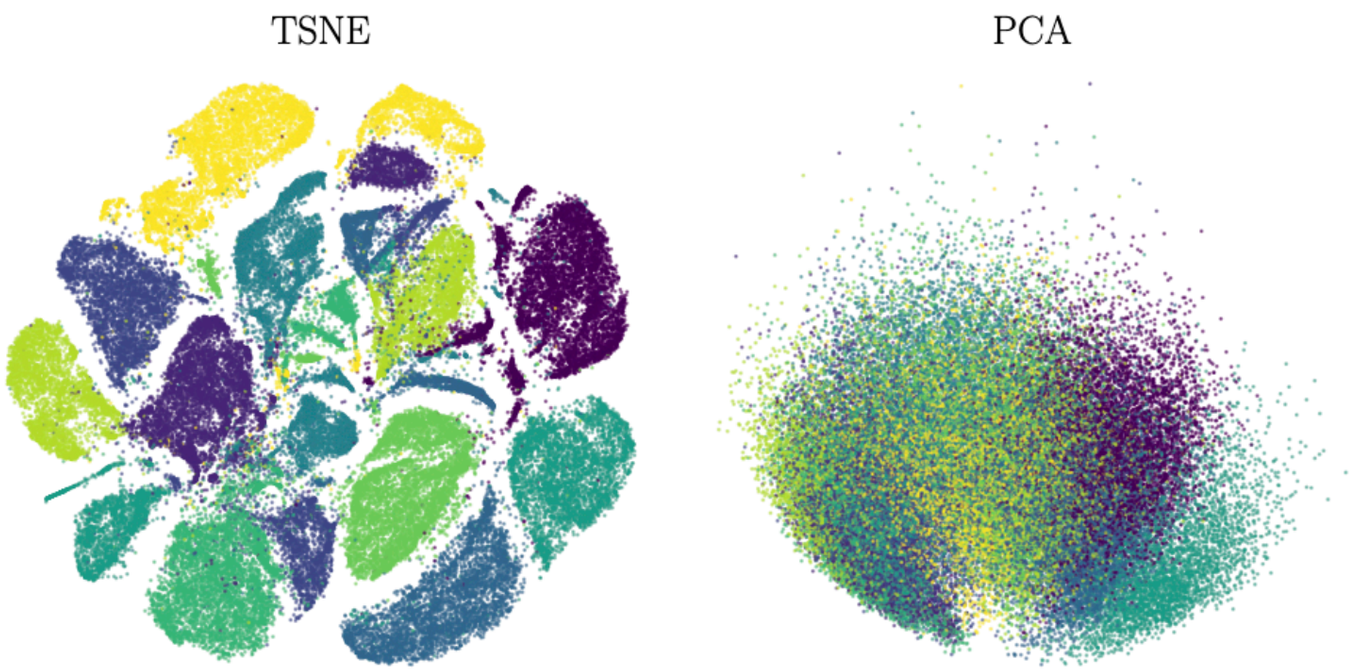


✦ Neighbor embedding methods

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} \mathrm{KL}\left([C_X]_{ij} \, , \, [C_Z]_{ij} \right)$$

✦ **Soft neighborhood graphs**

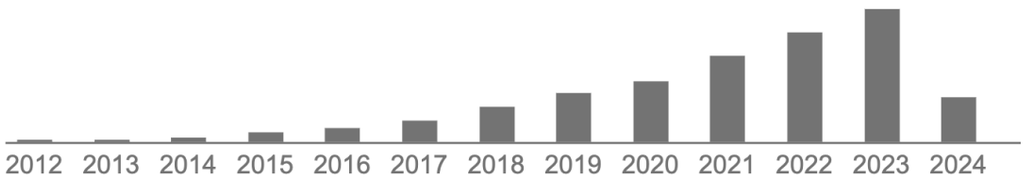
(We'll see in a few slides)

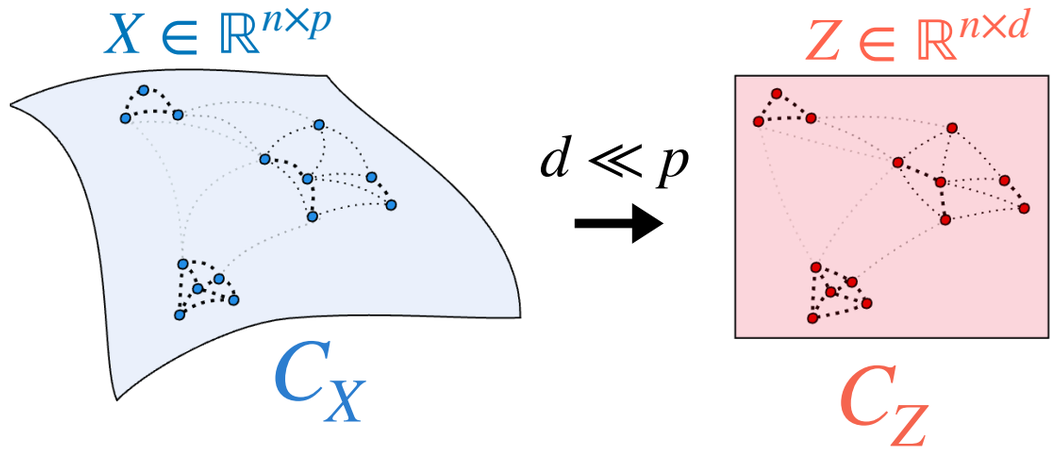


Kuzushiji - Japanese Letters

Dimension reduction

Total citations Cited by 36223





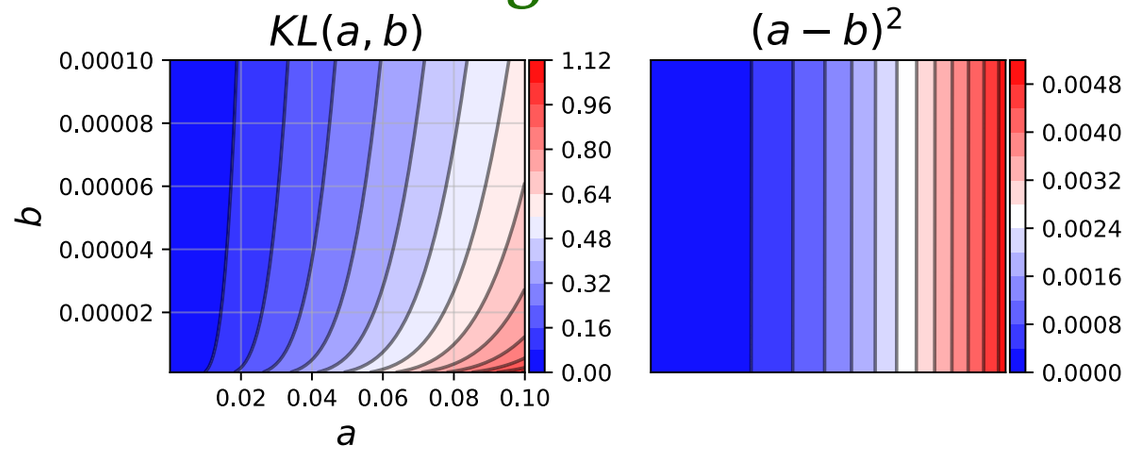
✦ Neighbor embedding methods

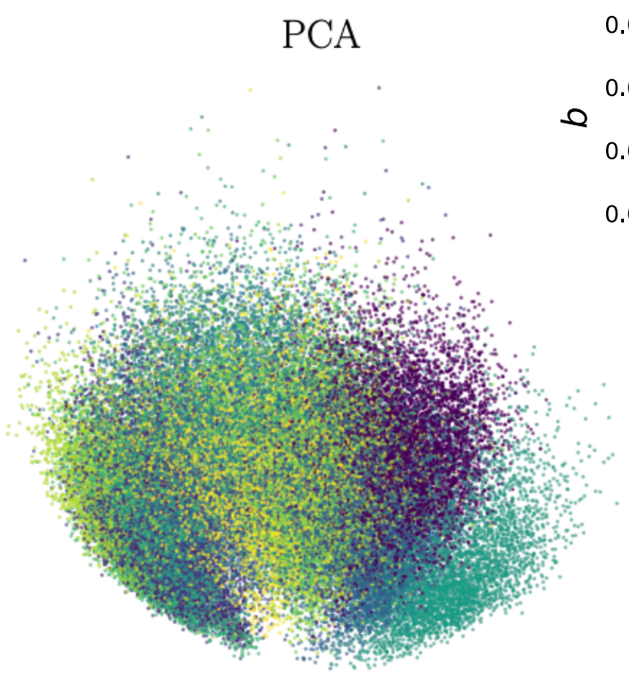
$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} \mathrm{KL}\left([C_X]_{ij} \, , \, [C_Z]_{ij} \right)$$

When $\sum_{i,j} [C_X]_{ij} = \sum_{i,j} [C_Z]_{ij}$ (same mass)

$$\sim \min_{Z \in \mathbb{R}^{n \times d}} \underbrace{\sum_{i,j=1}^{n} [C_X]_{ij} \, \log\left(\frac{[C_X]_{ij}}{[C_Z]_{ij}} \right)}_{\mathbf{KL}}$$

✦ Kullback-Leiber divergence

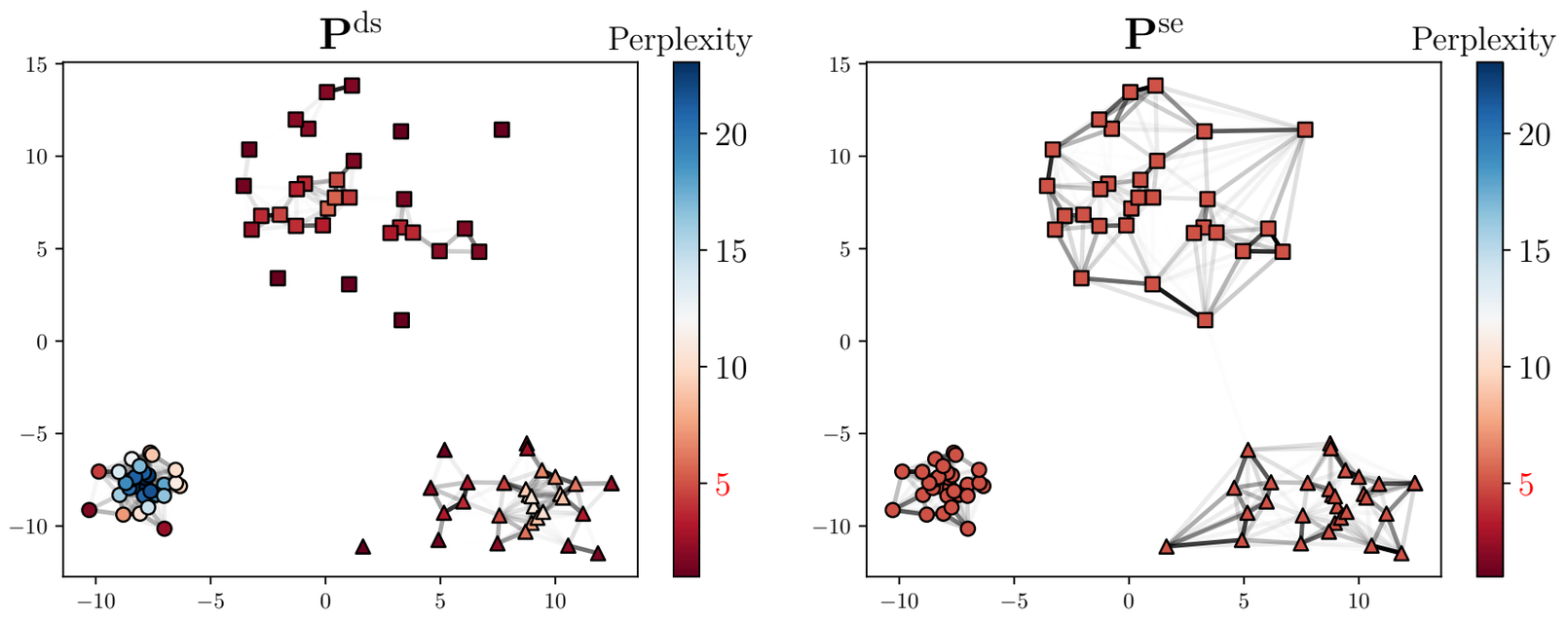




How to construct good affinities for neighbor embedding ?

Kuzushiji - Japanese Letters

Adaptive Affinities for DR



Gaussian Affinity (or Gibbs kernel)

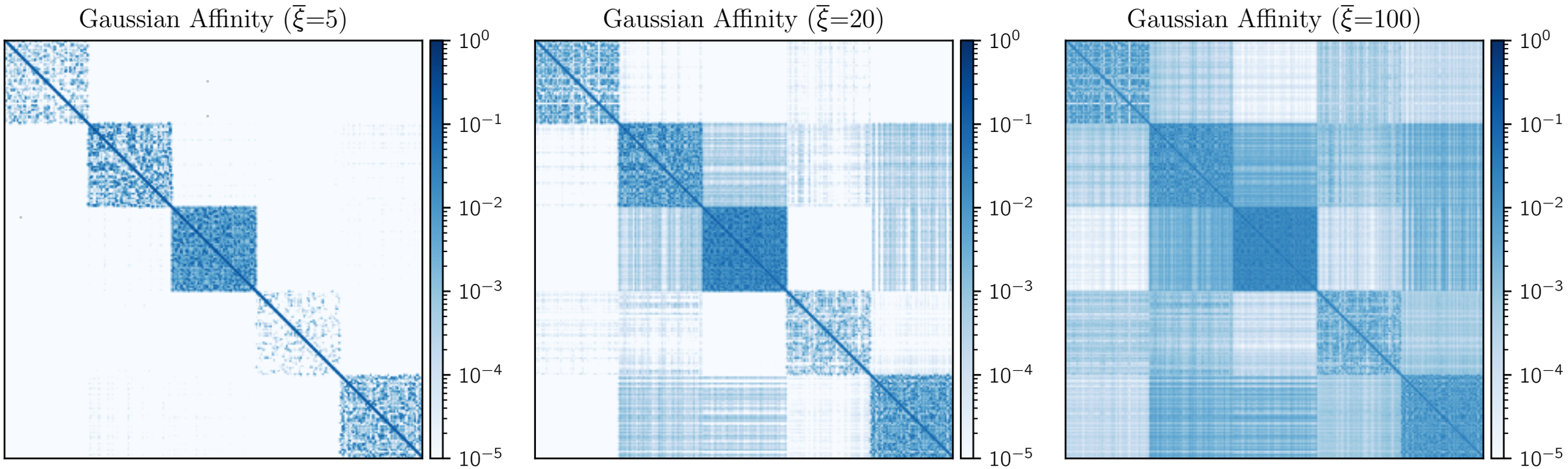


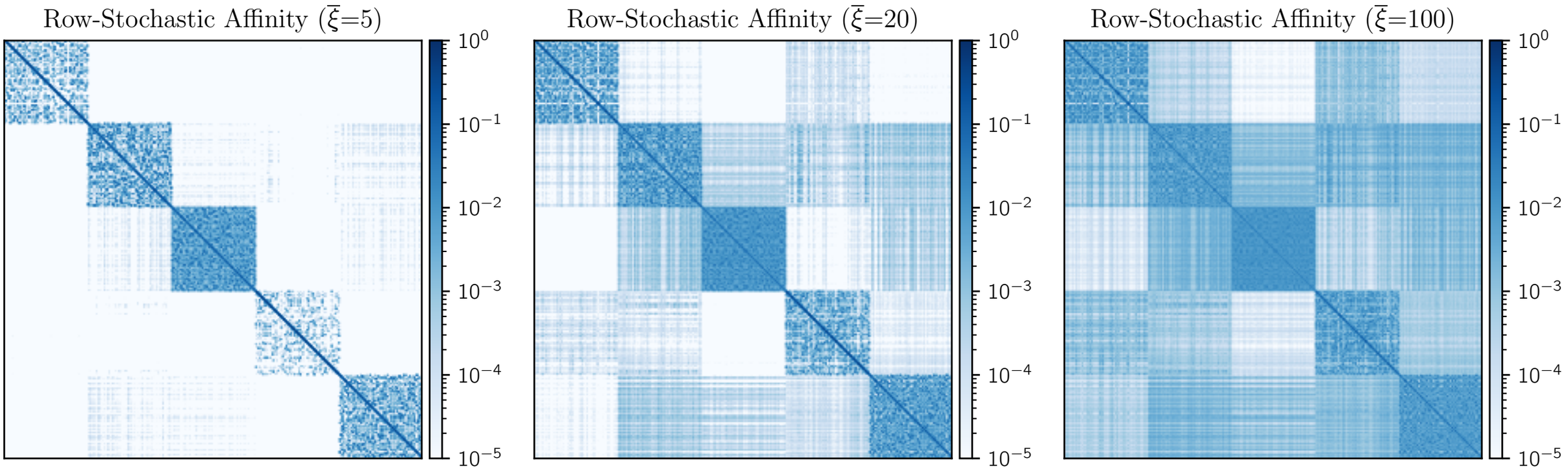
Fig : Affinity on 5 classes of the COIL Dataset [Nene et al., 1996]

Cost matrix: $\mathbf{C} \in \mathbb{R}_{+}^{n \times n}$ such that $\mathbf{C} = \mathbf{C}^{\top}$ and $C_{ij} = 0 \iff i = j$.

Example: $C_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$.

$$\text{Gibbs kernel} : \mathbf{K} = \exp(-\mathbf{C}/\sigma).$$

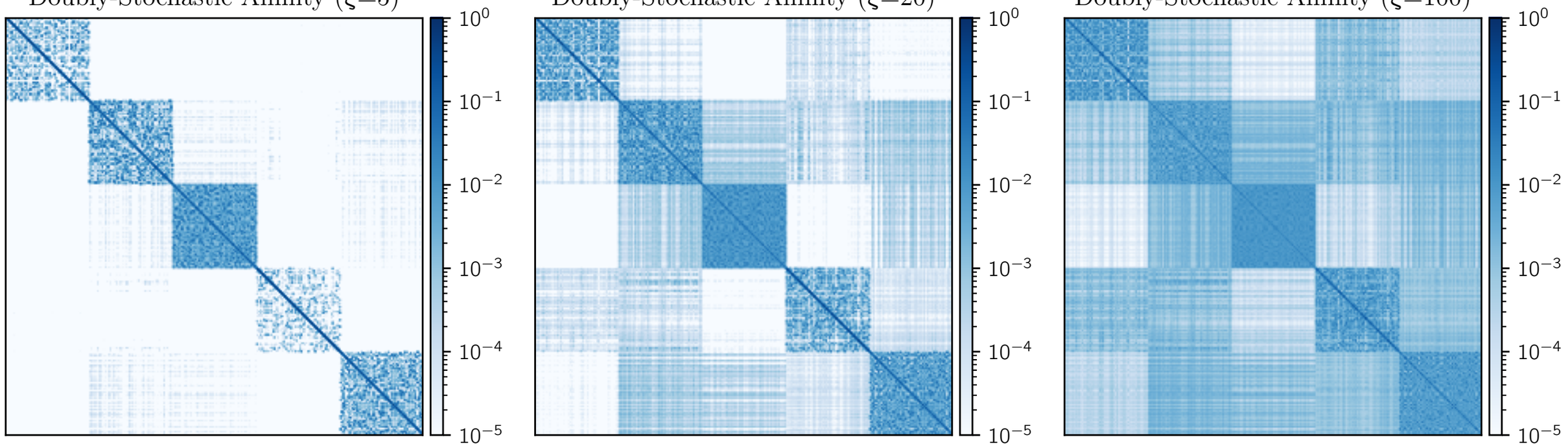
ℓ_1 Norm - Row Stochastic



$$\text{init } \mathbf{K} = \exp(-\mathbf{C}/\sigma)$$

$$\mathbf{K} \leftarrow \text{diag}(\mathbf{K1})^{-1}\mathbf{K} \qquad \# \text{ normalize rows}$$

ℓ_1 Norm - Doubly Stochastic



Sinkhorn Algorithm init $\mathbf{K} = \exp(-\mathbf{C}/\sigma)$

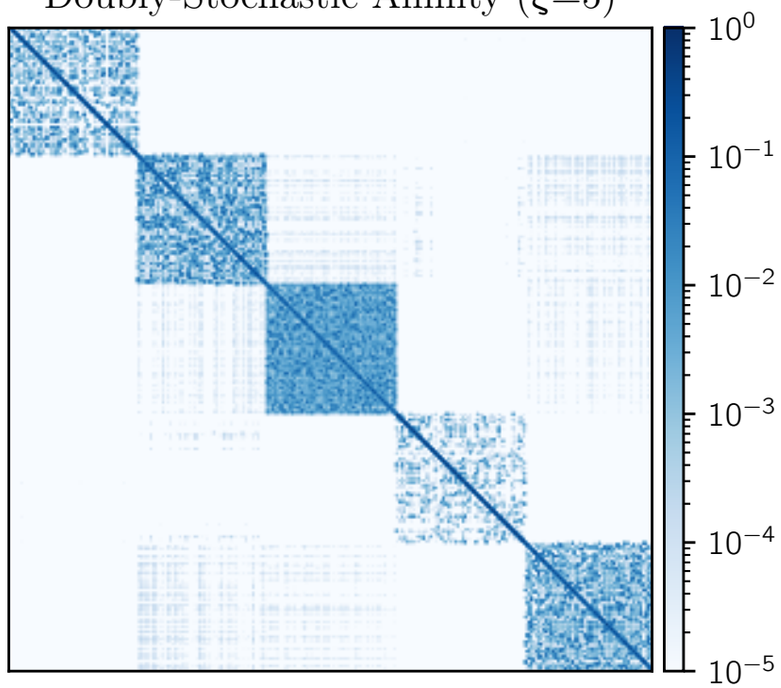
While not converged:

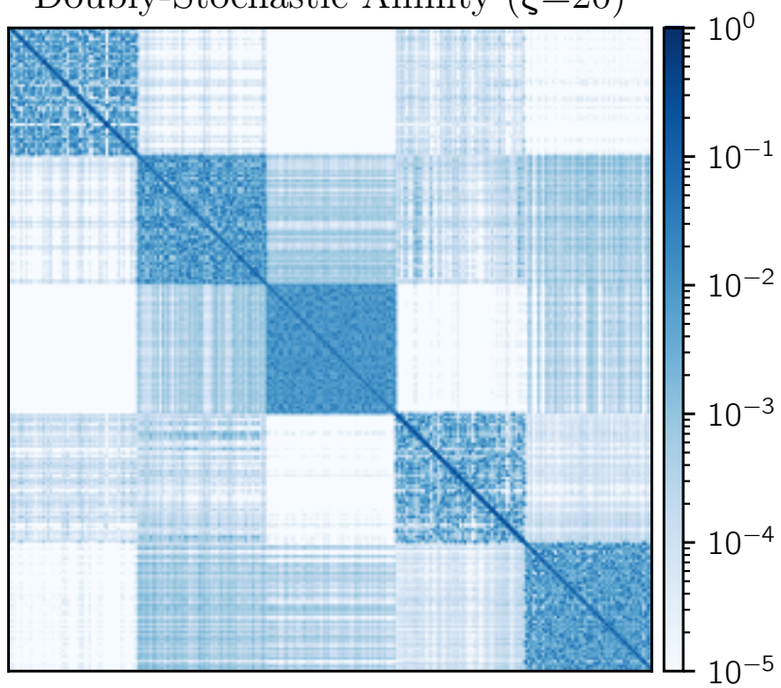
$$\mathbf{K} \leftarrow \mathrm{diag}(\mathbf{K1})^{-1}\mathbf{K} \qquad \# \text{ normalize rows}$$

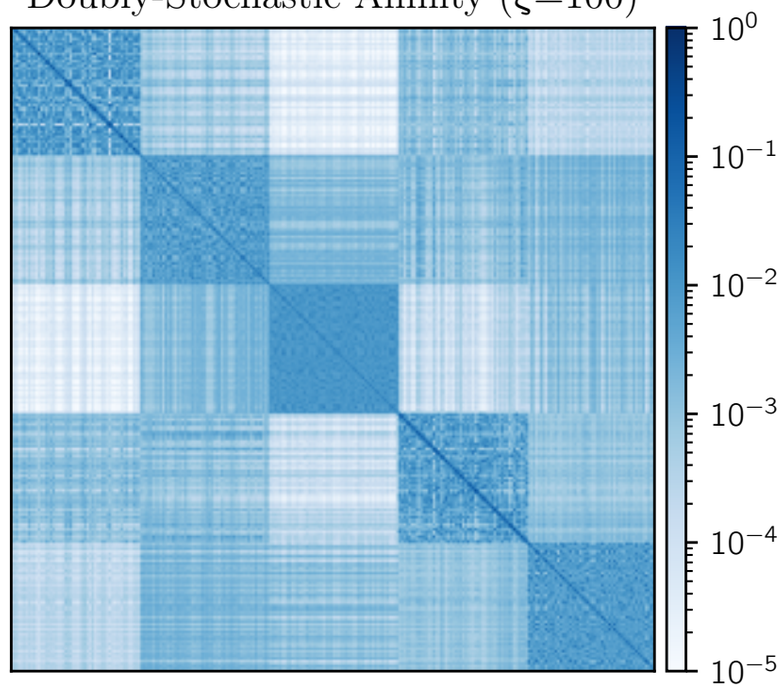
$$\mathbf{K} \leftarrow \mathbf{K}\,\mathrm{diag}(\mathbf{1K})^{-1} \qquad \# \text{ normalize columns}$$

converges to $\longrightarrow$ $\mathcal{DS} = \{\mathbf{P} \text{ s.t. } \mathbf{P1} = \mathbf{P}^\top\mathbf{1} = \mathbf{1}\}$.

ℓ_1 Norm - Doubly Stochastic







Sinkhorn Algorithm

$\text{init } \mathbf{K} = \exp(-\mathbf{C}/\sigma)$

While not converged:

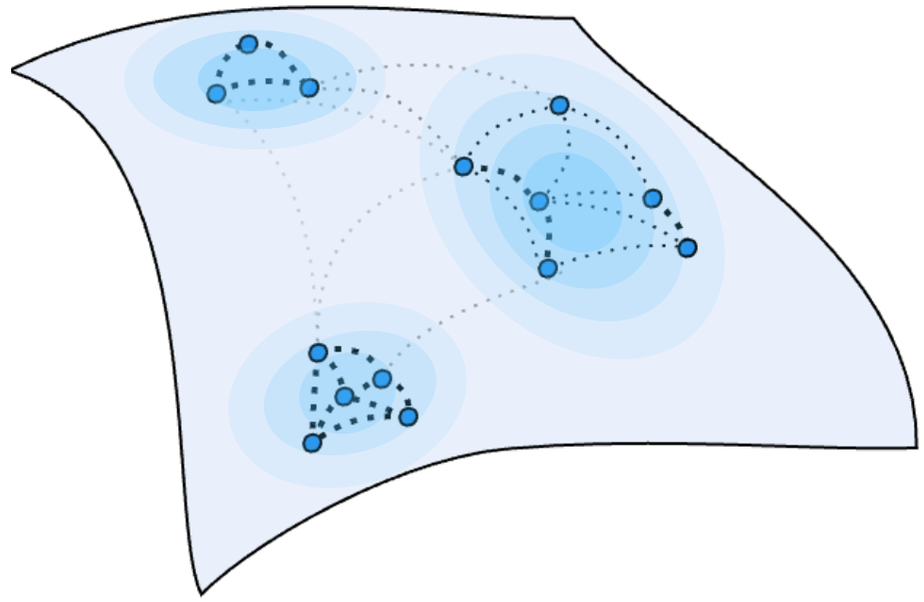
$\mathbf{K} \leftarrow \text{diag}(\mathbf{K1})^{-1}\mathbf{K}$ # normalize rows

$\mathbf{K} \leftarrow \mathbf{K}\,\text{diag}(\mathbf{1K})^{-1}$ # normalize columns

converges to $\longrightarrow$ $\mathcal{DS} = \{\mathbf{P} \text{ s.t. } \mathbf{P1} = \mathbf{P}^\top\mathbf{1} = \mathbf{1}\}$.

- Spectral clustering (Zass & Sashua, 2006)
- Transformers (Sander & al., 2021)
- Dim. Reduction (Yu & al., 2016)

Entropic Affinity



Data has **varying noise levels**.

We can control the entropy in each point with **adaptive bandwidths**.

Definition [Hinton, Roweis 2002]

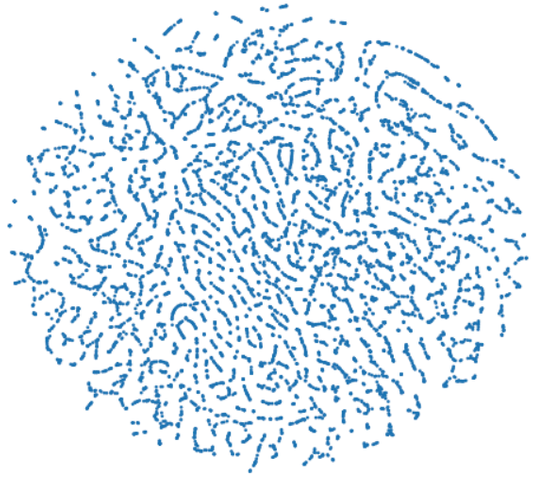
$$\forall i, \forall j,\ P^{\mathrm{e}}_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon^\star_i\right)}{\sum_\ell \exp\left(-C_{i\ell}/\varepsilon^\star_i\right)}$$

$$\text{with } \varepsilon^\star_i \in \mathbb{R}^*_+ \text{ s.t. } \mathrm{H}(\mathbf{P}^{\mathrm{e}}_{i:}) = \log \xi + 1\,.$$

$\mathrm{H}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \log \boldsymbol{p} - \mathbf{1} \rangle$ is the Shannon entropy.

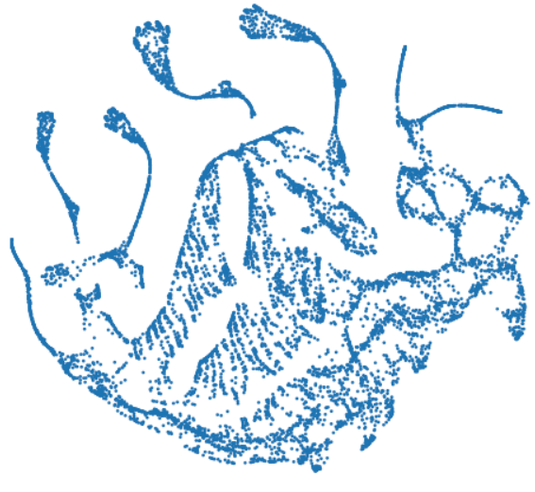
$\xi \in [\![1, n]\!]$ is the **perplexity** parameter.

Perplexity = 5

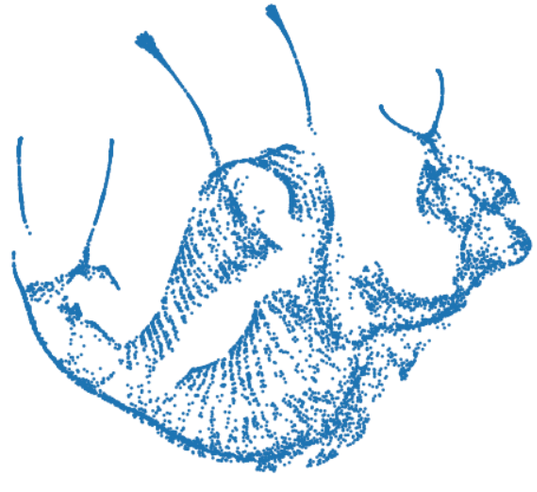


Perplexity = 30

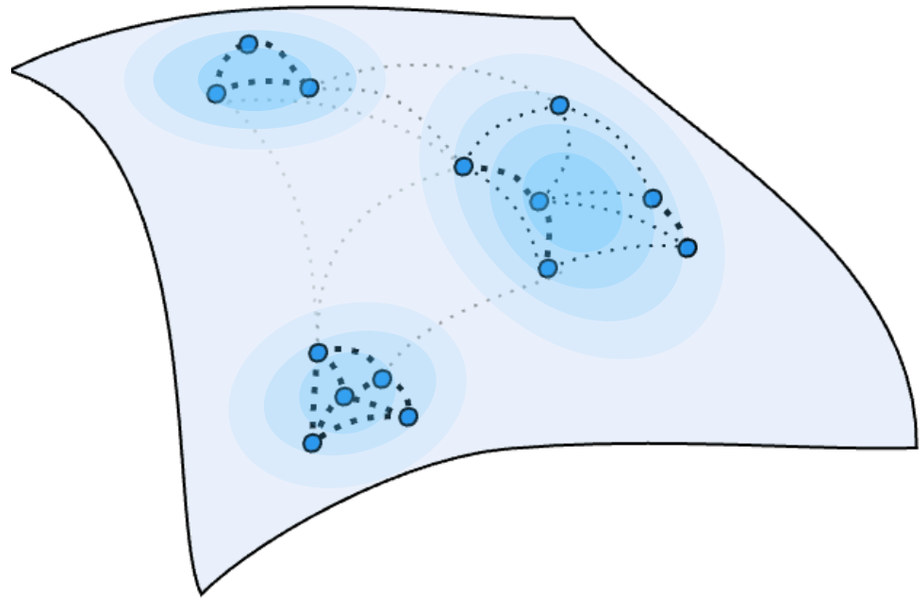
Perplexity = 100



Perplexity = 300



Entropic Affinity



Data has **varying noise levels**.

We can control the entropy in each point with **adaptive bandwidths**.

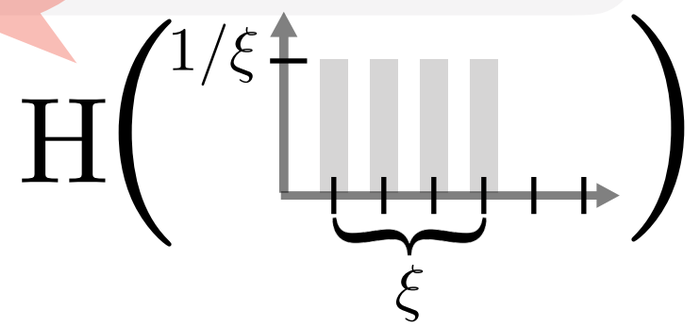
Definition [Hinton, Roweis 2002]

$$\forall i, \forall j,\ P^{\mathrm{e}}_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon_i^\star\right)}{\sum_\ell \exp\left(-C_{i\ell}/\varepsilon_i^\star\right)}$$

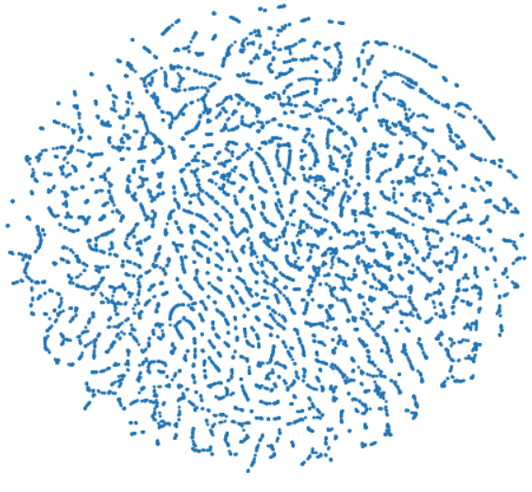
$$\text{with } \varepsilon_i^\star \in \mathbb{R}_+^* \text{ s.t. } \mathrm{H}(\mathbf{P}^{\mathrm{e}}_{i:}) = \log \xi + 1\,.$$

$\mathrm{H}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \log \boldsymbol{p} - \mathbf{1} \rangle$ is the Shannon entropy.

$\xi \in [\![1, n]\!]$ is the **perplexity** parameter.

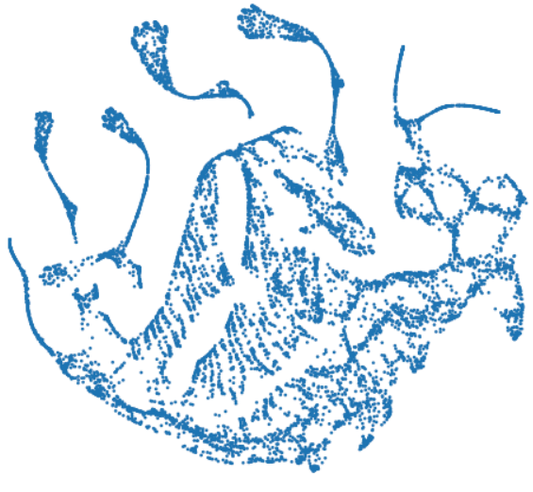


Perplexity = 5

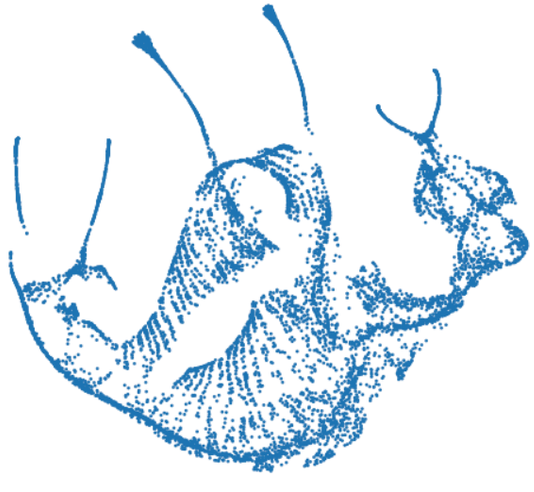


Perplexity = 30

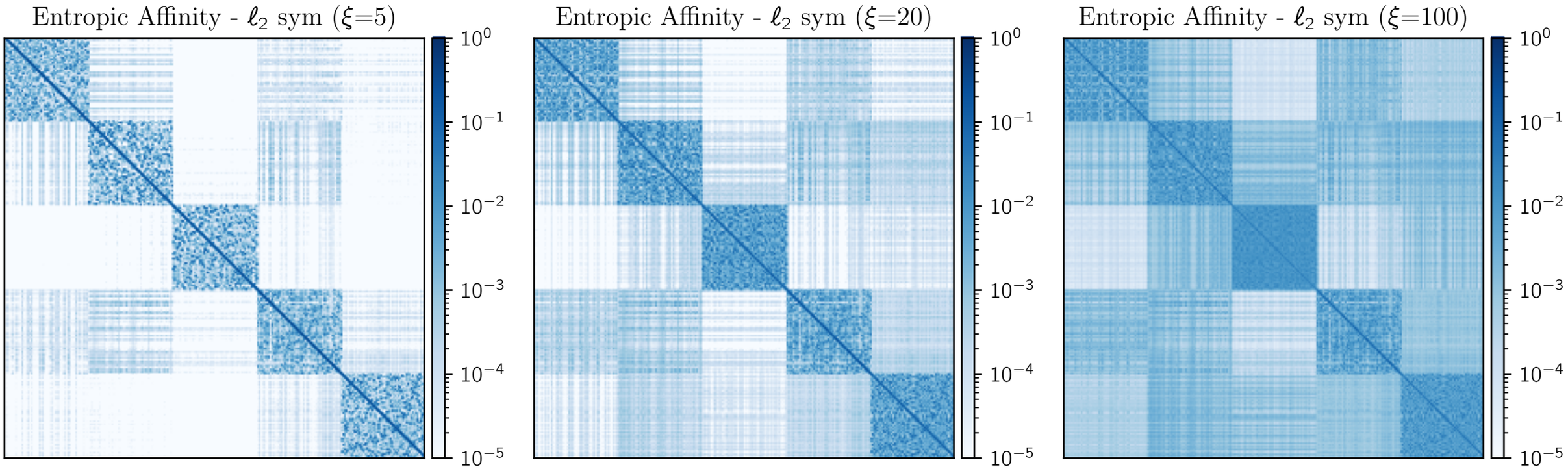
Perplexity = 100



Perplexity = 300



Entropic Affinity



Definition [Hinton, Roweis 2002]

$$\forall i, \forall j,\ P^{\mathrm{e}}_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon^\star_i\right)}{\sum_\ell \exp\left(-C_{i\ell}/\varepsilon^\star_i\right)}$$

$$\text{with } \varepsilon^\star_i \in \mathbb{R}^*_+ \text{ s.t. } \mathrm{H}(\mathbf{P}^{\mathrm{e}}_{i:}) = \log\xi + 1\,.$$

$\mathbf{P}^{\mathrm{e}}$ is **not symmetric**.

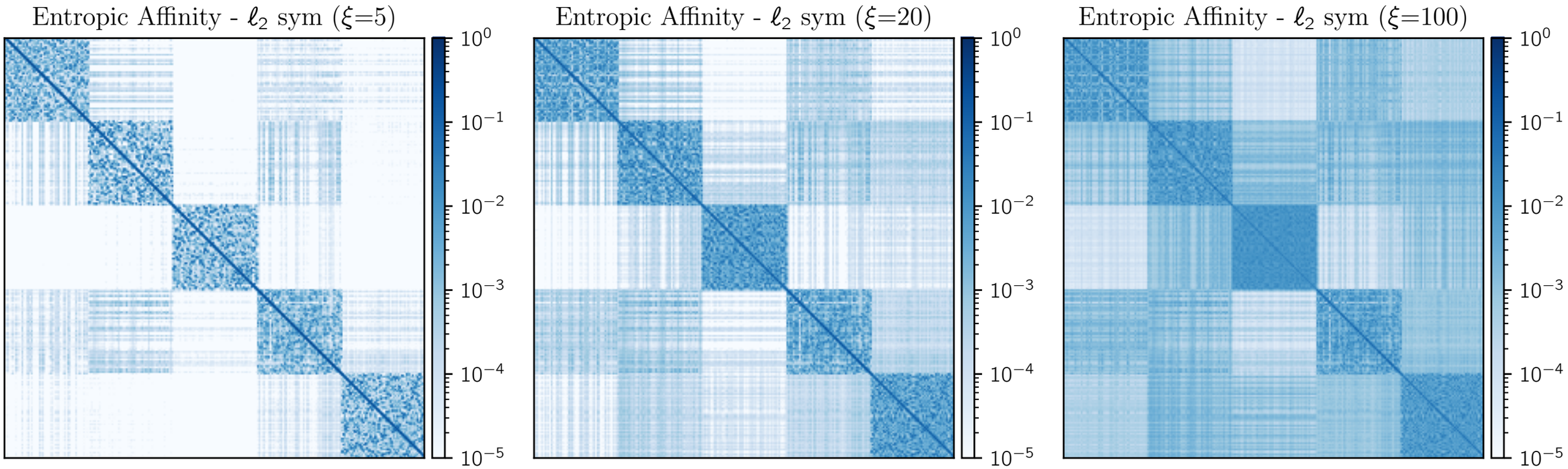
→ $\overline{\mathbf{P}^{\mathrm{e}}} = \frac{1}{2}(\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top})$ is used in practice. [Van der Maaten and Hinton, 2008]

t-SNE algorithm

↓

$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$

Entropic Affinity



Definition [Hinton, Roweis 2002]

$$\forall i, \forall j,\ P^{\mathrm{e}}_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon^\star_i\right)}{\sum_\ell \exp\left(-C_{i\ell}/\varepsilon^\star_i\right)}$$

$$\text{with } \varepsilon^\star_i \in \mathbb{R}^*_+ \text{ s.t. } \mathrm{H}(\mathbf{P}^{\mathrm{e}}_{i:}) = \log\xi + 1\,.$$

$\mathbf{P}^{\mathrm{e}}$ is **not symmetric**.

→ $\overline{\mathbf{P}^{\mathrm{e}}} = \frac{1}{2}(\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top})$ is used in practice. [Van der Maaten and Hinton, 2008]

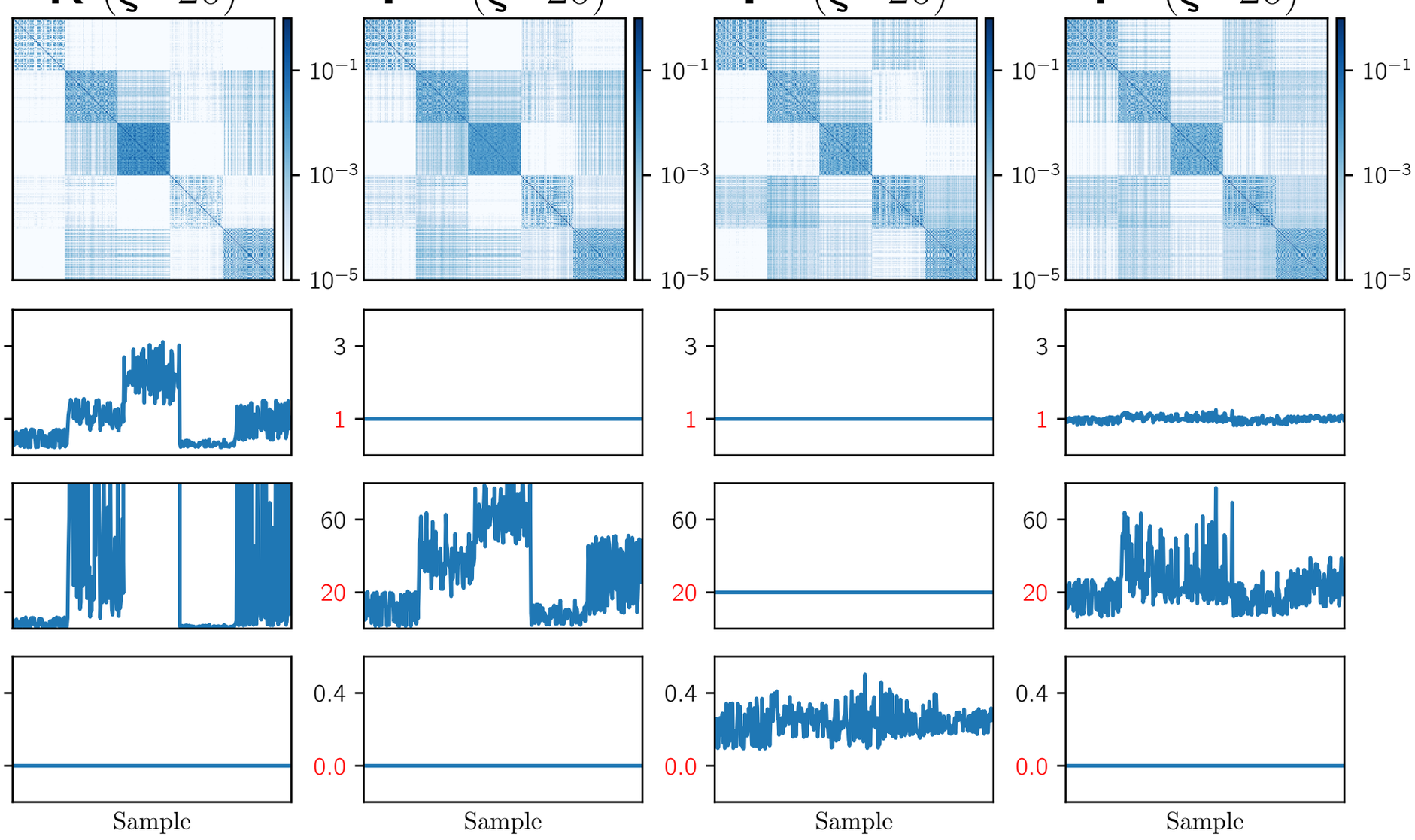
t-SNE algorithm

$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$

Breaks the construction of entropic affinities.

Affinity Panorama*

	Gibbs kernel	**Doubly-Sto**	**Entropic**	**Entropic (ℓ_2 Sym)**
	$\mathbf{K} = \exp(-\mathbf{C}/\sigma)$	$\mathbf{P}^{\mathrm{ds}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$	$\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K})$	$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$
Affinity	$\mathbf{K}$ $(\overline{\xi}=20)$**	$\mathbf{P}^{ds}$ $(\overline{\xi}=20)$	$\mathbf{P}^{e}$ $(\xi=20)$	$\overline{\mathbf{P}^{e}}$ $(\xi=20)$
ℓ_1 norm				
Perplexity				
Symmetry				

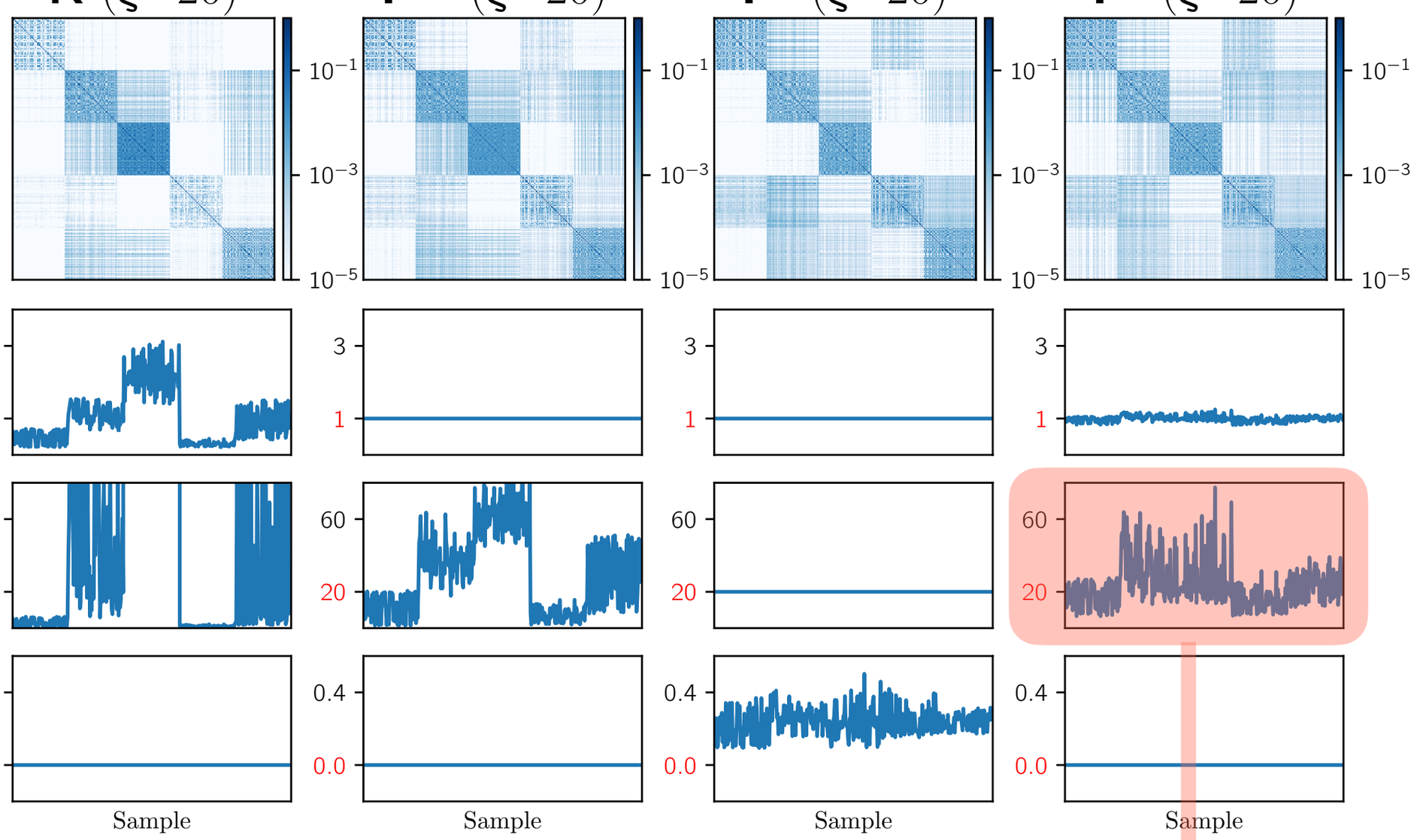


* On 5 classes of the COIL Dataset [Nene et al., 1996]

** $\overline{\xi}$ is average perplexity —> same global entropy as with $\xi = \overline{\xi}$.

Affinity Panorama*

	Gibbs kernel	**Doubly-Sto**	**Entropic**	**Entropic (ℓ_2 Sym)**
	$\mathbf{K} = \exp(-\mathbf{C}/\sigma)$	$\mathbf{P}^{\mathrm{ds}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$	$\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K})$	$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$
Affinity	$\mathbf{K}$ $(\overline{\xi}=20)$**	$\mathbf{P}^{ds}$ $(\overline{\xi}=20)$	$\mathbf{P}^{e}$ $(\xi=20)$	$\overline{\mathbf{P}^{e}}$ $(\xi=20)$
ℓ_1 norm				
Perplexity				
Symmetry				



* On 5 classes of the COIL Dataset [Nene et al., 1996]

** $\overline{\xi}$ is average perplexity —> same global entropy as with $\xi = \overline{\xi}$.

Entropies not controlled.

Affinity Panorama

	Gibbs kernel	**Doubly-Sto**	**Entropic**	**Entropic (ℓ_2 Sym)**
	$\mathbf{K} = \exp(-\mathbf{C}/\sigma)$	$\mathbf{P}^{\mathrm{ds}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$	$\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K})$	$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$
Affinity	$\mathbf{K}$ $(\overline{\xi}=20)$	$\mathbf{P}^{ds}$ $(\overline{\xi}=20)$	$\mathbf{P}^{e}$ $(\xi=20)$	$\overline{\mathbf{P}^{e}}$ $(\xi=20)$
ℓ_1 norm	3, 1	3, 1	3, 1	3, 1
Perplexity	60, 20	60, 20	60, 20	60, 20
Symmetry	0.4, 0.0	0.4, 0.0	0.4, 0.0	0.4, 0.0
	Sample	Sample	Sample	Sample

Can we control ℓ_1 norm, entropy and symmetry ?

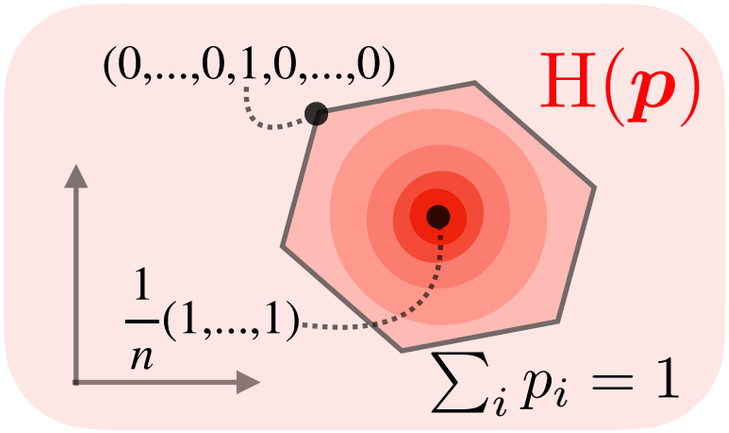
Symmetric Entropic Affinity

$$\mathcal{H}_\xi := \{\mathbf{P} \in \mathbb{R}_+^{n\times n} \text{ s.t. } \mathbf{P1} = \mathbf{1} \text{ and } \forall i,\ \mathrm{H}(\mathbf{P}_{i:}) \geq \log \xi + 1\}$$

Entropic Affinity as OT

$$\mathbf{P}^{\mathrm{e}} = \underset{\mathbf{P}\in\mathcal{H}_\xi}{\arg\min} \langle \mathbf{P}, \mathbf{C}\rangle.$$

- The constraints in $\mathcal{H}_\xi$ are saturated at the optimum.
- Symmetric matrices $\mathcal{S} = \{\mathbf{P} \text{ s.t. } \mathbf{P} = \mathbf{P}^\top\}$.



Definition **OURS**

$$\mathbf{P}^{\mathrm{se}} := \underset{\mathbf{P}\in\mathcal{H}_\xi\cap\mathcal{S}}{\arg\min} \langle \mathbf{P}, \mathbf{C}\rangle.$$

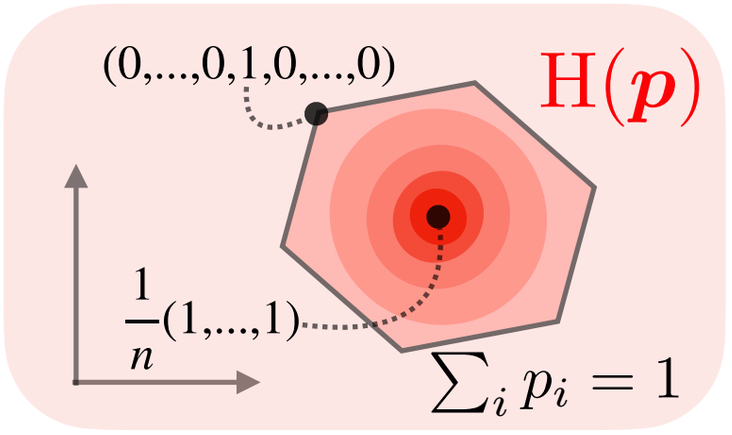
Symmetric Entropic Affinity

$$\mathcal{H}_\xi := \{\mathbf{P} \in \mathbb{R}_+^{n \times n} \text{ s.t. } \mathbf{P1} = \mathbf{1} \text{ and } \forall i,\ \mathrm{H}(\mathbf{P}_{i:}) \geq \log \xi + 1\}$$

Entropic Affinity as OT

$$\mathbf{P}^{\mathrm{e}} = \underset{\mathbf{P} \in \mathcal{H}_\xi}{\arg\min} \langle \mathbf{P}, \mathbf{C} \rangle.$$

- The constraints in $\mathcal{H}_\xi$ are saturated at the optimum.
- Symmetric matrices $\mathcal{S} = \{\mathbf{P} \text{ s.t. } \mathbf{P} = \mathbf{P}^\top\}$.



Definition **OURS**

$$\mathbf{P}^{\mathrm{se}} := \underset{\mathbf{P} \in \mathcal{H}_\xi \cap \mathcal{S}}{\arg\min} \langle \mathbf{P}, \mathbf{C} \rangle.$$

Enforce Symmetry

Symmetric Entropic Affinity

$$\mathcal{H}_\xi := \{\mathbf{P} \in \mathbb{R}_+^{n\times n} \text{ s.t. } \mathbf{P}\mathbf{1} = \mathbf{1} \text{ and } \forall i,\ \mathrm{H}(\mathbf{P}_{i:}) \geq \log \xi + 1\}$$

Definition **OURS**

$$\mathbf{P}^{\mathrm{se}} := \underset{\mathbf{P}\in\mathcal{H}_\xi \cap \mathcal{S}}{\arg\min} \langle \mathbf{P}, \mathbf{C}\rangle.$$

Enforce Symmetry

Property

For at least $n-1$ indices $i \in [\![n]\!]$, it holds $\mathrm{H}(\mathbf{P}^{\mathrm{se}}_{i:}) = \log \xi + 1$.

In practice, we have n saturated entropies.

Dual Ascent

$$\mathbf{P}^{\mathrm{se}} = \exp\left((\boldsymbol{\lambda}^\star \oplus \boldsymbol{\lambda}^\star - 2\mathbf{C}) \oslash (\boldsymbol{\gamma}^\star \oplus \boldsymbol{\gamma}^\star)\right)$$

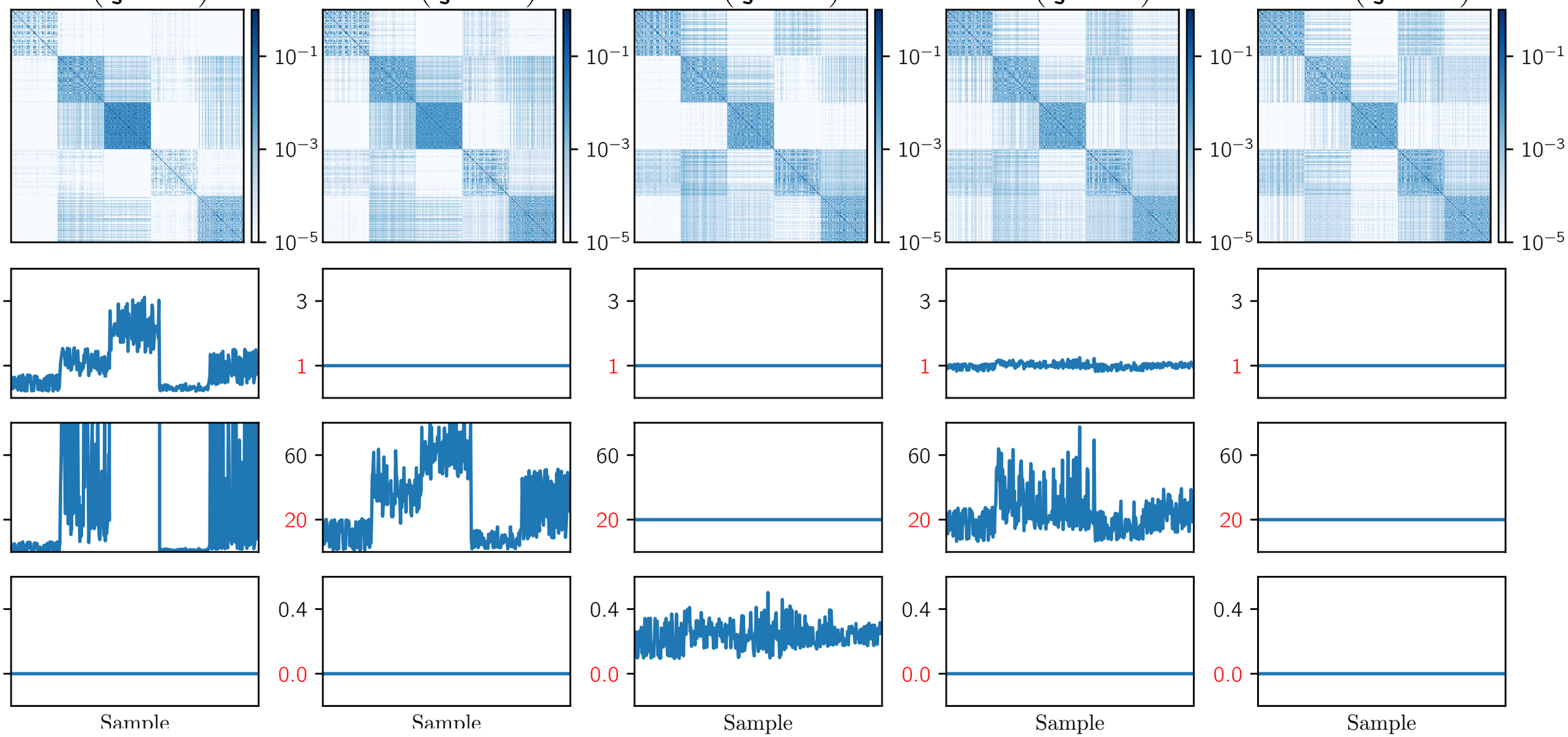
where $\boldsymbol{\lambda}^\star$ and $\boldsymbol{\gamma}^\star$ are computed using dual ascent.

Affinity Panorama *

t-SNE

OURS

Gibbs kernel	Doubly-Sto	Entropic	Entropic (ℓ_2 Sym)	Sym-Entropic
$\mathbf{K} = \exp(-\mathbf{C}/\sigma)$	$\mathbf{P}^{\mathrm{ds}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$	$\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K})$	$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$	$\mathbf{P}^{\mathrm{se}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi \cap \mathcal{S}}(\mathbf{K})$
$\mathbf{K}$ $(\overline{\xi}=20)$**	$\mathbf{P}^{ds}$ $(\overline{\xi}=20)$	$\mathbf{P}^{e}$ $(\xi=20)$	$\overline{\mathbf{P}^{e}}$ $(\xi=20)$	$\mathbf{P}^{se}$ $(\xi=20)$



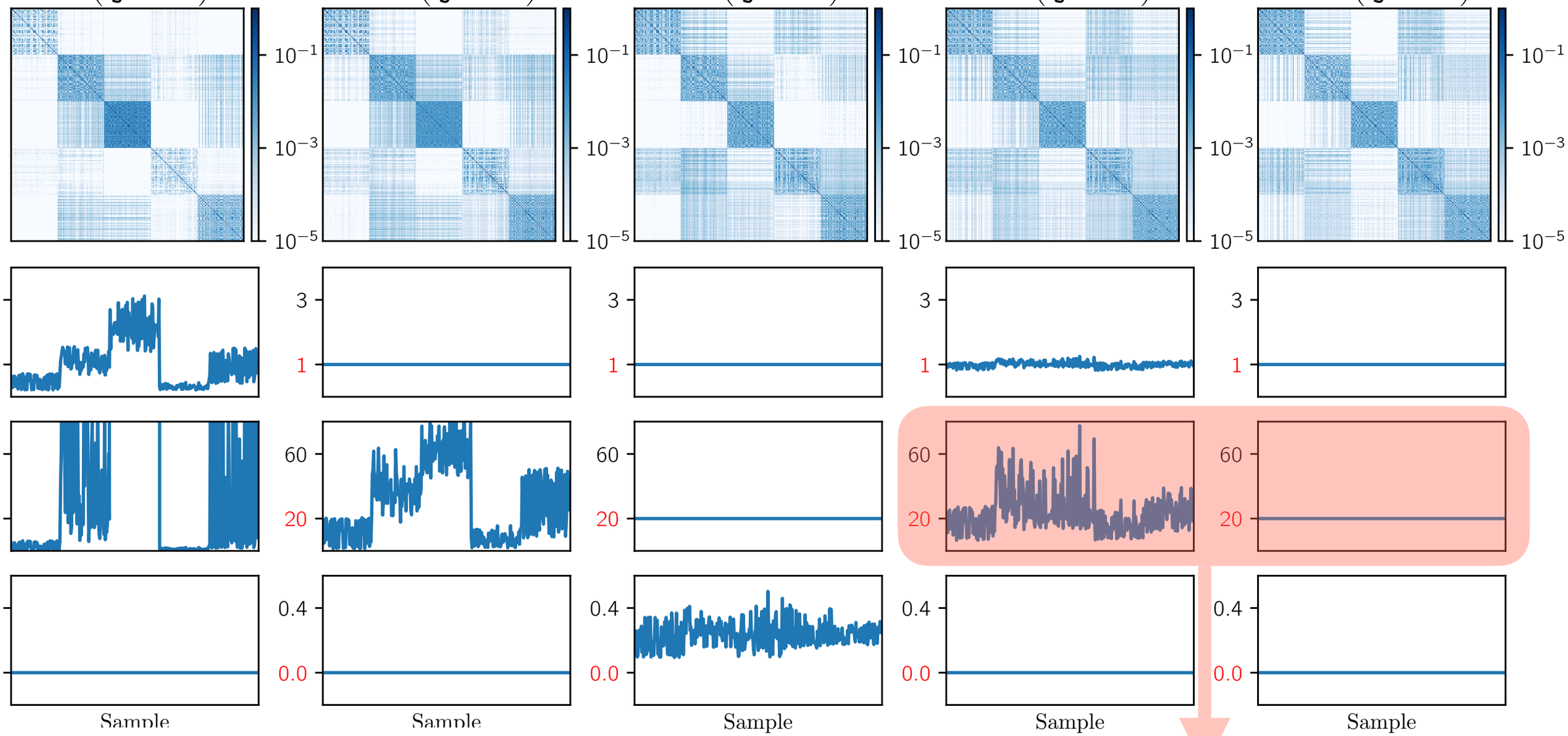
* On 5 classes of the COIL Dataset [Nene et al., 1996]

** $\overline{\xi}$ is average perplexity —> same global entropy as with $\xi = \overline{\xi}$.

Affinity Panorama *

t-SNE **OURS**

Gibbs kernel	**Doubly-Sto**	**Entropic**	**Entropic (ℓ_2 Sym)**	**Sym-Entropic**
$\mathbf{K} = \exp(-\mathbf{C}/\sigma)$	$\mathbf{P}^{\mathrm{ds}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$	$\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K})$	$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$	$\mathbf{P}^{\mathrm{se}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi \cap \mathcal{S}}(\mathbf{K})$
$\mathbf{K}$ $(\overline{\xi}=20)$**	$\mathbf{P}^{ds}$ $(\overline{\xi}=20)$	$\mathbf{P}^{e}$ $(\xi=20)$	$\overline{\mathbf{P}^{e}}$ $(\xi=20)$	$\mathbf{P}^{se}$ $(\xi=20)$



Effective control over entropies.

* On 5 classes of the COIL Dataset [Nene et al., 1996]

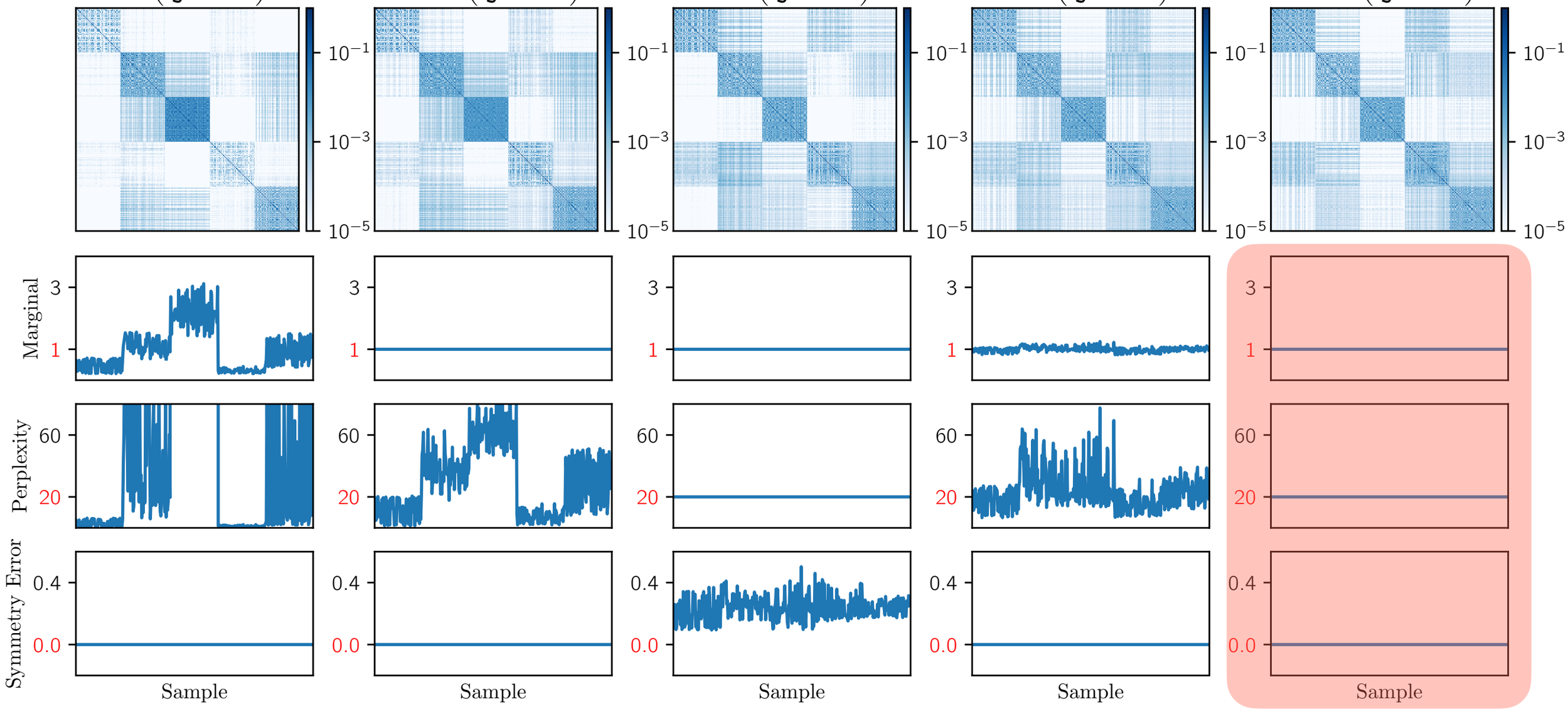
** $\overline{\xi}$ is average perplexity —> same global entropy as with $\xi = \overline{\xi}$.

Affinity Panorama *

t-SNE

OURS

Gibbs kernel	Doubly-Sto	Entropic	Entropic (ℓ_2 Sym)	Sym-Entropic
$\mathbf{K} = \exp(-\mathbf{C}/\sigma)$	$\mathbf{P}^{\mathrm{ds}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$	$\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K})$	$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$	$\mathbf{P}^{\mathrm{se}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi \cap \mathcal{S}}(\mathbf{K})$
$\mathbf{K}$ $(\overline{\xi}=20)$**	$\mathbf{P}^{ds}$ $(\overline{\xi}=20)$	$\mathbf{P}^{e}$ $(\xi=20)$	$\overline{\mathbf{P}^{e}}$ $(\xi=20)$	$\mathbf{P}^{se}$ $(\xi=20)$



➡ **Controls ℓ_1 norm, entropy and symmetry at the same time.**

Visualization on toy example

Doubly-Stochastic

$$\min_{\mathbf{P}\geq\mathbf{0},\mathbf{P}\mathbf{1}=\mathbf{1},\mathbf{P}\in\mathcal{S}} \langle \mathbf{P}, \mathbf{C} \rangle$$

$$\sum_i \mathrm{H}(\mathbf{P}_{i:}) \geq n(\log \xi + 1)$$

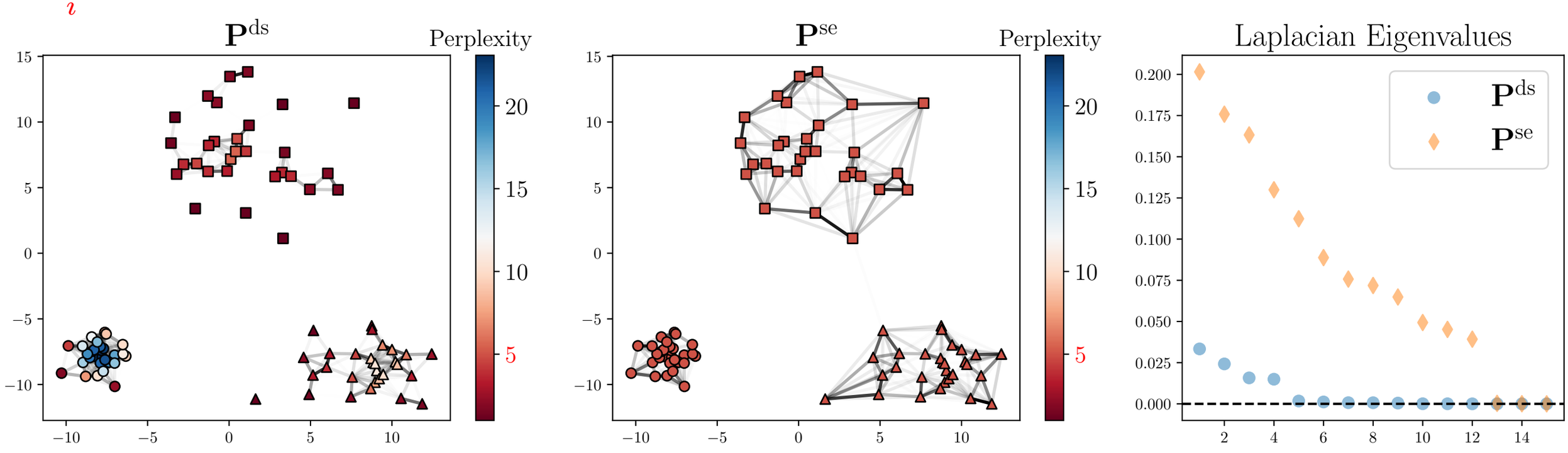
Symmetric-Entropic (OURS)

$$\min_{\mathbf{P}\geq\mathbf{0},\mathbf{P}\mathbf{1}=\mathbf{1},\mathbf{P}\in\mathcal{S}} \langle \mathbf{P}, \mathbf{C} \rangle$$

$$\forall i,\ \mathrm{H}(\mathbf{P}_{i:}) \geq \log \xi + 1$$

} Symmetric OT

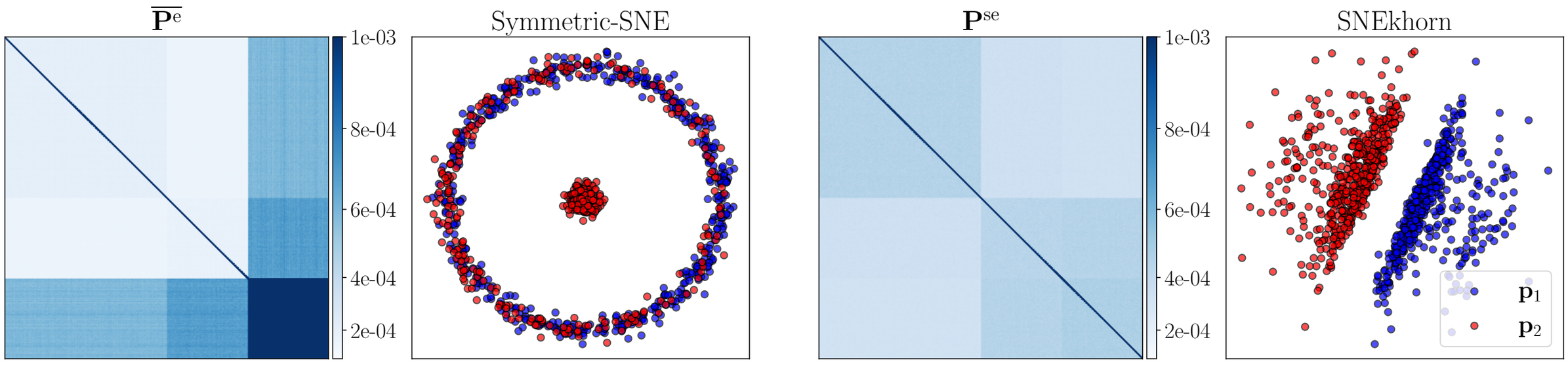
} Global vs Pointwise



- Perplexity is set to $\xi = 5$ (average for $\mathbf{P}^{\mathrm{ds}}$).
- $\mathbf{P}^{\mathrm{se}}$ can adapt to the varying noise levels.
- $\mathbf{P}^{\mathrm{ds}}$ retrieves many unwanted clusters.

null eigenvalues
=
clusters

Toy example: varying noise levels



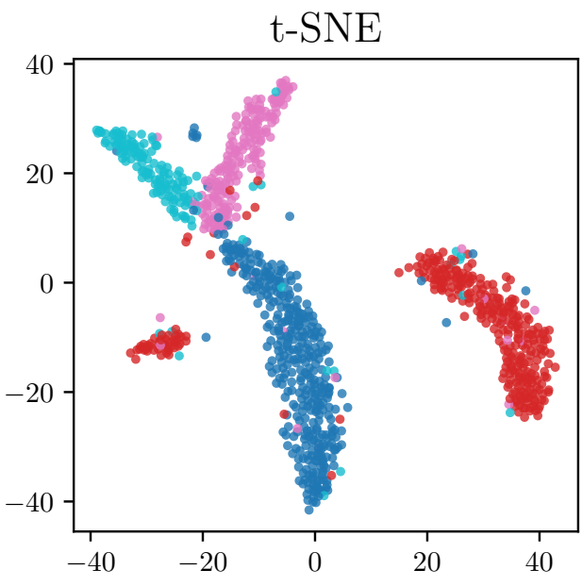
$\mathbf{p}_1$ and $\mathbf{p}_2$ taken in the 10^4-dimensional probability simplex.

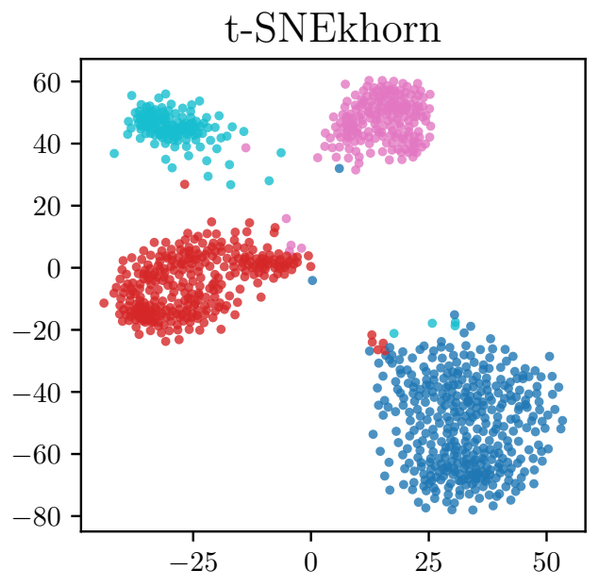
$$\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i / (\textstyle\sum_j \tilde{x}_{ij}), \qquad \tilde{\boldsymbol{x}}_i \sim \begin{cases} \mathcal{M}(1000, \mathbf{p}_1), & 1 \leq i \leq 500 \\ \mathcal{M}(1000, \mathbf{p}_2), & 501 \leq i \leq 750 \\ \mathcal{M}(2000, \mathbf{p}_2), & 751 \leq i \leq 1000\,. \end{cases}$$

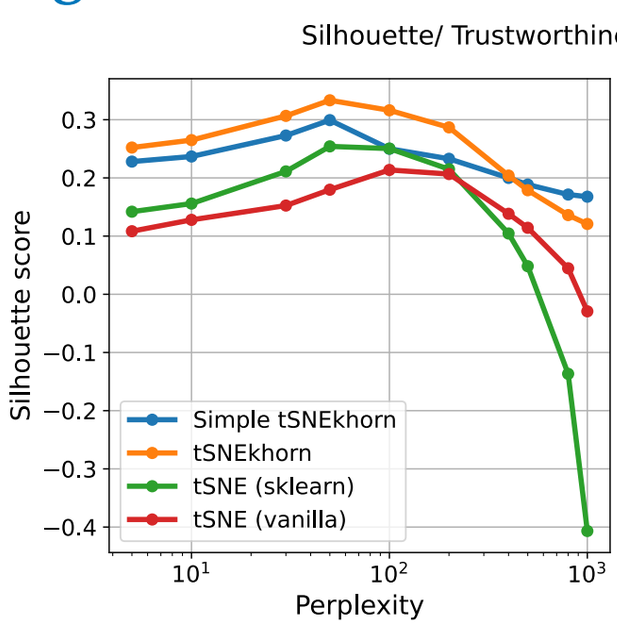
SNE is mislead by the batch effect unlike SNEkhorn.

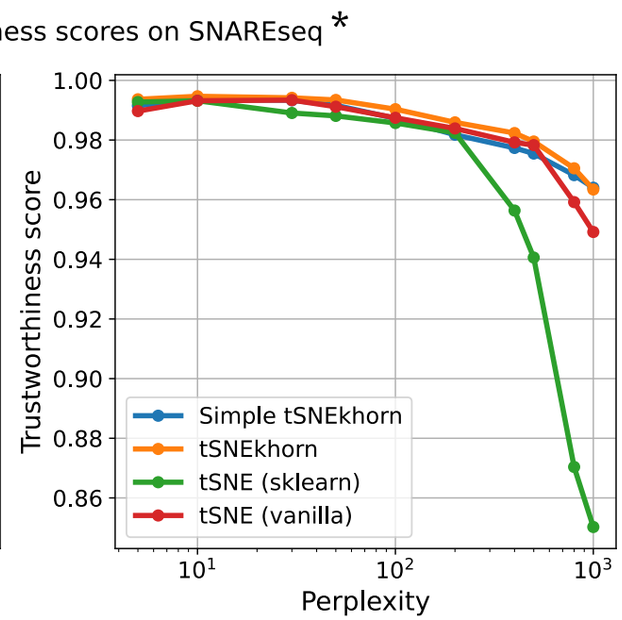
Dimension Reduction Results

SNAREseq Single Cell data

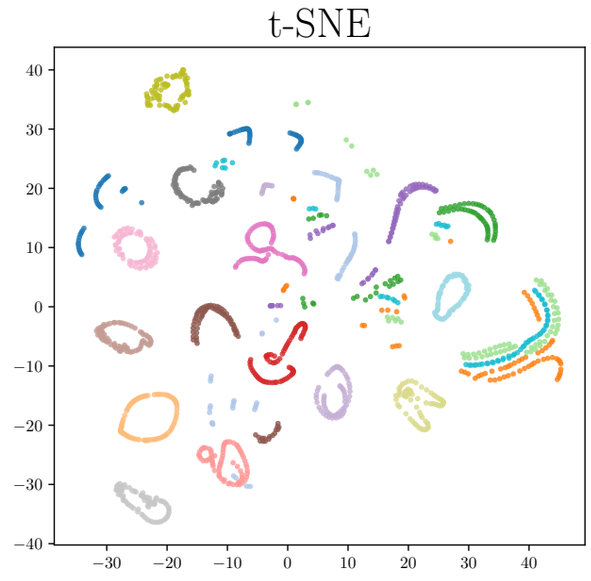


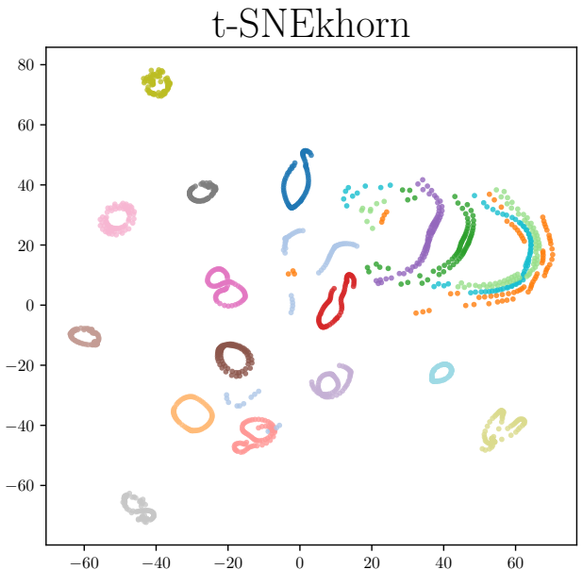


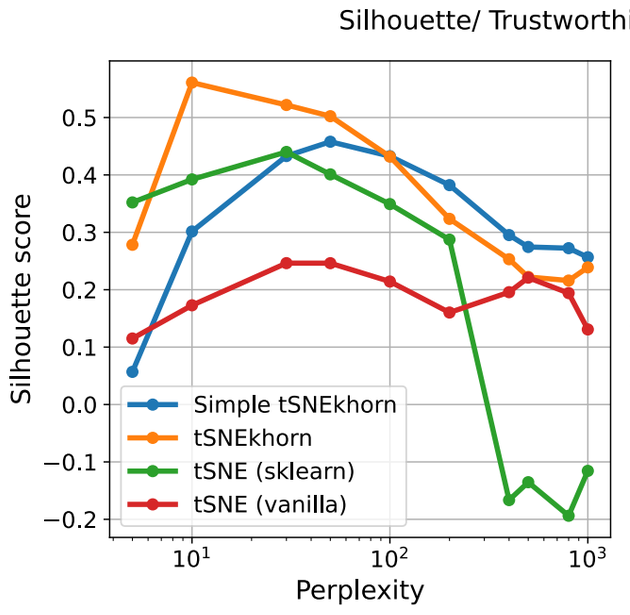


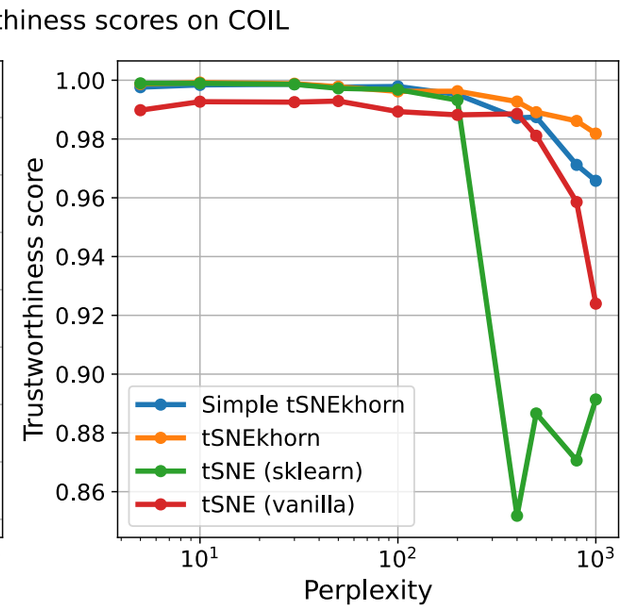


COIL-20 Image data



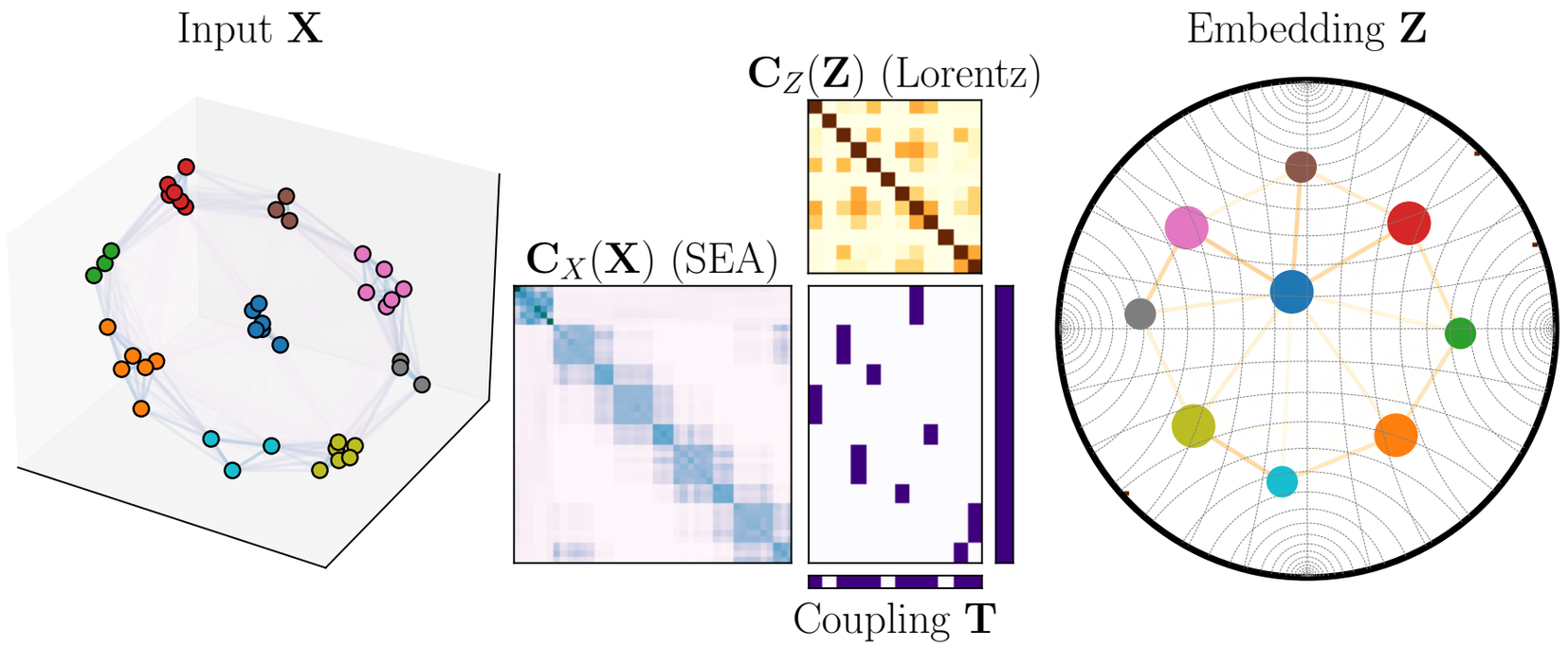






* Simple t-SNEkhorn —> t-SNEkhorn with same affinity as t-SNE for the embeddings $\mathbf{Z}$ (not doubly stochastic).

Distributional reduction



Motivation

✦ **Single-cell RNA-seq**

Technical noise due to partial sampling of RNA molecules within cells.

Baran *et al. Genome Biology* (2019) 20:206
https://doi.org/10.1186/s13059-019-1812-2

Genome Biology

METHOD Open Access

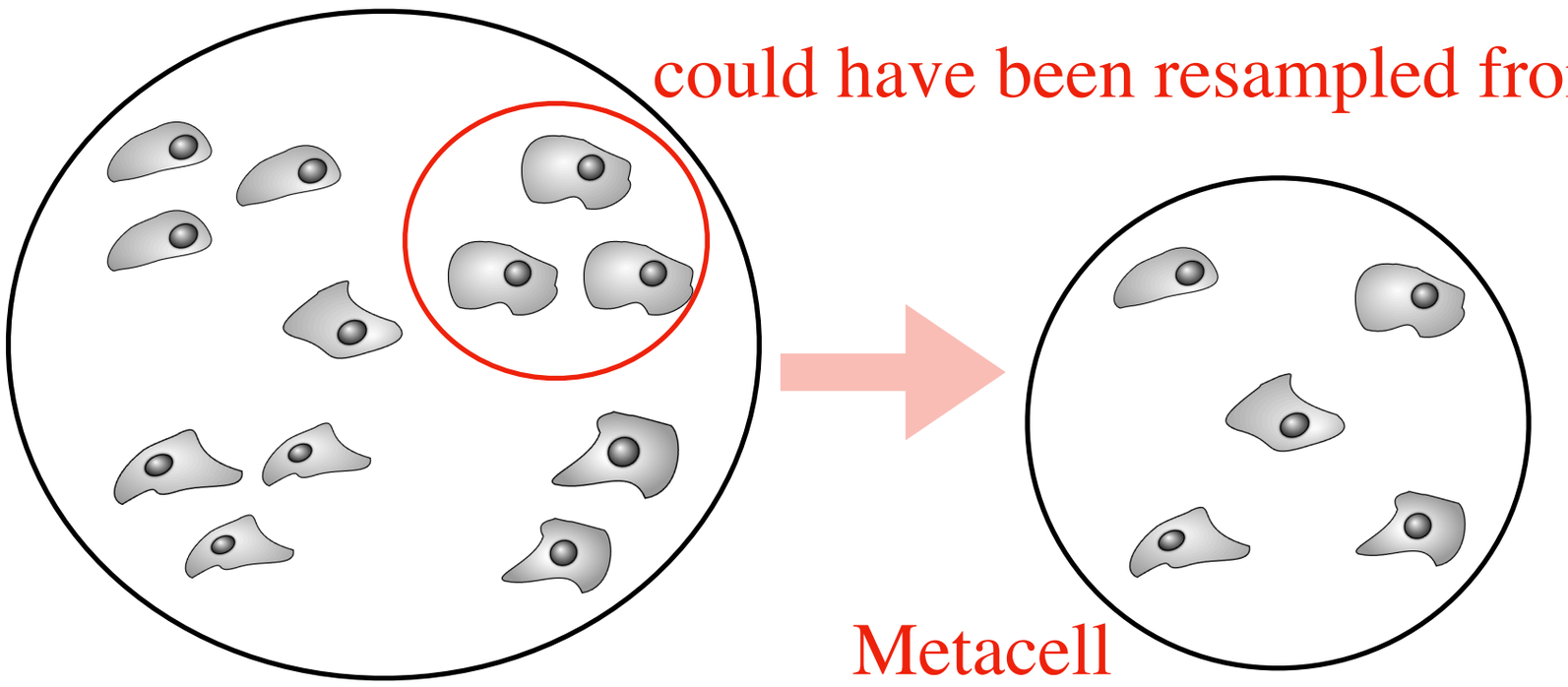
MetaCell: analysis of single-cell RNA-seq data using *K*-nn graph partitions

Yael Baran[1], Akhiad Bercovich[1], Arnau Sebe-Pedros[1], Yaniv Lubling[1], Amir Giladi[2], Elad Chomsky[1], Zohar Meir[1], Michael Hoichman[1], Aviezer Lifshitz[1] and Amos Tanay[1*]

Problem : **impossible to resample** a cell

→ integration of data from different cells

→ need to **separate the sampling effect from biological variance**

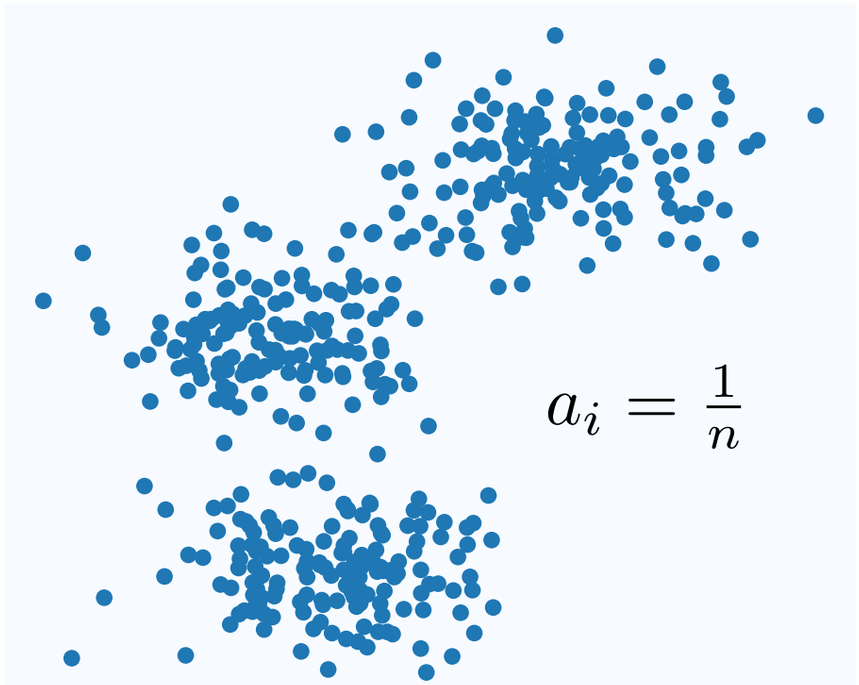


✦ **We would like to choose the granularity of the output data**

Framing DR in terms of distributions

Data: $(\mathbf{x}_i)_{i \in [\![n]\!]}$; $\mathbf{x}_i \in \mathbb{R}^d$ ⟶ A probability distribution describing the data

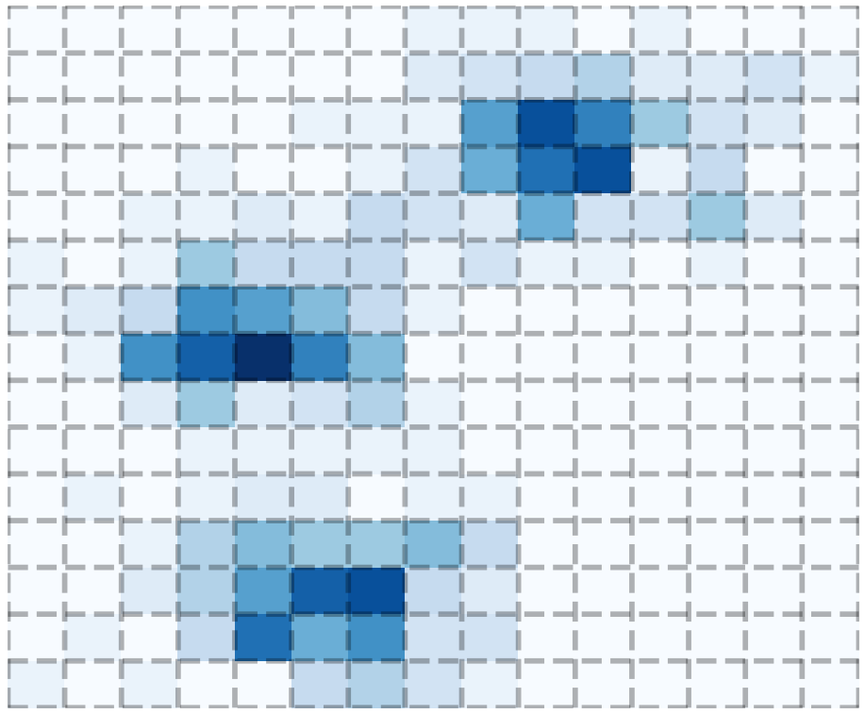
Lagrangian: $\sum_{i=1}^{n} a_i \delta_{x_i}$



(point clouds)

$\delta_{\mathbf{x}_i}(\mathbf{x}) = 1$ if $\mathbf{x} = \mathbf{x}_i$ else 0

Eulerian: $\sum_{i=1}^{N} a_i \delta_{\hat{x}_i}$



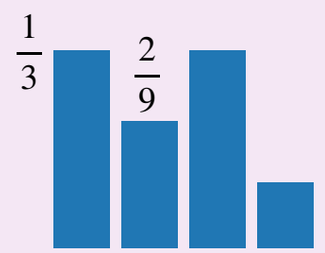
(histograms)

$\hat{x}_i$ fixed position (grid)

Probability simplex

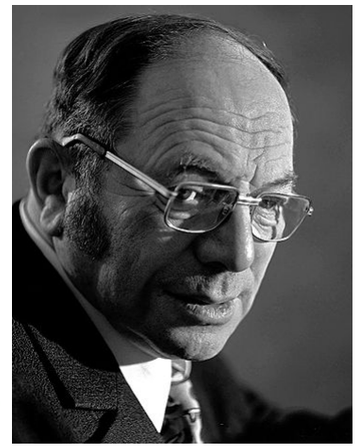
$\mathbf{a} = (a_i)_{i \in [\![n]\!]} \in \Sigma_n$

$a_i \geq 0, \sum_{i=1}^{n} a_i = 1$



From Wasserstein to Gromov-Wasserstein

✦ Classical optimal transport (in a nutshell)



Kantorovitch Formulation

Two probability distributions

$$\mu \in \mathcal{P}(\mathcal{X})$$

$$\nu \in \mathcal{P}(\mathcal{Y})$$

A cost function

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

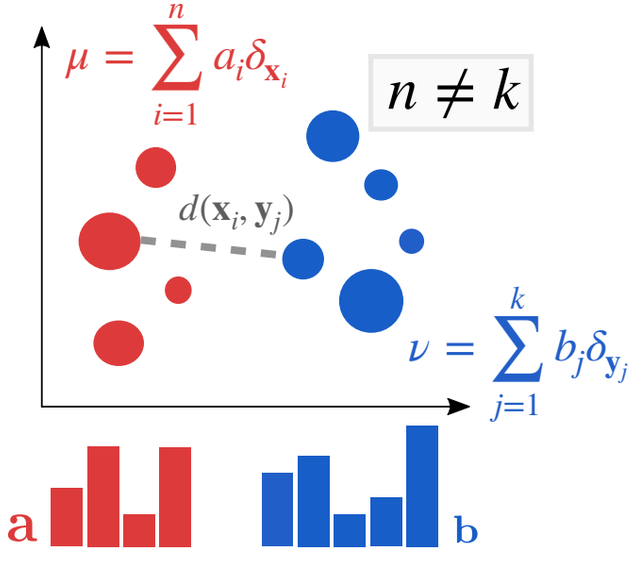
$\mu \in \mathcal{P}(\mathcal{X})$ is **transported** to $\nu \in \mathcal{P}(\mathcal{Y})$ by a **transport plan** $T \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$

We want to find the plan that **minimizes the overall cost** of moving all the points.

$$\inf_{T \in \Pi(\mu,\nu)} \int c(x, y) dT(x, y)$$

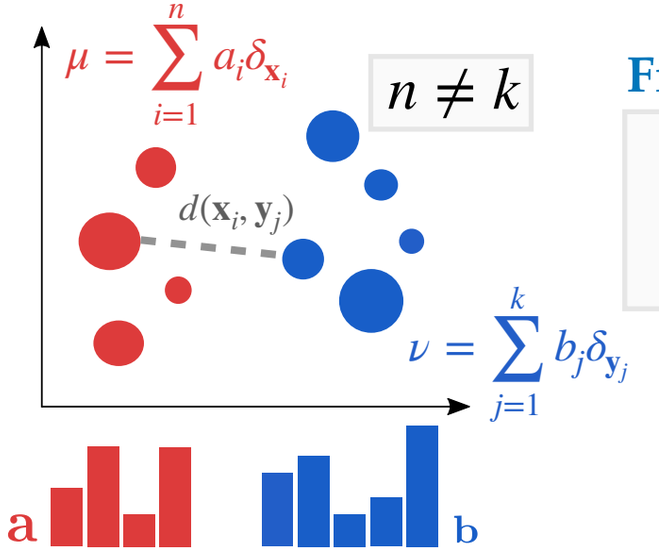
From Wasserstein to Gromov-Wasserstein

✦ Classical optimal transport (in a nutshell)



From Wasserstein to Gromov-Wasserstein

✦ Classical optimal transport (in a nutshell)

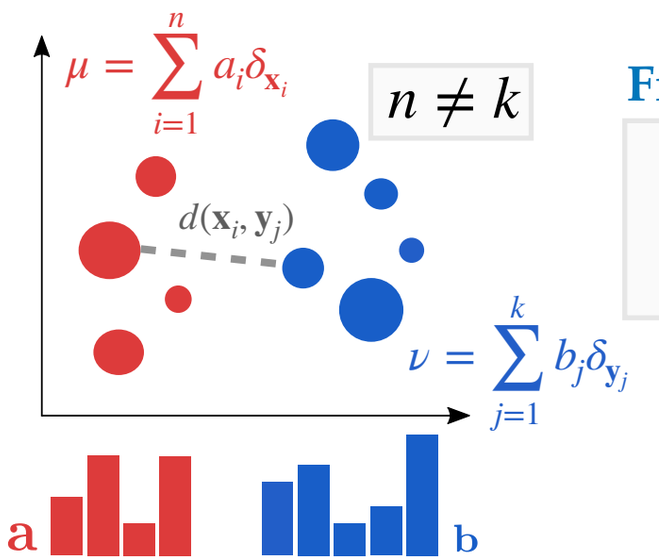


Find plan $T \in \mathbb{R}_+^{n \times k}$

$$\min_T \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

From Wasserstein to Gromov-Wasserstein

✦ Classical optimal transport (in a nutshell)



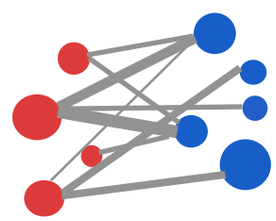
Find plan $T \in \mathbb{R}_+^{n \times k}$

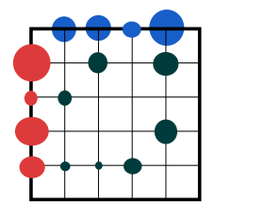
$$\min_T \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

which constraints ?

→ **Coupling** $\Pi(a, b)$

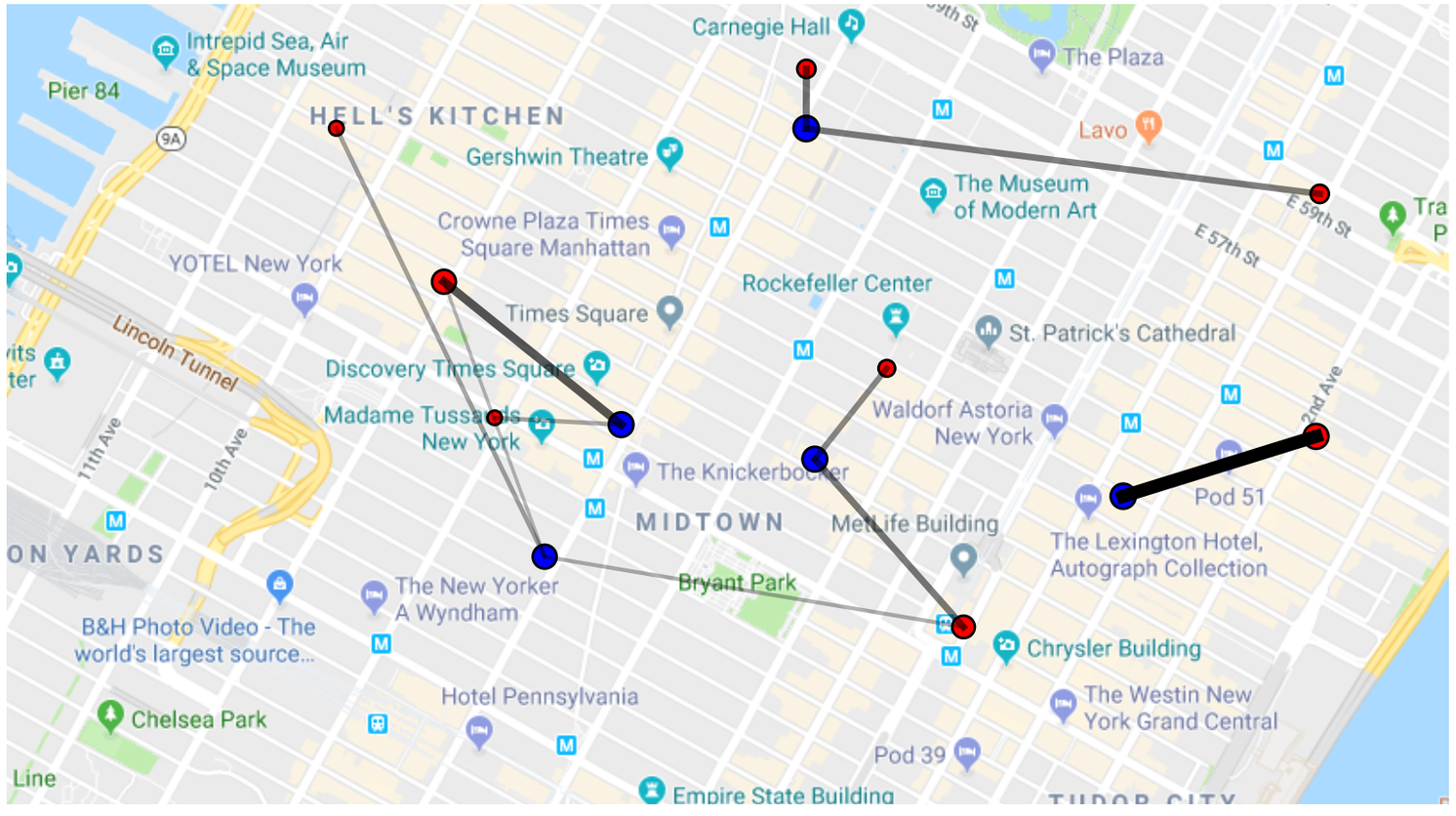
$$T^\top 1_n = a$$
$$T 1_k = b$$





$T \in [0,1]^{n \times k}$

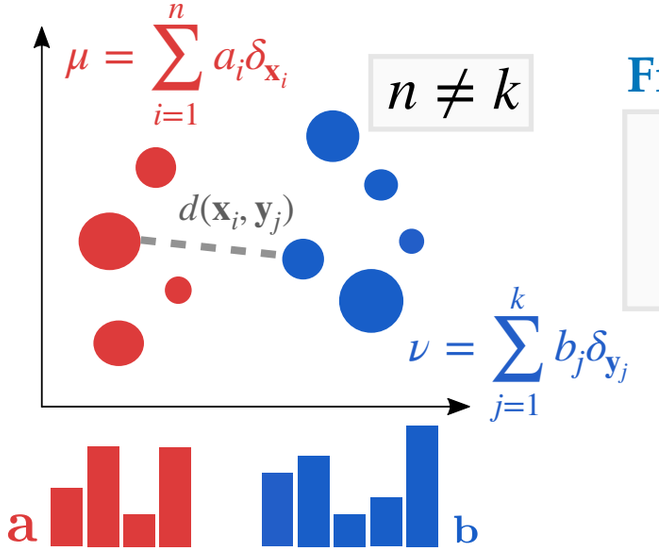
- Bakeries = quantity of breads
  loc: $\mathbf{x}_i$ quantity: a_i
- Cafés = demand of breads
  loc: $\mathbf{y}_j$ demand: b_j
- Distance between bakeries and cafés
  $c(\mathbf{x}_i, \mathbf{y}_j)$



We want to route all the breads from bakeries to cafés the cheapest way

From Wasserstein to Gromov-Wasserstein

✦ Classical optimal transport (in a nutshell)



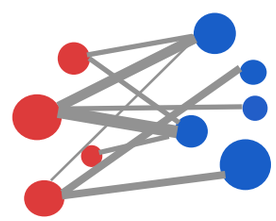
Find plan $T \in \mathbb{R}_+^{n \times k}$

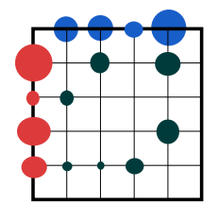
$$\min_T \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

which constraints ?

Coupling $\Pi(a, b)$

$$T^\top 1_n = a$$
$$T 1_k = b$$





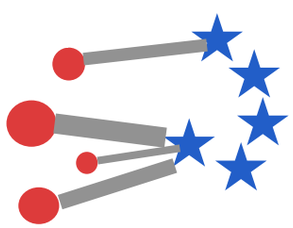
$T \in [0,1]^{n \times k}$

Semi-relaxed coupling

$$T^\top 1_n = a$$

$\sim$ `assign` in k-means

$$\sim \min_b \min_{T \in \Pi(a,b)} \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$



Bakeries = quantity of breads

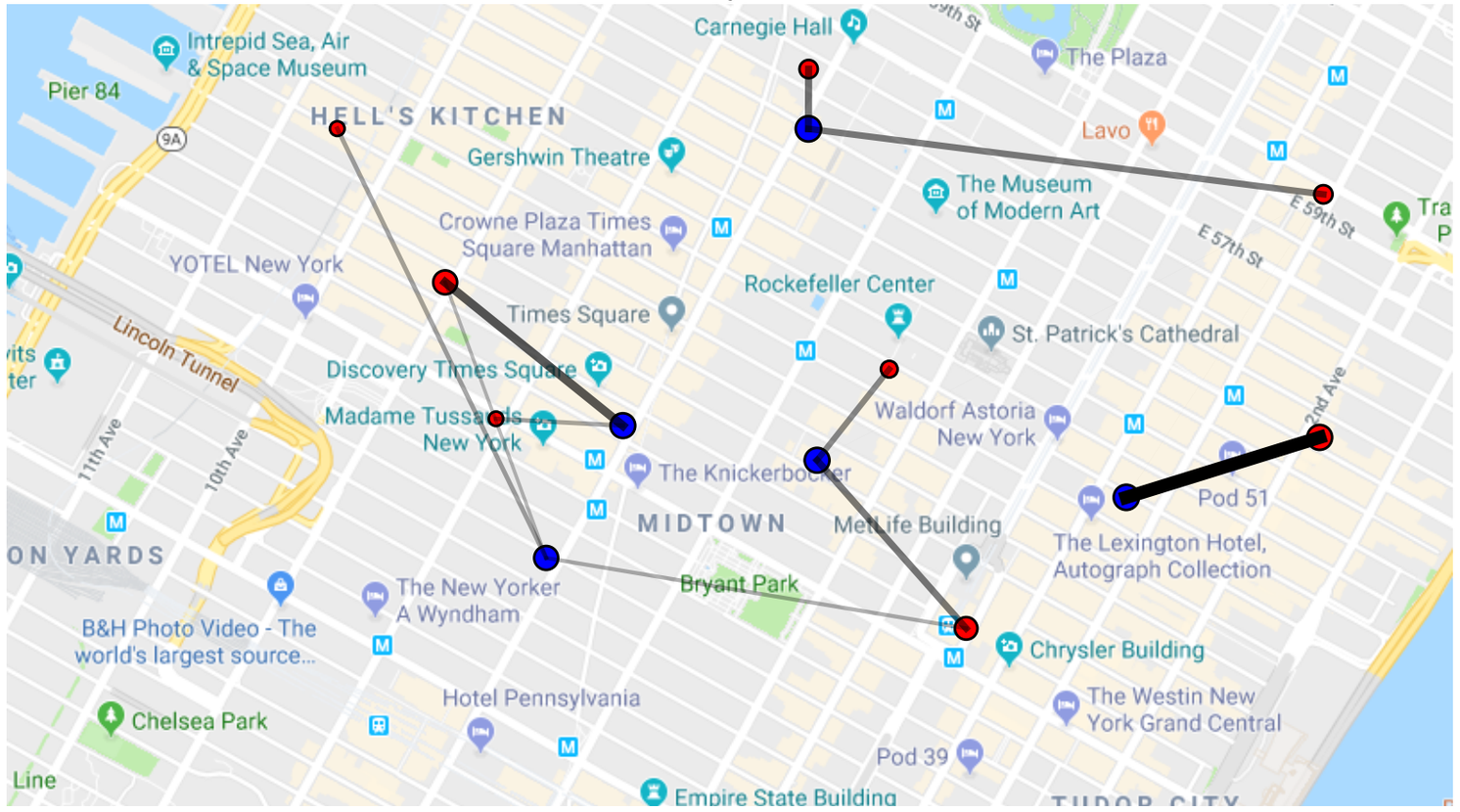
loc: $\mathbf{x}_i$ quantity: a_i

Cafés = demand of breads

loc: $\mathbf{y}_j$ ~~demand: b_j~~

Distance between bakeries and cafés

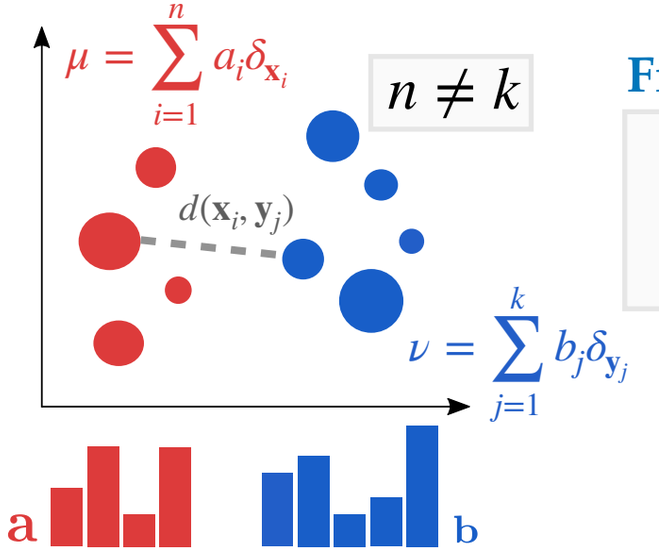
$$c(\mathbf{x}_i, \mathbf{y}_j)$$



We want to route all the breads from bakeries to cafés the cheapest way

From Wasserstein to Gromov-Wasserstein

✦ Classical optimal transport (in a nutshell)



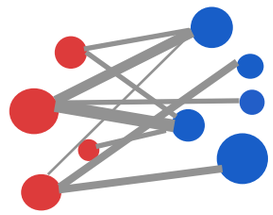
Find plan $T \in \mathbb{R}_+^{n \times k}$

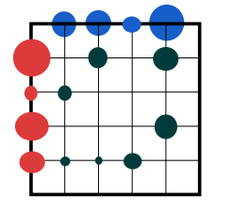
$$\min_T \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

which constraints ?

Coupling $\Pi(a, b)$

$$T^\top 1_n = a$$
$$T 1_k = b$$





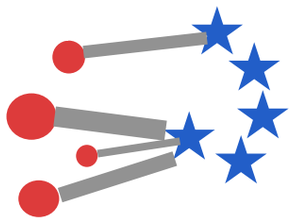
$T \in [0,1]^{n \times k}$

Semi-relaxed coupling

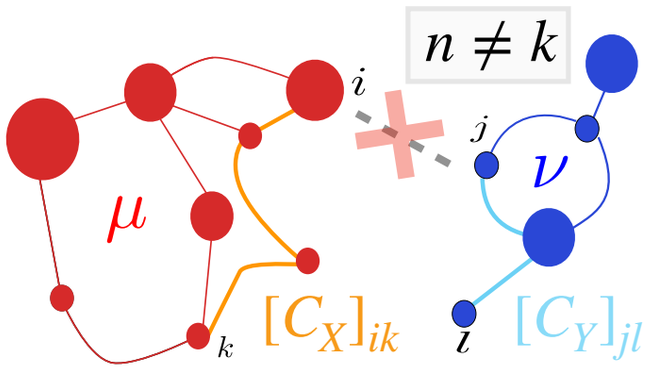
$$T^\top 1_n = a$$

$\sim$ `assign` in k-means

$$\sim \min_b \min_{T \in \Pi(a,b)} \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$



✦ Gromov-Wasserstein

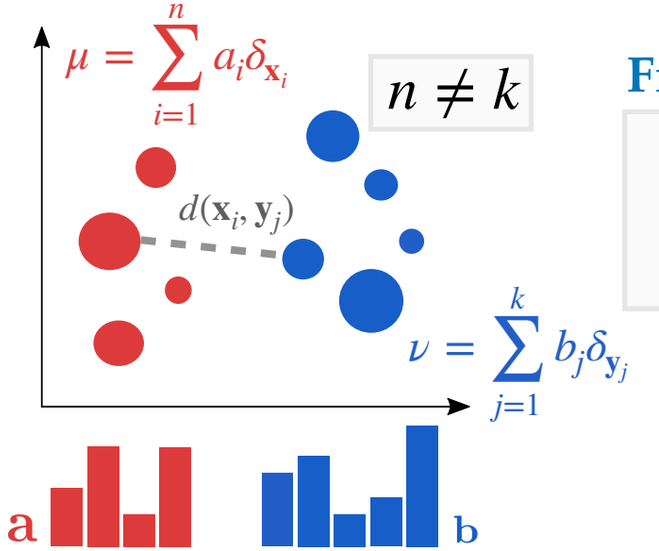


(Sturm, 2012)

(Mémoli, 2011)

From Wasserstein to Gromov-Wasserstein

✦ Classical optimal transport (in a nutshell)



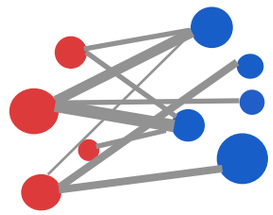
Find plan $T \in \mathbb{R}_+^{n \times k}$

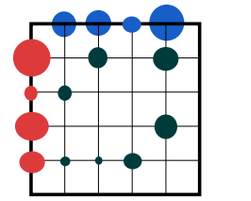
$$\min_T \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

which constraints ?

Coupling $\Pi(a, b)$

$$T^\top 1_n = a$$
$$T 1_k = b$$





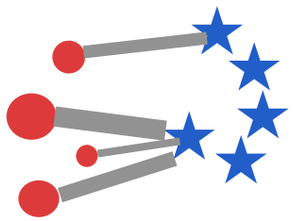
$T \in [0,1]^{n \times k}$

Semi-relaxed coupling

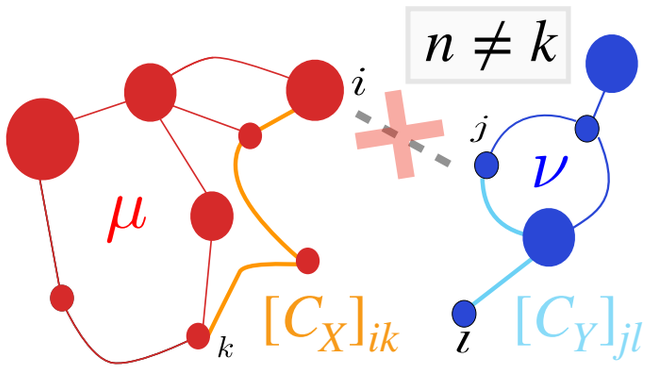
$$T^\top 1_n = a$$

$\sim$ `assign` in k-means

$$\sim \min_b \min_{T \in \Pi(a,b)} \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$



✦ Gromov-Wasserstein



(Sturm, 2012)

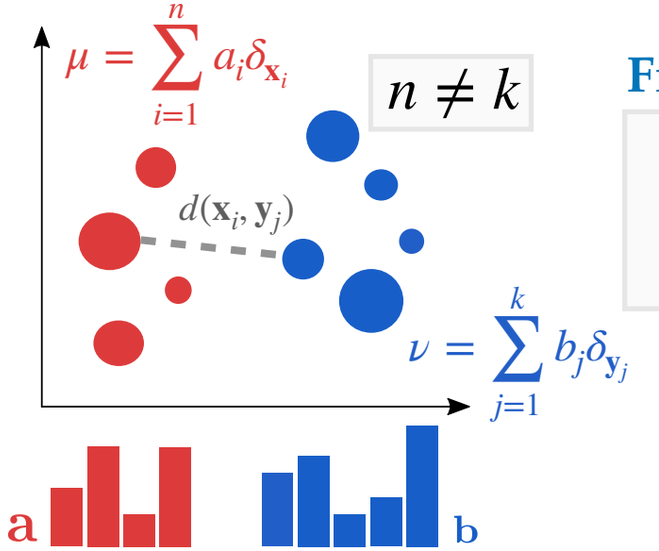
(Mémoli, 2011)

Quadratic OT: find the plan

$$\min_{T \in \Pi(a,b)} \sum_{ijkl} L\left([C_X]_{ik}, [C_Y]_{jl}\right) T_{ij} T_{kl}$$

From Wasserstein to Gromov-Wasserstein

✦ Classical optimal transport (in a nutshell)



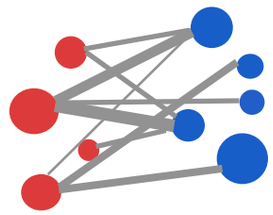
Find plan $T \in \mathbb{R}_+^{n \times k}$

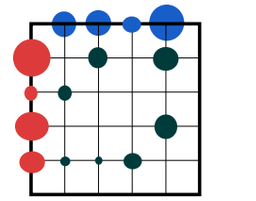
$$\min_T \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

which constraints ?

Coupling $\Pi(a, b)$

$$T^\top 1_n = a$$
$$T 1_k = b$$





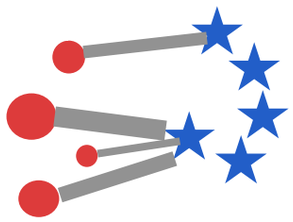
$T \in [0,1]^{n \times k}$

Semi-relaxed coupling

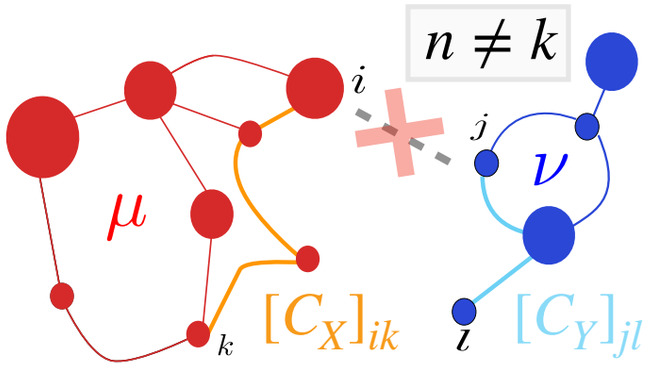
$$T^\top 1_n = a$$

$\sim$ `assign` in k-means

$$\sim \min_b \min_{T \in \Pi(a,b)} \sum_{ij} c(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$



✦ Gromov-Wasserstein



(Sturm, 2012)

(Mémoli, 2011)

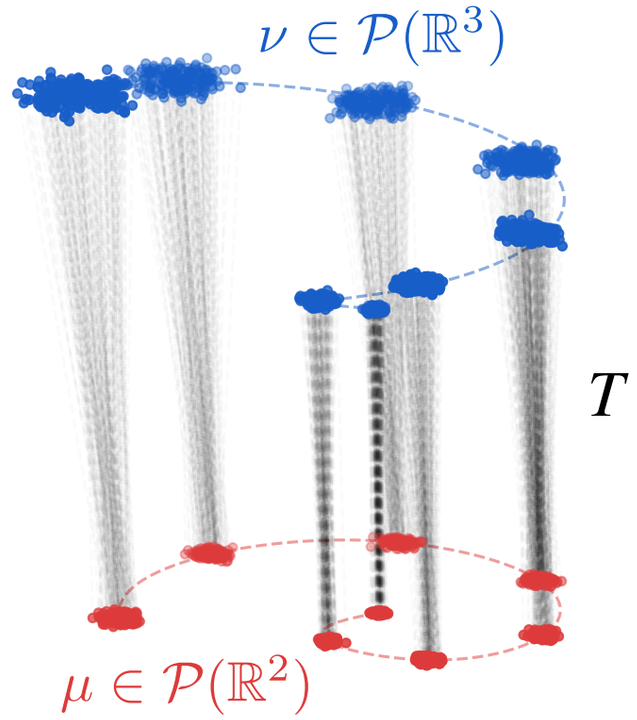
Quadratic OT: find the plan

$$\min_{T \in \Pi(a,b)} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Y]_{jl}\right) T_{ij} T_{kl}$$

- ✦ L measures distorsion

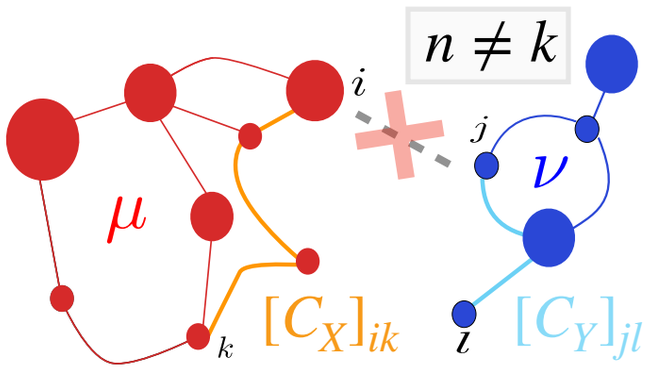
$$\left| [C_X]_{ik} - [C_Y]_{jl} \right|^2$$

- ✦ Goal : **preserving pairwise connectivity**
- ✦ Distance w.r.t. isomorphisms
- ✦ Difficult quadratic problem (NP-hard)



From Wasserstein to Gromov-Wasserstein

✦ Gromov-Wasserstein



(Sturm, 2012)

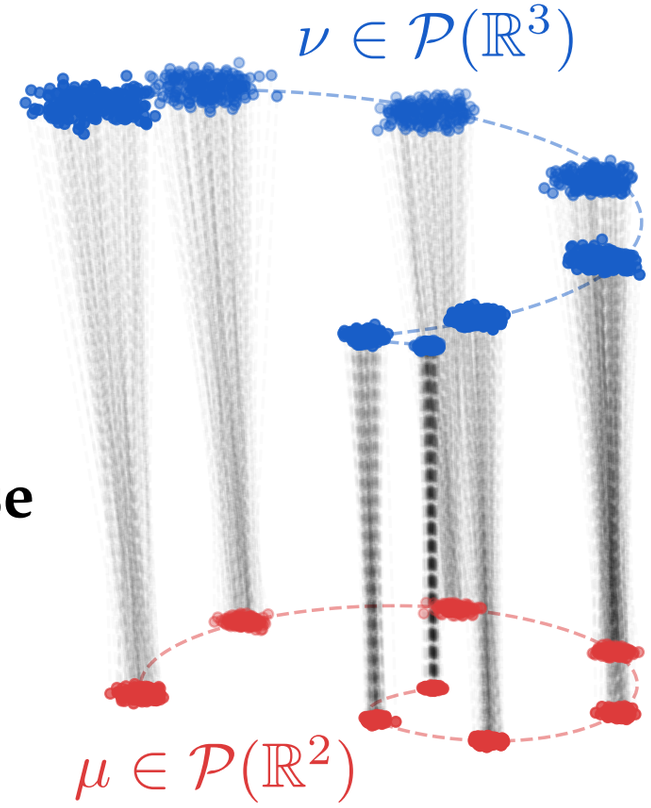
(Mémoli, 2011)

✦ L measures distorsion

$$\left| [C_X]_{ik} - [C_Y]_{jl} \right|^2$$

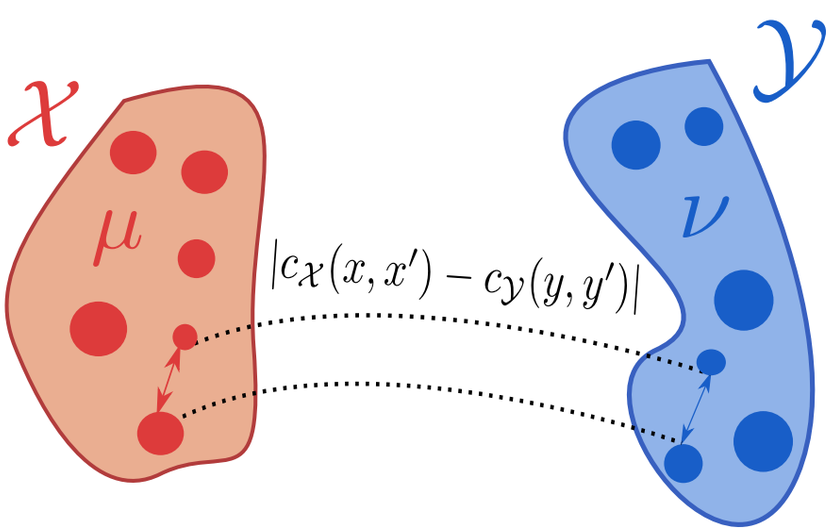
✦ Goal : **preserving pairwise connectivity**

✦ Non-convex quadratic problem (NP-hard)



Quadratic OT: find the plan

$$\min_{T \in \Pi(a,b)} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Y]_{jl}\right) T_{ij} T_{kl}$$



✦ Distance w.r.t. isomorphisms, on the space of metric measure spaces

$$\mathbb{X} = (\mathcal{X}, c_{\mathcal{X}}, \mu \in \mathcal{P}(\mathcal{X}))$$

$$\mathbb{Y} = (\mathcal{Y}, c_{\mathcal{Y}}, \nu \in \mathcal{P}(\mathcal{Y}))$$

$GW(\mathbb{X}, \mathbb{Y}) = 0$ iff

$\exists \phi : \mathcal{X} \to \mathcal{Y}$

Isometry

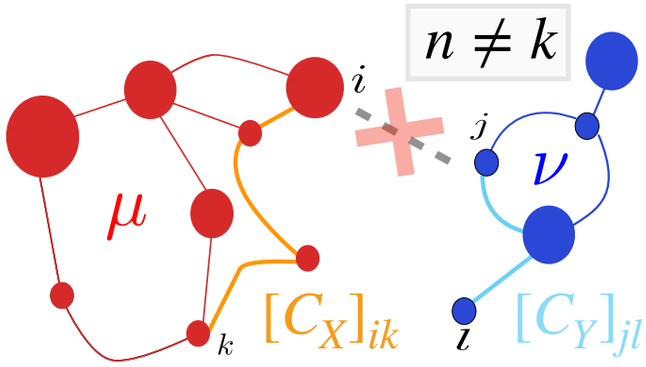
$c_{\mathcal{X}}(x, x') = c_{\mathcal{Y}}(\phi(x), \phi(x'))$

Measure preserving

$\phi \# \mu = \nu$

From Wasserstein to Gromov-Wasserstein

✦ Gromov-Wasserstein



(Sturm, 2012)

(Mémoli, 2011)

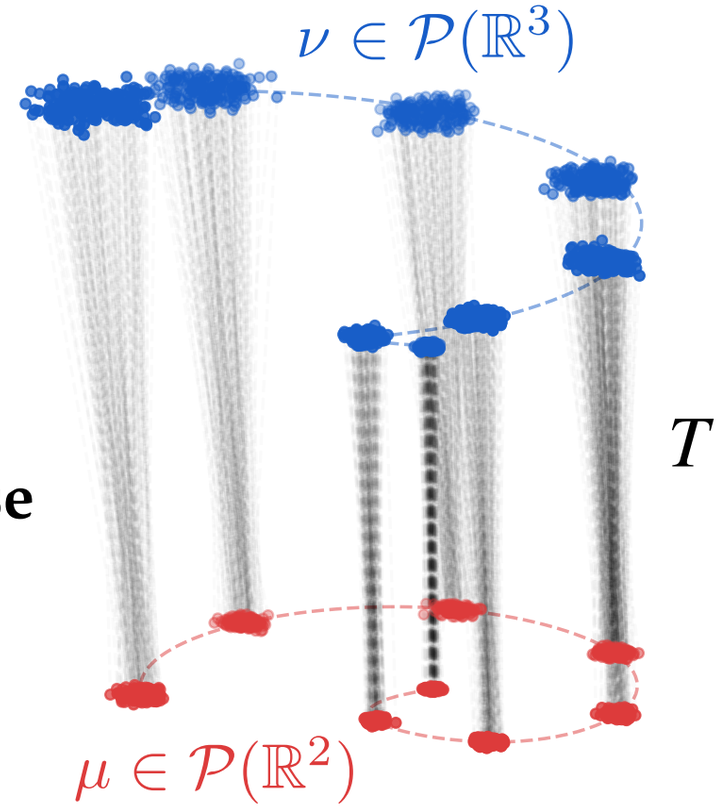
Quadratic OT: find the plan

$$\min_{T \in \Pi(a,b)} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Y]_{jl}\right) T_{ij} T_{kl}$$

- L measures distorsion

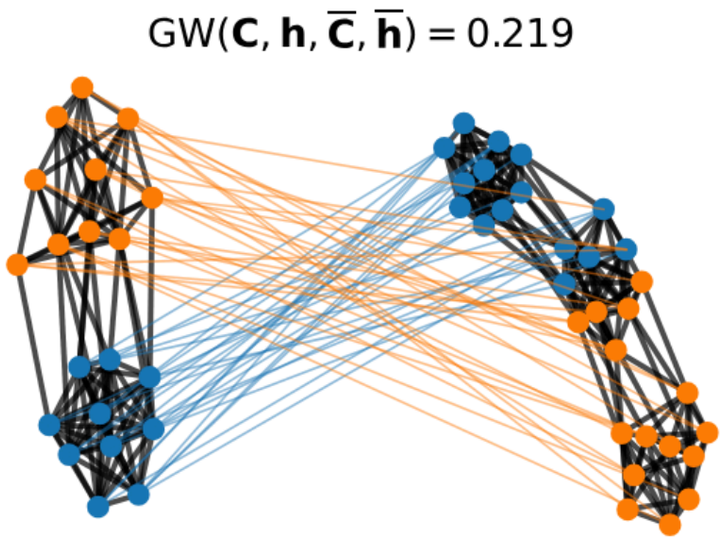
$$\left|[C_X]_{ik} - [C_Y]_{jl}\right|^2$$

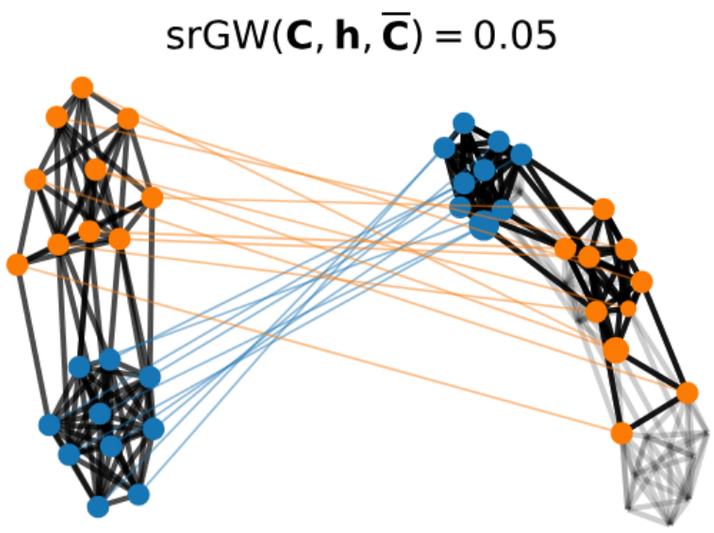
- Goal : **preserving pairwise connectivity**
- Non-convex quadratic problem (NP-hard)



✦ Semi relaxed Gromov-Wasserstein

(Vincent-Cuaz, 2022)



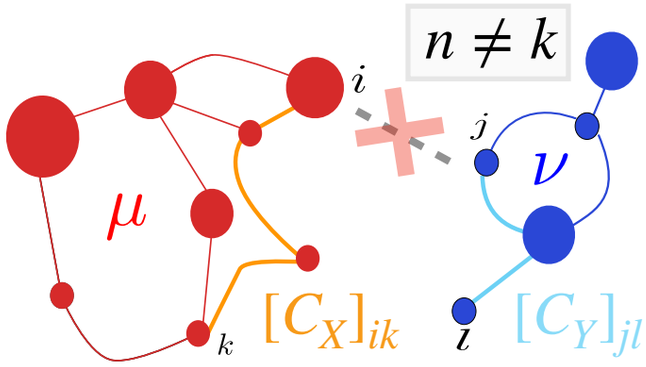


$$\min_{T} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Y]_{jl}\right) T_{ij} T_{kl}$$

$$\left| \quad T^\top 1_n = a \right.$$

From Wasserstein to Gromov-Wasserstein

✦ Gromov-Wasserstein




(Sturm, 2012)


(Mémoli, 2011)

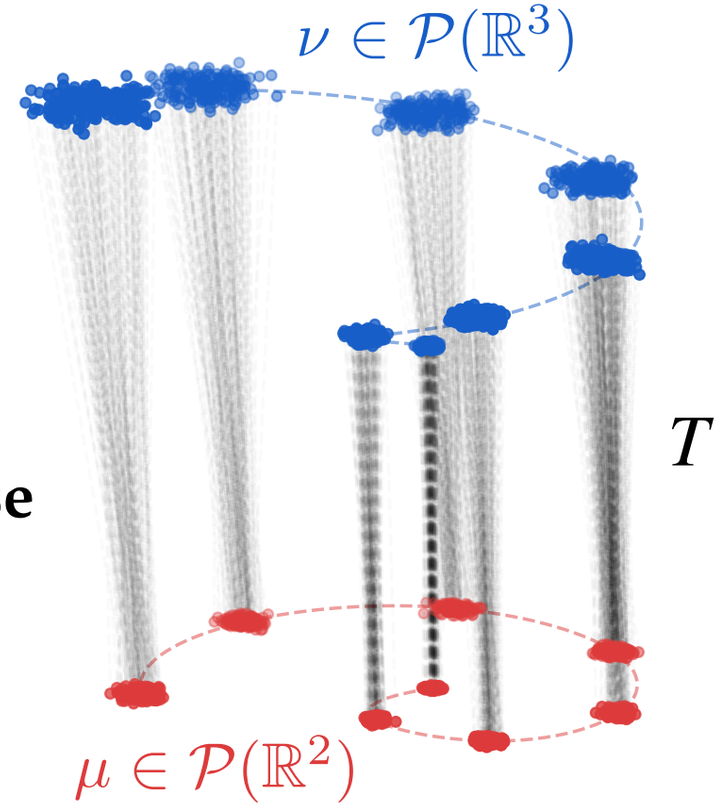
Quadratic OT: find the plan

$$\min_{T \in \Pi(a,b)} \sum_{ijkl} L\left([C_X]_{ik}\,,\,[C_Y]_{jl}\right) T_{ij} T_{kl}$$

- L measures distorsion

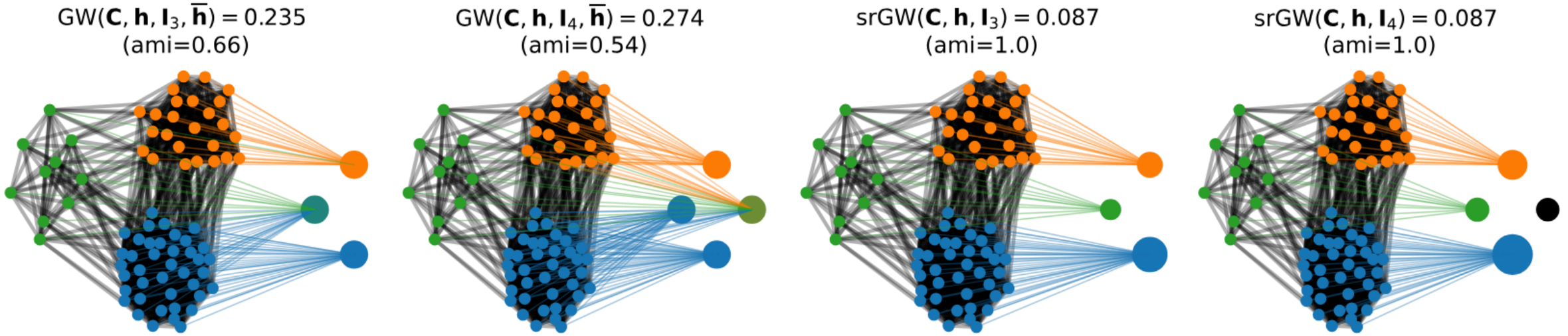
$$\left|[C_X]_{ik} - [C_Y]_{jl}\right|^2$$

- Goal : **preserving pairwise connectivity**
- Non-convex quadratic problem (NP-hard)



✦ Semi relaxed Gromov-Wasserstein

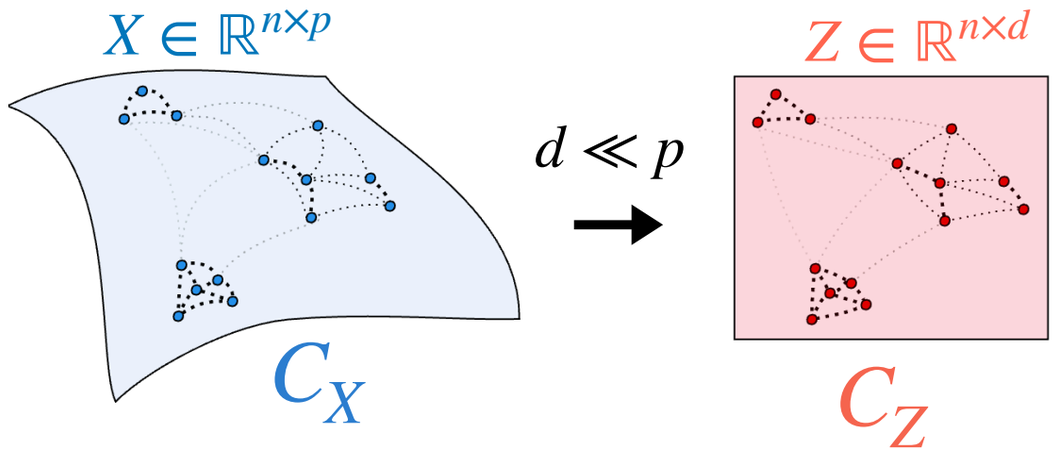
- Clustering of nodes



DR as OT in disguise

✦ Dimension reduction

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij} \, , \, [C_Z]_{ij} \right)$$



DR as OT in disguise

✦ Dimension reduction

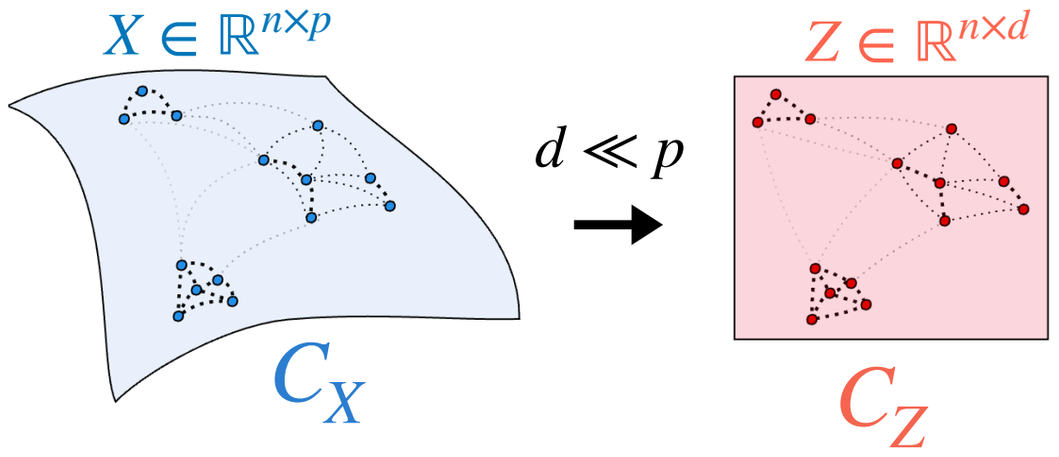
$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij} \, , \, [C_Z]_{ij} \right)$$

equiv

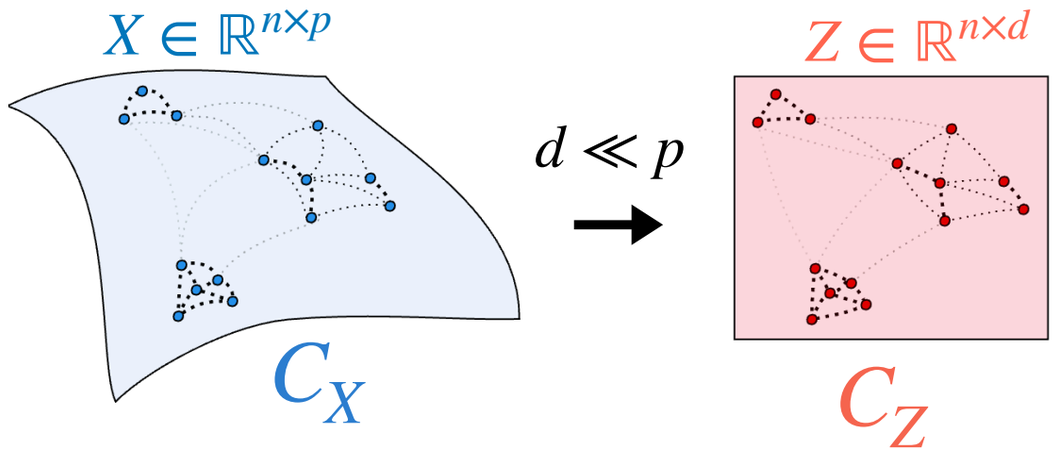
Permutation equivariance

$$\forall P, C_{PZ} = P C_Z P^\top$$

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{\sigma \in S_n} \sum_{i,j=1}^{n} L\left([C_X]_{ij} \, , \, [C_Z]_{\sigma(i)\sigma(j)} \right)$$



DR as OT in disguise



✦ Dimension reduction

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{ij}\right)$$

equiv

Permutation equivariance

$$\forall P, C_{PZ} = P C_Z P^\top$$

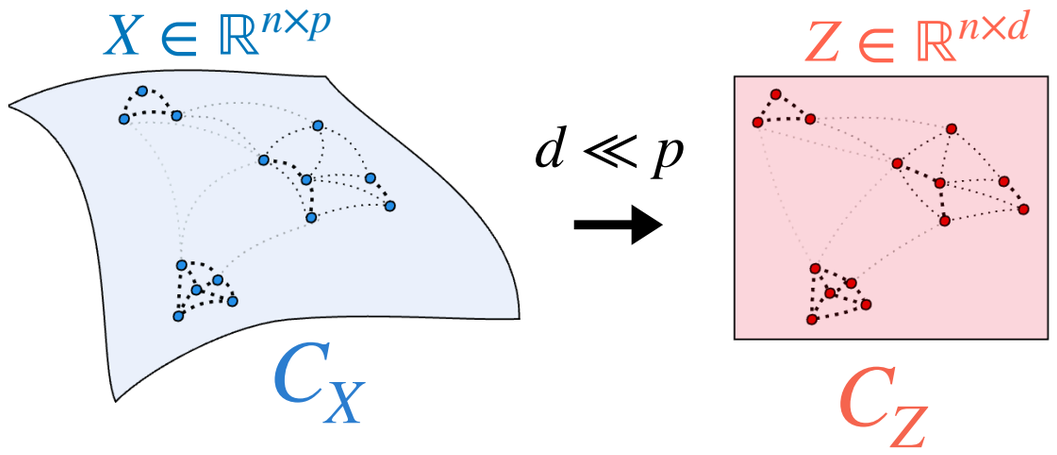
$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{\sigma \in S_n} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{\sigma(i)\sigma(j)}\right)$$

equiv

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{P} \sum_{i,j,k,l=1}^{n} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) P_{ij} P_{kl}$$

Gromov-Monge

DR as OT in disguise



✦ Dimension reduction

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{ij}\right)$$

equiv

Permutation equivariance

$$\forall P, C_{PZ} = P C_Z P^\top$$

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{\sigma \in S_n} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{\sigma(i)\sigma(j)}\right)$$

equiv

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{P} \sum_{i,j,k,l=1}^{n} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) P_{ij} P_{kl}$$

Gromov-Monge

?

✦ Gromov-Wasserstein projection

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{T \in \Pi(\frac{1_n}{n}, \frac{1_n}{n})} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) T_{ij} T_{kl}$$

Gromov-Wasserstein

DR as OT in disguise

✦ Dimension reduction

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{ij}\right)$$

equiv

Permutation equivariance

$$\forall P, C_{PZ} = P C_Z P^\top$$

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{\sigma \in S_n} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{\sigma(i)\sigma(j)}\right)$$

equiv

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{P} \sum_{i,j,k,l=1}^{n} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) P_{ij} P_{kl}$$

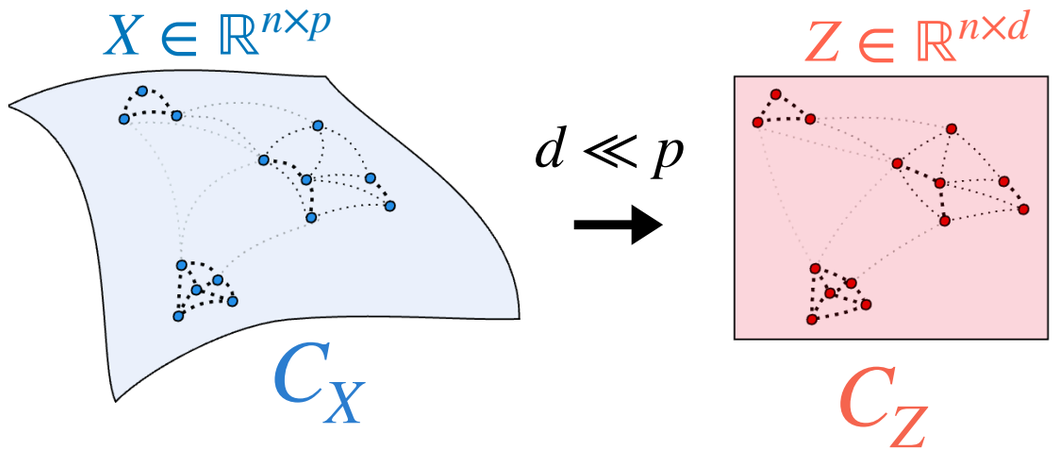
Gromov-Monge

?

✦ Gromov-Wasserstein projection

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{T \in \Pi(\frac{1_n}{n}, \frac{1_n}{n})} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) T_{ij} T_{kl}$$

Gromov-Wasserstein



✦ Equivalence holds for

Spectral methods

✦ C_X any matrix, $L = |\cdot|^2$, $C_Z = ZZ^\top$

DR as OT in disguise

✦ Dimension reduction

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{ij}\right)$$

equiv

Permutation equivariance

$$\forall P, C_{PZ} = P C_Z P^\top$$

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{\sigma \in S_n} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{\sigma(i)\sigma(j)}\right)$$

equiv

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{P} \sum_{i,j,k,l=1}^{n} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) P_{ij} P_{kl}$$

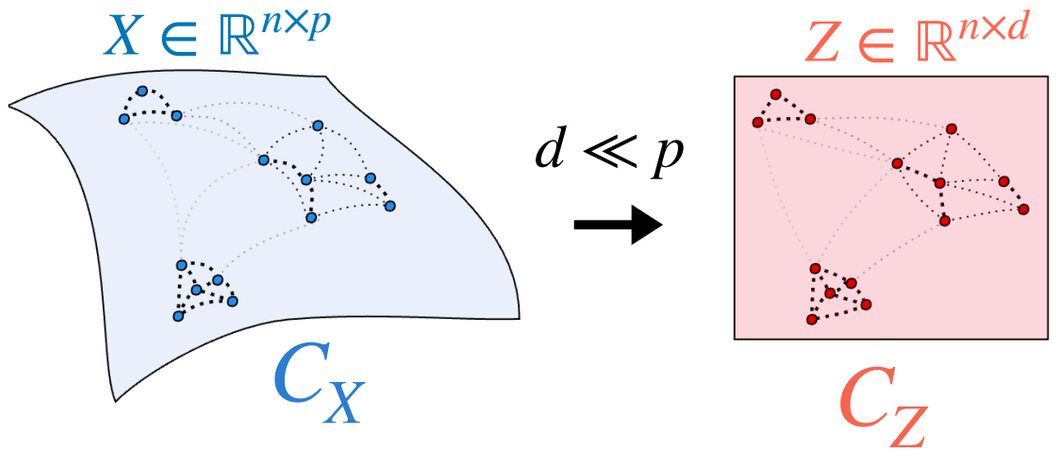
Gromov-Monge

?

✦ Gromov-Wasserstein projection

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{T \in \Pi(\frac{1_n}{n}, \frac{1_n}{n})} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) T_{ij} T_{kl}$$

Gromov-Wasserstein



✦ Equivalence holds for

Spectral methods

✦ C_X any matrix, $L = |\cdot|^2$, $C_Z = ZZ^\top$

A is CPD : $\forall x$ **s.t.** $x^\top 1 = 0,\ x^\top A x \geq 0$

DR as OT in disguise

✦ Dimension reduction

$$\min_{Z \in \mathbb{R}^{n \times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{ij}\right)$$

equiv

Permutation equivariance

$$\forall P, C_{PZ} = P C_Z P^\top$$

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{\sigma \in S_n} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{\sigma(i)\sigma(j)}\right)$$

equiv

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{P} \sum_{i,j,k,l=1}^{n} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) P_{ij} P_{kl}$$

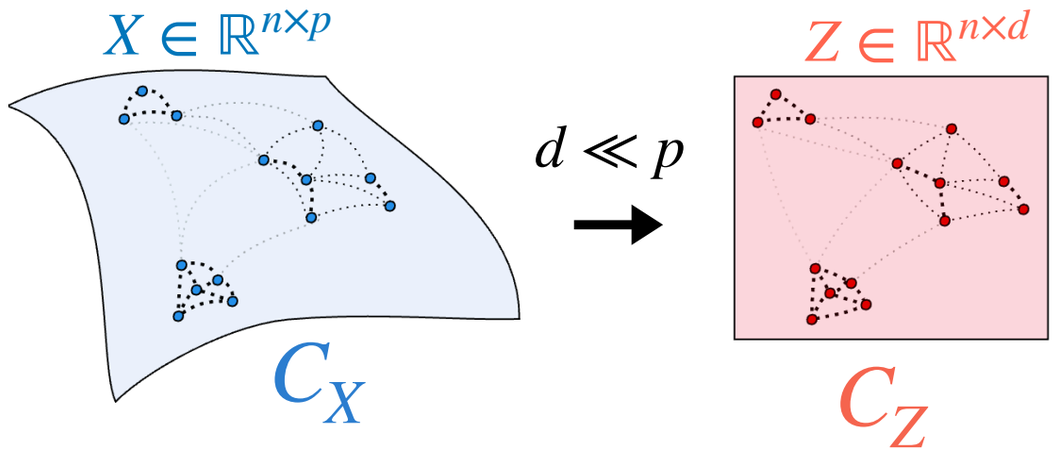
Gromov-Monge

?

✦ Gromov-Wasserstein projection

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{T \in \Pi(\frac{1_n}{n}, \frac{1_n}{n})} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) T_{ij} T_{kl}$$

Gromov-Wasserstein



✦ Equivalence holds for

Spectral methods

✦ C_X any matrix, $L = |\cdot|^2$, $C_Z = ZZ^\top$

A is CPD : $\forall x$ **s.t.** $x^\top 1 = 0,\ x^\top A x \geq 0$

Neighbor embedding methods

✦ C_X is CPD, $L = KL$

$C_Z = \mathrm{diag}(\alpha_Z)\ K_Z\ \mathrm{diag}(\beta_Z)$

where $\log(K_Z)$ is CPD

DR as OT in disguise

✦ Dimension reduction

$$\min_{Z\in\mathbb{R}^{n\times d}} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{ij}\right)$$

equiv

Permutation equivariance

$$\forall P, C_{PZ} = PC_ZP^\top$$

$$\min_{Z\in\mathbb{R}^{n\times d}} \min_{\sigma\in S_n} \sum_{i,j=1}^{n} L\left([C_X]_{ij}\ ,\ [C_Z]_{\sigma(i)\sigma(j)}\right)$$

equiv

$$\min_{Z\in\mathbb{R}^{n\times d}} \min_{P} \sum_{i,j,k,l=1}^{n} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) P_{ij}P_{kl}$$

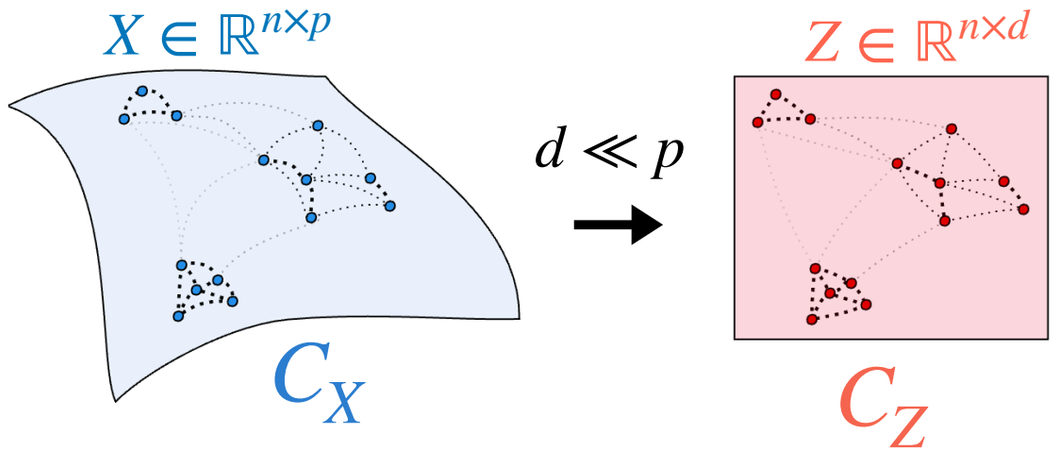
Gromov-Monge

?

✦ Gromov-Wasserstein projection

$$\min_{Z\in\mathbb{R}^{n\times d}} \min_{T\in\Pi(\frac{1_n}{n},\frac{1_n}{n})} \sum_{ijkl} L\left([C_X]_{ik}\ ,\ [C_Z]_{jl}\right) T_{ij}T_{kl}$$

Gromov-Wasserstein



✦ Equivalence holds for

Spectral methods

✦ C_X any matrix, $L = |\cdot|^2$, $C_Z = ZZ^\top$

A is CPD : $\forall x$ **s.t.** $x^\top 1 = 0,\ x^\top Ax \geq 0$

Neighbor embedding methods

✦ C_X is CPD, $L = KL$

$$C_Z = \mathrm{diag}(\alpha_Z)\ K_Z\ \mathrm{diag}(\beta_Z)$$

where $\log(K_Z)$ is CPD

e.g. $K_Z = \exp(-\|\mathbf{z}_i - \mathbf{z}_j\|_2^2)$
and its usual normalizations

$$1_n^\top K_Z 1_n = 1,\ K_Z 1_n = 1_n,\ K_Z^\top 1_n = 1_n$$

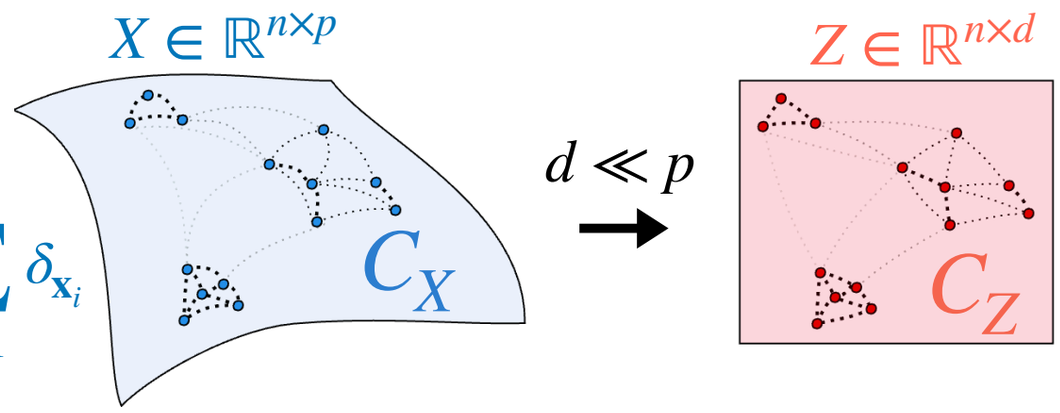
+ $K_Z 1_n = 1_n$

(Sinkhorn & Knopp, 1967)

Beware that C_X is not always CPD.

Distributional Reduction

$$\hat{\mu}_n = \sum_{i=1}^{n} \delta_{\mathbf{x}_i}$$



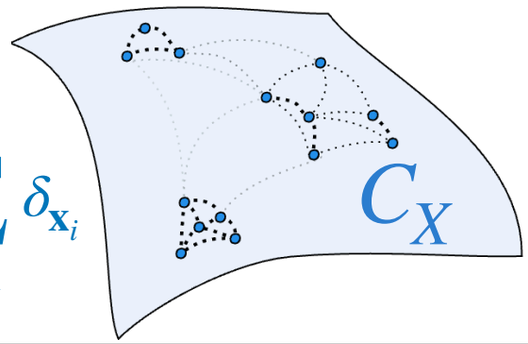
$d \ll p$

$$\sim \quad \min_{\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{z}_i}} \mathrm{GW}_L(\hat{\mu}_n, \nu)$$

Distributional Reduction

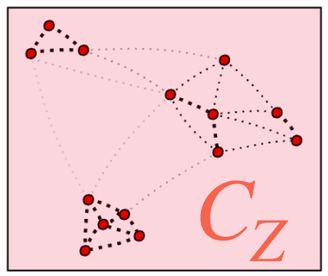
$X \in \mathbb{R}^{n \times p}$

$\hat{\mu}_n = \sum_{i=1}^{n} \delta_{\mathbf{x}_i}$



$d \ll p$

$Z \in \mathbb{R}^{n \times d}$



$\sim$ $\min_{\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{z}_i}} \mathrm{GW}_L(\hat{\mu}_n, \nu)$

$X \in \mathbb{R}^{n \times p}$

$k \ll n$

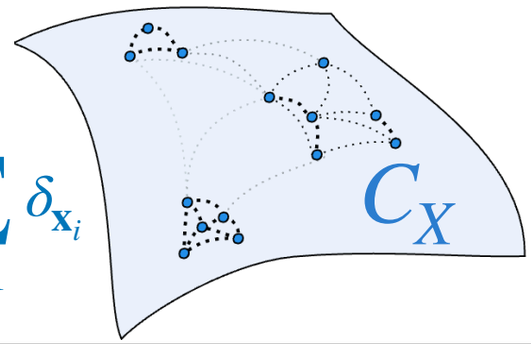
$d \ll p$

$Z \in \mathbb{R}^{k \times d}$

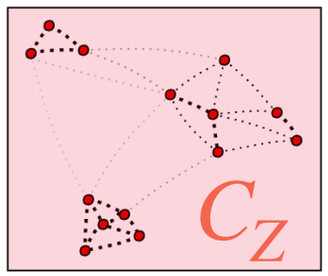
Distributional Reduction

$X \in \mathbb{R}^{n \times p}$ $\quad$ $Z \in \mathbb{R}^{n \times d}$

$\hat{\mu}_n = \sum_{i=1}^{n} \delta_{\mathbf{x}_i}$

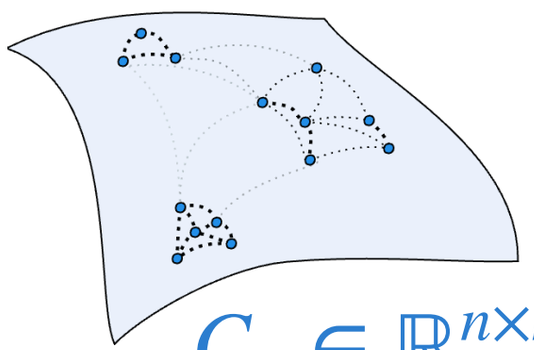


$d \ll p$



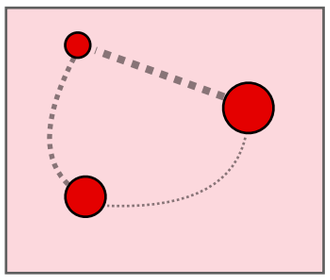
$\sim$ $\min_{\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{z}_i}} \mathrm{GW}_L(\hat{\mu}_n, \nu)$

$X \in \mathbb{R}^{n \times p}$ $\quad$ $Z \in \mathbb{R}^{k \times d}$



$k \ll n$

$d \ll p$



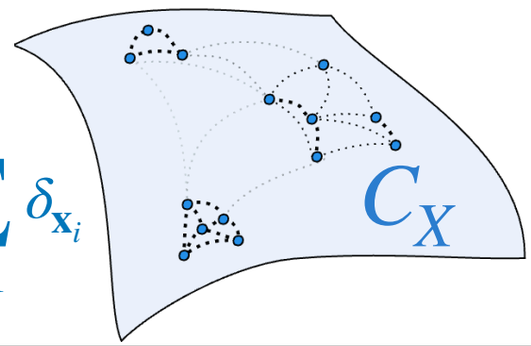
$C_X \in \mathbb{R}^{n \times n}$ $\quad$ $C_Z \in \mathbb{R}^{k \times k}$

✦ **GW projection**

$$\min_{\nu \in \mathscr{P}_k(\mathbb{R}^d)} \mathrm{GW}(\hat{\mu}_n, \nu) \qquad \nu = \sum_{j=1}^{k} b_j \delta_{\mathbf{z}_j}$$

Distributional Reduction

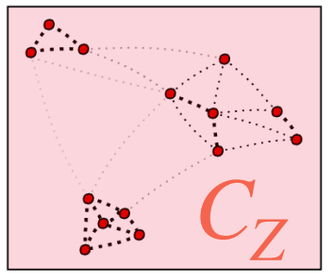
$X \in \mathbb{R}^{n \times p}$



$\hat{\mu}_n = \sum_{i=1}^{n} \delta_{\mathbf{x}_i}$

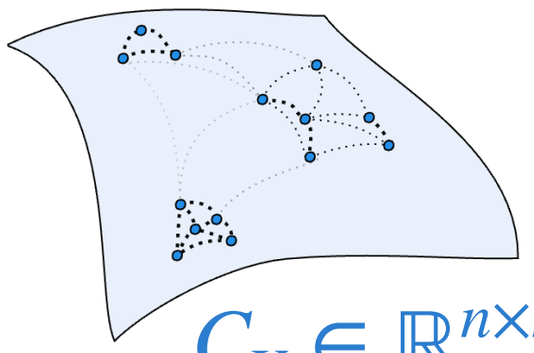
$d \ll p$ →

$Z \in \mathbb{R}^{n \times d}$



$\sim \min_{\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{z}_i}} \mathrm{GW}_L(\hat{\mu}_n, \nu)$

$X \in \mathbb{R}^{n \times p}$

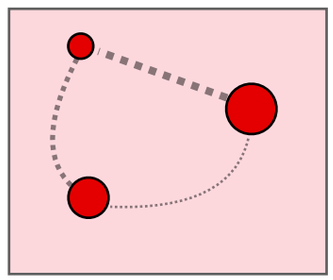


$C_X \in \mathbb{R}^{n \times n}$

$k \ll n$

$d \ll p$ →

$Z \in \mathbb{R}^{k \times d}$



$C_Z \in \mathbb{R}^{k \times k}$

✦ Optimization problem

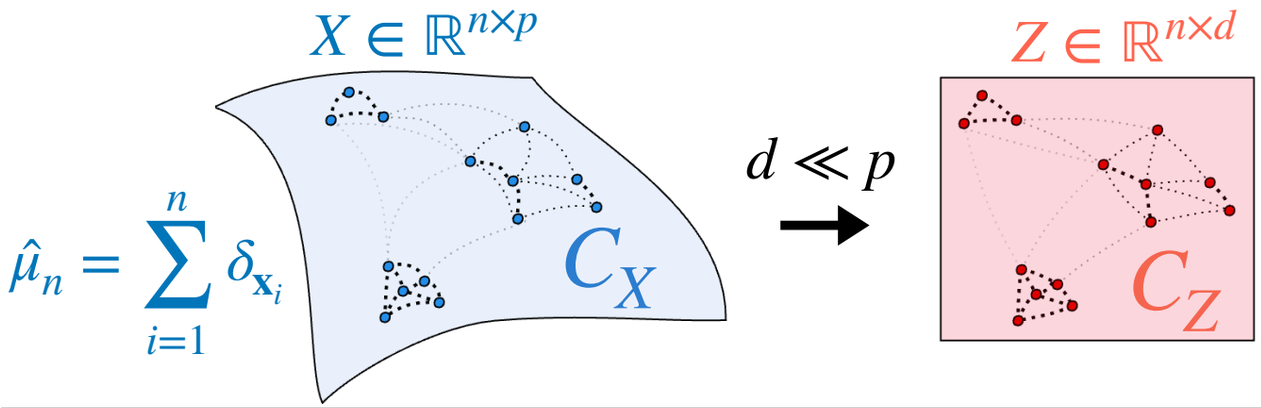
$$\min_{Z \in \mathbb{R}^{k \times d}} \min_{b \in \Sigma_k} \mathrm{GW}_L(C_X, C_Z, \frac{1_n}{n}, b)$$

- ✦ Find **few prototypes** in **low dim.**
- ✦ Find the weights / cluster size

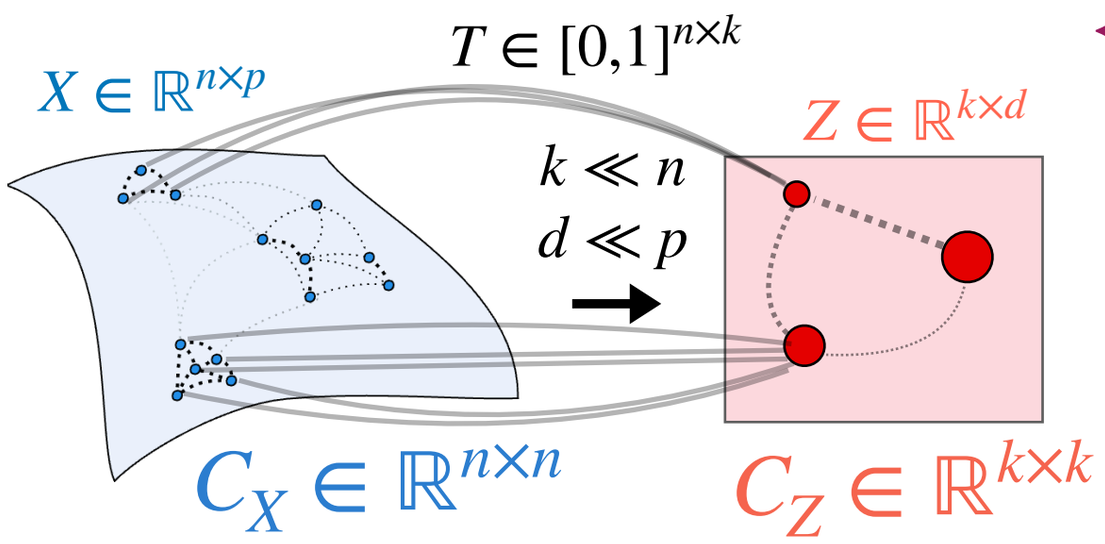
✦ GW projection

$$\min_{\nu \in \mathscr{P}_k(\mathbb{R}^d)} \mathrm{GW}(\hat{\mu}_n, \nu) \qquad \nu = \sum_{j=1}^{k} b_j \delta_{\mathbf{z}_j}$$

Distributional Reduction



$$\sim \min_{\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{z}_i}} \mathrm{GW}_L(\hat{\mu}_n, \nu)$$



✦ Optimization problem

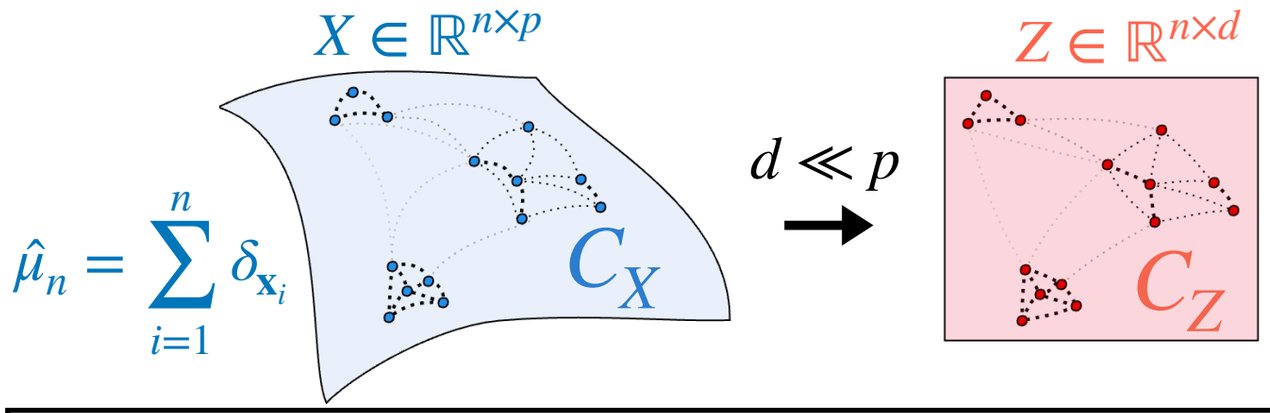
$$\min_{Z \in \mathbb{R}^{k \times d}} \min_{b \in \Sigma_k} \mathrm{GW}_L(C_X, C_Z, \frac{1_n}{n}, b)$$

- ✦ Find **few prototypes** in **low dim.**
- ✦ Find the weights / cluster size
- ✦ **Clustering via the coupling** T (soft-assignment)
- ✦ Sufficient conditions for hard assignment (see paper)

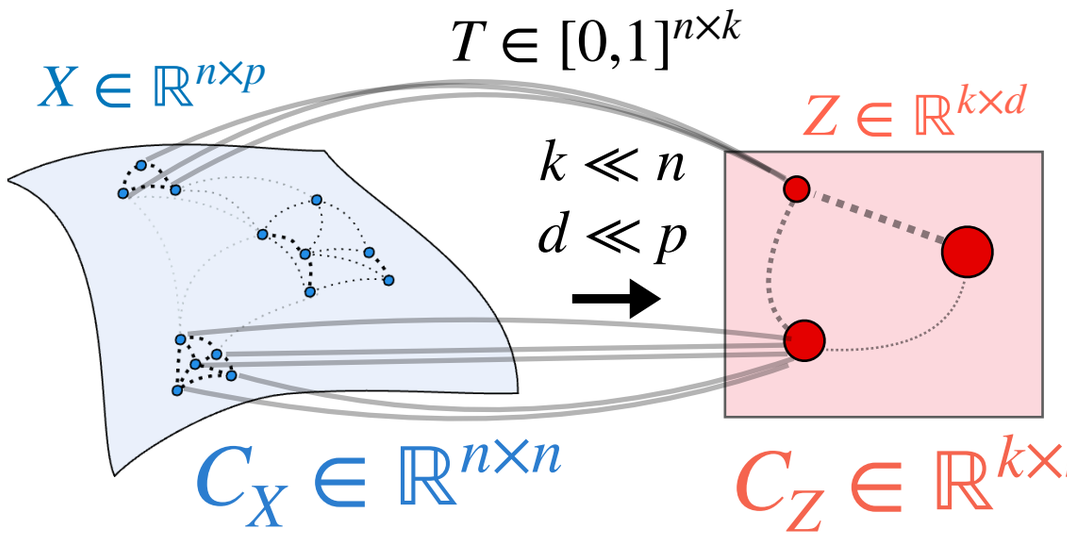
✦ GW projection

$$\min_{\nu \in \mathscr{P}_k(\mathbb{R}^d)} \mathrm{GW}(\hat{\mu}_n, \nu) \qquad \nu = \sum_{j=1}^{k} b_j \delta_{\mathbf{z}_j}$$

Distributional Reduction



$$\sim \quad \min_{\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{z}_i}} \mathrm{GW}_L(\hat{\mu}_n, \nu)$$



✦ Optimization problem

$$\min_{Z \in \mathbb{R}^{k \times d}} \min_{b \in \Sigma_k} \mathrm{GW}_L(C_X, C_Z, \frac{1_n}{n}, b)$$

- ✦ Find **few prototypes** in **low dim.**
- ✦ Find the weights / cluster size
- ✦ **Clustering via the coupling** T (soft-assignment)
- ✦ Sufficient conditions for hard assignment (see paper)

✦ GW projection

$$\min_{\nu \in \mathscr{P}_k(\mathbb{R}^d)} \mathrm{GW}(\hat{\mu}_n, \nu) \qquad \nu = \sum_{j=1}^{k} b_j \delta_{\mathbf{z}_j}$$

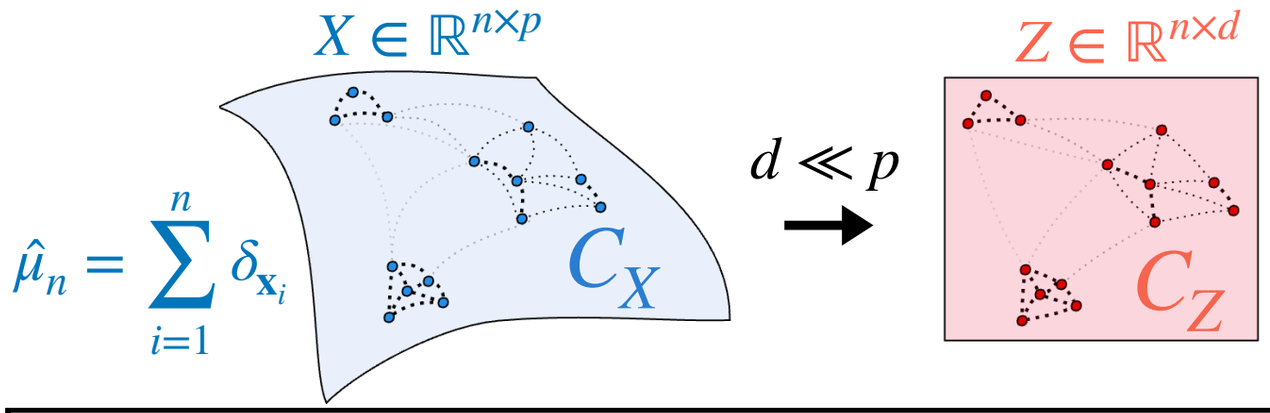
✦ A semi-relaxed objective

(Vincent-Cuaz et al., 2022)

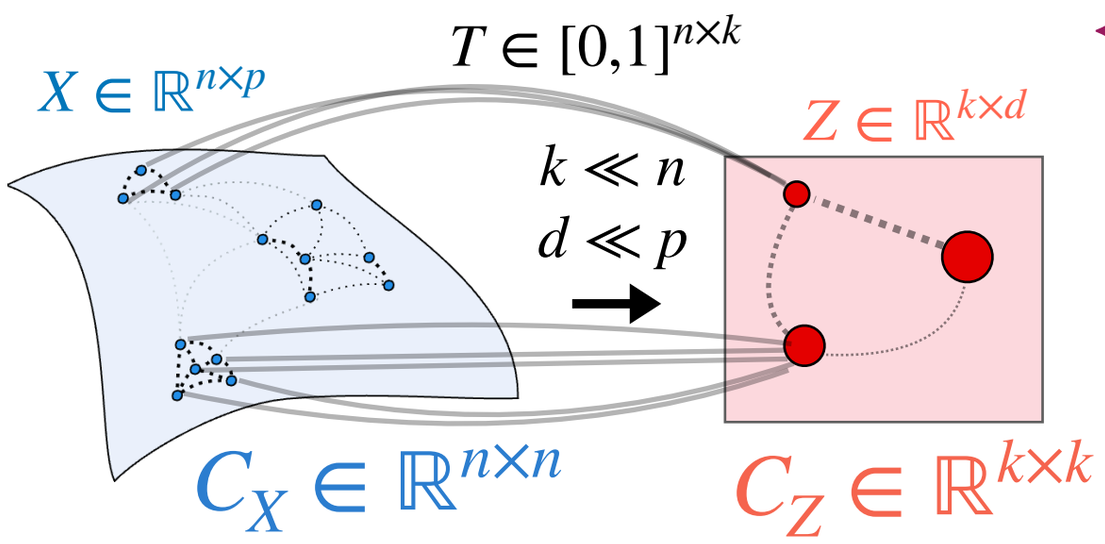
$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{T : T1_k = \frac{1_n}{n}} \sum_{ijkl} L\left([C_X]_{ik}, [C_Z]_{jl}\right) T_{ij} T_{kl}$$

→ easier than GW

Distributional Reduction



$$\sim \quad \min_{\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{z}_i}} \mathrm{GW}_L(\hat{\mu}_n, \nu)$$



✦ Optimization problem

$$\min_{Z \in \mathbb{R}^{k \times d}} \min_{b \in \Sigma_k} \mathrm{GW}_L(C_X, C_Z, \frac{1_n}{n}, b)$$

- ✦ Find **few prototypes** in **low dim.**
- ✦ Find the weights / cluster size
- ✦ **Clustering via the coupling** T (soft-assignment)
- ✦ Sufficient conditions for hard assignment (see paper)

✦ GW projection

$$\min_{\nu \in \mathscr{P}_k(\mathbb{R}^d)} \mathrm{GW}(\hat{\mu}_n, \nu) \qquad \nu = \sum_{j=1}^{k} b_j \delta_{\mathbf{z}_j}$$

✦ A semi-relaxed objective

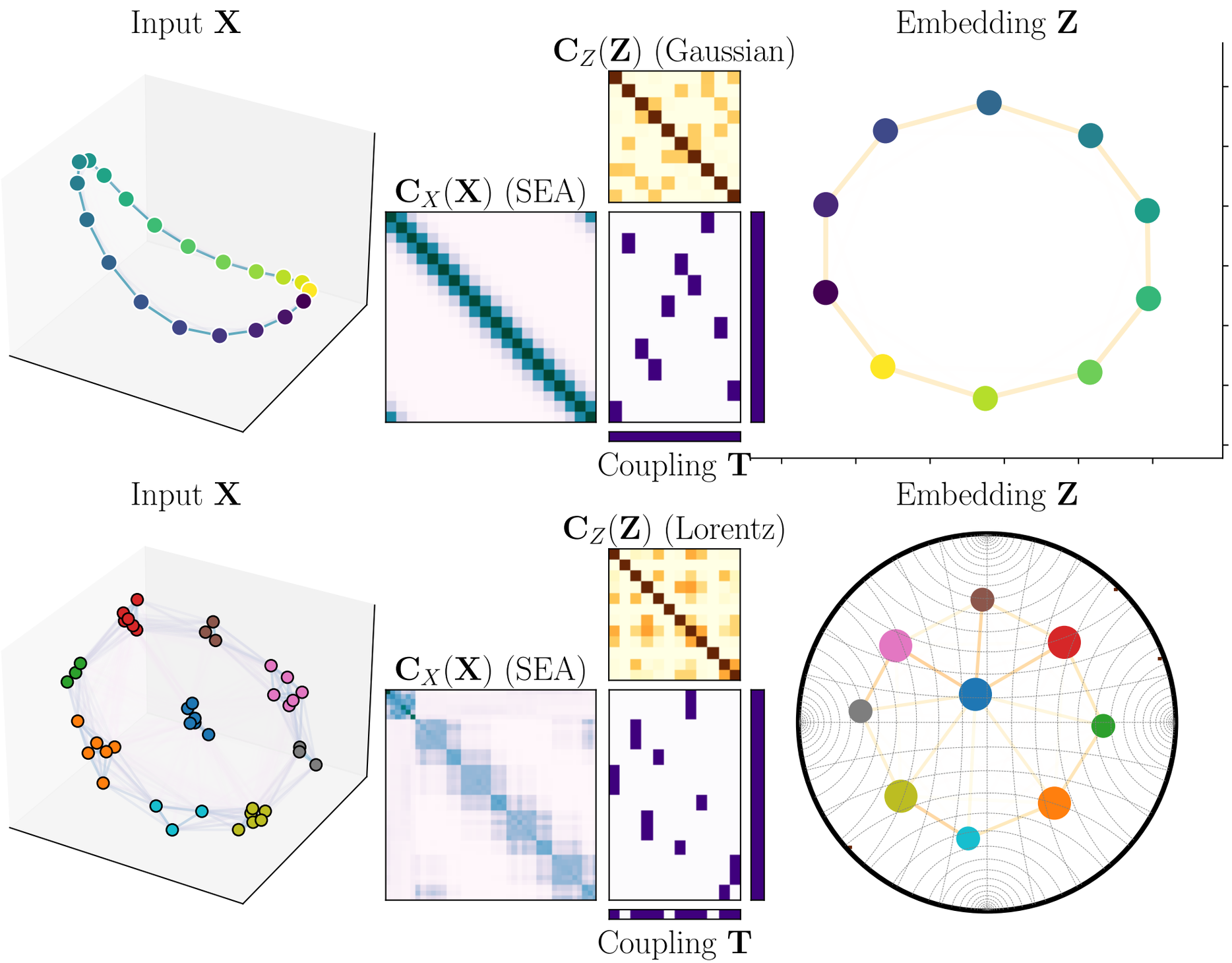
(Vincent-Cuaz et al., 2022)

$$\min_{Z \in \mathbb{R}^{n \times d}} \min_{T : T1_k = \frac{1_n}{n}} \sum_{ijkl} L\left([C_X]_{ik}, [C_Z]_{jl}\right) T_{ij} T_{kl}$$

→ easier than GW

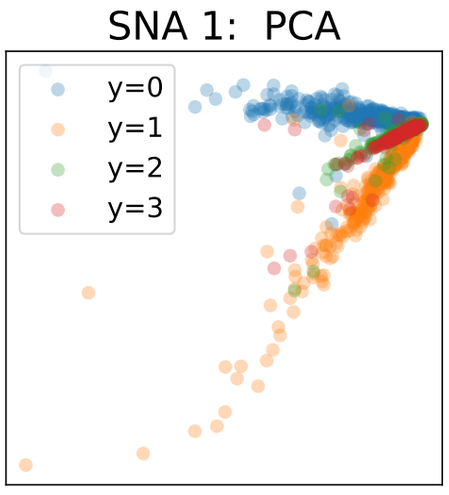
- ✦ Non-convex problem
- ✦ BCD: alternates optim in Z, in T
- ✦ Optim in T: CG solver in $O(n^2 k)$ for $L \in \{\mathrm{KL}, |\cdot|^2\}$
- ✦ With low-rank structures $O(nkr + n^2)$

Distributional Reduction



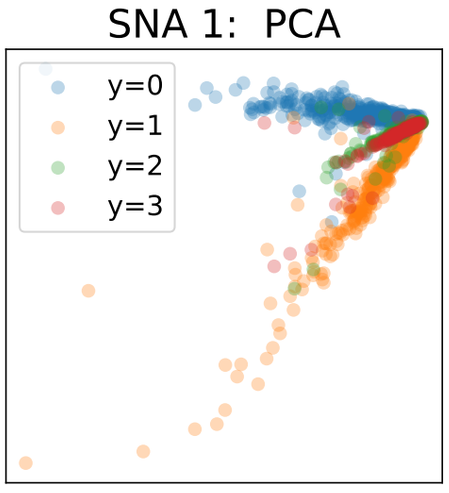
Distributional Reduction

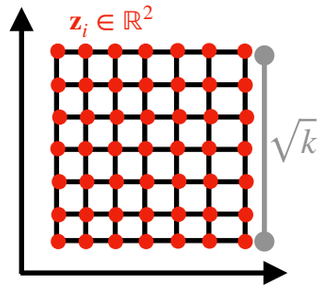
✦ **Single-cell dataset** (Chen et al., 2019) Only solve $\min_{b \in \Sigma_k} \mathrm{GW}_L(C_X, C_Z, \frac{1_n}{n}, b)$



Distributional Reduction

✦ **Single-cell dataset** (Chen et al., 2019) Only solve $\min_{b \in \Sigma_k} \mathrm{GW}_L(C_X, C_Z, \frac{1_n}{n}, b)$ with $C_X = XX^\top$ and $C_Z = ZZ^\top$ **fixed on a grid** $Z \in \mathbb{R}^{k \times 2}$

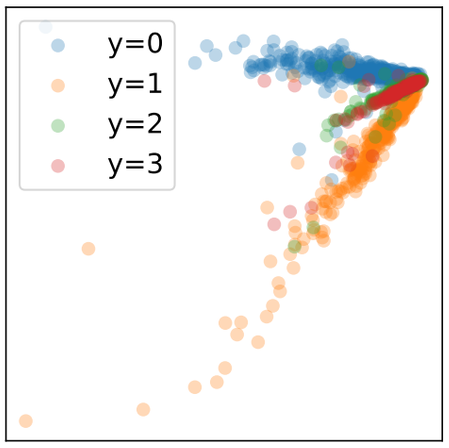




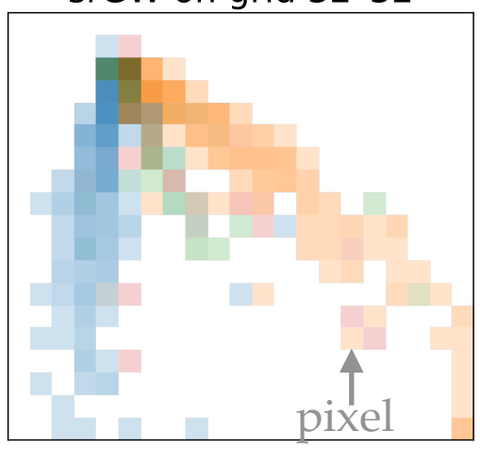
Distributional Reduction

✦ **Single-cell dataset** (Chen et al., 2019) Only solve $\min_{b \in \Sigma_k} \mathrm{GW}_L(C_X, C_Z, \frac{1_n}{n}, b)$ with $C_X = XX^\top$ and $C_Z = ZZ^\top$ **fixed on a grid** $Z \in \mathbb{R}^{k \times 2}$

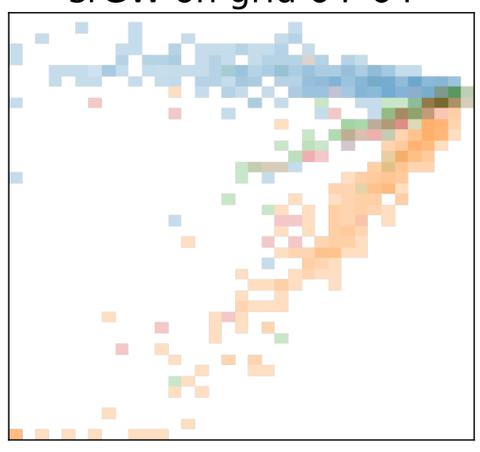
SNA 1: PCA



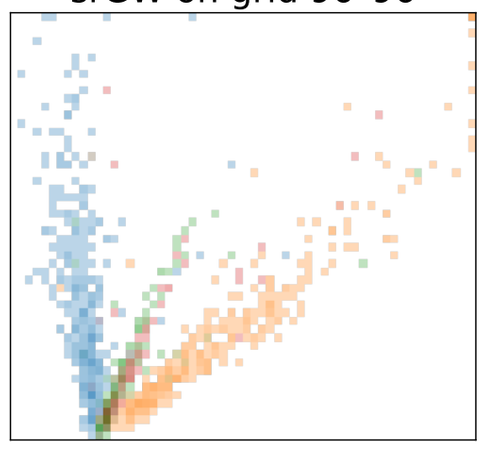
srGW on grid 32*32



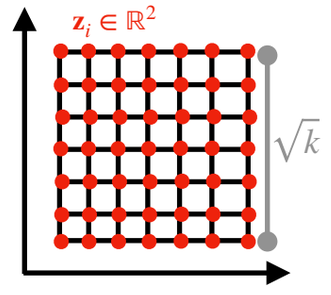
srGW on grid 64*64



srGW on grid 96*96



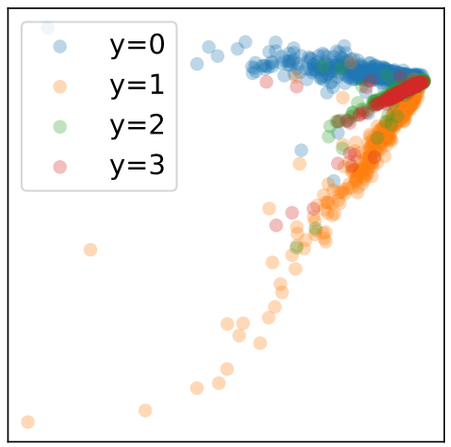
dataset as an image

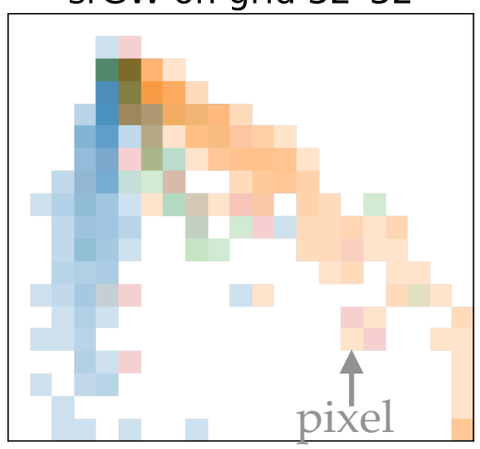


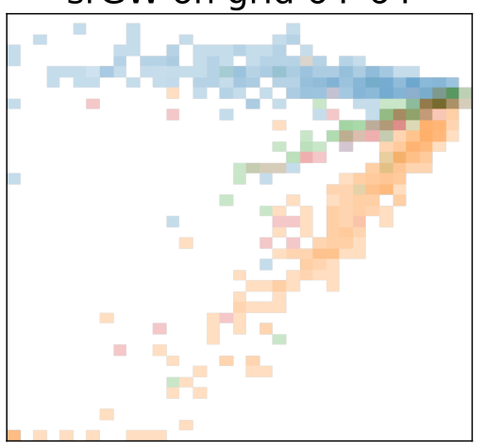
Distributional Reduction

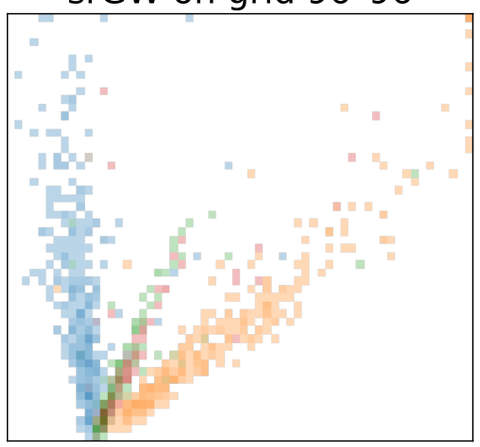
✦ Single-cell dataset (Chen et al., 2019)

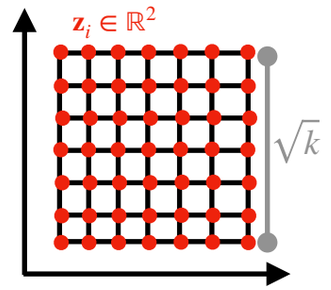
Only solve $\min_{b \in \Sigma_k} \mathrm{GW}_L(C_X, C_Z, \frac{1_n}{n}, b)$ with $C_X = XX^\top$ and $C_Z = ZZ^\top$ **fixed on a grid** $Z \in \mathbb{R}^{k \times 2}$







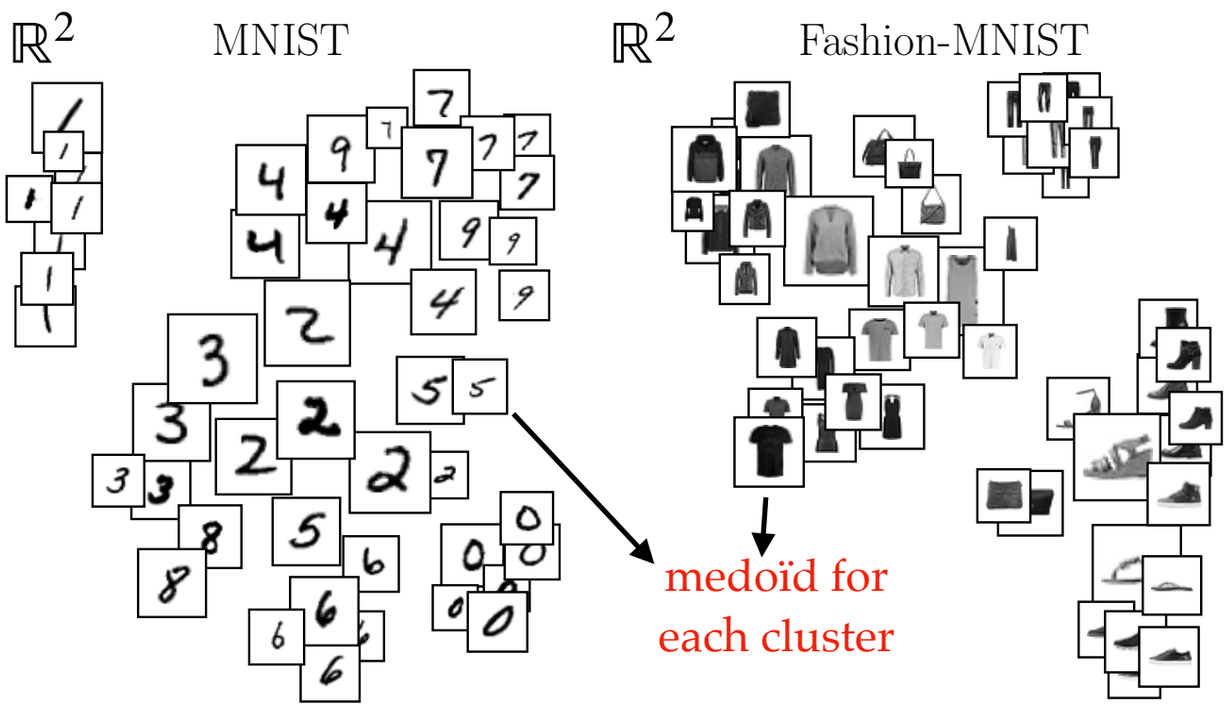




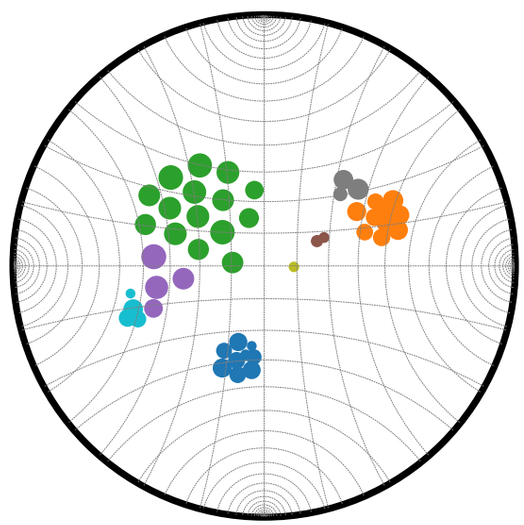
dataset as an image

✦ Image datasets

C_X symmetric entropic aff. (Van Assel et al., 2023) C_Z Student t-kernel

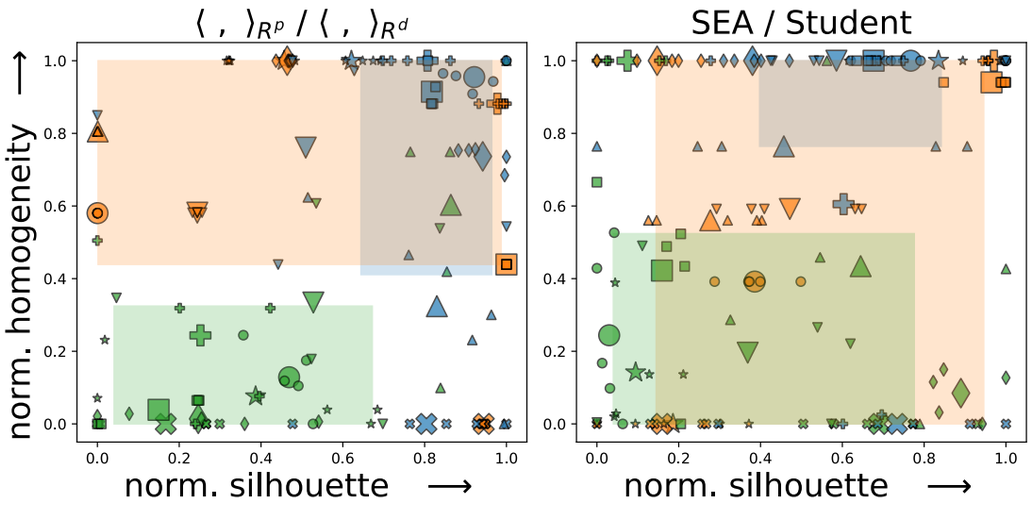


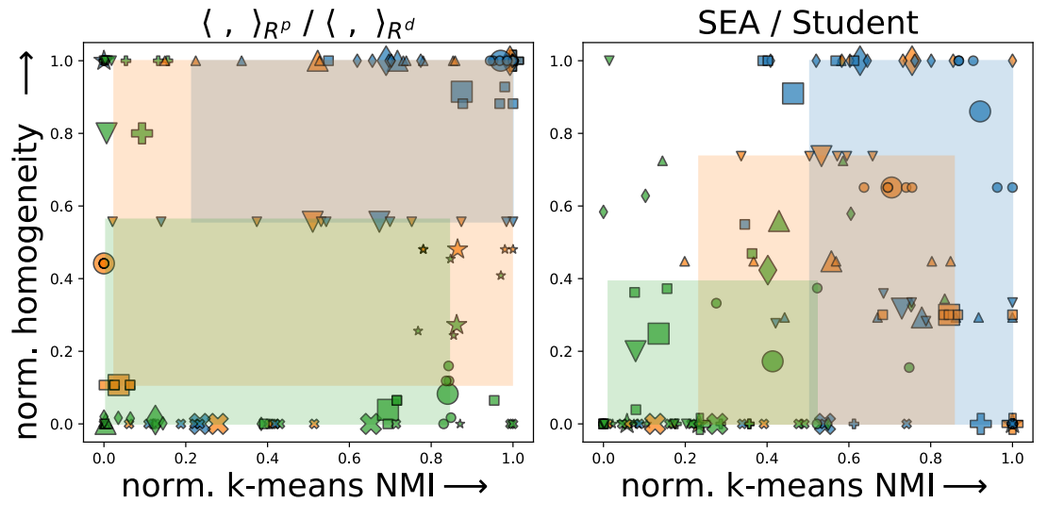
Hyperbolic geometry

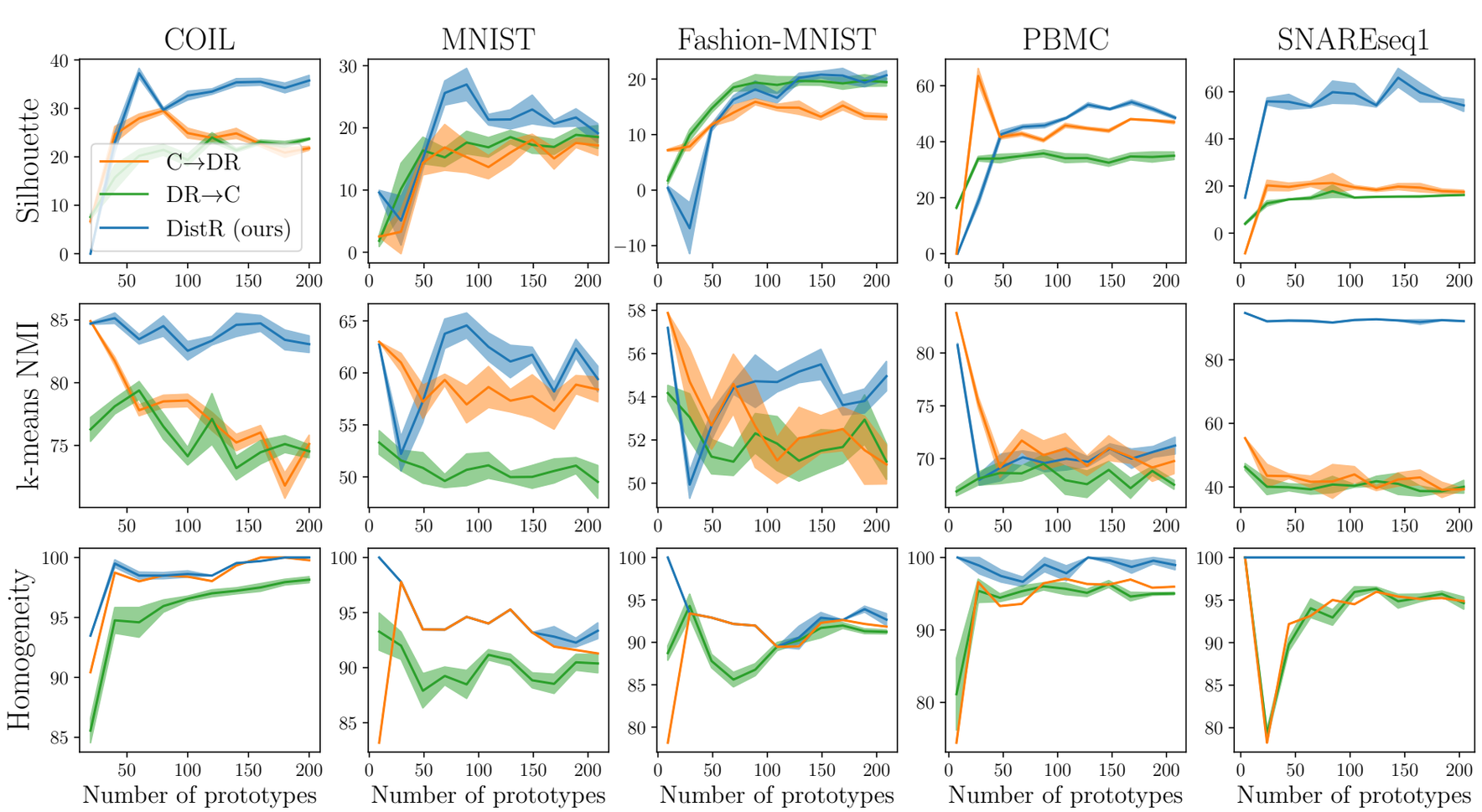


Distributional Reduction

✦ Comparison with DR then clustering or clustering then DR







Thank you!

Open-source implementations are available here:

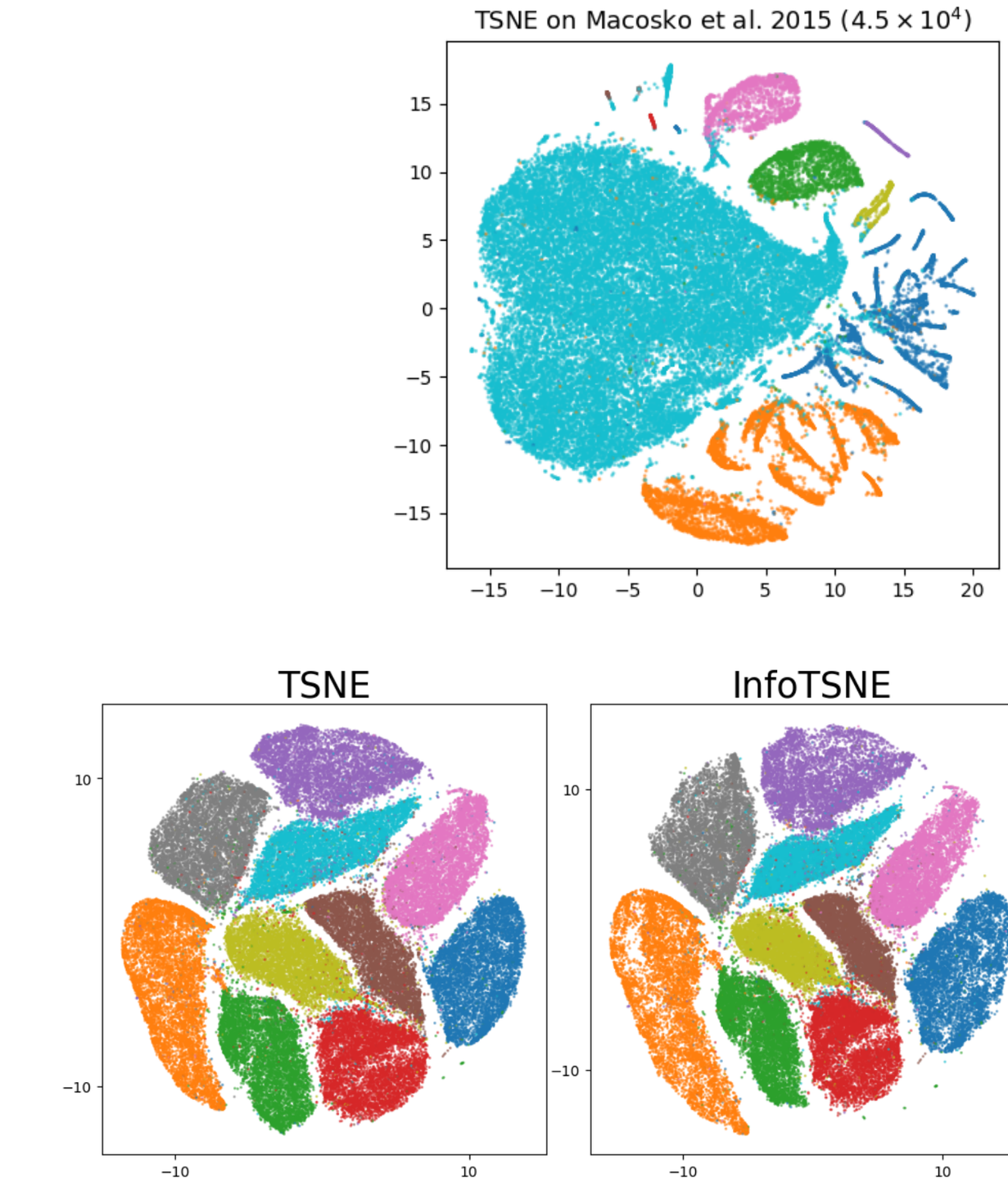
TorchDR

```python
from sklearn.datasets import fetch_openml
from torchdr import PCA, TSNE

x = fetch_openml("mnist_784").data.astype("float32")

x_ = PCA(n_components=50).fit_transform(x)
z = TSNE(perplexity=30).fit_transform(x_)
```

Modularity	All of it is written in **python** in a **highly modular** way, making it easy to create or transform components.
Speed	Supports **GPU acceleration**, leverages **sparsity** and **batching** strategies with **contrastive learning** techniques.
Memory efficiency	Relies on **sparsity** and/or `pykeops` [19] symbolic tensors to **avoid memory overflows**.
Compatibility	Implemented methods are fully **compatible** with the `sklearn` [21] API and `torch` [20] ecosystem.

Thank you!

TorchDR



Appendix

Solving OT

A linear problem

Discrete probability measures

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

Linear Program:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \sum_{ij} c_{i,j} \pi_{i,j} = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \boldsymbol{\pi} \rangle$$

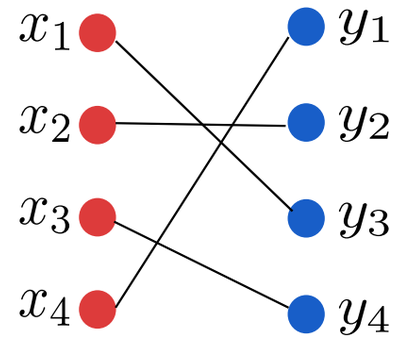
Simplex, Network flow, Hungarian algorithms $\sim O(n^3 \log(n))$

Uniform weights

$$\mathbf{a} = \mathbf{b} = \frac{\mathbf{1}_n}{n}$$

Monge Problem

$$\min_{\sigma \in S_n} \sum_{i=1}^{n} c_{i,\sigma(i)}$$



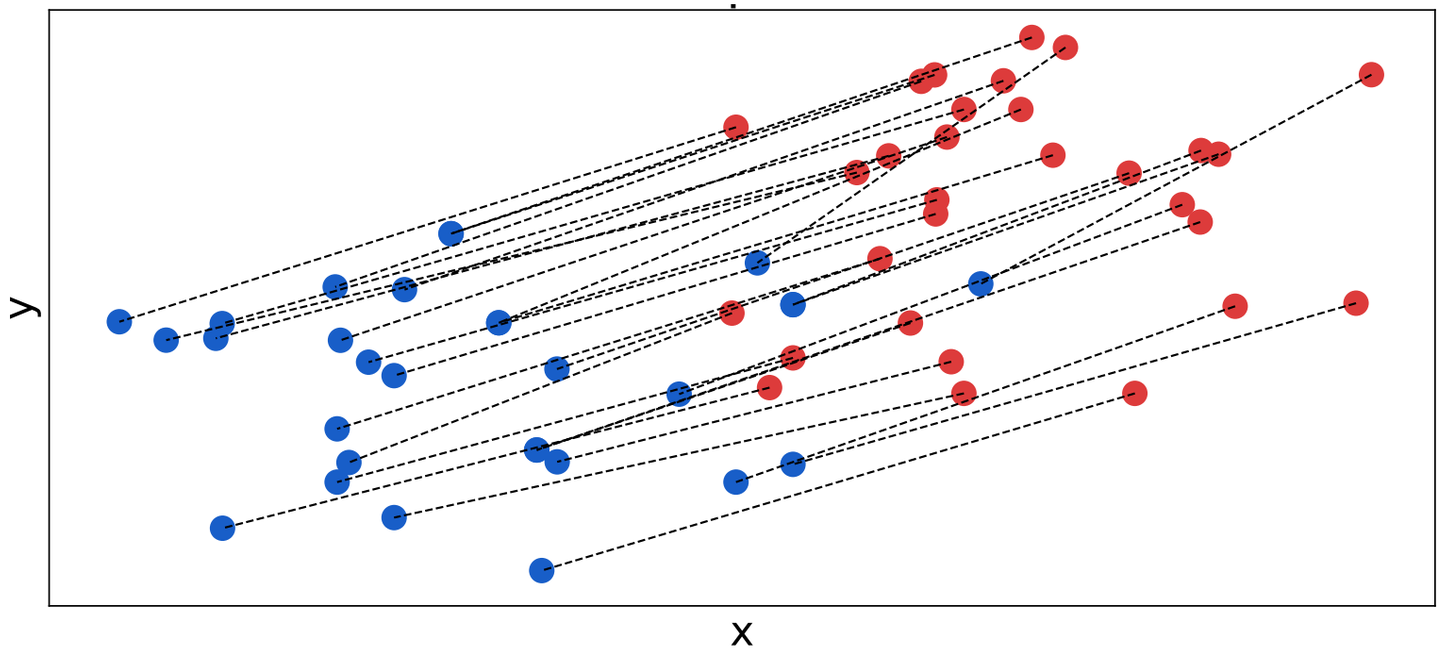
One-to-one

Fundamental theorem LP:

$$\boldsymbol{\pi}^* \leftrightarrow \sigma^* \in S_n$$

Optimal coupling is a permutation

Solves the Monge Problem



Solving OT

Entropic regularization

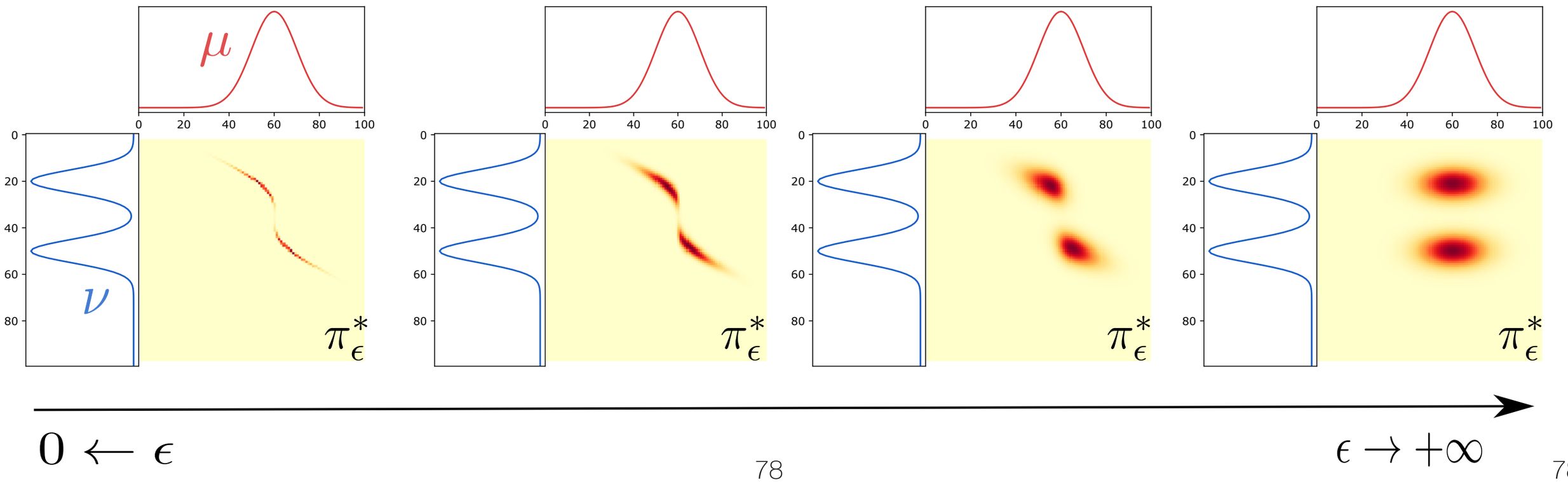
Discrete probability measures

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

Strongly convex problem:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \boldsymbol{\pi} \rangle - \varepsilon H(\boldsymbol{\pi})$$

- Entropy term $H(\boldsymbol{\pi}) = -\sum_{ij} (\log(\pi_{ij}) - 1) \pi_{ij}$
- Sinkhorn-Knopp algorithm: 1) fast 2) based on matrix multiplication
- Complexity $O(n^2)$



78

Solving OT

Computing GW

Solving FGW: a non convex QP

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \sum_{ijkl} |C_1(i,k) - C_2(j,l)|^p \pi_{ij} \pi_{kl}$$

Quadratic function over polytope -> Conditional Gradient algorithm (a.k.a Frank-Wolfe)

Non convex but converges to a **local optimal solution** [Lacoste-Julien 2016]

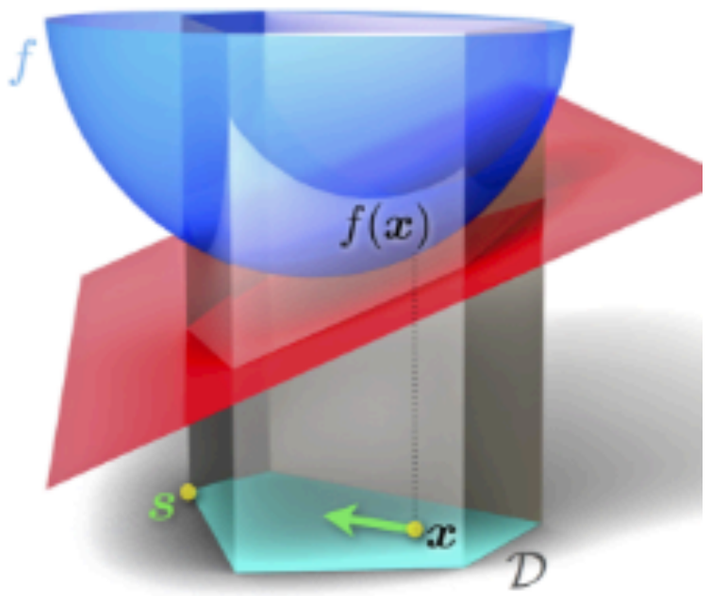
Find a **sparse** solution. FW gap = $O(\frac{1}{\sqrt{n_{iter}}})$

Algorithm 1 Conditional Gradient (CG) for FGW

$\pi^{(0)} \leftarrow \mathbf{h}\mathbf{g}^\top$
for $i = 1, \ldots,$ **do**
  $\mathbf{G} \leftarrow$ Gradient from GW loss $w.r.t.$ $\boldsymbol{\pi}^{(i-1)}$
  $\tilde{\boldsymbol{\pi}}^{(i)} \leftarrow$ Solve OT with ground loss $\mathbf{G}$
  $\tau^{(i)} \leftarrow$ Line-search for GW loss with $\tau \in (0, 1)$ (closed-form)
  $\boldsymbol{\pi}^{(i)} \leftarrow (1 - \tau^{(i)})\boldsymbol{\pi}^{(i-1)} + \tau^{(i)}\tilde{\boldsymbol{\pi}}^{(i)}$
end for

Complexity

$$O(n_{iter}\ n^3)$$



79