SNEkhorn : Dimension Reduction with Symmetric Entropic Affinities

Hugues Van Assel PhD student ENS Lyon

Titouan Vayer Rémi Flamary Nicolas Courty

Overview of the talk

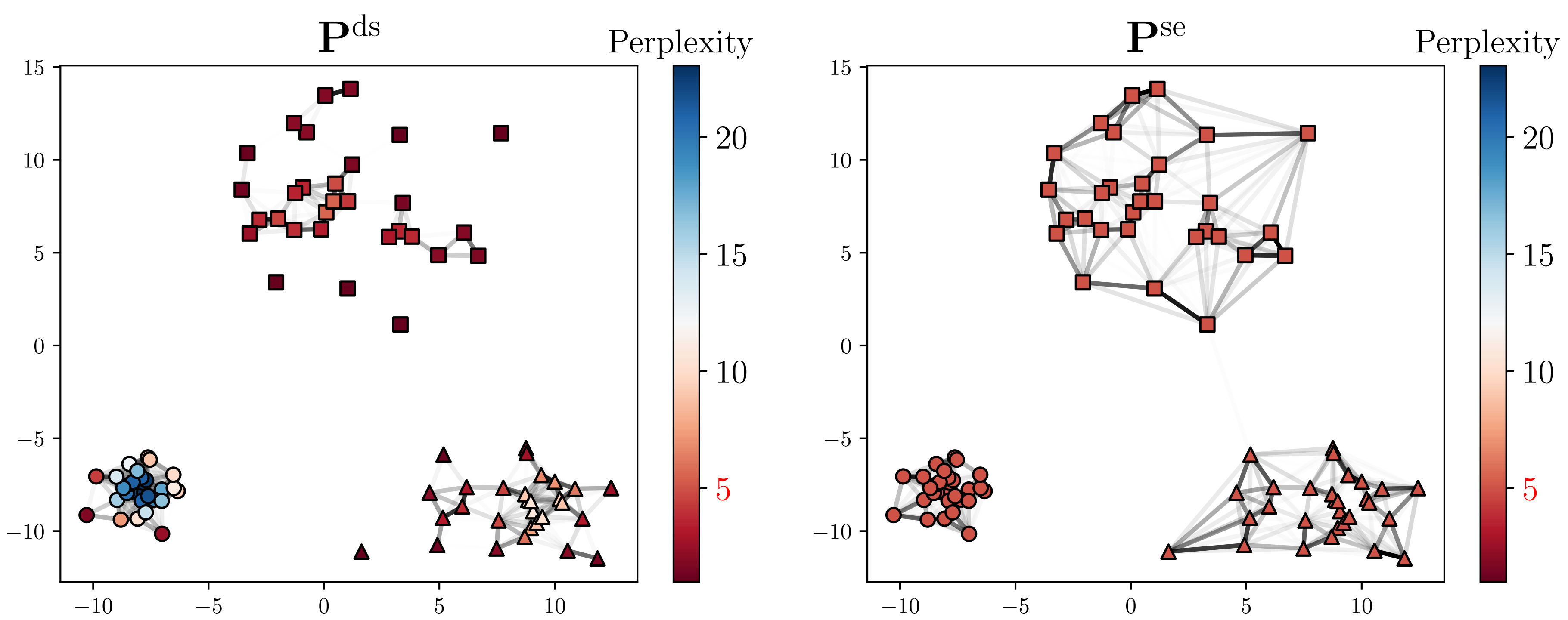
- **Part I: Symmetric Entropic Affinities**
- **Part II: Application to dimensionality reduction**
- **Part III: Future works**

Overview of the talk

Part I: Symmetric Entropic Affinities

Part II: Application to dimensionality reduction

Part III: Future works

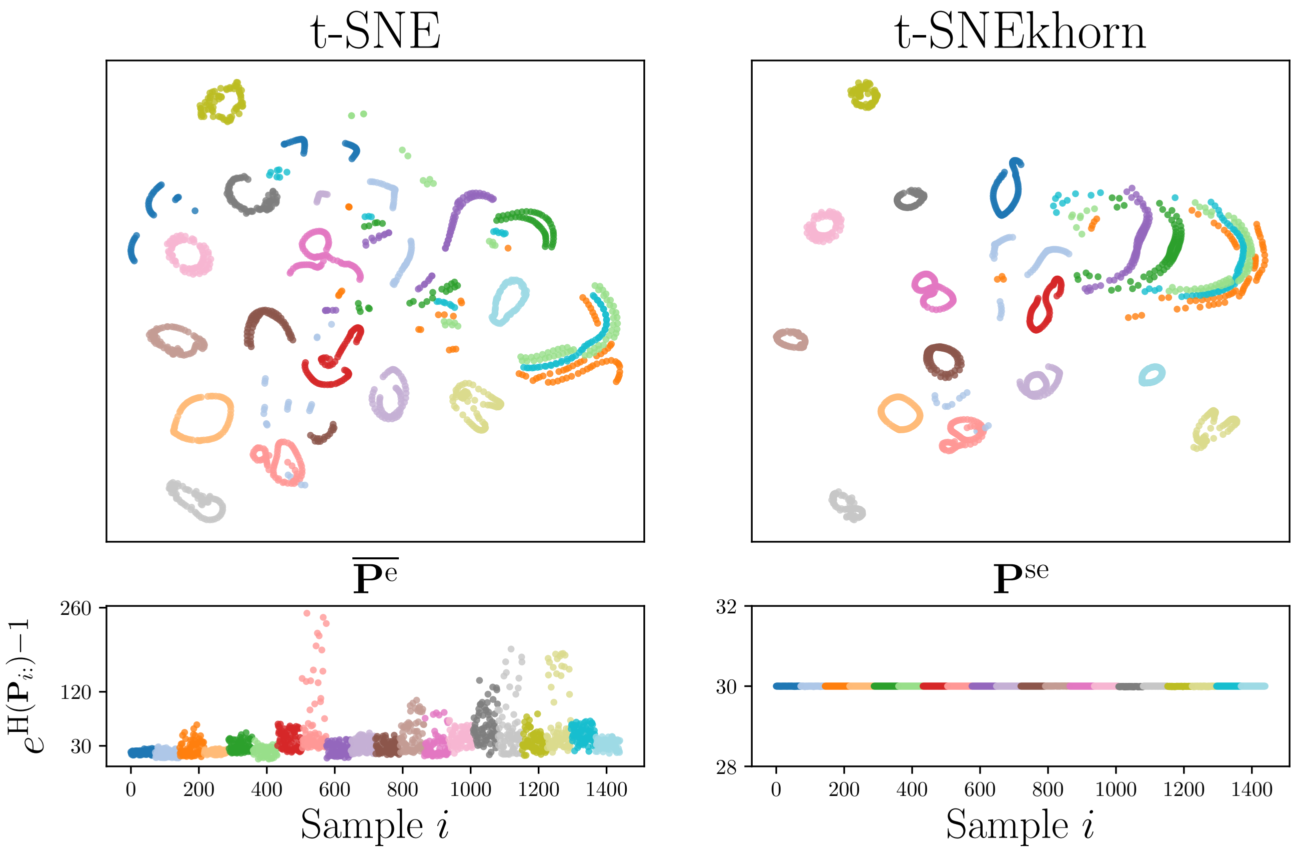


Overview of the talk

Part I: Symmetric Entropic Affinities

Part II: Application to dimensionality reduction

Part III: Future works

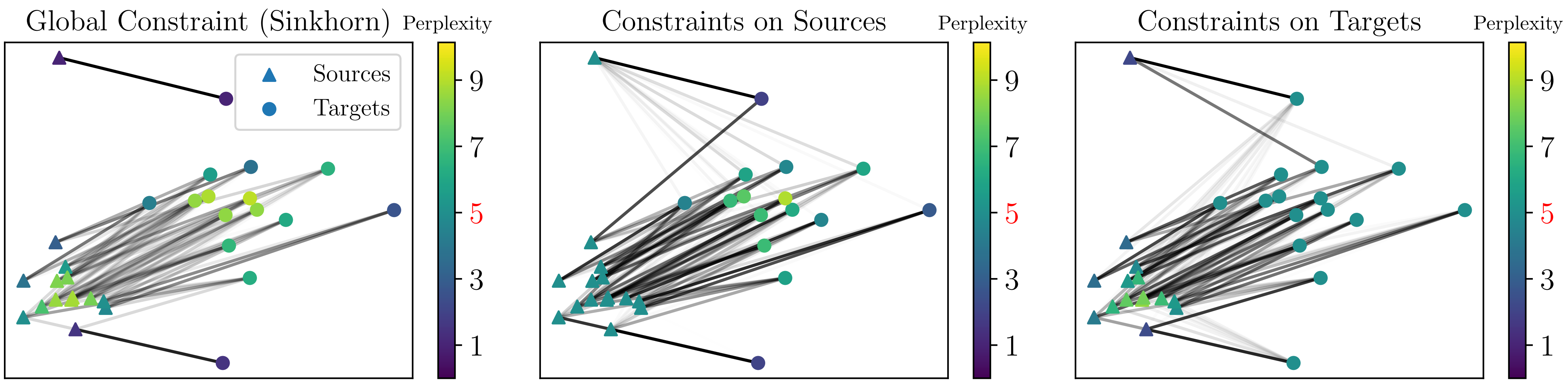


Overview of the talk

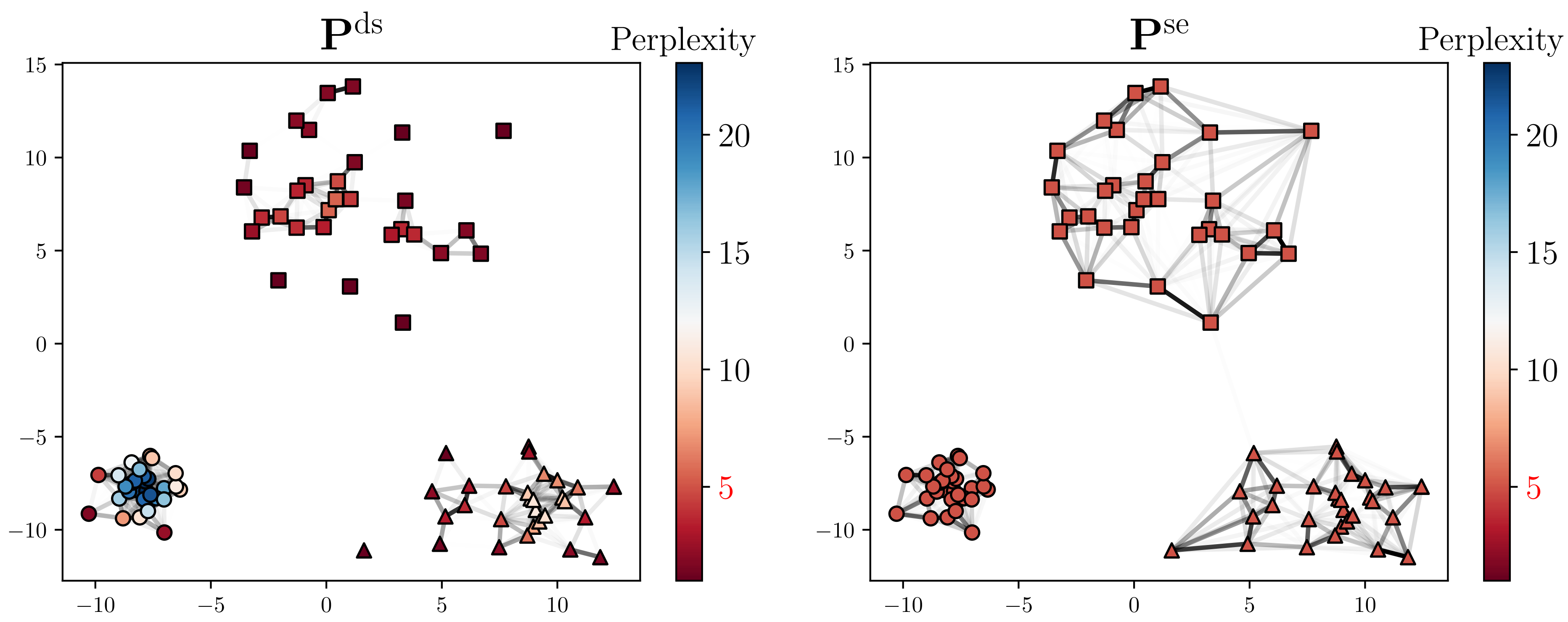
Part I: Symmetric Entropic Affinities

Part II: Application to dimensionality reduction

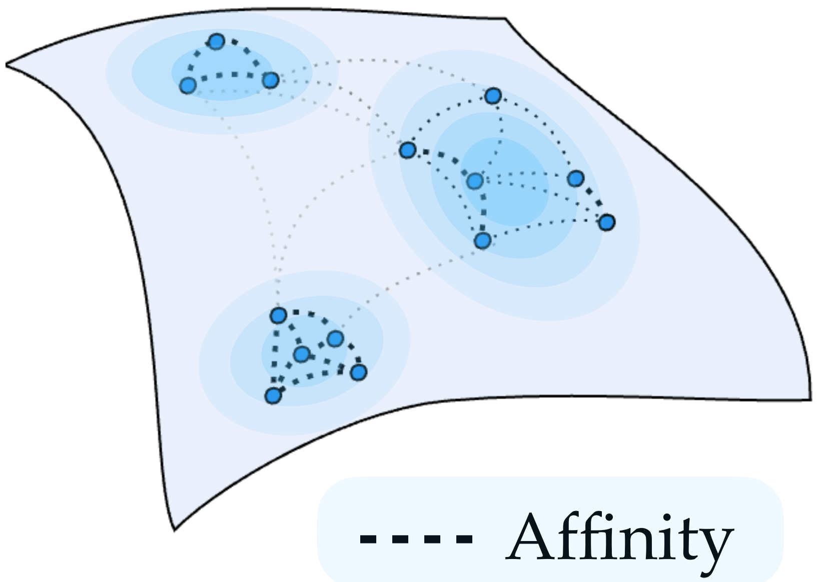
Part III: Future works

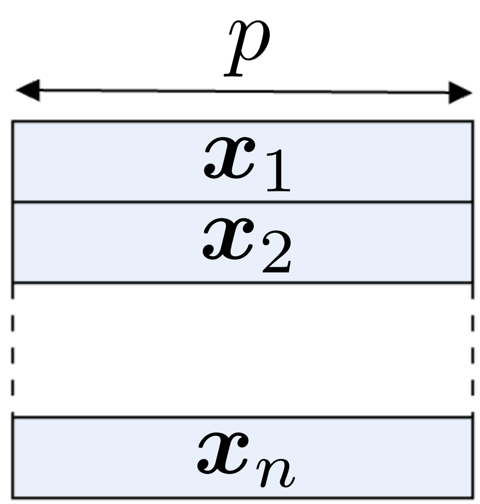


Part I: Symmetric Entropic Affinities



Affinity Matrices



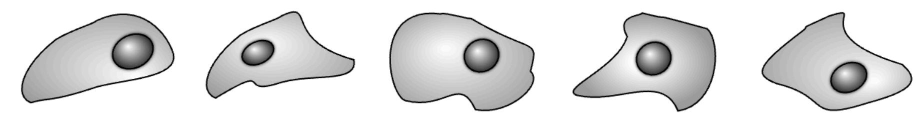


Dataset $\mathbf{X}$

Images

COIL Dataset [Nene et al., 1996]

Cells

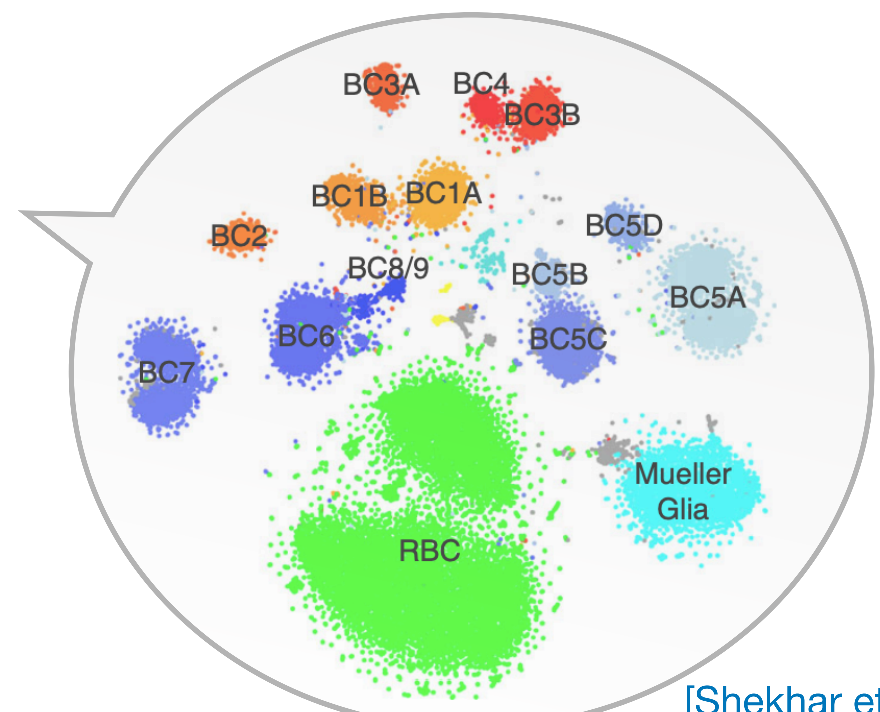


Symmetric matrix with **non-negative coefficients.**

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

Useful in many ML methods :

- **Dimensionality reduction** [Van der Maaten and Hinton, 2008]
- Spectral clustering [Von Luxburg, 2007]
- Kernel machines [Schölkopf and Smola, 2002]
- Semi-supervised learning [Zhou et al., 2003]
- Self-supervised learning [Zbontar et al., 2021]



[Shekhar et al., 2016]

Gaussian Affinity (or Gibbs kernel)

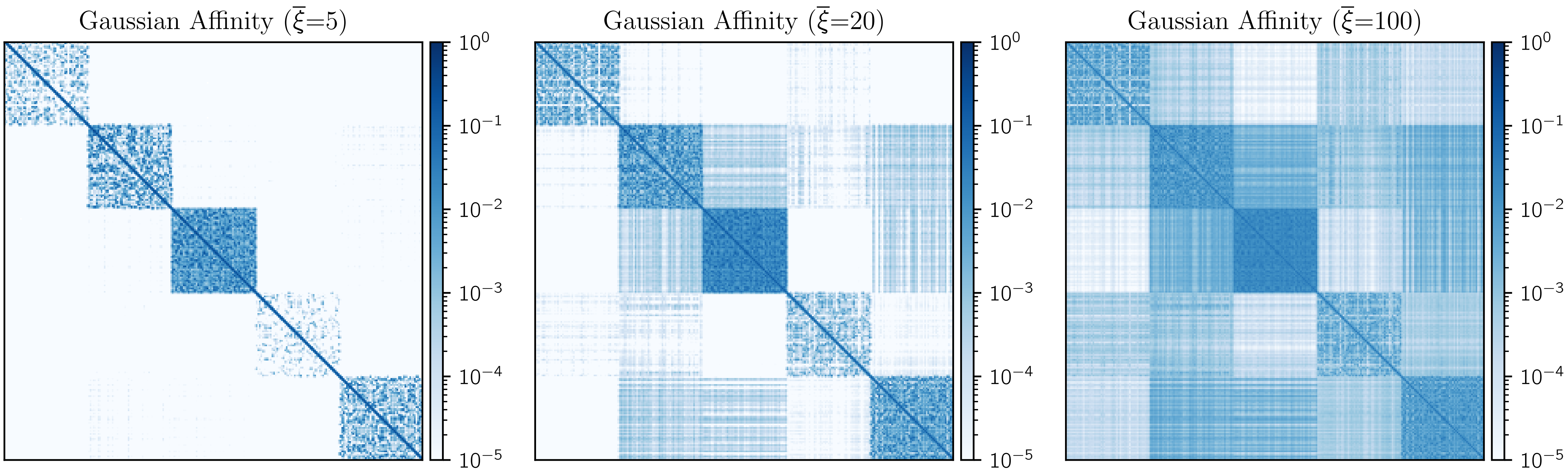


Fig : Affinity on 5 classes of the COIL Dataset [Nene et al., 1996]

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

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

Gaussian Affinity (or Gibbs kernel)

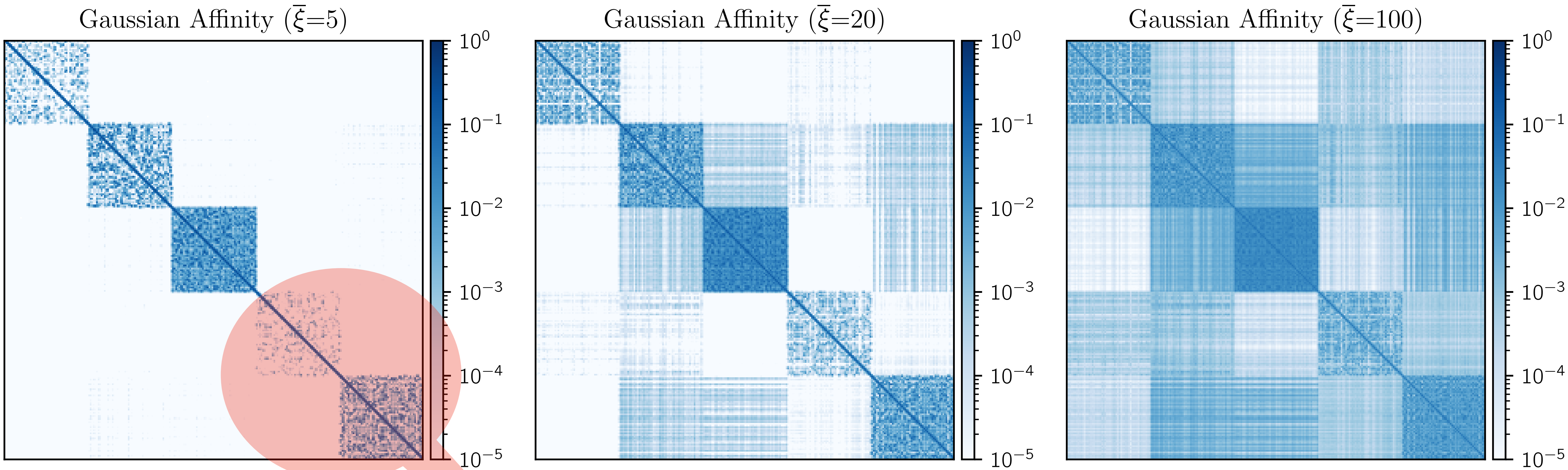


Fig : Affinity on 5 classes of the COIL Dataset [Nene et al., 1996]

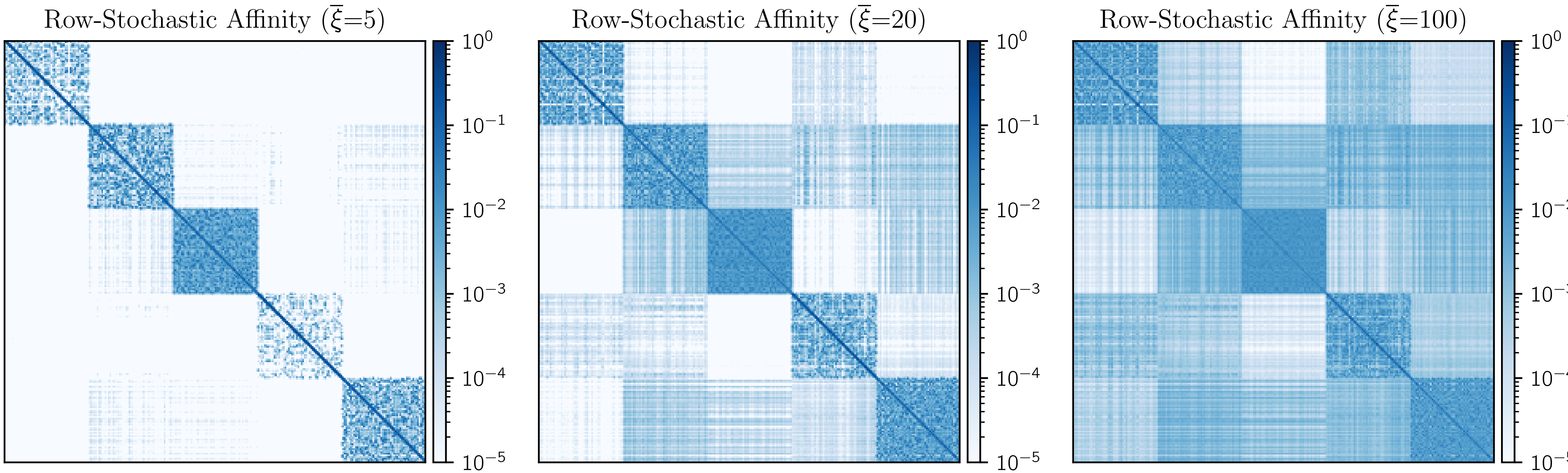
Cost matrix: $\mathbf{C} \in \mathbb{R}_+^{n \times n}$ such that $\mathbf{C} = \mathbf{C}^\top$ and $C_{ij} = 0 \iff i = j$.

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

How to better reveal classes ?

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

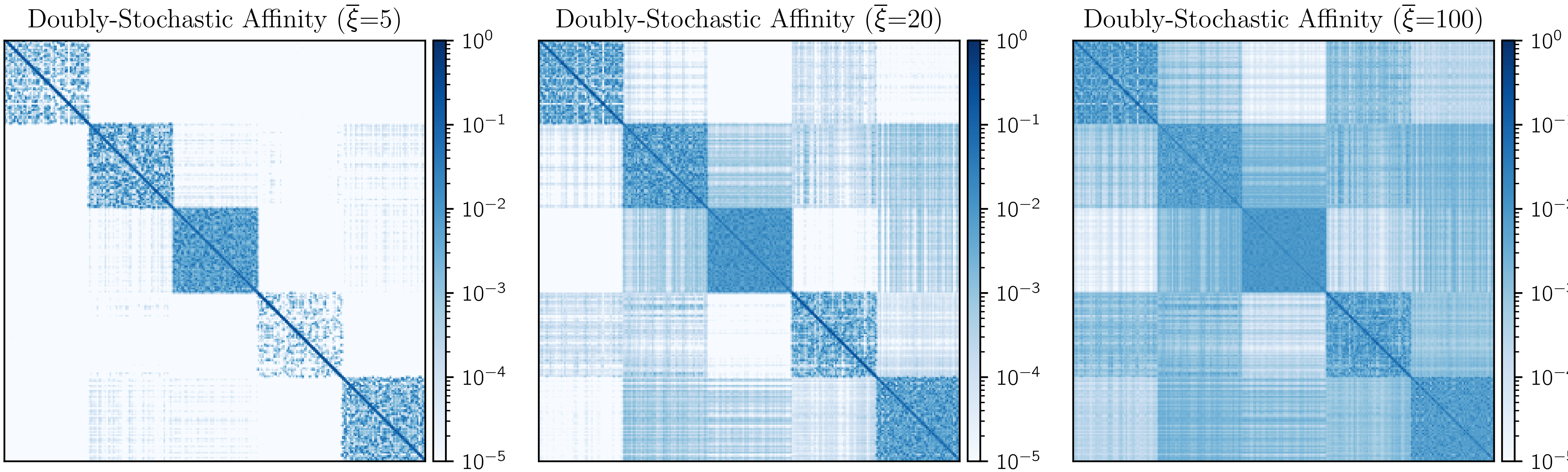
ℓ_1 Norm - Row Stochastic



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

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

ℓ_1 Norm - Doubly Stochastic



Sinkhorn Algorithm

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

While not converged:

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

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

$\xrightarrow{\textbf{converges to}}$ $\mathcal{DS} = \{\mathbf{P} \text{ s.t. } \mathbf{P}\mathbf{1} = \mathbf{P}^\top\mathbf{1} = \mathbf{1}\}\,.$

Doubly Stochastic Affinity

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

While not converged:

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

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

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

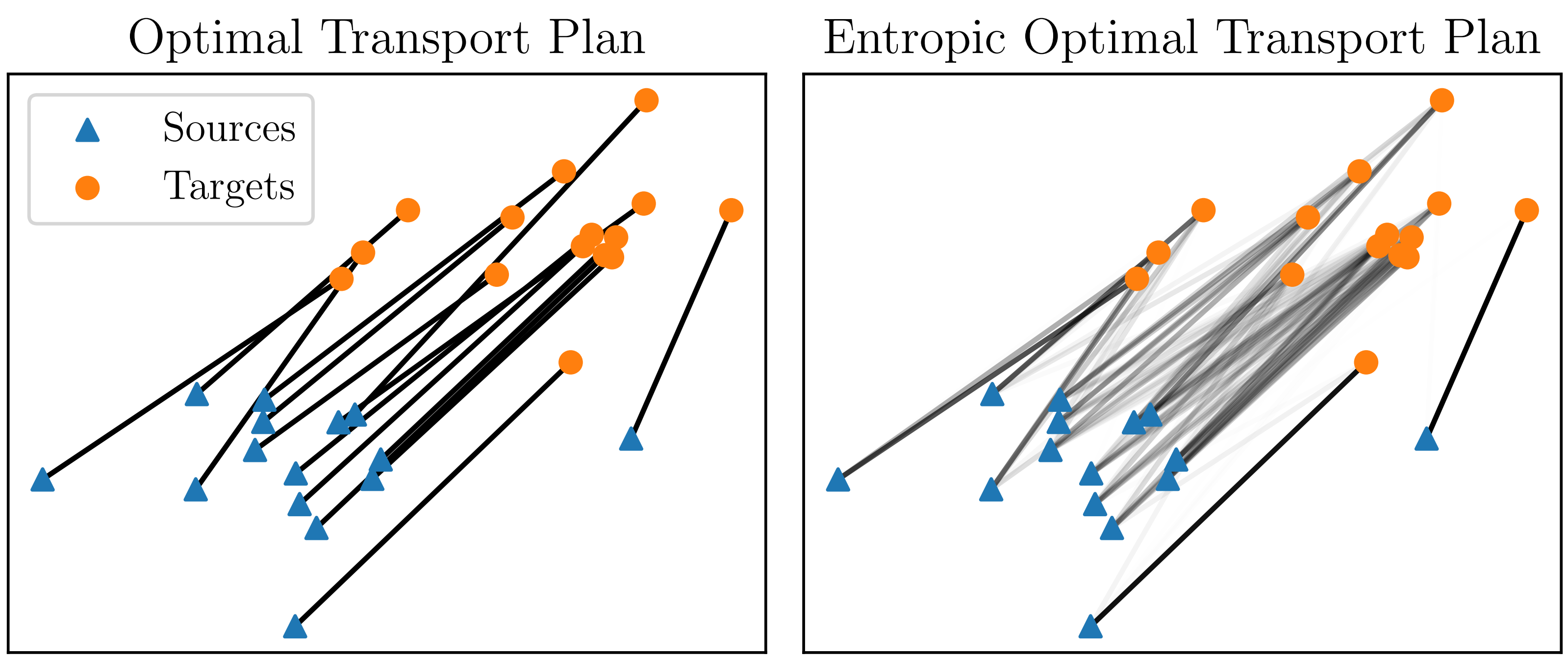
For any set $\mathcal{E}$: $\operatorname{Proj}^{\mathrm{KL}}_{\mathcal{E}}(\mathbf{K}) = \arg\min_{\mathbf{P}\in\mathcal{E}} \langle \mathbf{P}, \log(\mathbf{P} \oslash \mathbf{K})\rangle$.

Sinkhorn iterations compute $\mathbf{P}^{\mathrm{ds}} := \operatorname{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$.

$\mathbf{P}^{\mathrm{ds}}$ solves the optimal transport problem:

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times n}} \langle \mathbf{P}, \mathbf{C}\rangle - \sigma \sum_i \mathrm{H}(\mathbf{P}_{i:}) \quad \text{s.t.} \quad \mathbf{P} \in \mathcal{DS}\,.$$

Entropic Optimal Transport



OT

OT plan — **Pairwise cost between sources and targets**

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times n}} \langle \mathbf{P}, \mathbf{C}\rangle \quad \text{s.t.} \quad \mathbf{P}\mathbf{1} = \alpha, \ \mathbf{P}^\top\mathbf{1} = \beta.$$

Marginals

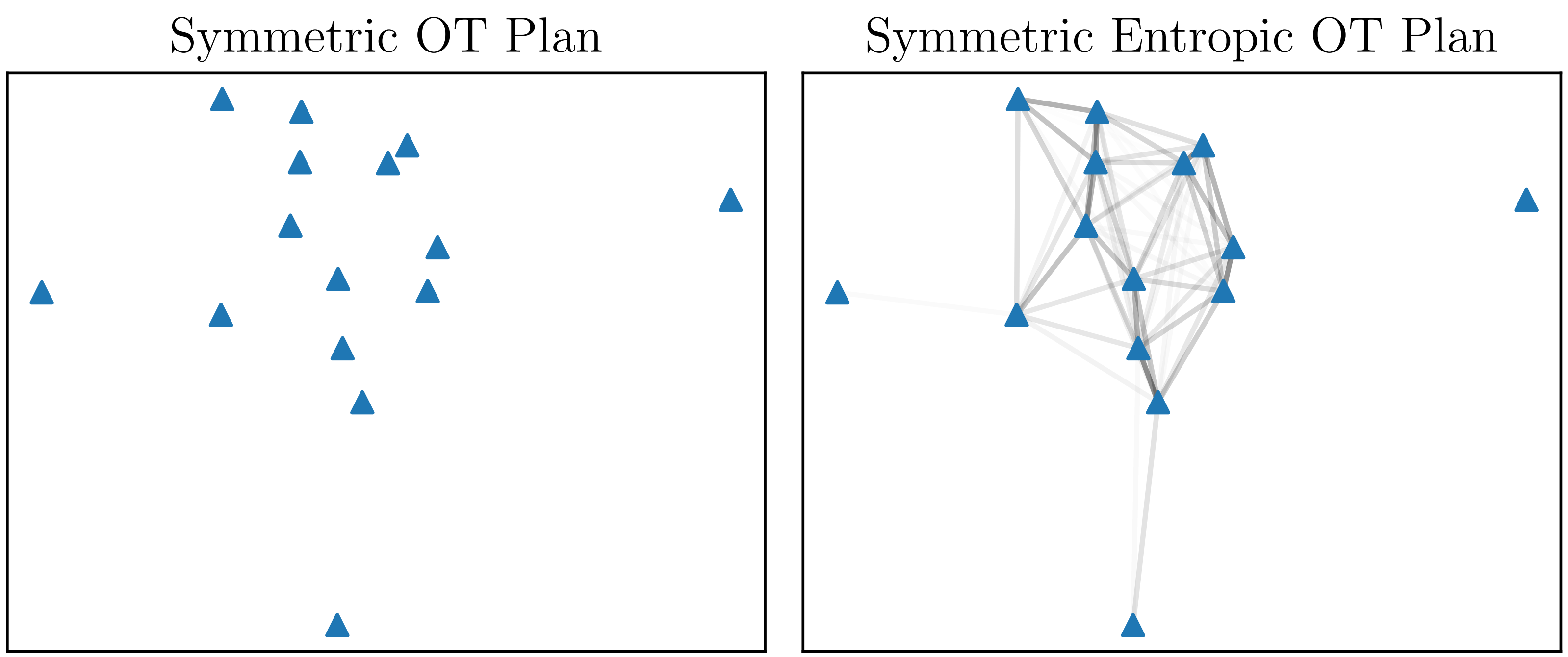
$\mathrm{H}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \log \boldsymbol{p} - \mathbf{1}\rangle$ is the Shannon entropy.

Entropic OT [Cuturi, 2013]

Entropic regularizer

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times n}} \langle \mathbf{P}, \mathbf{C}\rangle - \sigma \sum_i \mathrm{H}(\mathbf{P}_{i:}) \quad \text{s.t.} \quad \mathbf{P}\mathbf{1} = \alpha, \ \mathbf{P}^\top\mathbf{1} = \beta.$$

Symmetric Entropic OT



Sym. OT OT plan — Pairwise cost among points (SYMMETRIC)

$$\min_{\mathbf{P} \in \mathbb{R}_+^{n \times n}} \langle \mathbf{P}, \mathbf{C} \rangle \quad \text{s.t.} \quad \mathbf{P}\mathbf{1} = \mathbf{1}, \; \mathbf{P}^\top \mathbf{1} = \mathbf{1}.$$

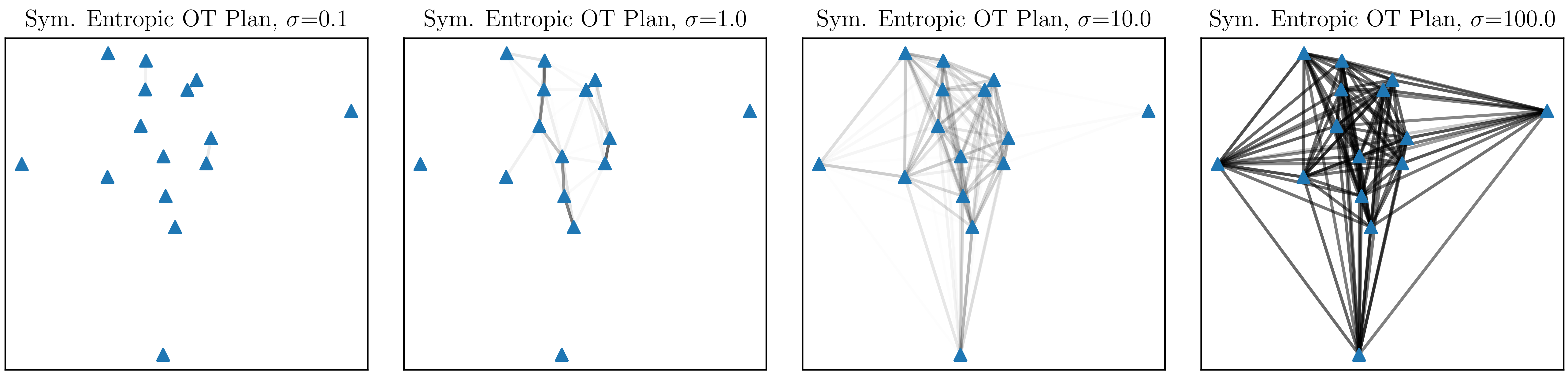
Marginals

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

Sym. Entropic OT Entropic regularizer

$$\min_{\mathbf{P} \in \mathbb{R}_+^{n \times n}} \langle \mathbf{P}, \mathbf{C} \rangle - \sigma \sum_i \mathrm{H}(\mathbf{P}_{i:}) \quad \text{s.t.} \quad \mathbf{P}\mathbf{1} = \mathbf{1}, \; \mathbf{P}^\top \mathbf{1} = \mathbf{1}.$$

Symmetric Entropic OT



Sym. Entropic OT

Entropic regularizer

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times n}} \langle \mathbf{P}, \mathbf{C} \rangle - \sigma \sum_i \mathrm{H}(\mathbf{P}_{i:}) \quad \text{s.t.} \quad \mathbf{P}\mathbf{1} = \mathbf{1}, \ \mathbf{P}^\top \mathbf{1} = \mathbf{1}.$$

Interpolates between:

- identity $\mathbf{I}_n$ for $\sigma \to 0$.
- uniform $\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ for $\sigma \to \infty$.

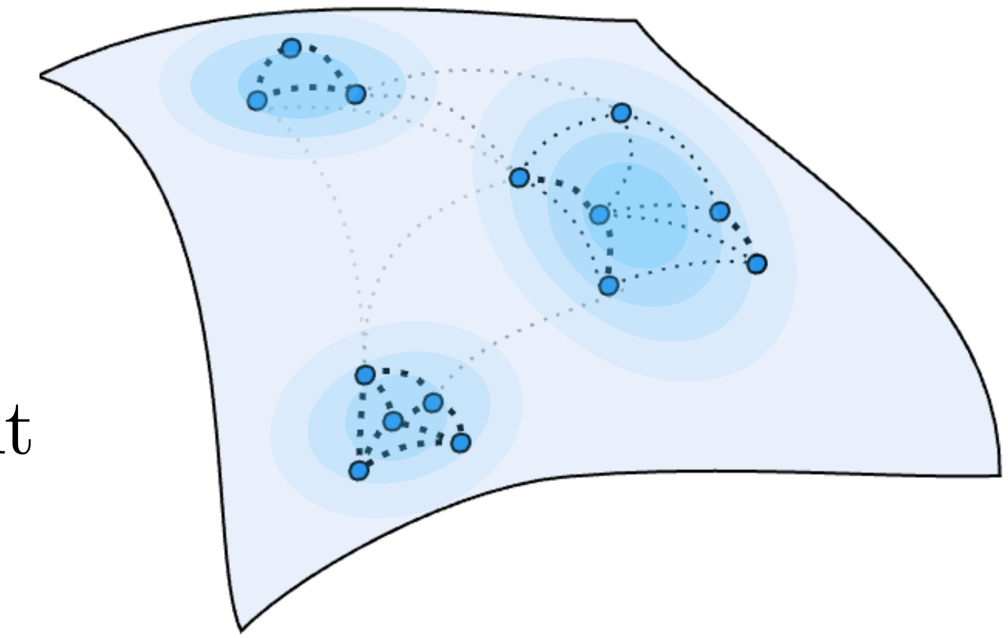
Constrained Formulation

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times n}, \mathbf{P}\mathbf{1}=\mathbf{1}, \mathbf{P}^\top\mathbf{1}=\mathbf{1}} \langle \mathbf{P}, \mathbf{C} \rangle$$

Entropy OT plan

$$\text{s.t.} \quad \sum_i \mathrm{H}(\mathbf{P}_{i:}) \geq \eta.$$

Entropic Affinity



Data has **varying noise levels**.

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

Definition [Hinton, Roweis 2002]

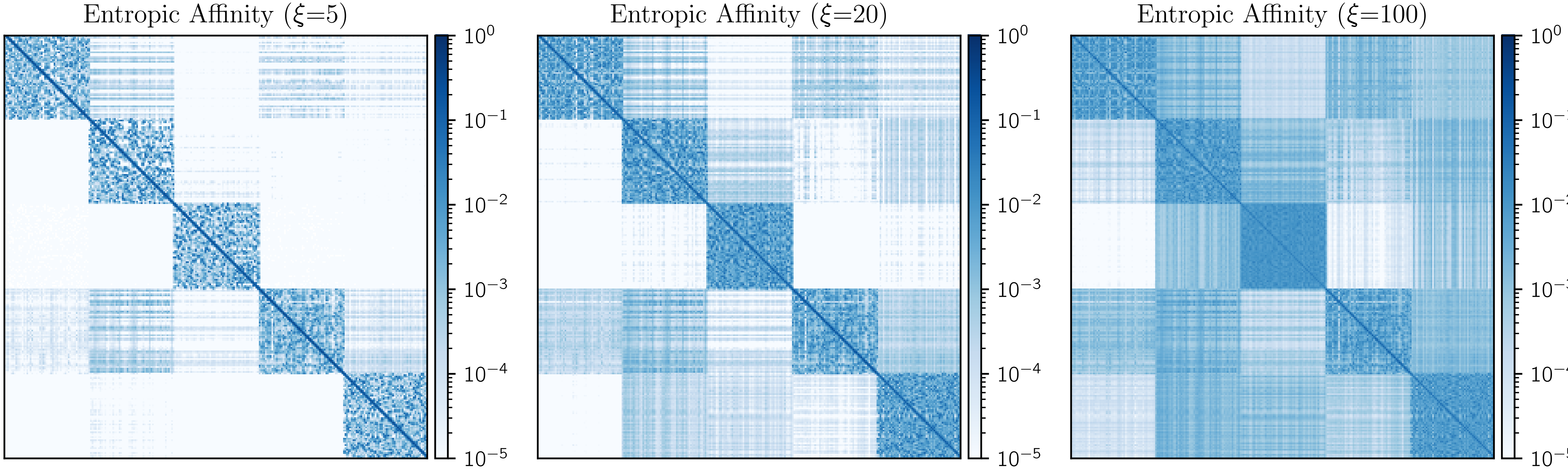
$$\forall i, \forall j, \; P^{\mathrm{e}}_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon^\star_i\right)}{\sum_\ell \exp\left(-C_{i\ell}/\varepsilon^\star_i\right)}$$

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

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

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

Entropic Affinity



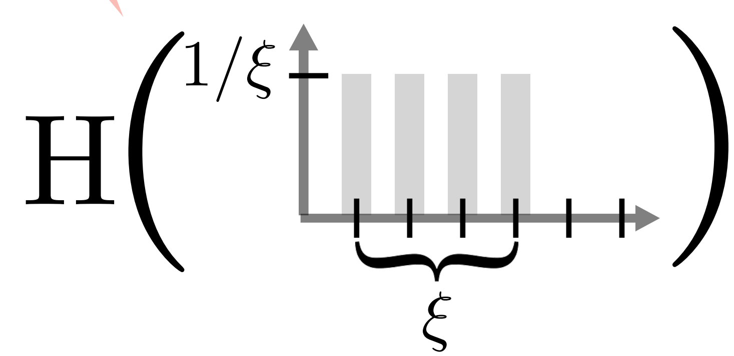
Definition [Hinton, Roweis 2002]

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

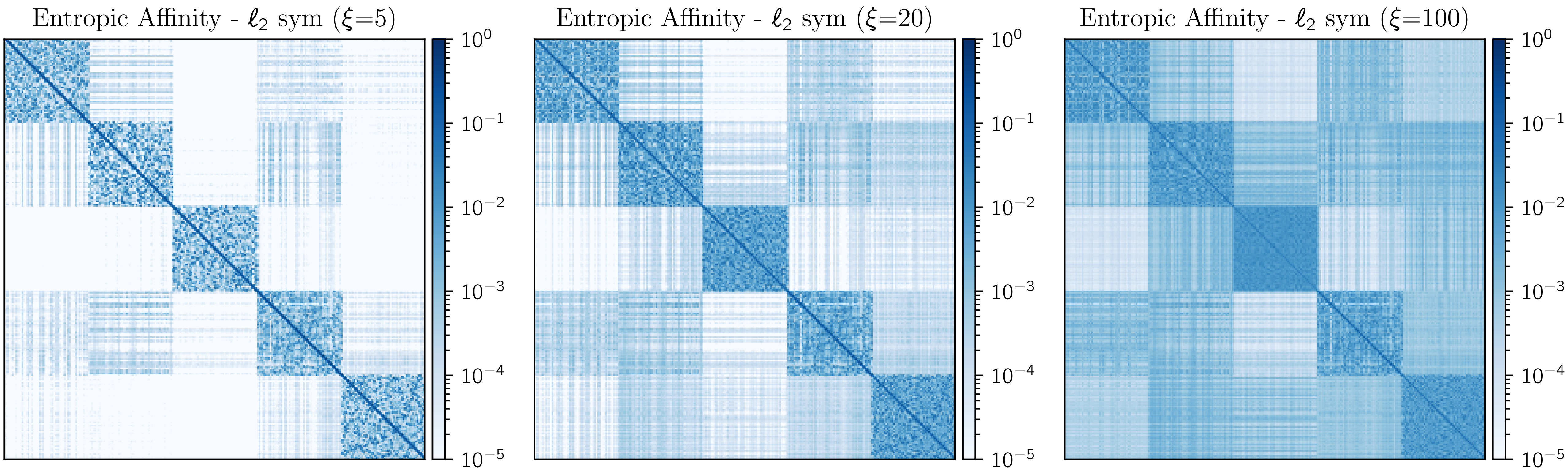
ξ effective neighbors

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

- $\mathrm{H}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \log \boldsymbol{p} - \mathbf{1} \rangle$ is the Shannon entropy.
- $\xi \in [\![1, n]\!]$ is the **perplexity** parameter.



Entropic Affinity



Definition [Hinton, Roweis 2002]

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

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

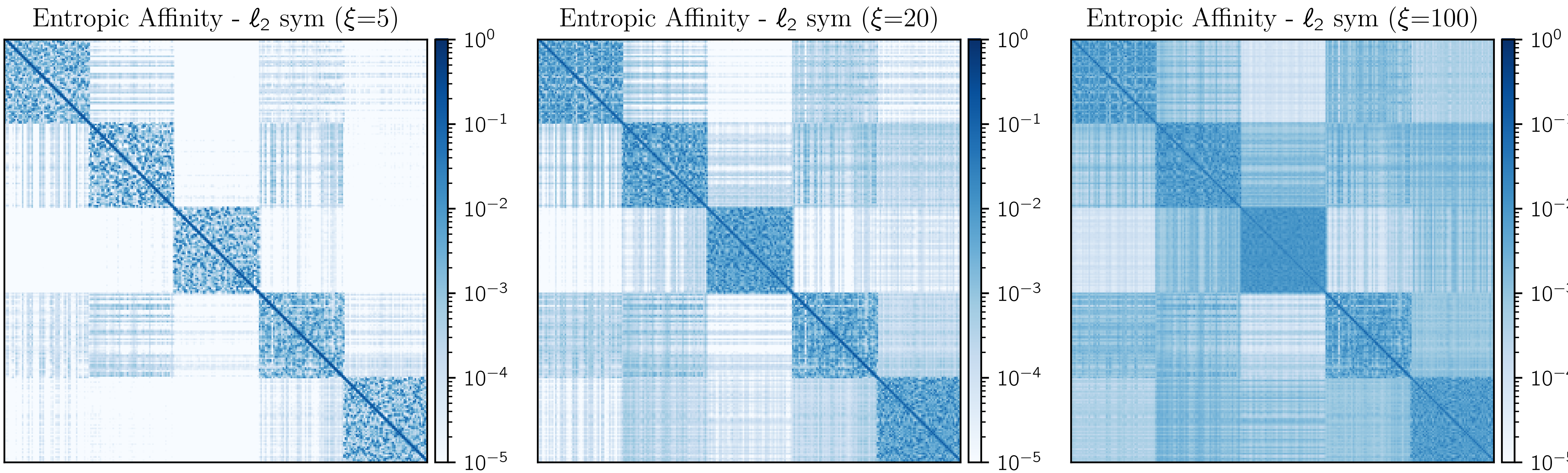
$\mathbf{P}^{\mathrm{e}}$ is **not symmetric**.

→ $\overline{\mathbf{P}^{\mathrm{e}}} = \frac{1}{2}(\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top})$ 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]

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

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

$\mathbf{P}^{\mathrm{e}}$ 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]

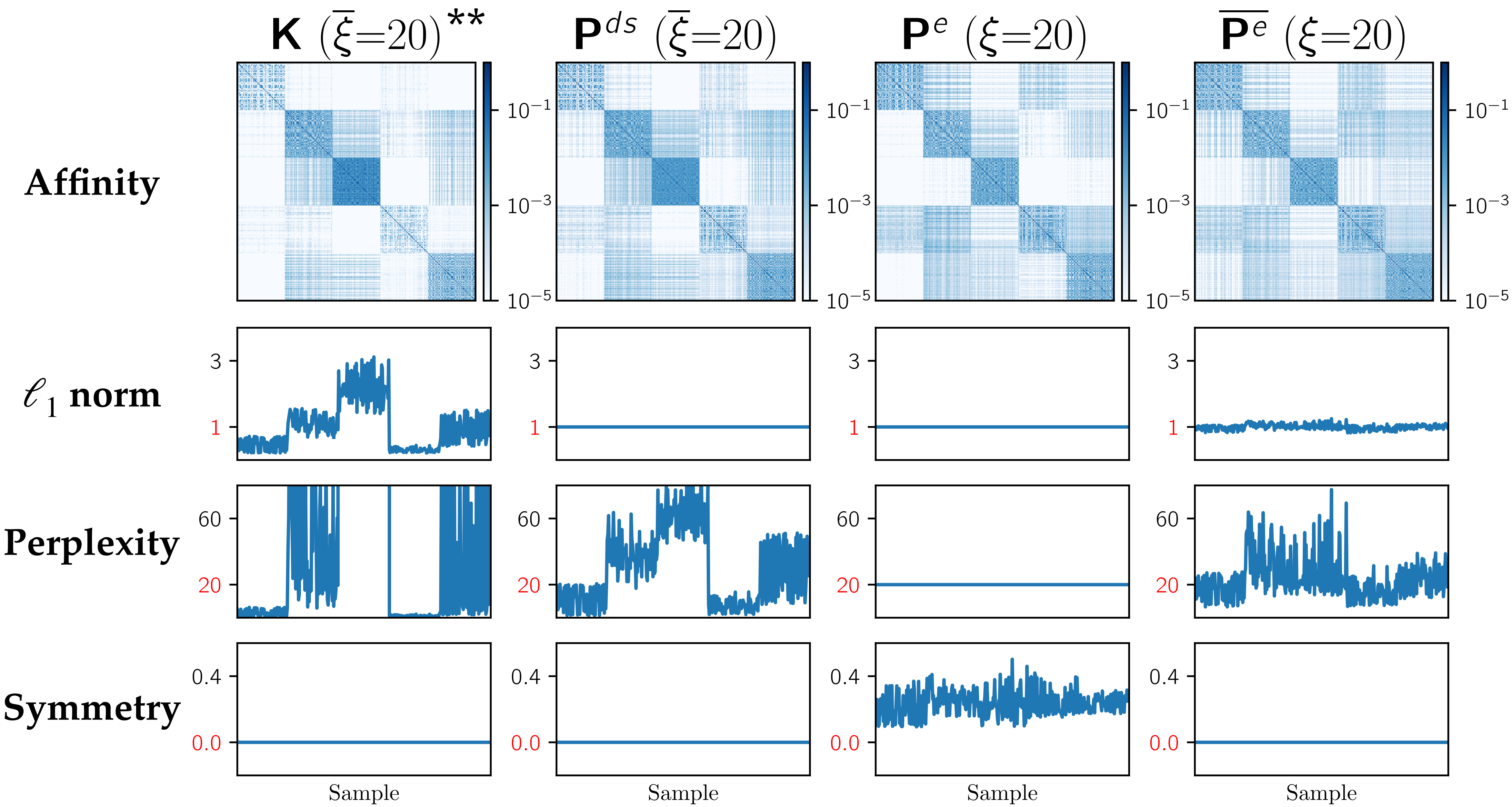
t-SNE algorithm

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

Breaks the construction of entropic affinities.

Affinity Panorama*

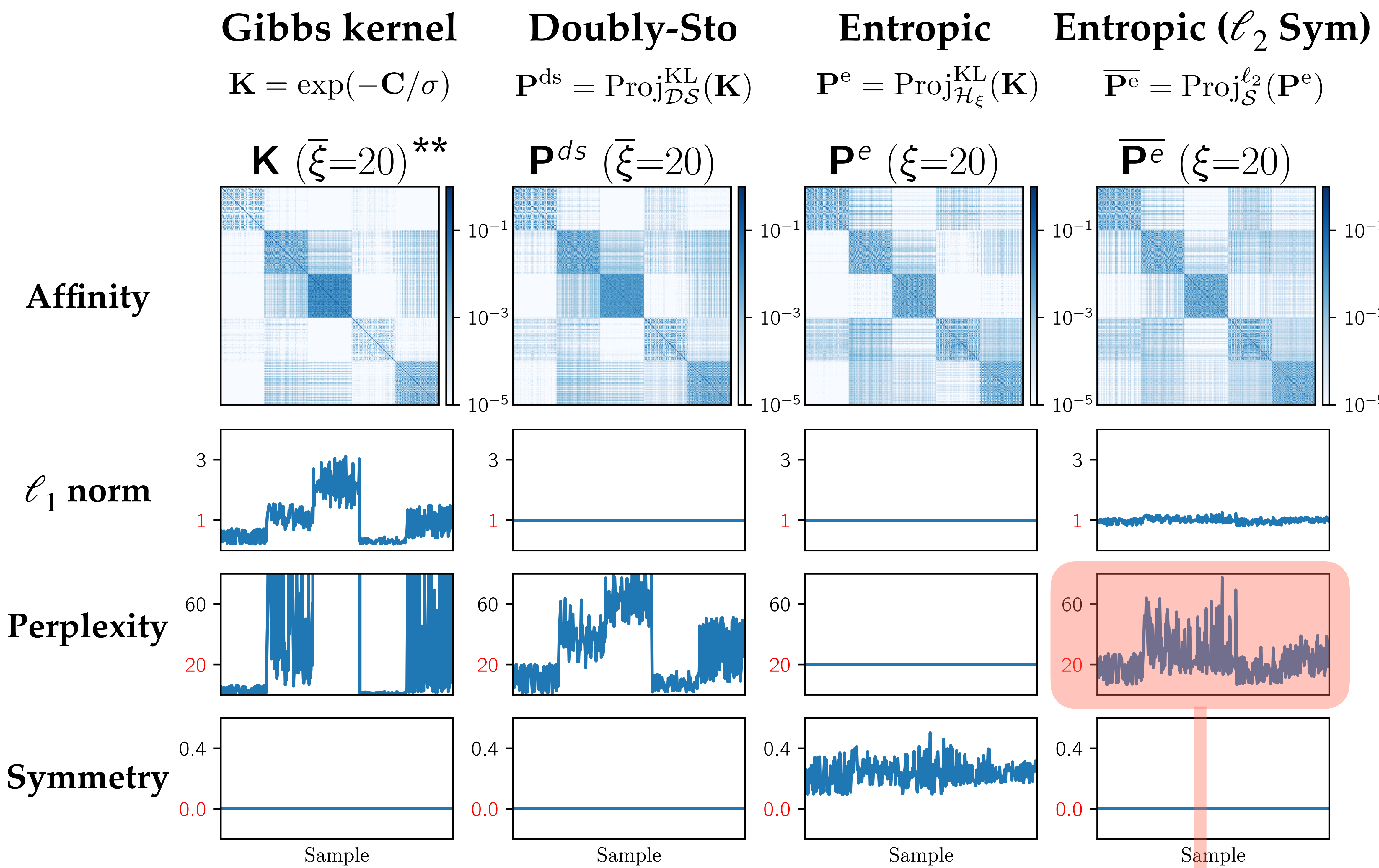
Gibbs kernel	Doubly-Sto	Entropic	Entropic (ℓ_2 Sym)
$\mathbf{K} = \exp(-\mathbf{C}/\sigma)$	$\mathbf{P}^{\mathrm{ds}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$	$\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K})$	$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$



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

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

Affinity Panorama*



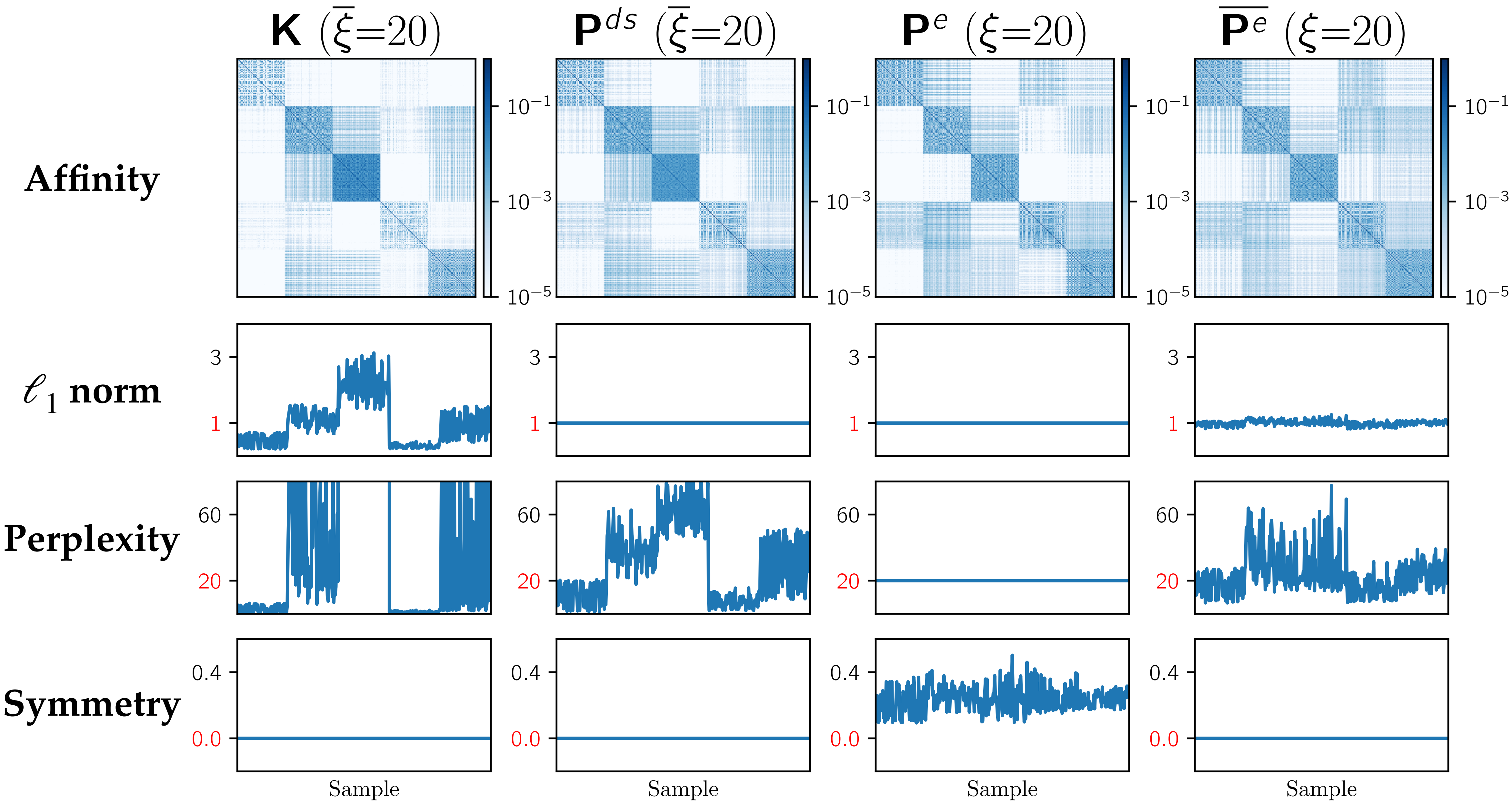
* On 5 classes of the COIL Dataset [Nene et al., 1996]

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

Entropies not controlled.

Affinity Panorama

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



Can we control ℓ_1 norm, entropy and symmetry ?

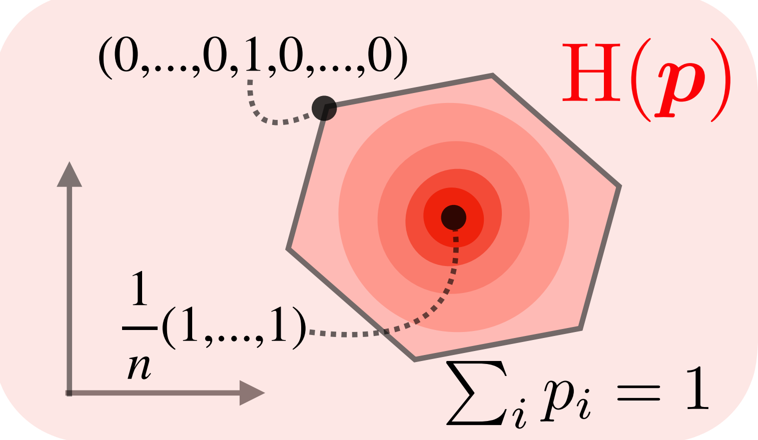
Symmetric Entropic Affinity

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

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**

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

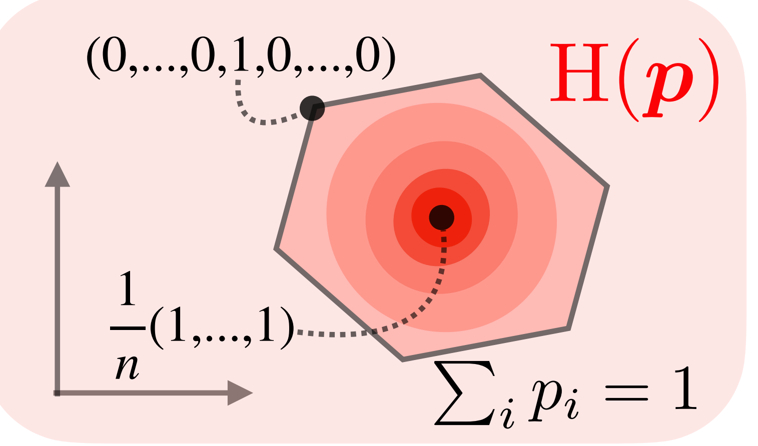
Symmetric Entropic Affinity

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

Entropic Affinity as OT

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

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**

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

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.

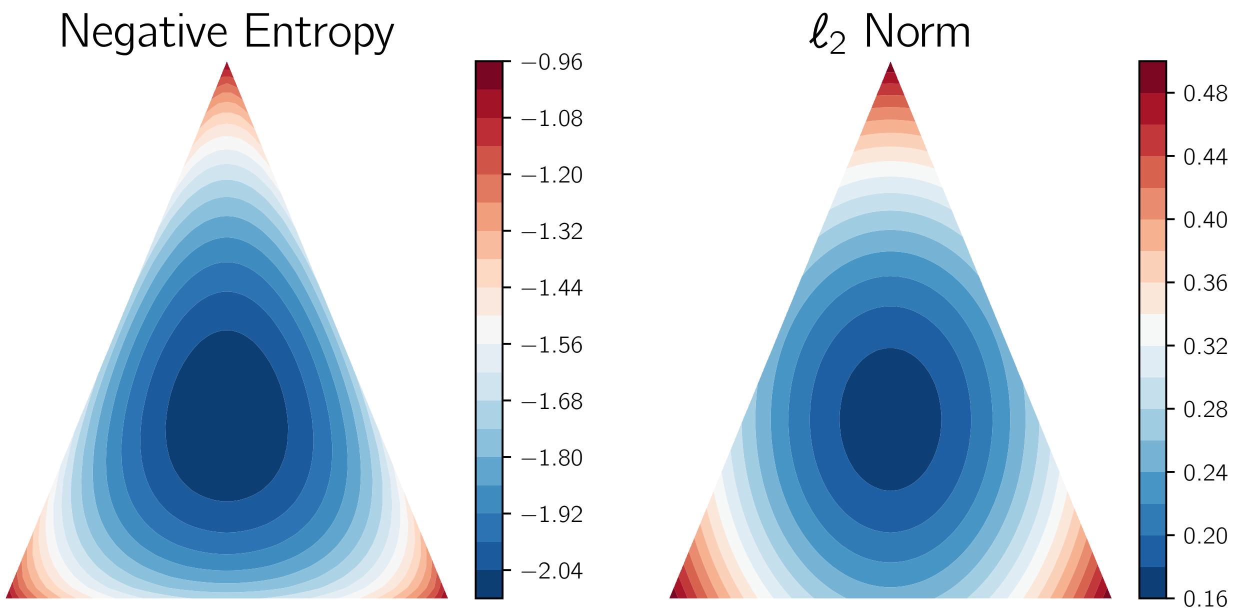
Formulation via Projection

For any set $\mathcal{E}$:

$$\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{E}}(\mathbf{K}) = \arg\min_{\mathbf{P}\in\mathcal{E}} \langle \mathbf{P}, \log(\mathbf{P} \oslash \mathbf{K})\rangle$$

$$\mathrm{Proj}^{\ell_2}_{\mathcal{E}}(\mathbf{K}) = \arg\min_{\mathbf{P}\in\mathcal{E}} \|\mathbf{P} - \mathbf{K}\|_2$$

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



Entropic Affinity as Projection

$$\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K}_\sigma) \text{ for any } 0 < \sigma \leq \min_i \varepsilon^\star_i .$$

$\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$ is a mixture of KL and ℓ_2 projections.

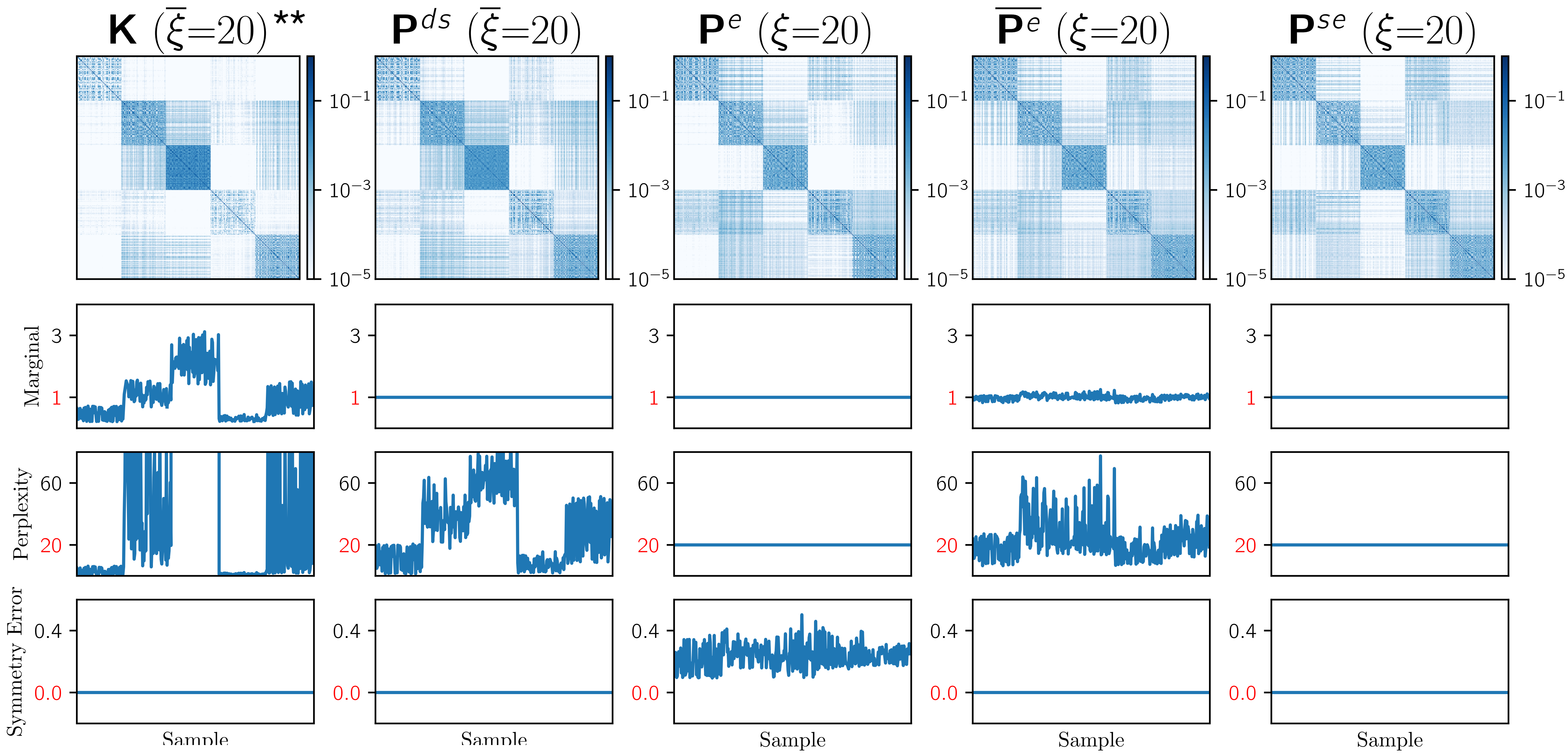
Symmetric Entropic Affinity as Projection

$$\mathbf{P}^{\mathrm{se}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi \cap \mathcal{S}}(\mathbf{K}_\sigma) \text{ for any } 0 < \sigma \leq \min_i \gamma^\star_i .$$

Affinity Panorama*

t-SNE | OURS

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



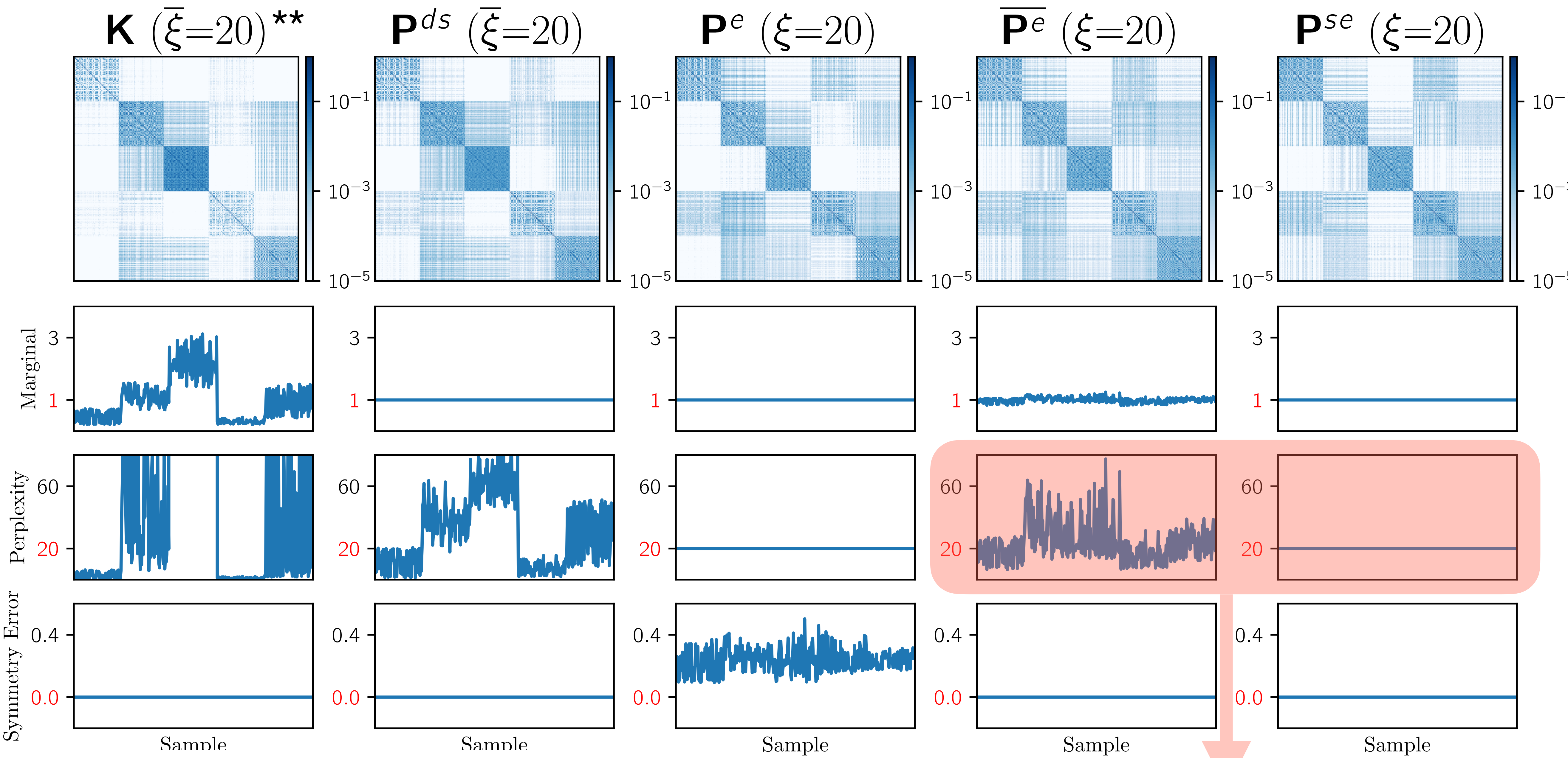
* On 5 classes of the COIL Dataset [Nene et al., 1996]

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

Affinity Panorama*

OURS

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



Effective control over entropies.

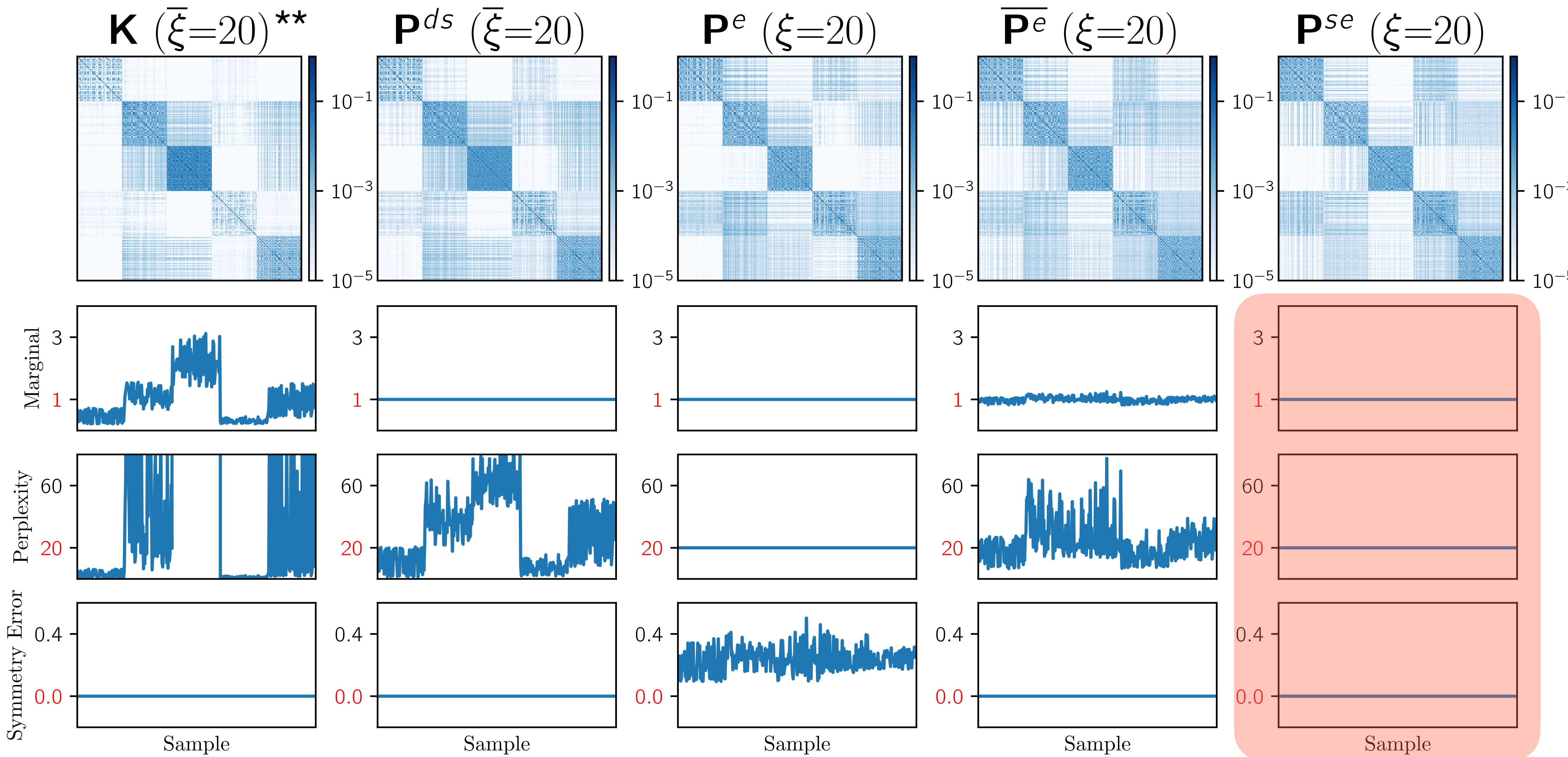
* On 5 classes of the COIL Dataset [Nene et al., 1996]

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

Affinity Panorama*

OURS

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



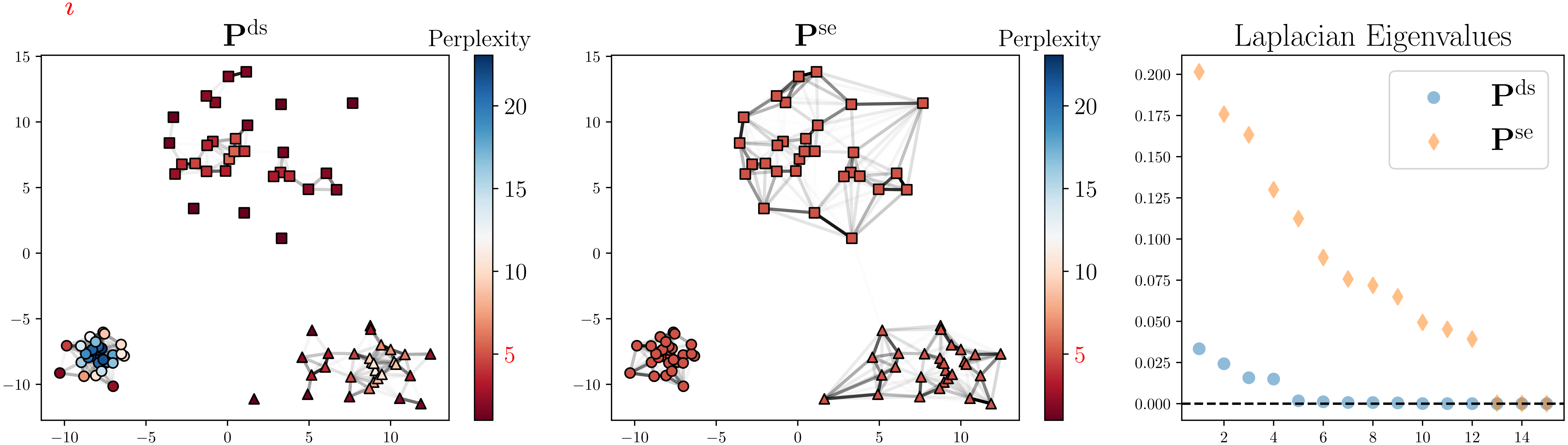
Controls ℓ_1 norm, entropy and symmetry at the same time.

Visualization on toy example

Doubly-Stochastic | **Symmetric-Entropic (OURS)**

$$\min_{\mathbf{P}\geq 0, \mathbf{P}\mathbf{1}=\mathbf{1}, \mathbf{P}\in\mathcal{S}} \langle \mathbf{P}, \mathbf{C} \rangle \qquad \min_{\mathbf{P}\geq 0, \mathbf{P}\mathbf{1}=\mathbf{1}, \mathbf{P}\in\mathcal{S}} \langle \mathbf{P}, \mathbf{C} \rangle \quad \Big\} \quad \text{Symmetric OT}$$

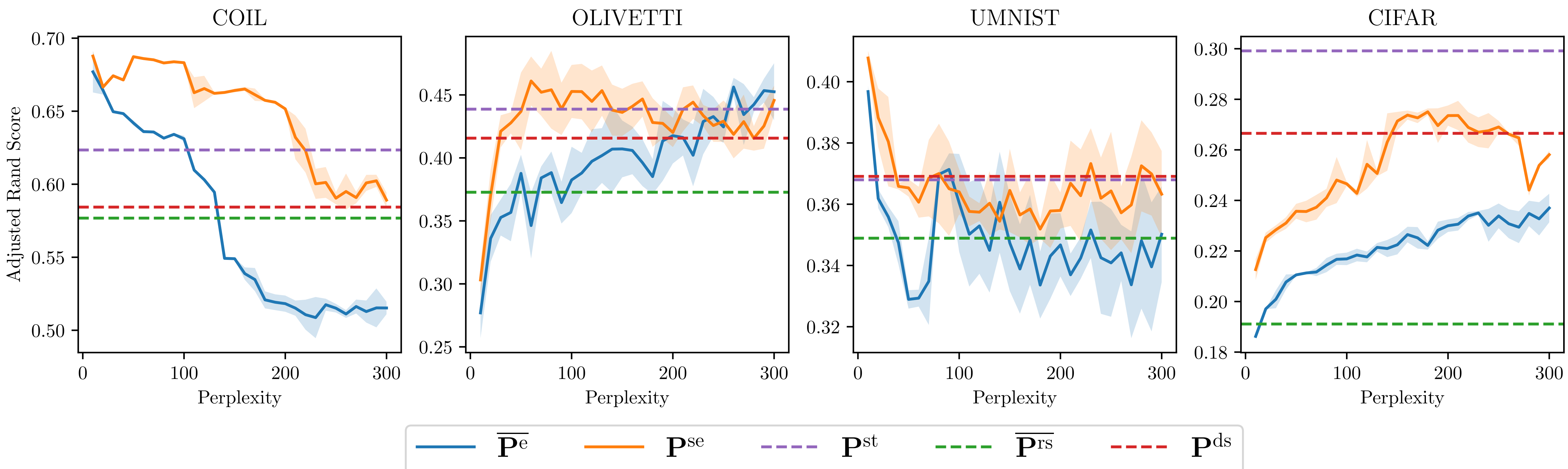
$$\sum_i \mathrm{H}(\mathbf{P}_{i:}) \geq n(\log \xi + 1) \qquad \forall i, \ \mathrm{H}(\mathbf{P}_{i:}) \geq \log \xi + 1 \quad \Big\} \quad \text{Global vs Pointwise}$$



- 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.

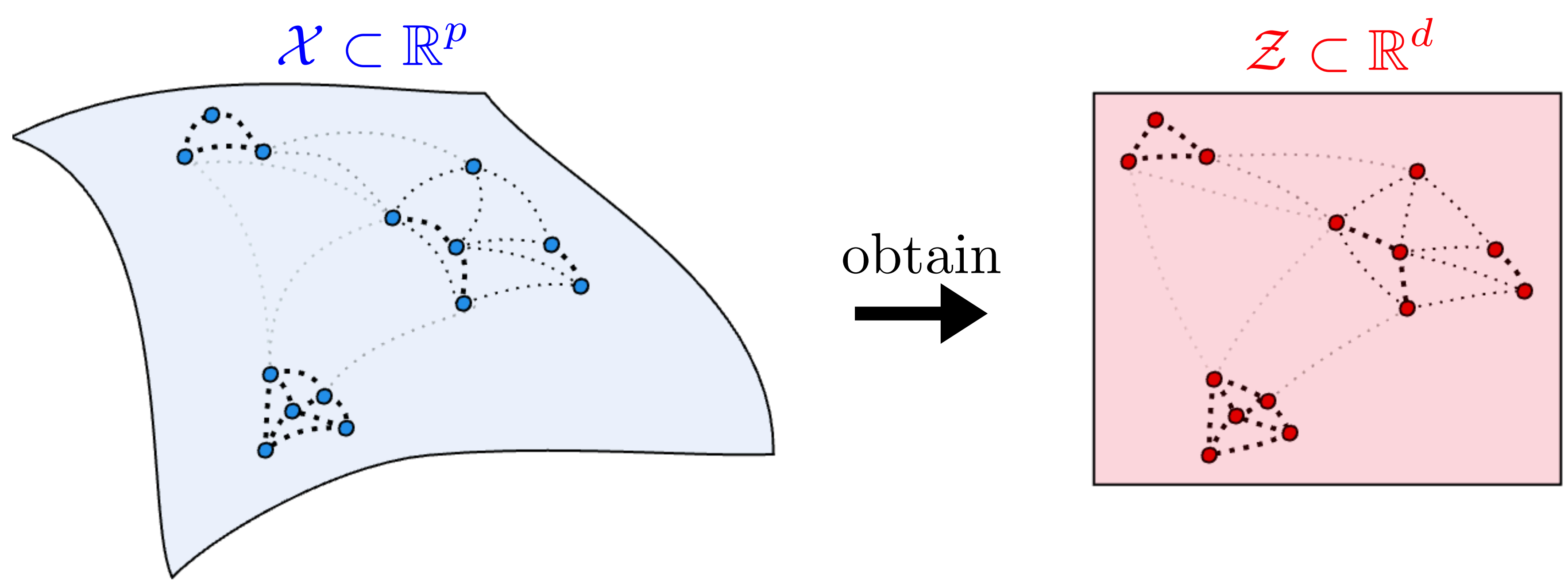
null eigenvalues
=
clusters

Spectral Clustering Results

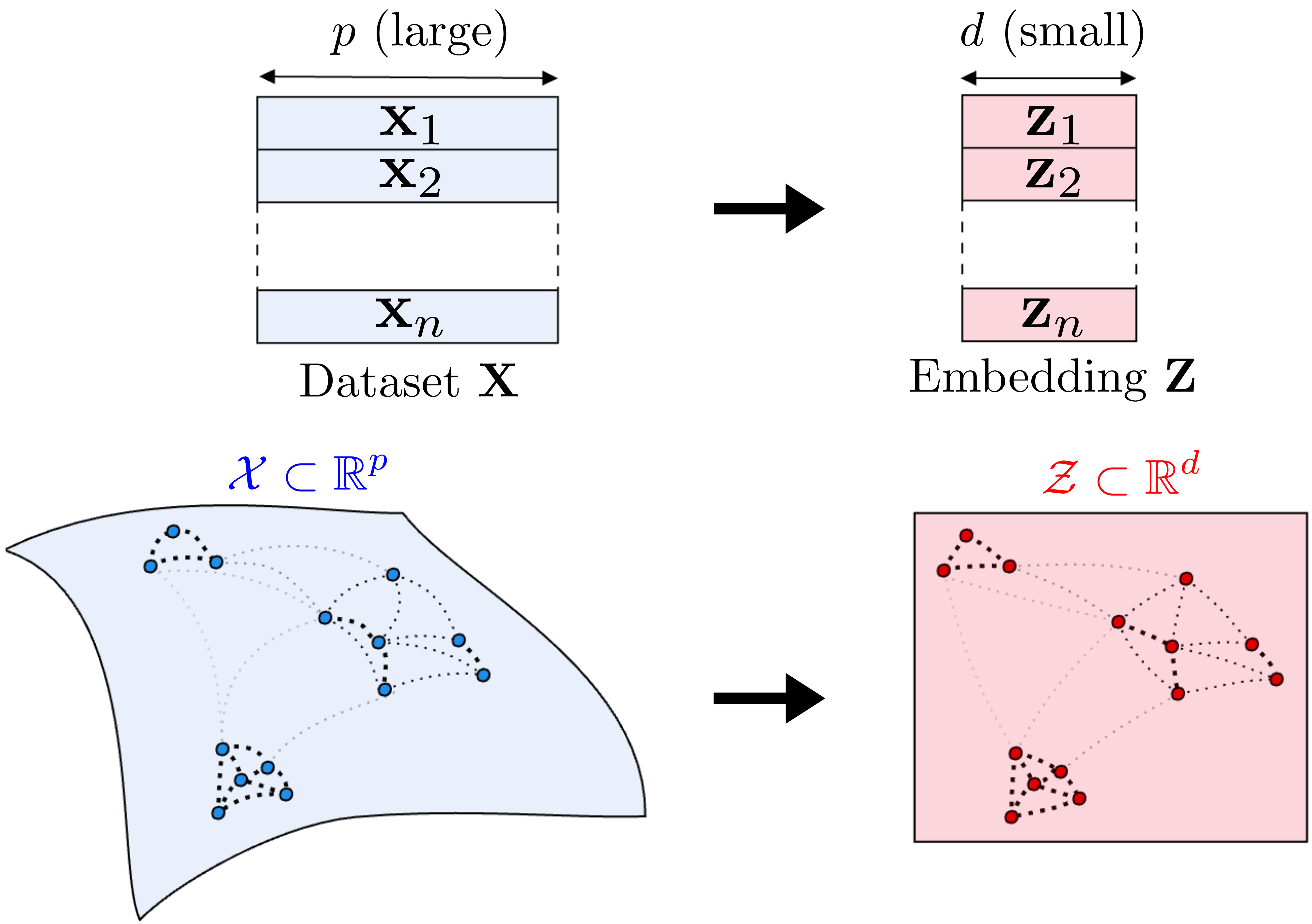


- $\overline{\mathbf{P}^{\mathrm{rs}}}$ is the ℓ_2 symmetrized row-stochastic Gaussian kernel with constant bandwidth.
- $\mathbf{P}^{\mathrm{st}}$ is the self-tuning affinity. [Zelnik-Manor et al., 2004]
- $\mathbf{P}^{\mathrm{se}}$ **consistently outperforms other affinities** (except $\mathbf{P}^{\mathrm{st}}$ on CIFAR).

Part II: Application to Dimensionality Reduction

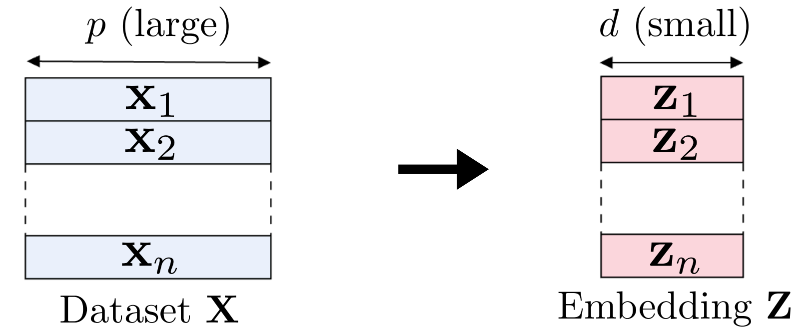


Dimensionality Reduction



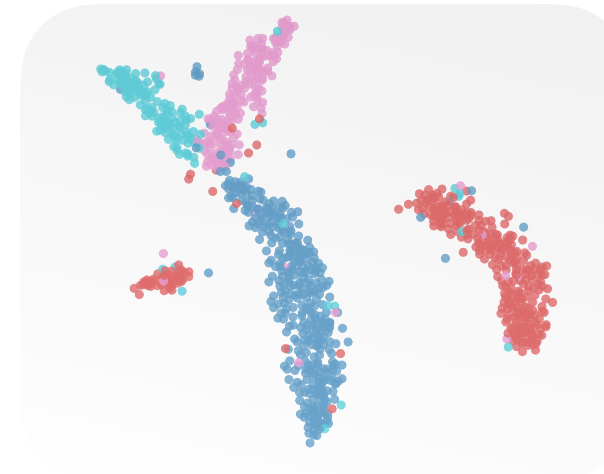
The goal is to capture the geometry in $\mathcal{X}$ and reproduce it in $\mathcal{Z}$.

SNE & SNEkhorn



Cost matrix between embeddings: $[\mathbf{C_Z}]_{ij} = \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2$.

SNAREseq Single Cell data



Stochastic Neighbor Embedding (SNE)

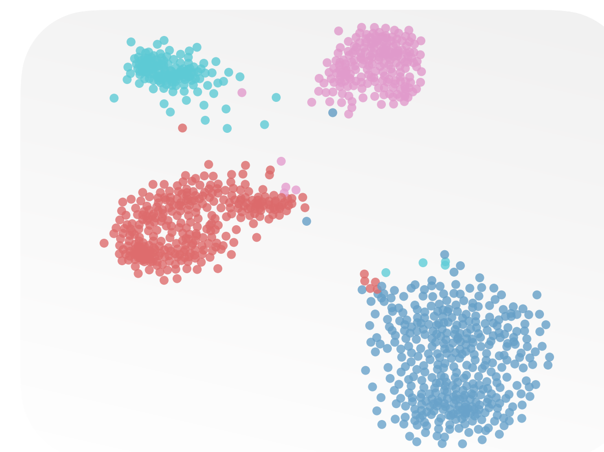
$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times d}} \mathrm{KL}(\overline{\mathbf{P}^{\mathrm{e}}} | \widetilde{\mathbf{Q}}_{\mathbf{Z}})$$

where $[\widetilde{\mathbf{Q}}_{\mathbf{Z}}]_{ij} = \exp(-[\mathbf{C_Z}]_{ij}) / \sum_{\ell,t} \exp(-[\mathbf{C_Z}]_{\ell t})$.

SNEkhorn **OURS**

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times d}} \mathrm{KL}\left(\mathbf{P}^{\mathrm{se}} | \mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}}\right)$$

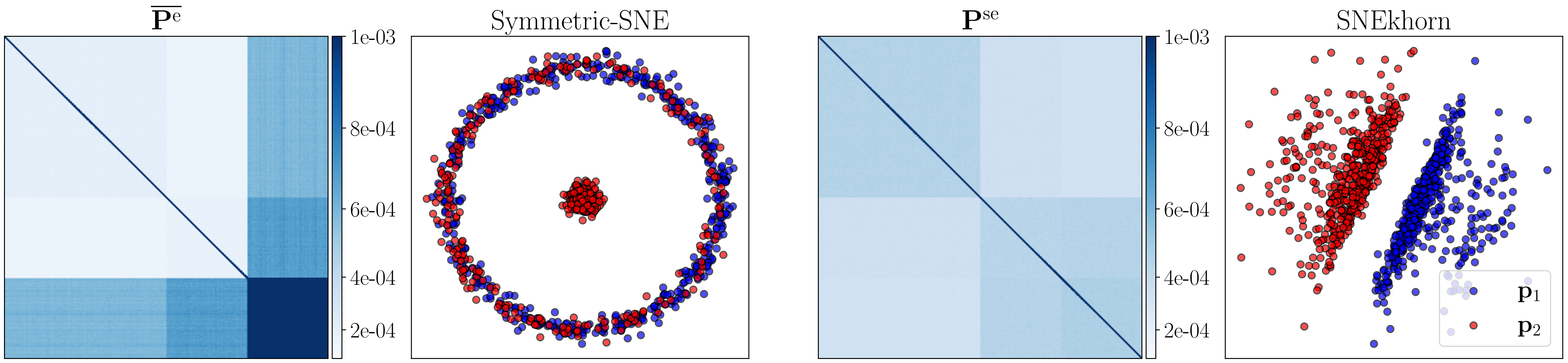
where $\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}} = \exp\left(\mathbf{f_Z} \oplus \mathbf{f_Z} - \mathbf{C_Z}\right)$ is the DS affinity.



Extension to t-SNE / t-SNEkhorn with heavy-tailed kernels:
[Van der Maaten and Hinton, 2008]

$$[\mathbf{C_Z}]_{ij} = \left(\log(1 + \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2)\right)_{ij}.$$

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.

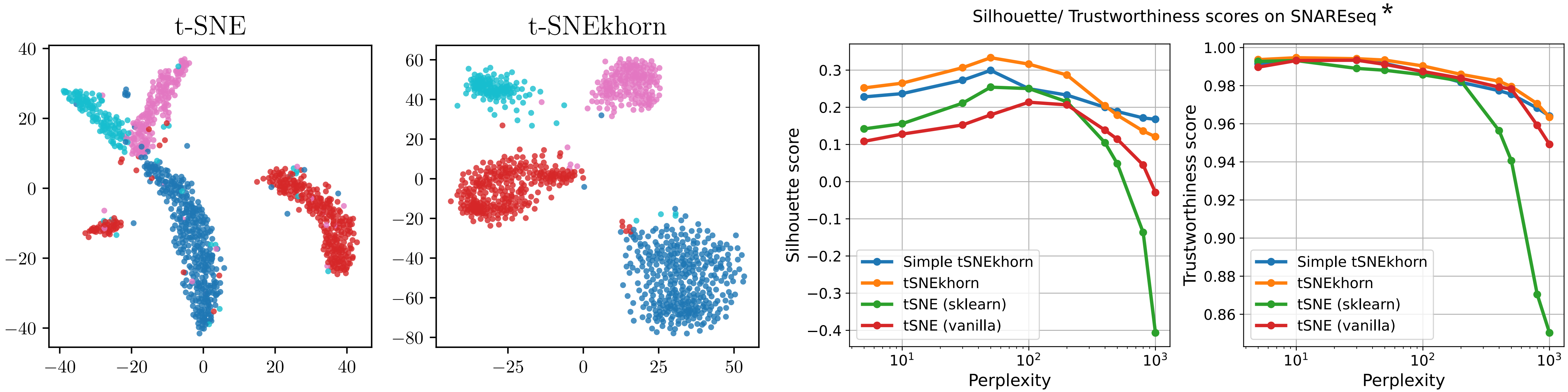
Dimension Reduction Results

	Silhouette ($\times 100$)			Trustworthiness ($\times 100$)		
	UMAP	t-SNE	t-SNEkhorn	UMAP	t-SNE	t-SNEkhorn
COIL	20.4 ± 3.3	30.7 ± 6.9	$\mathbf{52.3 \pm 1.1}$	99.6 ± 0.1	99.6 ± 0.1	$\mathbf{99.9 \pm 0.1}$
OLIVETTI	6.4 ± 4.2	4.5 ± 3.1	$\mathbf{15.7 \pm 2.2}$	96.5 ± 1.3	96.2 ± 0.6	$\mathbf{98.0 \pm 0.4}$
UMNIST	-1.4 ± 2.7	-0.2 ± 1.5	$\mathbf{25.4 \pm 4.9}$	93.0 ± 0.4	99.6 ± 0.2	$\mathbf{99.8 \pm 0.1}$
CIFAR	13.6 ± 2.4	18.3 ± 0.8	$\mathbf{31.5 \pm 1.3}$	90.2 ± 0.8	90.1 ± 0.4	$\mathbf{92.4 \pm 0.3}$
Liver (14520)	49.7 ± 1.3	50.9 ± 0.7	$\mathbf{61.1 \pm 0.3}$	89.2 ± 0.7	90.4 ± 0.4	$\mathbf{92.3 \pm 0.3}$
Breast (70947)	28.6 ± 0.8	29.0 ± 0.2	$\mathbf{31.2 \pm 0.2}$	90.9 ± 0.5	91.3 ± 0.3	$\mathbf{93.2 \pm 0.4}$
Leukemia (28497)	22.3 ± 0.7	20.6 ± 0.7	$\mathbf{26.2 \pm 2.3}$	90.4 ± 1.1	92.3 ± 0.8	$\mathbf{94.3 \pm 0.5}$
Colorectal (44076)	67.6 ± 2.2	69.5 ± 0.5	$\mathbf{74.8 \pm 0.4}$	93.2 ± 0.7	93.7 ± 0.5	$\mathbf{94.3 \pm 0.6}$
Liver (76427)	39.4 ± 4.3	38.3 ± 0.9	$\mathbf{51.2 \pm 2.5}$	85.9 ± 0.4	89.4 ± 1.0	$\mathbf{92.0 \pm 1.0}$
Breast (45827)	35.4 ± 3.3	39.5 ± 1.9	$\mathbf{44.4 \pm 0.5}$	93.2 ± 0.4	94.3 ± 0.2	$\mathbf{94.7 \pm 0.3}$
Colorectal (21510)	38.0 ± 1.3	$\mathbf{42.3 \pm 0.6}$	35.1 ± 2.1	85.6 ± 0.7	$\mathbf{88.3 \pm 0.9}$	88.2 ± 0.7
Renal (53757)	44.4 ± 1.5	45.9 ± 0.3	$\mathbf{47.8 \pm 0.1}$	93.9 ± 0.2	$\mathbf{94.6 \pm 0.2}$	94.0 ± 0.2
Prostate (6919)	5.4 ± 2.7	8.1 ± 0.2	$\mathbf{9.1 \pm 0.1}$	77.6 ± 1.8	$\mathbf{80.6 \pm 0.2}$	73.1 ± 0.5
Throat (42743)	26.7 ± 2.4	28.0 ± 0.3	$\mathbf{32.3 \pm 0.1}$	$\mathbf{91.5 \pm 1.3}$	88.6 ± 0.8	86.8 ± 1.0
scGEM	26.9 ± 3.7	33.0 ± 1.1	$\mathbf{39.3 \pm 0.7}$	95.0 ± 1.3	96.2 ± 0.6	$\mathbf{96.8 \pm 0.3}$
SNAREseq	6.8 ± 6.0	35.8 ± 5.2	$\mathbf{67.9 \pm 1.2}$	93.1 ± 2.8	99.1 ± 0.1	$\mathbf{99.2 \pm 0.1}$

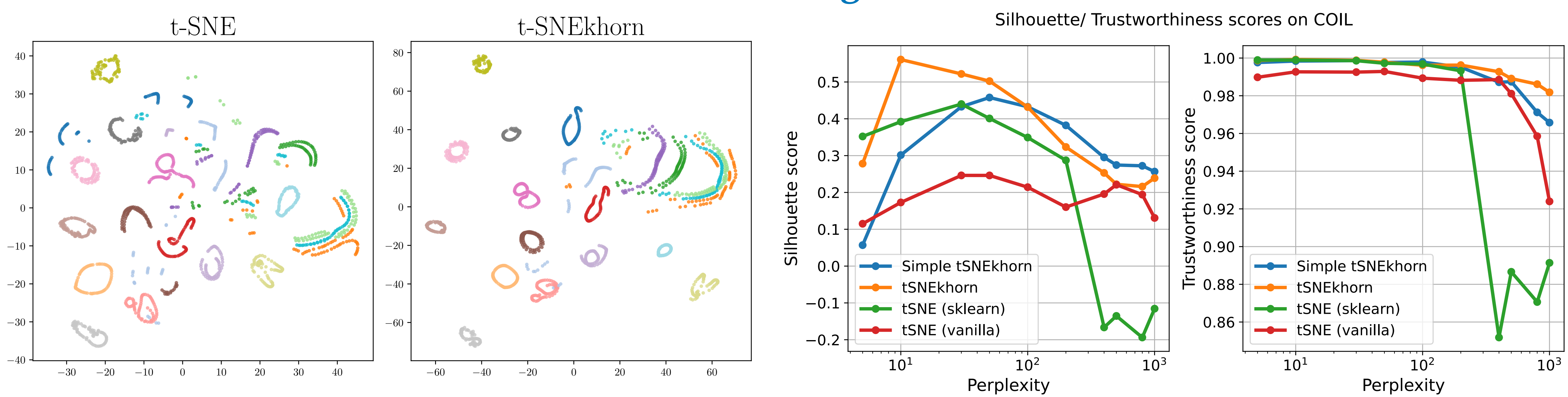
t-SNEkhorn outperforms **t-SNE** and **UMAP** on various real-world datasets.

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 **Z** (not doubly stochastic).

Conclusion

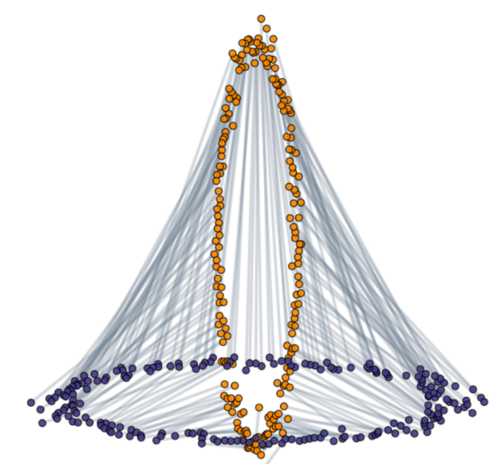
- We provide a new **symmetric** affinity matrix, **controlling for each row/column**:
  - ℓ_1 **norm** **&** **Shannon entropy.**

- We show its **robustness to heteroscedastic noise** (crucial for single cell data).

- Based on this affinity, we propose a **new DR method** : **SNEkhorn**.

- Python code available at :

  https://github.com/PythonOT/SNEkhorn

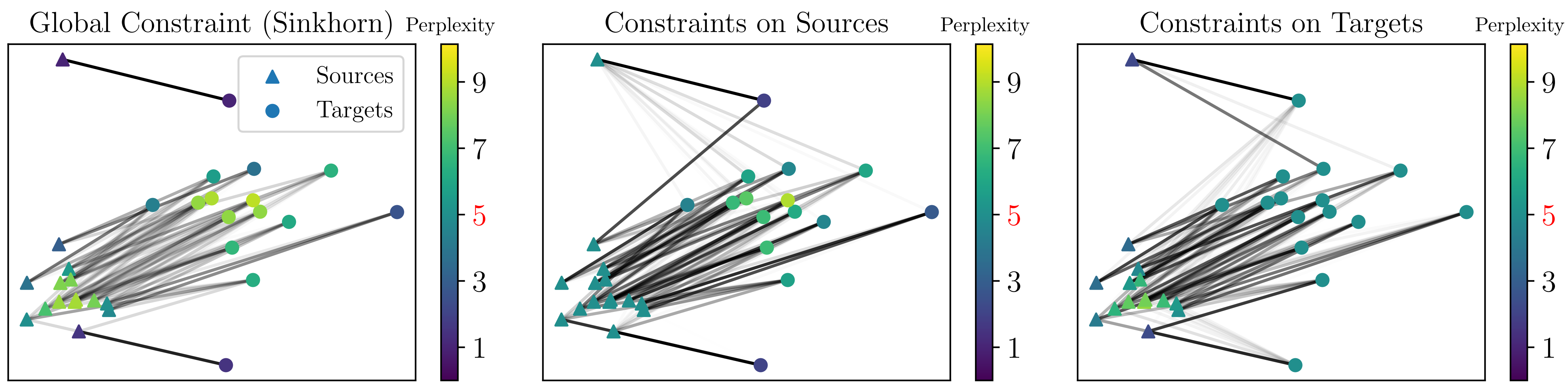


NEURAL INFORMATION PROCESSING SYSTEMS

SNEkhorn : Dimension Reduction with Symmetric Entropic Affinities NeurIPS 2023

Part III: Future Works : OT with Adaptive Regularisation

OT with Adaptive Regularisation



Application to Domain Adaptation