SNEkhorn

Dimension Reduction with Symmetric Entropic Affinities

Hugues Van Assel, Titouan Vayer, **Rémi Flamary**, Nicolas Courty

September 15 2023

New Monge Problems and Applications, University Gustave Eiffel


H. Van Assel


T. Vayer

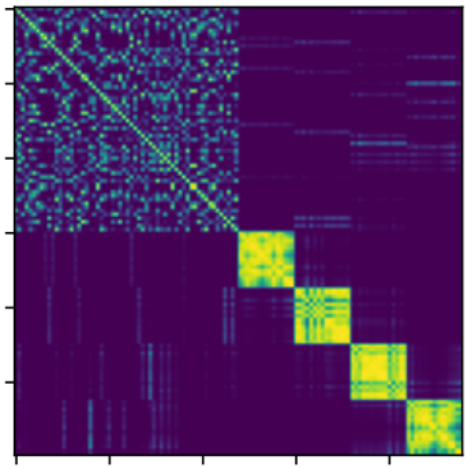

R. Flamary

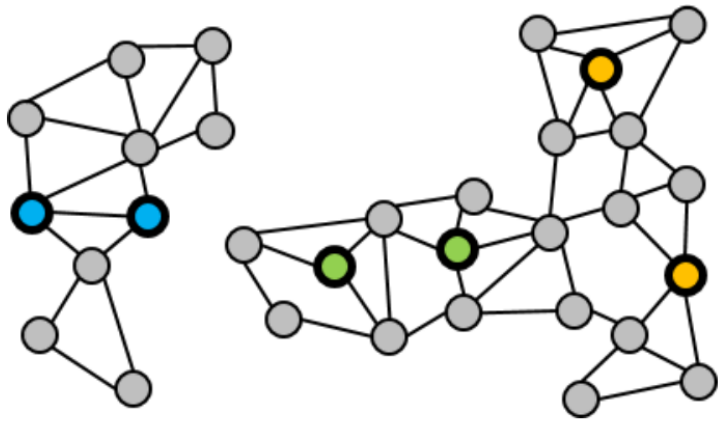

N. Courty

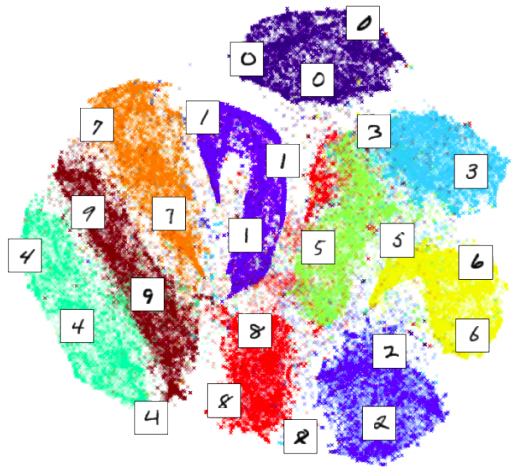
Table of content

Affinity matrices in machine learning

Affinity matrices in machine learning

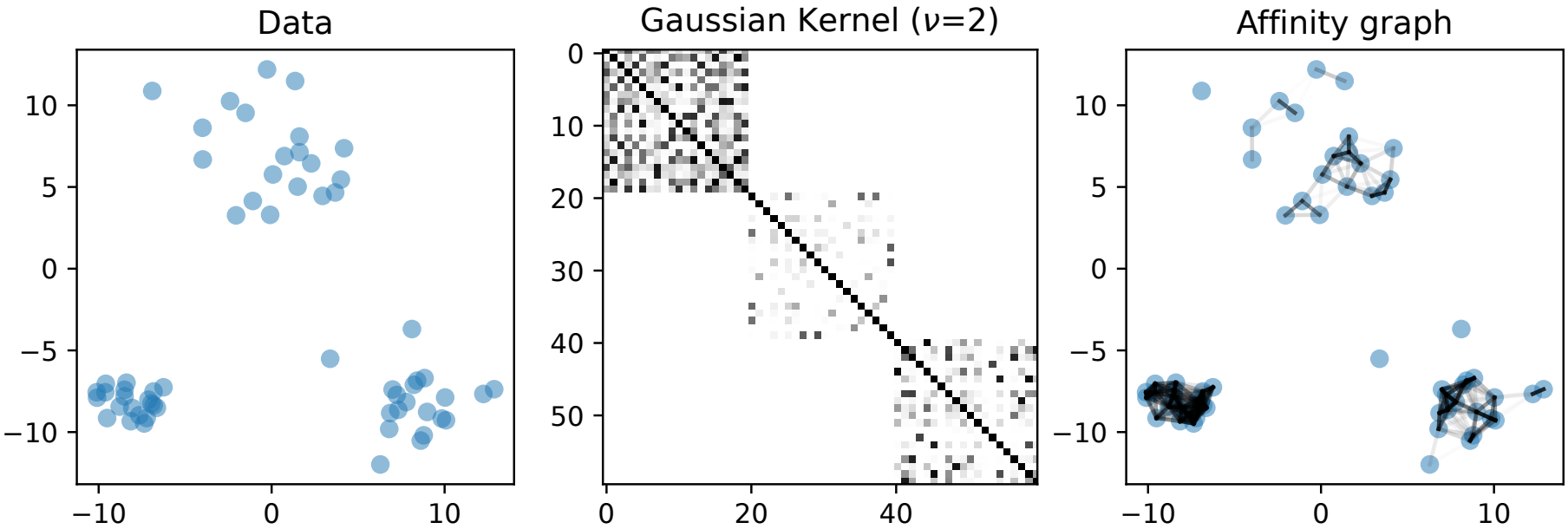






- Many ML methods rely on similarity/affinity matrices to encode the relationship between data points (graph, kernel or similarity).
- Examples :
  - Kernel machines [Schölkopf and Smola, 2002].
  - Clustering (spectral, kernel) [Von Luxburg, 2007].
  - Dimensionality reduction (TNSE [Van der Maaten and Hinton, 2008]).
  - Semi-supervised learning[Zhou et al., 2003]
  - Self-supervised learning (Barlow twins [Zbontar et al., 2021]).

Kernel matrices

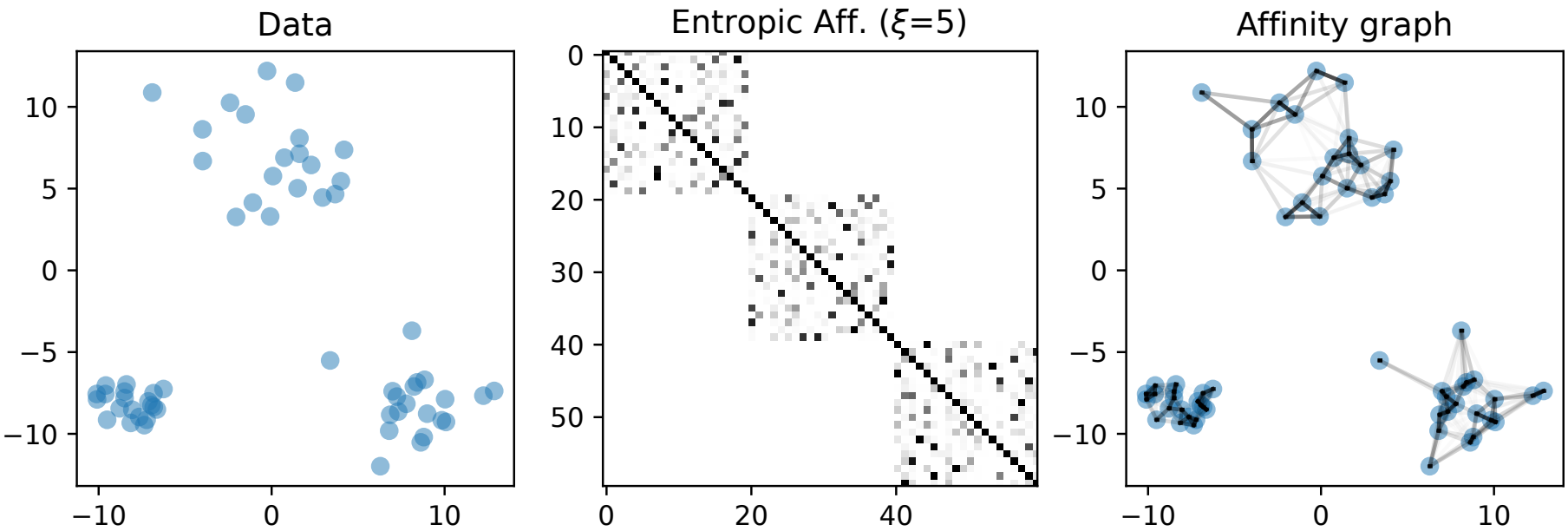


Gaussian kernel

$$K_{ij}^{g} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\nu) = \exp(-C_{ij}/\nu) \qquad \text{(Gaussian kernel)}$$

- $\{\mathbf{x}_i\}_{i=1,\ldots,n}$ are data points in $\mathbb{R}^d$.
- $\mathbf{K}$ is the kernel matrix of components K_{ij}.
- ν is a parameter tuning the neighborhood size.
- Used in support vector machines, spectral clustering, etc.

Entropic Affinity matrices

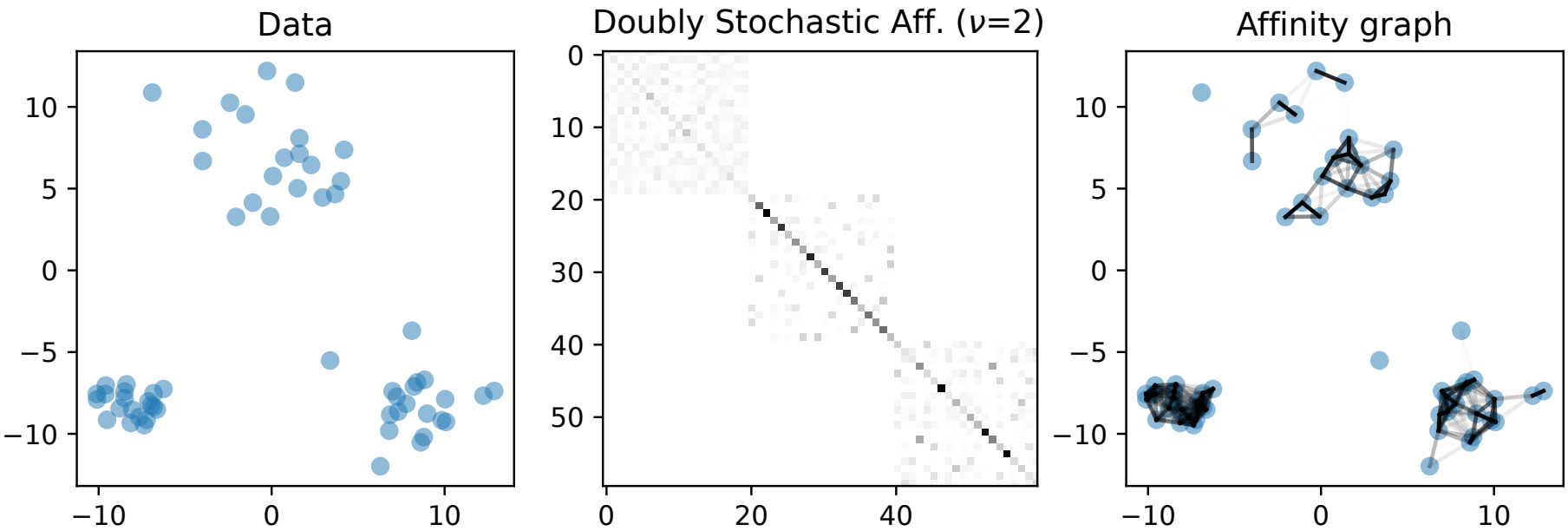


Entropic affinities (perplexity ξ)

$$P^{\mathrm{e}}_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon^\star_i\right)}{\sum_\ell \exp\left(-C_{i\ell}/\varepsilon^\star_i\right)} \quad \text{with } \varepsilon^\star_i \in \mathbb{R}^*_+ \text{ s.t. } \mathrm{H}(\mathbf{P}^{\mathrm{e}}_{i:}) = \log\xi + 1\,. \qquad \text{(EA)}$$

- $H(\mathbf{v}) = -\sum_i v_i(\log(v_i) - 1)$ is the entropy and $C_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$.
- Adaptive scaling $\varepsilon^\star_i$ per point to ensure an equivalent "spread" of the mass.
- $\mathbf{P}^{\mathrm{se}}$ is not symmetric. **Symmetric variant** : $\overline{\mathbf{P}^{\mathrm{se}}} = (\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top})/2$.
- Used in Stochastic Neighbor Embedding (SNE) [Hinton and Roweis, 2002] and tSNE [Van der Maaten and Hinton, 2008] (symmetric variant).

Doubly Stochastic affinity matrix



DS affinity matrix

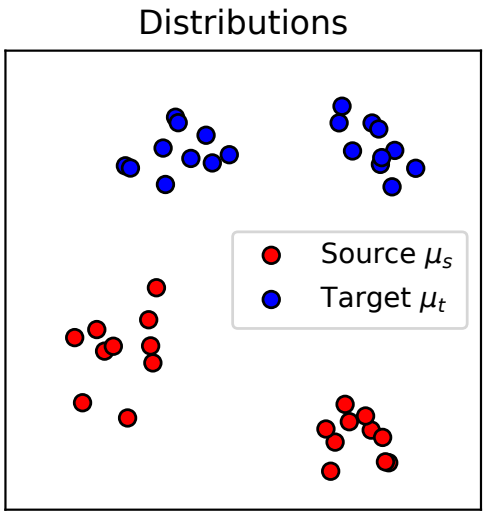
$$P_{ij}^{\mathrm{ds}} = \exp\left((f_i + f_j - C_{ij})/\nu\right) \text{ where } \mathbf{f} \in \mathbb{R}^n \,. \tag{DS}$$

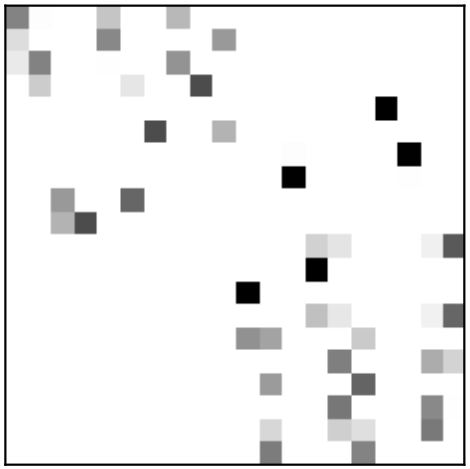
- Can be solved (estimation of $\mathbf{f}$) using the Sinkhorn-Knopp algorithm with iterations : $f_i^{t+1} \leftarrow \frac{1}{2}\left(f_i^t - \log \sum_k \exp\left(f_k^t - C_{ki}\right)\right) \ \forall i.$
- Projection of $\mathbf{K}$ on the set of doubly stochastic matrices Π:

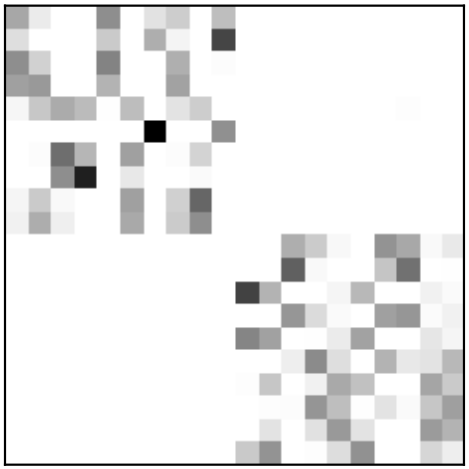
$$\mathbf{P}^{\mathrm{ds}} = \min_{\mathbf{P} \in \Pi} \ \mathrm{KL}(\mathbf{P}|\mathbf{K})$$

- Equivalent to self entropic regularized optimal transport [Cuturi, 2013].

Entropic regularized optimal transport







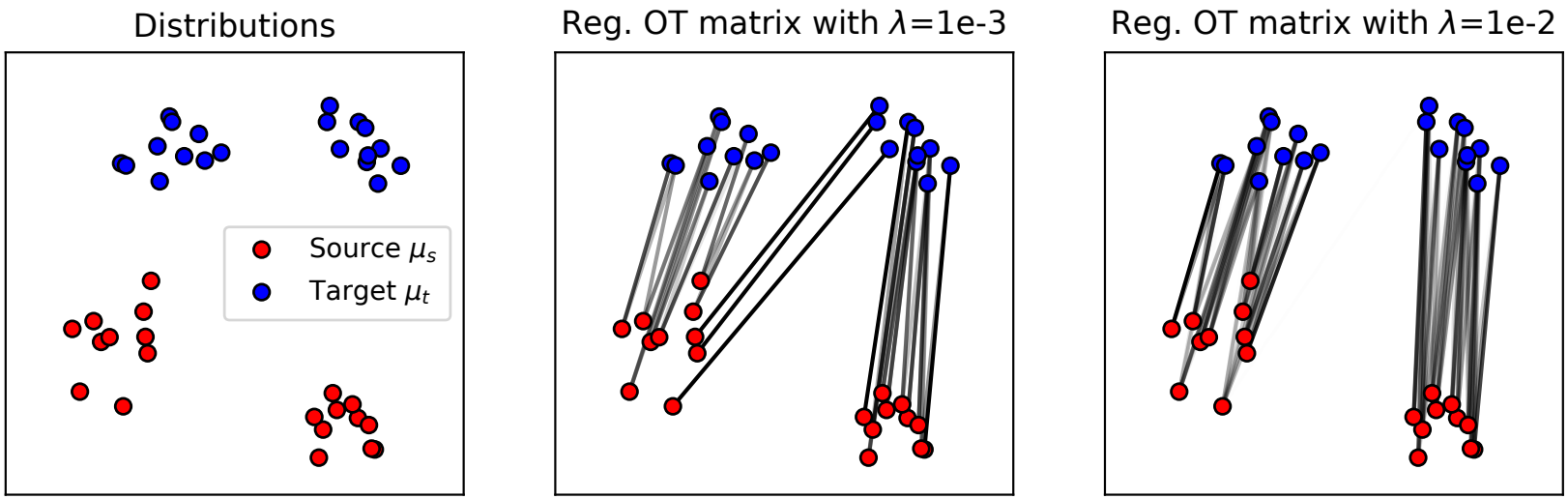
Entropic regularized OT [Cuturi, 2013]

$$\mathbf{P}^{ds} = \underset{\mathbf{P}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{P}, \mathbf{C} \rangle_F - \nu H(\mathbf{P})$$

- Regularization with the entropy of $H(\mathbf{P}) = -\sum_{i,j} P_{i,j}(\log P_{i,j} - 1)$.
- Looses sparsity, gains stability, strictly convex, solved with Sinkhorn.
- Equivalent to the following problem (global constraint on entropy)

$$\mathbf{P}^{ds} = \underset{\mathbf{P}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{P}, \mathbf{C} \rangle_F \quad \text{s.t.} \quad H(\mathbf{P}) \geq \eta$$

Entropic regularized optimal transport



Entropic regularized OT [Cuturi, 2013]

$$\mathbf{P}^{ds} = \underset{\mathbf{P}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{P}, \mathbf{C} \rangle_F - \nu H(\mathbf{P})$$

- Regularization with the entropy of $H(\mathbf{P}) = -\sum_{i,j} P_{i,j}(\log P_{i,j} - 1)$.
- Looses sparsity, gains stability, strictly convex, solved with Sinkhorn.
- Equivalent to the following problem (global constraint on entropy)

$$\mathbf{P}^{ds} = \underset{\mathbf{P}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{P}, \mathbf{C} \rangle_F \quad \text{s.t.} \quad H(\mathbf{P}) \geq \eta$$

Symmetric Entropic Affinities (SEA)

Entropic affinities seen as an OT problem

Set of affinity matrices with row entropic constraints

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

EA matrices are in this set $\mathbf{P}^e \in \mathcal{H}_\xi$

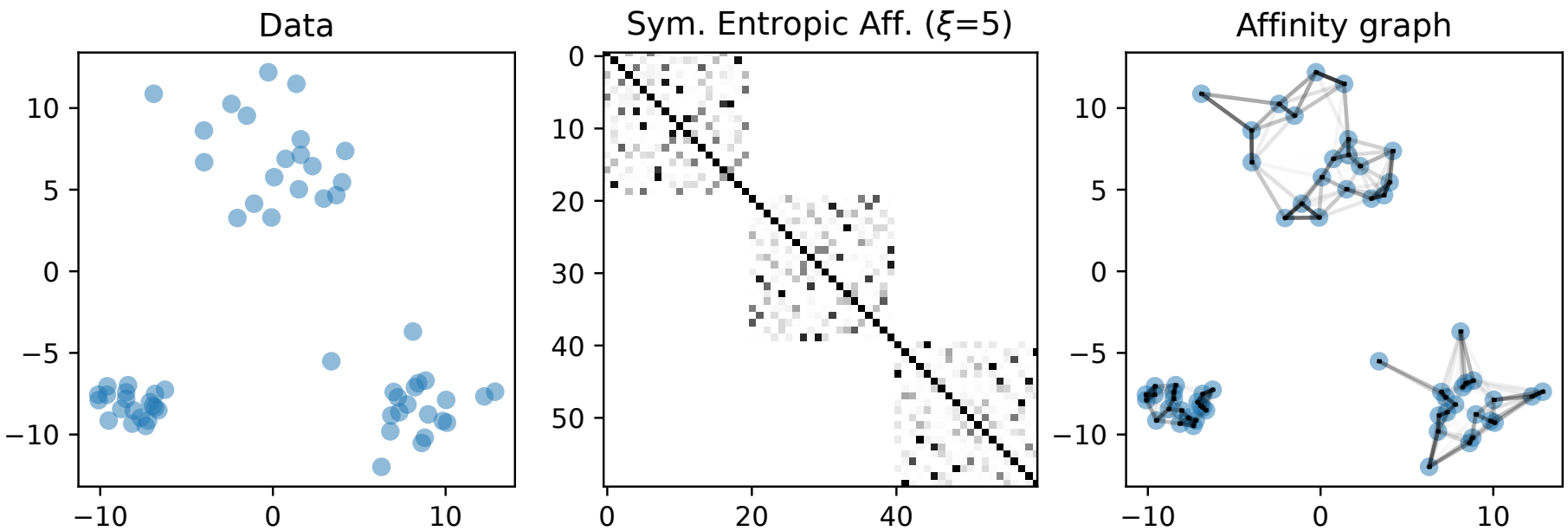
EA matrices seen as an OT problem

Let $\mathbf{C} \in \mathbb{R}^{n\times n}$ without constant rows. Then $\mathbf{P}^{\mathrm{e}}$ solves the entropic affinity problem (EA) with cost $\mathbf{C}$ if and only if $\mathbf{P}^{\mathrm{e}}$ is the unique solution of the convex problem

$$\mathbf{P}^e = \underset{\mathbf{P}\in\mathcal{H}_\xi}{\operatorname{argmin}} \langle \mathbf{P}, \mathbf{C}\rangle. \tag{EA as OT}$$

- EA matrix computation is a semi-relaxed OT with line entropy constraints.
- The solution $\mathbf{P}^e$ has saturated entropy with equality in the constraints.
- Can be solved with n independent root-finding algorithms.

Symmetric Entropic Affinities (SEA)

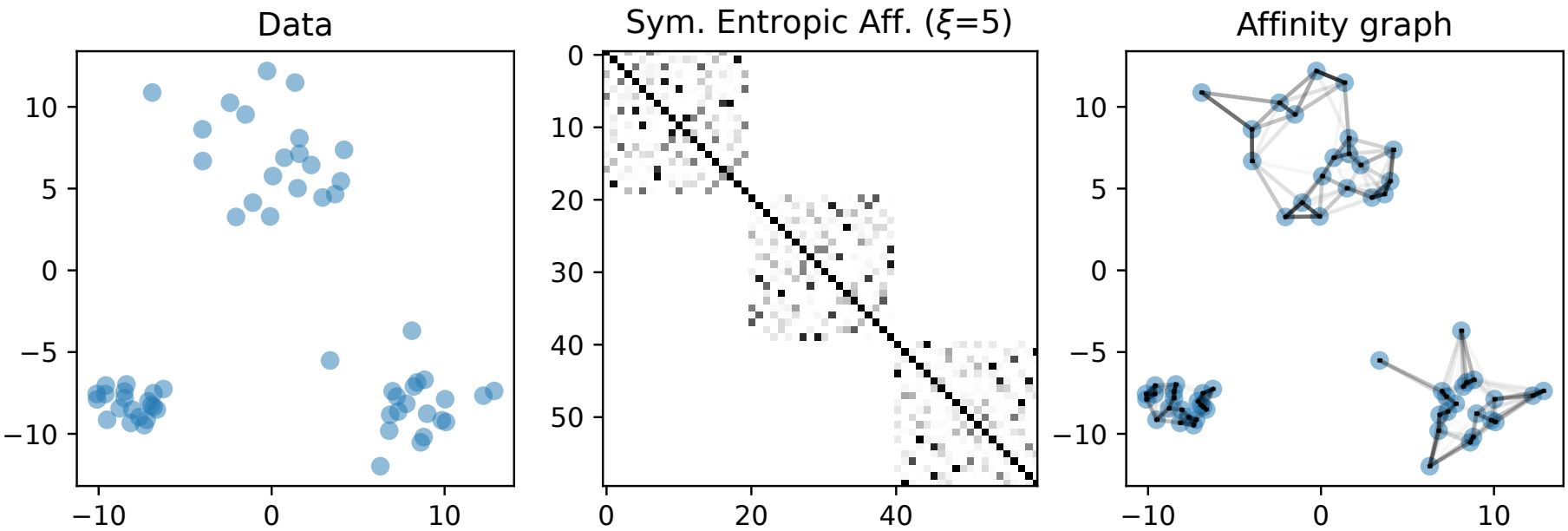


Problem formulation for SEA

$$\mathbf{P}^{se} = \underset{\mathbf{P} \in \mathcal{H}_\xi \cap \mathcal{S}}{\operatorname{argmin}} \langle \mathbf{P}, \mathbf{C} \rangle \,. \qquad \text{(SEA)}$$

- $\mathcal{S} = \{\mathbf{P} \in \mathbb{R}_+^{n \times n} \text{ s.t. } \mathbf{P} = \mathbf{P}^\top\}$ is the set of symmetric matrices.
- $\mathbf{P}^{se}$ is the unique solution of the convex problem (SEA) and has at least $n-1$ saturated entropy constraints (in practice we have n).

Optimizing Symmetric Entropic Affinities



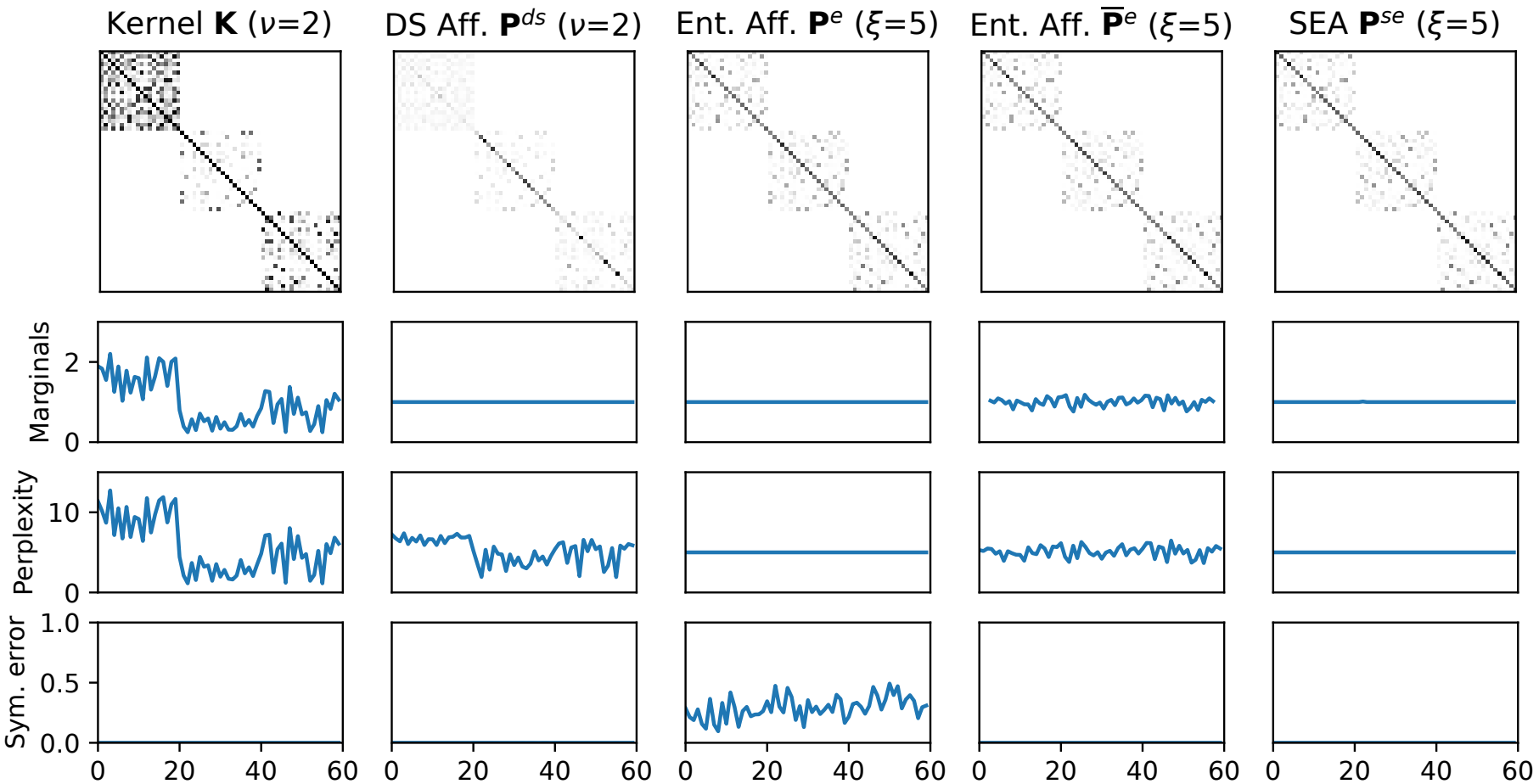
Solving for SEA

- Strong duality holds and the dual problem is

$$\max_{\boldsymbol{\gamma}>0,\boldsymbol{\lambda}} \quad \langle \mathbf{P}(\boldsymbol{\gamma},\boldsymbol{\lambda}), \mathbf{C}\rangle + \langle \boldsymbol{\gamma}, (\log \xi + 1)\mathbf{1} - \mathrm{H_r}(\mathbf{P}(\boldsymbol{\gamma},\boldsymbol{\lambda}))\rangle + \langle \boldsymbol{\lambda}, \mathbf{1} - \mathbf{P}(\boldsymbol{\gamma},\boldsymbol{\lambda})\mathbf{1}\rangle$$

  where $\mathbf{P}(\boldsymbol{\gamma},\boldsymbol{\lambda}) = \exp\left((\boldsymbol{\lambda} \oplus \boldsymbol{\lambda} - 2\mathbf{C}) \oslash (\boldsymbol{\gamma} \oplus \boldsymbol{\gamma})\right)$.
- Solution is $\mathbf{P}^{se} = \mathbf{P}(\boldsymbol{\gamma}^\star, \boldsymbol{\lambda}^\star)$ for optimal dual variables $\boldsymbol{\gamma}^\star, \boldsymbol{\lambda}^\star$.
- Dual optimizer (L-BFGS, ADAM, etc.) is used to solve the dual problem in practice.

Comparison between all affinity matrices



- For all affinities we compute the marginals, the entropy and the TV symmetry error for all samples.

Illustration of symmetric entropic affinities

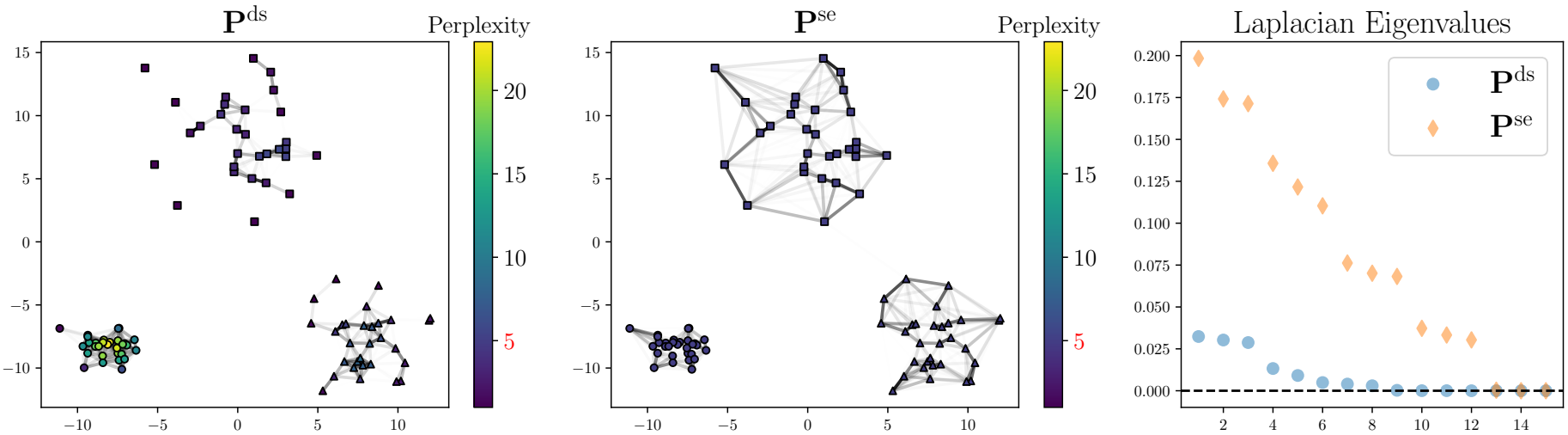
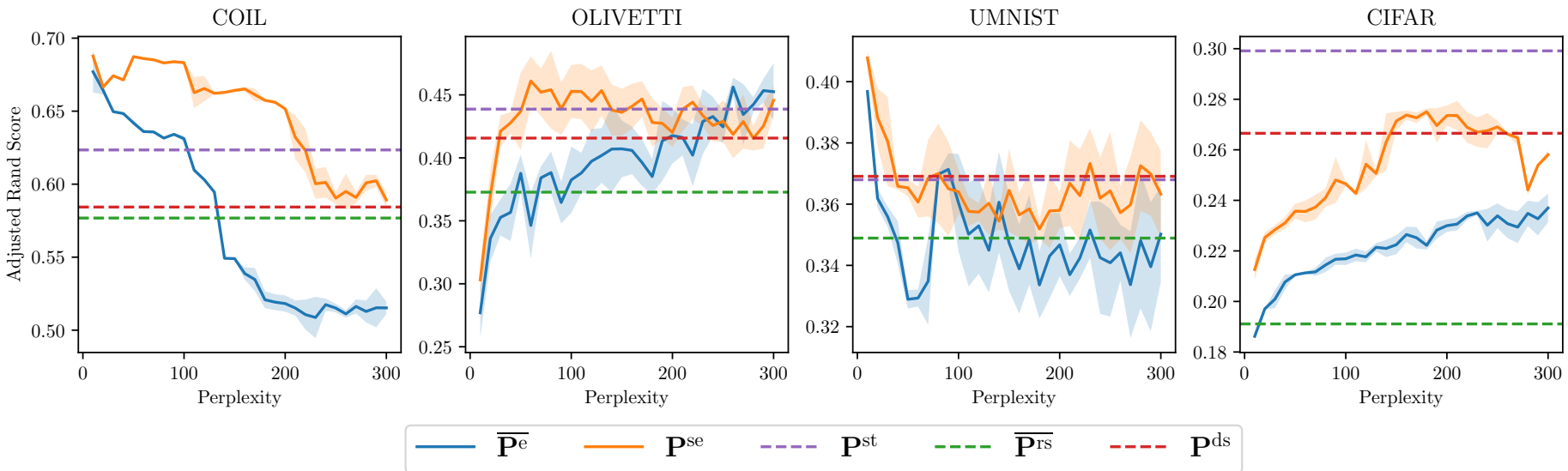


Illustration on 2D example (mixture of Gaussians)

- Comparison between Doubly Stochastic affinity and symmetric entropic affinity.
- Symmetric entropic affinity has a constant perplexity.
- Fixed perplexity adapt better to cluster of different sizes (local density).
- Eigenvalues of the Laplacian are much more separated (better clustering).

SEA for spectral clustering on image data



Experiment on image datasets

- Compute Adjusted Rand Index (ARI) for different affinities.
- Plot evolution of ARI as a function of perplexity.
- SEA is state of the art except on CIFAR.

SEA for spectral clustering on Curated Microarray Database (CuMiDa)

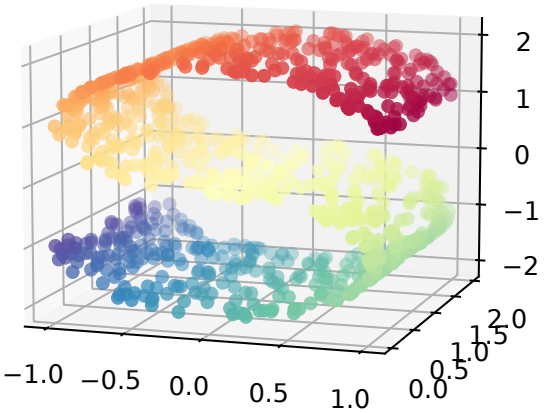
Data set	$\overline{\mathbf{P}^{\mathrm{rs}}}$	$\mathbf{P}^{\mathrm{ds}}$	$\mathbf{P}^{\mathrm{st}}$	$\overline{\mathbf{P}^{\mathrm{e}}}$	$\mathbf{P}^{\mathrm{se}}$
Liver (14520)	75.8	75.8	84.9	80.8	**85.9**
Breast (70947)	**30.0**	**30.0**	26.5	23.5	28.5
Leukemia (28497)	43.7	44.1	49.7	42.5	**50.6**
Colorectal (44076)	**95.9**	**95.9**	93.9	**95.9**	**95.9**
Liver (76427)	76.7	76.7	**83.3**	81.1	81.1
Breast (45827)	43.6	53.8	74.7	71.5	**77.0**
Colorectal (21510)	57.6	57.6	54.7	**94.0**	79.3
Renal (53757)	47.6	47.6	**49.5**	**49.5**	**49.5**
Prostate (6919)	12.0	13.0	13.2	16.3	**17.4**
Throat (42743)	9.29	9.29	11.4	11.8	**44.2**
scGEM	57.3	58.5	**74.8**	69.9	71.6
SNAREseq	8.89	9.95	46.3	55.4	**96.6**

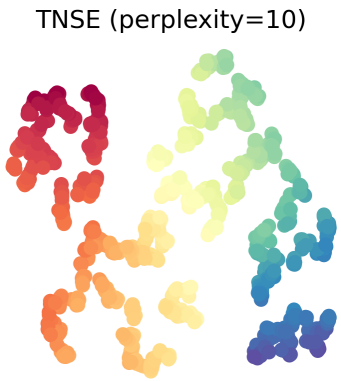
Numerical experiments

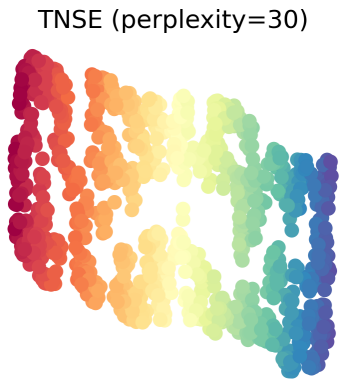
- ARI ($\times 100$) for spectral clustering reported on CuMiDa datasets.
- Curated Microarray Database [Feltes et al., 2019].
- SEA is state of the art on 8/12 datasets.

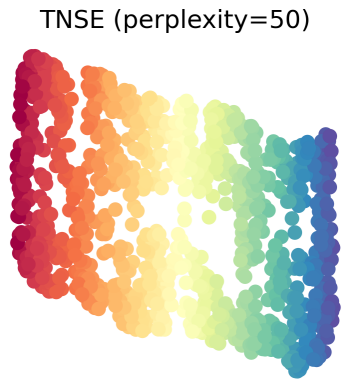
Dimensionality reduction with SNEkhorn

Stochastic Neighbor Embedding (SNE) and tSNE









Symmetric SNE [Van der Maaten and Hinton, 2008]

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times q}} \mathrm{KL}(\overline{\mathbf{P}^{\mathrm{e}}}|\widetilde{\mathbf{Q}}_{\mathbf{Z}}) \quad \text{where} \quad \overline{\mathbf{P}^{\mathrm{e}}} = \frac{1}{2}(\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top}). \qquad \text{(Symmetric-SNE)}$$

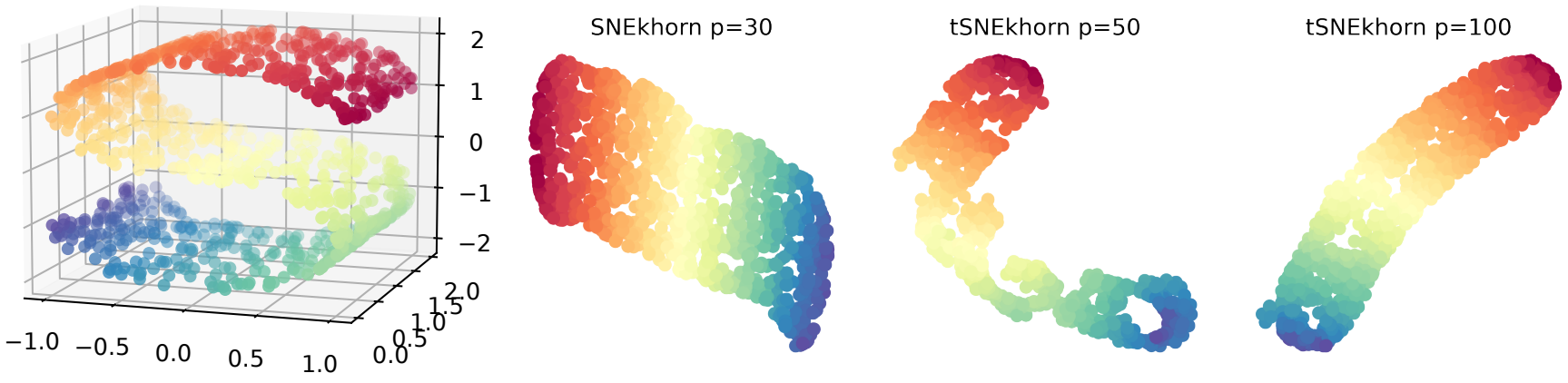
with $[\widetilde{\mathbf{Q}}_{\mathbf{Z}}]_{ij} = \exp(-[\mathbf{C}_{\mathbf{Z}}]_{ij})/(\sum_{\ell t}\exp(-[\mathbf{C}_{\mathbf{Z}}]_{\ell t}))$ and $q \leq d$.

- Minimize the Kullback-Leibler divergence between the affinities of the data in the original space and the affinities of the embedded data.
- Embedding $\mathbf{Z}$ computed by gradient descent.

Other variants

- Original **SNE** [Hinton and Roweis, 2002] uses $\mathbf{P}^{\mathrm{e}}$ and $\mathbf{Q}_{\mathbf{Z}}$ normalized by row.
- **tSNE** uses $\overline{\mathbf{P}^{\mathrm{e}}}$ and $[\widetilde{\mathbf{Q}}_{\mathbf{Z}}]_{ij} = (1+[\mathbf{C}_{\mathbf{Z}}]_{ij})^{-1}/\sum_{\ell,t}(1+[\mathbf{C}_{\mathbf{Z}}]_{\ell t})^{-1}$

SNEkhorn optimization problem

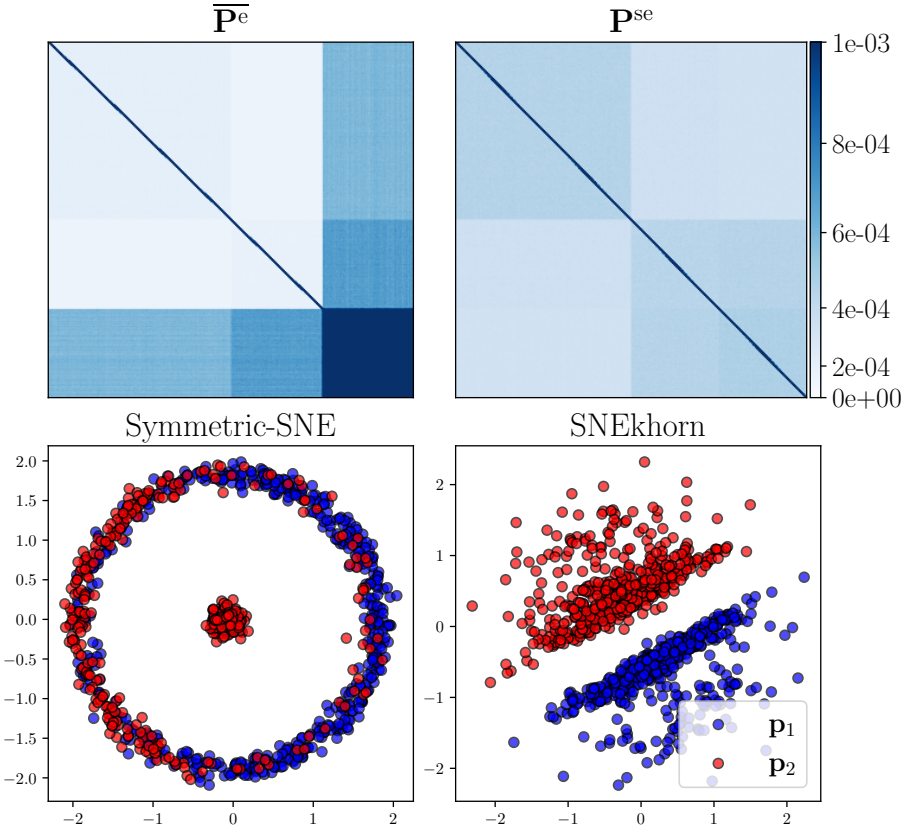


SNEkhorn

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times q}} \mathrm{KL}\big(\mathbf{P}^{\mathrm{se}}|\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}}\big)\,, \qquad \text{(SNEkhorn)}$$

- $\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}} = \exp\left(\mathbf{f}_{\mathbf{Z}} \oplus \mathbf{f}_{\mathbf{Z}} - \mathbf{C}_{\mathbf{Z}}\right)$ is computed with Sinkhorn algorithm.
- We set the bandwidth to $\nu = 1$ in $\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}}$.
- Optimized with gradient descent and fast computation/update of Sinkhorn dual variables $\mathbf{f}_{\mathbf{Z}}$ with warm starting strategy.
- Variants:
  - **SNEkhorn** : $[\mathbf{C}_{\mathbf{Z}}]_{ij} = \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2$
  - **tSNEkhorn** : $[\mathbf{C}_{\mathbf{Z}}]_{ij} = \big(\log(1 + \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2)\big)$

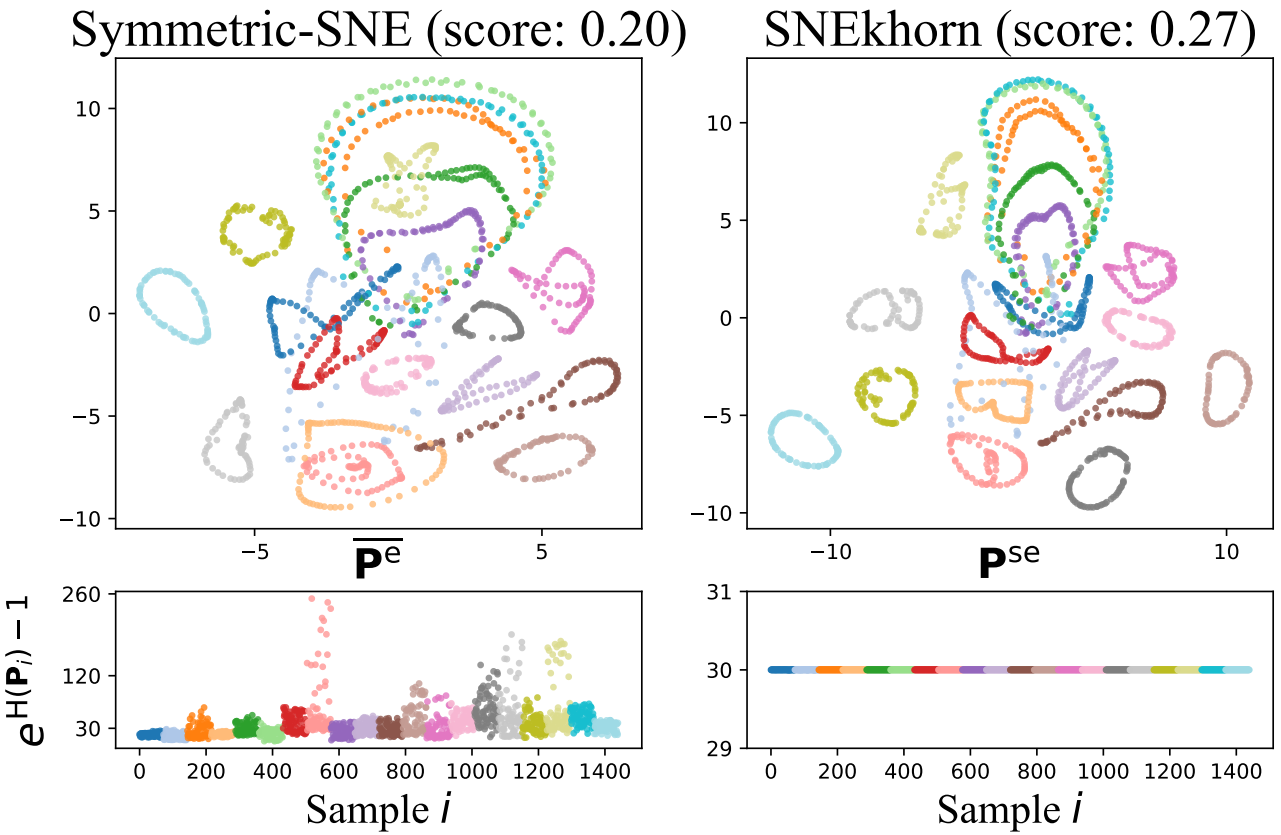
SNE vs SNEkhorn (1)



Experiment on simulated data

- Simulated data with heteroscedastic noise.
- Two classes from multinomial distribution with different probability vectors.
- Second class has samples with either low or high variance.
- Model used for biomedical data.

SNE vs SNEkhorn (2)



Experiment on COIL (20) dataset [Nene et al., 1996]

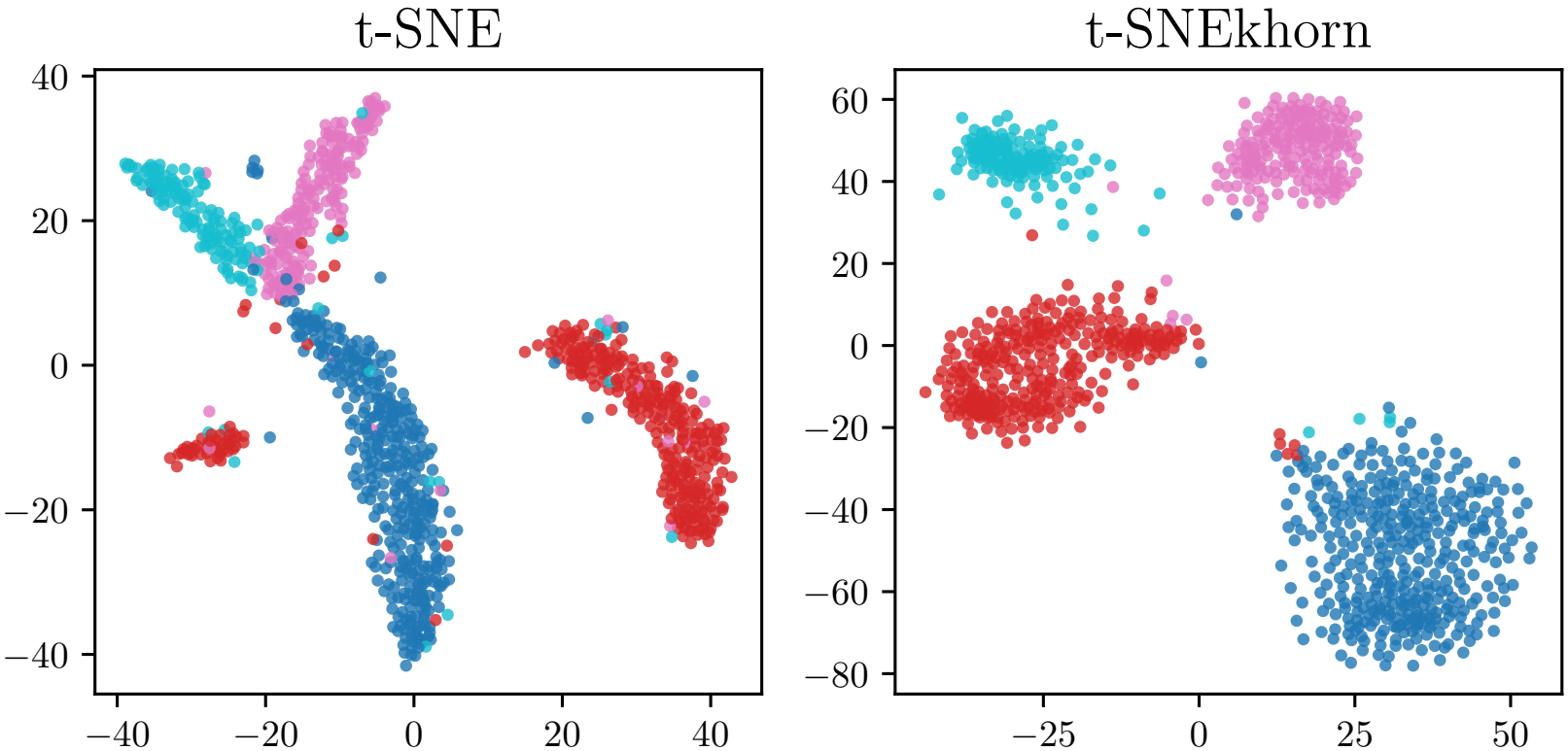
- Dataset of images of 20 objects with different orientations.
- Silhouette score of the embeddings en sample-wise perplexity reported.
- SNEkhorn has a constant perplexity for all samples
- SNEkhorn separates better visually and has a better score.

Dimensionality reduction performance

	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	$\mathbf{96.8ß \pm 0.3}$
SNAREseq	6.8 ± 6.0	35.8 ± 5.2	$\mathbf{67.9ß \pm 1.2}$	93.1 ± 2.8	99.1 ± 0.1	$\mathbf{99.2 \pm 0.1}$

- Comparison for different DR methods.
- Silhouette and Trustworthiness scores reported.
- t-SNEkhorn is state of the art on majority of criterion/datasets.

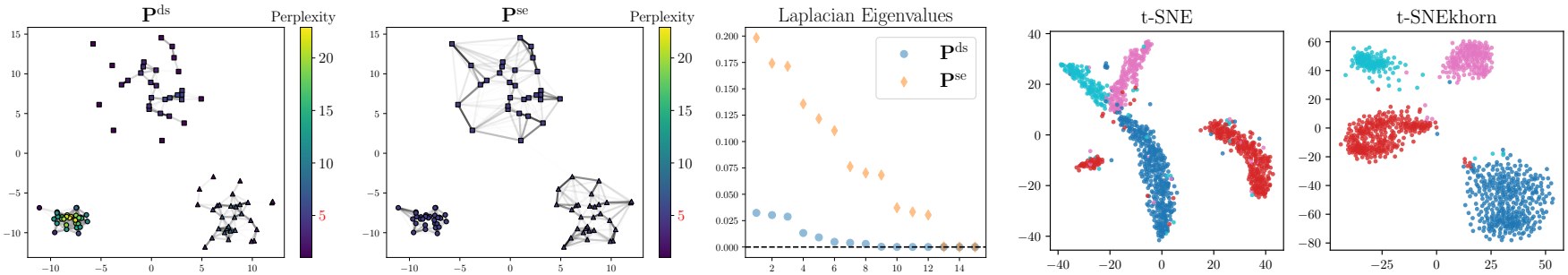
DR vizualisation on SNAREsed dataset



- Comparison of 2D visualization on SNAREseq dataset.
- tSNEkhorn has better class separability that tSNE.

Conclusion

Conclusion



Symmetric entropic affinities and SNEkhorn

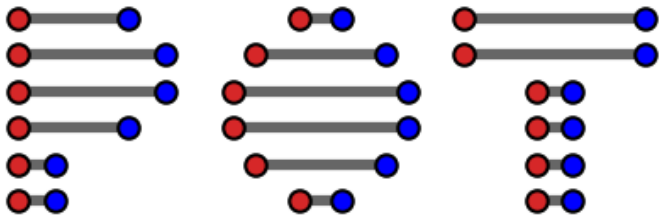
- SEA are adaptive, symmetric and doubly stochastic.
- SEA have good performances for spectral clustering.
- (t)SNEkhorn is a new algorithm for nonlinear dimensionality reduction.
- (t)SNEkhorn is more stable and has better performance.

Future works

- OT with point-wize entropy constraint (between different distributions).
- Implement SNEkhorn with all the (t)SNE accelerations.
- relations between Gromov-Wasserstein and DR.

Thank you

Python code available on GitHub:



Python code available on GitHub:

`https://github.com/PythonOT/POT`

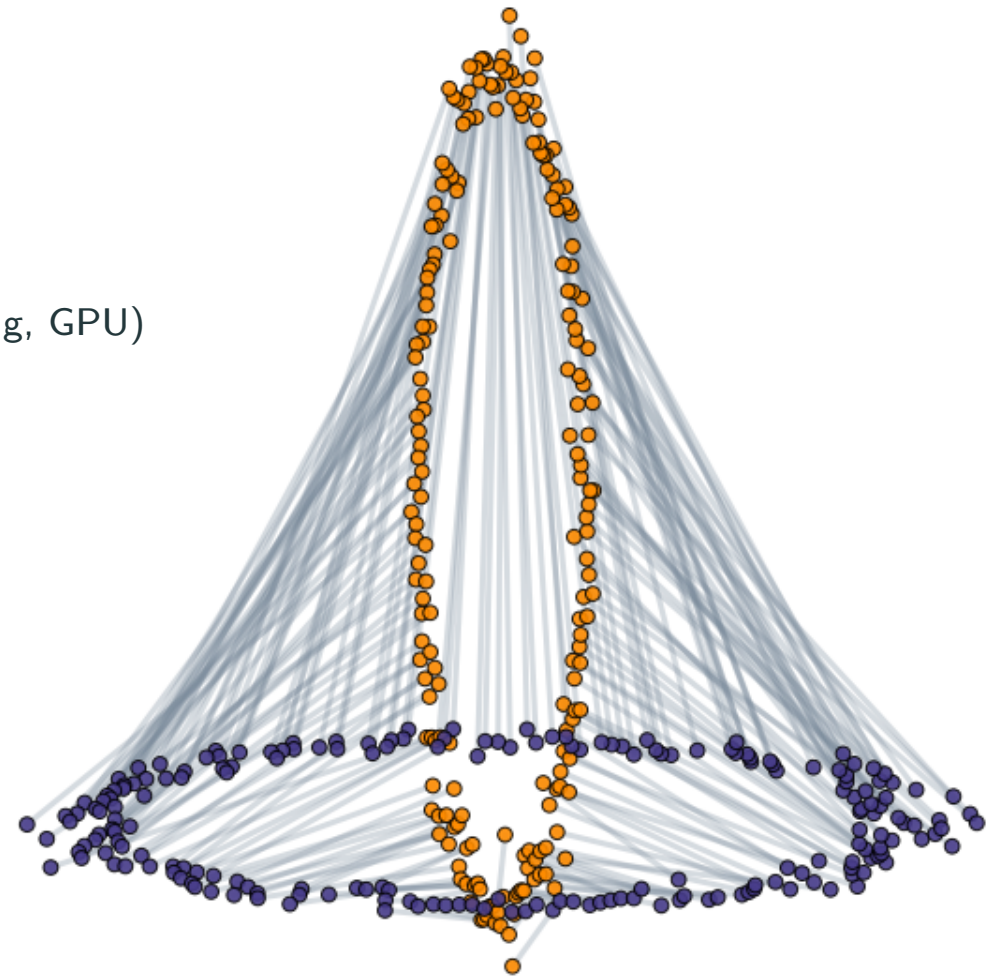
- OT LP solver, Sinkhorn (stabilized, $\epsilon-$scaling, GPU)
- Domain adaptation with OT.
- Barycenters, Wasserstein unmixing.
- Wasserstein Discriminant Analysis.

Tutorial on OT for ML:

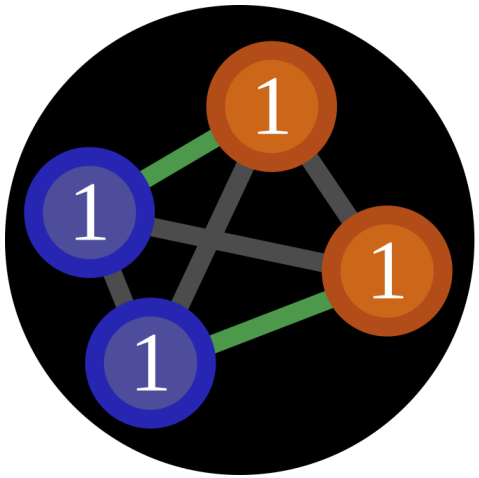
`http://tinyurl.com/otml-isbi`

Papers available on my website:

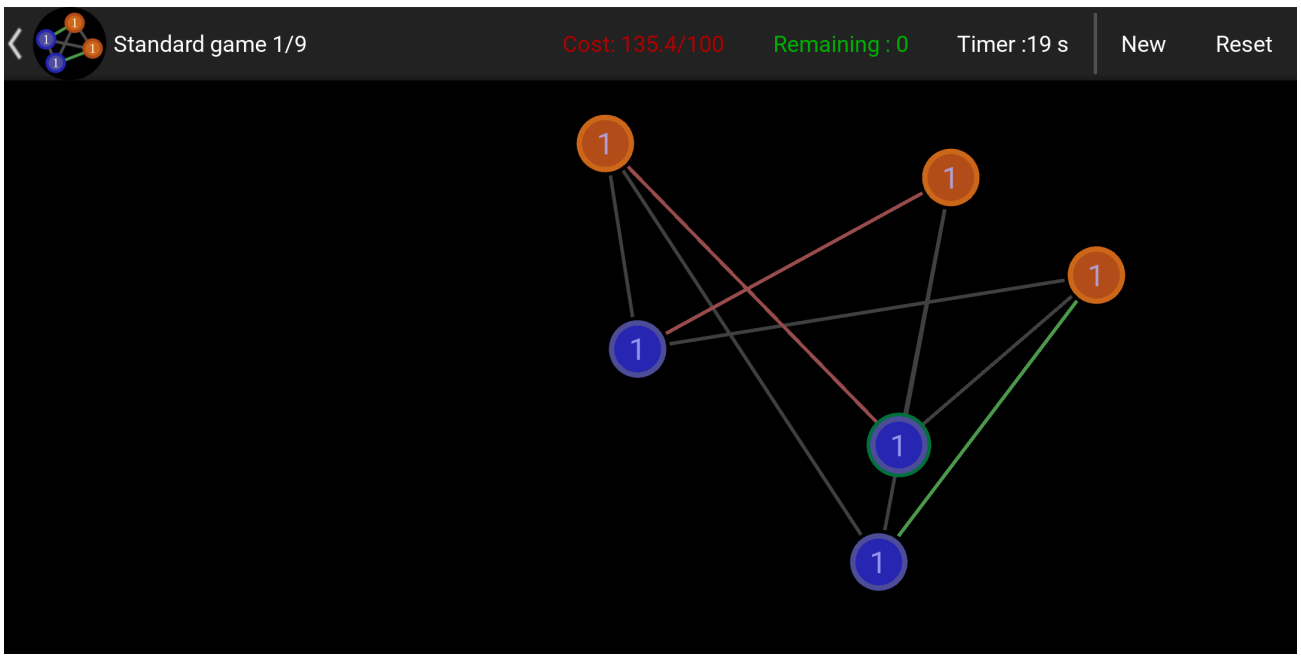
`https://remi.flamary.com/`



OTGame (OT Puzzle game on android)



OTGame



https://play.google.com/store/apps/details?id=com.flamary.otgame

References i

Cuturi, M. (2013).

Sinkhorn distances: Lightspeed computation of optimal transportation.

In *Neural Information Processing Systems (NIPS)*, pages 2292–2300.

Feltes, B. C., Chandelier, E. B., Grisci, B. I., and Dorn, M. (2019).

Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research.

Journal of Computational Biology, 26(4):376–386.

PMID: 30789283.

Hinton, G. E. and Roweis, S. (2002).

Stochastic neighbor embedding.

Advances in neural information processing systems, 15.

Nene, S. A., Nayar, S. K., Murase, H., et al. (1996).

Columbia object image library (coil-20).

References ii

Schölkopf, B. and Smola, A. J. (2002).

Learning with kernels: Support vector machines, regularization, optimization, and beyond.

MIT press.

Van der Maaten, L. and Hinton, G. (2008).

Visualizing data using t-sne.

Journal of Machine Learning Research, 9(2579-2605):85.

Von Luxburg, U. (2007).

A tutorial on spectral clustering.

Statistics and computing, 17:395–416.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021).

Barlow twins: Self-supervised learning via redundancy reduction.

In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

References iii

Zhou, D., Bousquet, O., Lal, T., Weston, J., and Schölkopf, B. (2003).

Learning with local and global consistency.

Neural Information Processing Systems (NeurIPS), 16.