SNEkhorn: Dimension Reduction with Symmetric Entropic Affinities
Hugues Van Assel, Titouan Vayer, Rémi Flamary, Nicolas Courty

To cite this version:
Hugues Van Assel, Titouan Vayer, Rémi Flamary, Nicolas Courty. SNEkhorn: Dimension Reduction with Symmetric Entropic Affinities. Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS), Dec 2023, New Orleans, United States. hal-04103326v2

HAL Id: hal-04103326
https://hal.science/hal-04103326v2
Submitted on 27 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Abstract

Many approaches in machine learning rely on a weighted graph to encode the similarities between samples in a dataset. Entropic affinities (EAs), which are notably used in the popular Dimensionality Reduction (DR) algorithm t-SNE, are particular instances of such graphs. To ensure robustness to heterogeneous sampling densities, EAs assign a kernel bandwidth parameter to every sample in such a way that the entropy of each row in the affinity matrix is kept constant at a specific value, whose exponential is known as perplexity. EAs are inherently asymmetric and row-wise stochastic, but they are used in DR approaches after undergoing heuristic symmetrization methods that violate both the row-wise constant entropy and stochasticity properties. In this work, we uncover a novel characterization of EA as an optimal transport problem, allowing a natural symmetrization that can be computed efficiently using dual ascent. The corresponding novel affinity matrix derives advantages from symmetric doubly stochastic normalization in terms of clustering performance, while also effectively controlling the entropy of each row thus making it particularly robust to varying noise levels. Following, we present a new DR algorithm, SNEkhorn, that leverages this new affinity matrix. We show its clear superiority to existing approaches with several indicators on both synthetic and real-world datasets.

1 Introduction

Exploring and analyzing high-dimensional data is a core problem of data science that requires building low-dimensional and interpretable representations of the data through dimensionality reduction (DR). Ideally, these representations should preserve the data structure by mimicking, in the reduced representation space (called latent space), a notion of similarity between samples. We call affinity the weight matrix of a graph that encodes this similarity. It has positive entries and the higher the weight in position \((i, j)\), the higher the similarity or proximity between samples \(i\) and \(j\). Seminal approaches relying on affinities include Laplacian eigenmaps \([2]\), spectral clustering \([50]\) and semi-supervised learning \([55]\). Numerous methods can be employed to construct such affinities. A common choice is to use a kernel (e.g., Gaussian) derived from a distance matrix normalized by a bandwidth parameter that usually has a large influence on the outcome of the algorithm. Indeed, excessively small kernel
bandwidth can result in solely capturing the positions of closest neighbors, at the expense of large-scale dependencies. Inversely, setting too large a bandwidth blurs information about close-range pairwise relations. Ideally, one should select a different bandwidth for each point to accommodate varying sampling densities and noise levels. One approach is to compute the bandwidth of a point based on the distance from its \(k\)-th nearest neighbor [53]. However, this method fails to consider the entire distribution of distances. In general, selecting appropriate kernel bandwidths can be a laborious task, and many practitioners resort to greedy search methods. This can be limiting in some settings, particularly when dealing with large sample sizes.

**Entropic Affinities and SNE/t-SNE.** Entropic affinities (EAs) were first introduced in the seminal paper *Stochastic Neighbor Embedding* (SNE) [16]. It consists in normalizing each row \(i\) of a distance matrix by a bandwidth parameter \(\varepsilon_i\) such that the distribution associated with each row of the corresponding stochastic (i.e., row-normalized) Gaussian affinity has a fixed entropy. The value of this entropy, whose exponential is called the *perplexity*, is then the only hyperparameter left to tune and has an intuitive interpretation as the number of effective neighbors of each point [59]. EAs are notoriously used to encode pairwise relations in a high-dimensional space for the DR algorithm t-SNE [46], among other DR methods including [6]. t-SNE is increasingly popular in many applied fields [20, 32] mostly due to its ability to represent clusters in the data [27, 5]. Nonetheless, one major flaw of EAs is that they are inherently directed and often require post-processing symmetrization.

**Doubly Stochastic Affinities.** Doubly stochastic (DS) affinities are non-negative matrices whose rows and columns have unit \(\ell_1\) norm. In many applications, it has been demonstrated that DS affinity normalization (i.e., determining the nearest DS matrix to a given affinity matrix) offers numerous benefits. First, it can be seen as a relaxation of k-means [51] and it is well-established that it enhances spectral clustering performances [10, 52, 1]. Additionally, DS matrices present the benefit of being invariant to the various Laplacian normalizations [50]. Recent observations indicate that the DS projection of the Gaussian kernel under the KL geometry is more resilient to heteroscedastic noise compared to its stochastic counterpart [23]. It also offers a more natural analog to the heat kernel [50]. These properties have led to a growing interest in DS affinities, with their use expanding to various applications such as smoothing filters [53], subspace clustering [25] and transformers [41].

**Contributions.** In this work, we study the missing link between EAs, which are easy to tune and adaptable to data with heterogeneous density, and DS affinities which have interesting properties in practical applications as aforementioned. Our main contributions are as follows. We uncover the convex optimization problem that underpins classical entropic affinities, exhibiting novel links with entropy-regularized Optimal Transport (OT) (Section 3.1). We then propose in Section 3.2 a principled symmetrization of entropic affinities. The latter enables controlling the entropy in each point, unlike t-SNE’s post-processing symmetrization, and produces a genuinely doubly stochastic affinity. We show how to compute this new affinity efficiently using a dual ascent algorithm. In Section 4, we introduce SNEkhorn: a DR algorithm that couples this new symmetric entropic affinity with a doubly stochastic kernel in the low-dimensional embedding space, without sphere normalization (i.e., invariant to the various Laplacian normalizations [50]). Recent observations indicate that the DS geometry is more resilient to heteroscedastic noise compared to its stochastic counterpart [23]. It also offers a more natural analog to the heat kernel [50]. These properties have led to a growing interest in DS affinities, with their use expanding to various applications such as smoothing filters [53], subspace clustering [25] and transformers [41].

**Notations.** \([n]\) denotes the set \(\{1, \ldots, n\}\). \(\exp\) and \(\log\) applied to vectors/matrices are taken element-wise. \(1 = (1, \ldots, 1)^\top\) is the vector of 1. \(\langle \cdot, \cdot \rangle\) is the standard inner product for matrices/vectors. \(S\) is the space of \(n \times n\) symmetric matrices. \(P_{ij}\) denotes the \(i\)-th row of a matrix \(P\). \(\odot\) (resp. \(\otimes\)) stands for element-wise multiplication (resp. division) between vectors/matrices. For \(\alpha, \beta \in \mathbb{R}^n\), \(\alpha \odot \beta \in \mathbb{R}^{n \times n}\) is \((\alpha_i + \beta_j)_{ij}\). The entropy of \(\mathbf{p} \in \mathbb{R}^n_+\) is \(H(p) = -\sum_i p_i (\log(p_i) - 1) = -\langle \mathbf{p}, \log \mathbf{p} - \mathbf{1} \rangle\). The Kullback-Leibler divergence between two matrices \(P, Q\) with nonnegative entries such that \(Q_{ij} = 0 \implies P_{ij} = 0\) is \(KL(P|Q) = \sum_{ij} P_{ij} \left( \log(\frac{P_{ij}}{Q_{ij}}) - 1 \right) = \langle P, \log (P \odot Q) - 11^\top \rangle\).

2 Entropic Affinities, Dimensionality Reduction and Optimal Transport

Given a dataset \(X \in \mathbb{R}^{n \times p}\) of \(n\) samples in dimension \(p\), most DR algorithms compute a representation of \(X\) in a lower-dimensional latent space \(Z \in \mathbb{R}^{n \times q}\) with \(q \ll p\) that faithfully captures and represents pairwise dependencies between the samples (or rows) in \(X\). This is generally achieved

\footnote{With the convention \(0 \log 0 = 0\).}
by optimizing $Z$ such that the corresponding affinity matrix matches another affinity matrix defined from $X$. These affinities are constructed from a matrix $C \in \mathbb{R}^{n \times n}$ that encodes a notion of “distance” between the samples, e.g., the squared Euclidean distance $C_{ij} = \|X_i - X_j\|^2_2$ or more generally any cost matrix $C \in D := \{C \in \mathbb{R}^{n \times n} : C = C^T \text{ and } C_{ij} = 0 \iff i = j\}$. A commonly used option is the Gaussian affinity that is obtained by performing row-wise normalization of the kernel $\exp(-C/\varepsilon)$, where $\varepsilon > 0$ is the bandwidth parameter.

**Entropic Affinities (EAs).** Another frequently used approach to generate affinities from $C \in D$ is to employ entropic affinities [16]. The main idea is to consider adaptive kernel bandwidths $(\varepsilon^*_i)_{i \in [n]}$ to capture finer structures in the data compared to constant bandwidths [47]. Indeed, EAs rescale distances to account for the varying density across regions of the dataset. Given $\xi \in [n - 1]$, the goal of EAs is to build a Gaussian chain transition matrix $P^\xi$ with prescribed entropy as

$$\forall i, \forall j, P^\xi_{ij} = \frac{\exp(-C_{ij}/\varepsilon^*_i)}{\sum_l \exp(-C_{il}/\varepsilon^*_i)} \quad \text{(EA)}$$

with $\varepsilon^*_i \in \mathbb{R}^+_*$ s.t. $H(P^\xi_i) = \log \xi + 1$.

The hyperparameter $\xi$, which is also known as perplexity, can be interpreted as the effective number of neighbors for each data point [49]. Indeed, a perplexity of $\xi$ means that each row of $P^\xi$ (which is a discrete probability since $P^\xi$ is row-wise stochastic) has the same entropy as a uniform distribution over $\xi$ neighbors. Therefore, it provides the practitioner with an interpretable parameter specifying which scale of dependencies the affinity matrix should faithfully capture. In practice, a root-finding algorithm is used to find the bandwidth parameters $(\varepsilon^*_i)_{i \in [n]}$ that satisfy the constraints [49]. Hereafter, with a slight abuse of language, we call $e^{H(P_i)}$ the perplexity of the point $i$.

**Dimension Reduction with SNE/t-SNE.** One of the main applications of EAs is the DR algorithm SNE [16]. We denote by $C_X = (\|X_i - X_j\|^2_{ij})_{ij}$ and $C_Z = (\|Z_i - Z_j\|^2_{ij})_{ij}$ the cost matrices derived from the rows (i.e., the samples) of $X$ and $Z$ respectively. SNE focuses on minimizing the following expectations in the latent coordinates $Z \in \mathbb{R}^{n \times q}$, where $P^\xi$ is the optimal transition kernel of a Markov chain with prescribed entropy.

$$\min_{Z\in\mathbb{R}^{n \times q}} KL(P^z,Q) \quad \text{where} \quad P^z = \frac{1}{2}(P^\xi + P^\xi^T).$$

In other words, the affinity matrix $P^z$ is the Euclidean projection of $P^\xi$ on the space of symmetric matrices $S$: $P^z = \text{Proj}_S(P^\xi) = \arg\min_{P \in S} \|P - P^\xi\|_2$ (see Appendix A.1). Instead of the Gaussian kernel, the popular extension t-SNE [46] considers a different distribution in the latent space $Q_Z = (1 + |C_{Zij}|)^{-1} / \sum_{t,i} (1 + |C_{Zit}|)^{-1}$. In this formulation, $Q_Z$ is a joint Student $t$-distribution that accounts for crowding effects: a relatively small distance in a high-dimensional space can be accurately represented by a significantly greater distance in the low-dimensional space.

Considering symmetric similarities is appealing since the proximity between two points is inherently symmetric. Nonetheless, the Euclidean projection in (Symmetric-SNE) does not preserve the construction of entropic affinities. In particular, $P^z$ is not stochastic in general and $H(P^z_i) \neq \log(\xi + 1)$ thus the entropy associated with each point is no longer controlled after symmetrization (see the bottom left plot of Figure 1). This is arguably one of the main drawbacks of the approach. By contrast, the $P^\xi$ affinity that will be introduced in Section 3 can accurately set the entropy in each point to the desired value $\log(\xi + 1)$. As shown in Figure 1, this leads to more faithful embeddings with better separation of the classes when combined with the t-SNEkhorn algorithm (Section 4).

**Figure 1:** Top: COIL [34] embeddings with silhouette scores produced by t-SNE and t-SNEkhorn (our method introduced in Section 3) for $\xi = 30$. Bottom: $e^{H(P_i)}$ (perplexity) for each point $i$. 

3
Symmetric Entropy-Constrained Optimal Transport. Entropy-regularized OT [37] and its connection to affinity matrices are crucial components in our solution. In the special case of uniform marginals, and for $\nu > 0$, entropic OT computes the minimum of $\nu \sum_i H(P_{c_i})$ over the space of doubly stochastic matrices $\{P \in \mathbb{R}_+^{n \times n} : P1 = P^T1 = 1\}$. The optimal solution is the unique doubly stochastic matrix $P_{ds}$ of the form $P_{ds} = \text{diag}(u)K\text{diag}(v)$ where $K = \exp(-C/\nu)$ is the Gibbs energy derived from $C$ and $u, v$ are positive vectors that can be found with the celebrated Sinkhorn-Knopp’s algorithm [8, 42]. Interestingly, when the cost $C$ is symmetric (e.g., $C \in \mathbb{D}$) we can take $u = v$ [17] Section 5.2 so that the unique optimal solution is itself symmetric and writes

$$P_{ds} = \exp ((f \oplus f - C)/\nu) \quad \text{where} \quad f \in \mathbb{R}^n.$$  

In this case, by relying on convex duality as detailed in Appendix A.2 an equivalent formulation for the symmetric entropic OT problem is

$$\min_{P \in \mathbb{R}_+^{n \times n}} \langle P, C \rangle \quad \text{s.t.} \quad P1 = 1, \ P = P^T \ \text{and} \ \sum_i H(P_{c_i}) \geq \eta,$$  

(1)

where $0 \leq \eta \leq n(\log n + 1)$ is a constraint on the global entropy $\sum_i H(P_{c_i})$ of the OT plan $P$ which happens to be saturated at optimum (Appendix A.2). This constrained formulation of symmetric entropic OT will provide new insights into entropic affinities, as detailed in the next sections.

3 Symmetric Entropic Affinities

In this section, we present our first major contribution: symmetric entropic affinities. We begin by providing a new perspective on EAs through the introduction of an equivalent convex problem.

3.1 Entropic Affinities as Entropic Optimal Transport

We introduce the following set of matrices with row-wise stochasticity and entropy constraints:

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

(1)

This space is convex since $p \in \mathbb{R}_+^{n \times n} \rightarrow H(p)$ is concave, thus its superlevel set is convex. In contrast to the entropic constraints utilized in standard entropic optimal transport which set a lower-bound on the global entropy, as demonstrated in the formulation (EOT), $\mathcal{H}_\xi$ imposes a constraint on the entropy of each row of the matrix $P$. Our first contribution is to prove that EAs can be computed by solving a specific problem involving $\mathcal{H}_\xi$ (see Appendix A for the proof).

Proposition 1. Let $C \in \mathbb{R}^{n \times n}$ without constant rows. Then $P^*$ solves the entropic affinity problem (EA) with cost $C$ if and only if $P^*$ is the unique solution of the convex problem

$$\min_{P \in \mathcal{H}_\xi} \langle P, C \rangle.$$  

(EA as OT)

Interestingly, this result shows that EAs boil down to minimizing a transport objective with cost $C$ and row-wise entropy constraints $\mathcal{H}_\xi$ where $\xi$ is the desired perplexity. As such, (EA as OT) can be seen as a specific semi-relaxed OT problem [39, 14] (i.e., without the second constraint on the marginal $P1 = 1$) but with entropic constraints on the rows of $P$. We also show that the optimal solution $P^*$ of (EA as OT) has saturated entropy i.e., $\forall i, \ H(P_{c_i}) = \log \xi + 1$. In other words, relaxing the equality constraint in (EA) as an inequality constraint in $P \in \mathcal{H}_\xi$ does not affect the solution while it allows reformulating entropic affinity as a convex optimization problem. To the best of our knowledge, this connection between OT and entropic affinities is novel and is an essential key to the method proposed in the next section.

Remark 2. The kernel bandwidth parameter $\epsilon$ from the original formulation of entropic affinities (EA) is the Lagrange dual variable associated with the entropy constraint in (EA as OT). Hence computing $\epsilon^*$ in (EA) exactly corresponds to solving the dual problem of (EA as OT).

Remark 3. Let $K_\sigma = \exp(-C/\sigma)$. As shown in Appendix A.5, if $\sigma^*$ solves (EA) and $\sigma \leq \min(\epsilon^*)$, then $P^* = \text{Proj}_{\mathcal{H}_\xi}(K_{\sigma^*}) = \arg \min_{P \in \mathcal{H}_\xi} \text{KL}(P \mid K_{\sigma^*})$. Therefore $P^*$ can be seen as a KL Bregman projection [3] of a Gaussian kernel onto $\mathcal{H}_\xi$. Hence the input matrix in (Symmetric-SNE) is $P^* = \text{Proj}_{\mathcal{H}_\xi}(\text{Proj}_{\mathcal{H}_\xi}(K_{\sigma}))$ which corresponds to a surprising mixture of KL and orthogonal projections.
where we recall that $S$ (for $P$ vector of row-wise entropies of $P$) version of EAs that enables keeping the entropy associated with each row (or equivalently column) to concave $L > \gamma P$. By defining the symmetric matrix $\gamma$ entropy and marginal constraints. Then, for at least $n - 1$ points have exactly a perplexity of $H\left( \lambda^\star \right) = \log \xi + 1$. Right plot: smallest eigenvalues of the Laplacian for the two affinities.

### 3.2 Symmetric Entropic Affinity Formulation

Based on the previous formulation we now propose symmetric entropic affinities: a symmetric version of EAs that enables keeping the entropy associated with each row (or equivalently column) to the desired value of $\log \xi + 1$ while producing a symmetric doubly stochastic matrix $P^\text{se}$ with zero diagonal, then $H\left( \lambda^\star \right) = \log \xi + 1$.

In other words, the unique solution $P^\text{se}$ has at least $n - 1$ saturated entropies i.e., the corresponding $n - 1$ points have exactly a perplexity of $\xi$. In practice, with the algorithmic solution detailed below, we have observed that all $n$ entropies are saturated. Therefore, we believe that this proposition can be extended with a few more assumptions on $C$. Accordingly, problem (SEA) allows accurate control over the point-wise entropies while providing a symmetric doubly stochastic matrix, unlike $P^\text{ds}$ defined in (Symmetric-SNE), as summarized in Table 1. In the sequel, we denote by $H_i\left( P \right) = \left( H\left( P_i \right) \right)$, the vector of row-wise entropies of $P$. We rely on the following result to compute $P^\text{se}$.

**Proposition 4** (Saturation of the entropies). Let $C \in S$ with zero diagonal, then $H\left( P \right)$ with cost $C$ has a unique solution that we denote by $P^\text{se}$. If moreover $C \in D$, then for at least $n - 1$ indices $i \in [n]$ the solution satisfies $H\left( P^\text{se}_i \right) = \log \xi + 1$.

![Figure 2: Samples from a mixture of three Gaussians with varying standard deviations. The edges’ strength is proportional to the weights in the affinities $P^\text{ds}$ (DS) and $P^\text{se}$ (SEA) computed with $\xi = 5$ (for $P^\text{ds}$, $\xi$ is the average perplexity such that $\sum_i H\left( P^\text{ds}_i \right) = \sum_i H\left( P^\text{se}_i \right)$). Points’ color represents the perplexity $e^{H\left( P \right)} - 1$.](image)

In other words, the unique solution $P^\text{se}$ has at least $n - 1$ saturated entropies i.e., the corresponding $n - 1$ points have exactly a perplexity of $\xi$. In practice, with the algorithmic solution detailed below, we have observed that all $n$ entropies are saturated. Therefore, we believe that this proposition can be extended with a few more assumptions on $C$. Accordingly, problem (SEA) allows accurate control over the point-wise entropies while providing a symmetric doubly stochastic matrix, unlike $P^\text{ds}$ defined in (Symmetric-SNE), as summarized in Table 1. In the sequel, we denote by $H_i\left( P \right) = \left( H\left( P_i \right) \right)$, the vector of row-wise entropies of $P$. We rely on the following result to compute $P^\text{se}$.

**Proposition 5** (Solving for $P^\text{se}$). Let $C \in D$, $L\left( P, \gamma, \lambda \right) = \left( P, C \right) + \left( \gamma, \left( \log \xi + 1 \right) \right)$, $q\left( \gamma, \lambda \right) = \min_{P \in \mathbb{R}^{n \times n}} L\left( P, \gamma, \lambda \right)$, Strong duality holds for (SEA). Moreover, let $\gamma^\star, \lambda^\star \in \arg\max_{\gamma \geq 0, \lambda} q\left( \gamma, \lambda \right)$ be the optimal dual variables respectively associated with the entropy and marginal constraints. Then, for at least $n - 1$ indices $i \in [n], \gamma_i^\star > 0$. When $\forall i \in [n], \gamma_i^\star > 0$ then $H_i\left( P^\text{se} \right) = \left( \log \xi + 1 \right)1$ and $P^\text{se}$ has the form

$$P^\text{se} = \exp \left( \left( \lambda^\star \oplus \lambda^\star - 2C \right) \otimes \left( \gamma^\star \oplus \gamma^\star \right) \right).$$

By defining the symmetric matrix $P\left( \gamma, \lambda \right) = \exp \left( \left( \lambda \oplus \lambda - 2C \right) \otimes \left( \gamma \oplus \gamma \right) \right)$, we prove that, when $\gamma > 0$, $\min_{P \in S} L\left( P, \gamma, \lambda \right)$ has a unique solution given by $P\left( \gamma, \lambda \right)$ which implies $q\left( \gamma, \lambda \right) = L\left( P\left( \gamma, \lambda \right), \gamma, \lambda \right)$. Thus the proposition shows that when $\gamma^\star > 0$, $P^\text{se} = P\left( \gamma^\star, \lambda^\star \right)$ where $\gamma^\star, \lambda^\star$ solve the following convex problem (as maximization of a concave objective)

$$\max_{\gamma > 0, \lambda} L\left( P\left( \gamma, \lambda \right), \gamma, \lambda \right).$$  \hspace{1cm} \text{(Dual-SEA)}
Consequently, to find \( P^{\text{se}} \) we solve the problem \( \text{(Dual-SEA)} \). Although the form of \( P^{\text{se}} \) presented in Proposition 5 is only valid when \( \gamma^* \) is positive and we have only proved it for \( n - 1 \) indices, we emphasize that if \( \text{(Dual-SEA)} \) has a finite solution, then it is equal to \( P^{\text{se}} \). Indeed in this case the solution satisfies the KKT system associated with \( \text{(SEA)} \).

**Numerical optimization.** The dual problem \( \text{(Dual-SEA)} \) is concave and can be solved with guarantees through a dual ascent approach with closed-form gradients (using e.g., SGD, BFGS [28] or ADAM [18]). At each gradient step, one can compute the current estimate \( P(\gamma, \lambda) \) while the gradients of the loss \( \gamma \) and \( \lambda \) are given respectively by the constraints \((\log \xi + 1)1 - H_i(P(\gamma, \lambda)) \) and \( 1 - P(\gamma, \lambda)1 \) (see e.g., [4] Proposition 6.1.1). Concerning time complexity, each step can be performed with \( O(n^2) \) algebraic operations. From a practical perspective, we found that using a change of variable \( \gamma \leftarrow \gamma^2 \) and optimize \( \gamma \in \mathbb{R}^n \) leads to enhanced numerical stability.

**Remark 6.** In the same spirit as Remark 3, one can express \( P^{\text{se}} \) as a KL projection of \( K_{se} = \exp(-C/\sigma) \). Indeed, we show in Appendix A.5 that if \( 0 < \sigma \leq \min_i \gamma_i^* \), then \( P^{\text{se}} = \text{Proj}_{H_i \leq \xi}(K_{\sigma}) \).

**Comparison between \( P^{\text{ds}} \) and \( P^{\text{se}} \).** In Figure 2 we illustrate the ability of our proposed affinity \( P^{\text{se}} \) adapt to varying noise levels. In the OT problem that we consider, each sample is given a mass of one that is distributed over its neighbors (including itself since self-loops are allowed). For each sample, we refer to the entropy of the distribution over its neighbors as the spreading of its mass. One can notice that for \( P^{\text{ds}} \) (OT problem with global entropy constraint \( \text{(EOT)} \)), the samples do not spread their mass evenly depending on the density around them. On the contrary, the per-row entropy constraints of \( P^{\text{se}} \) force equal spreading among samples. This can have benefits, particularly for clustering, as illustrated in the rightmost plot, which shows the eigenvalues of the associated Laplacian matrices (recall that the number of connected components equals the dimension of the null space of its Laplacian [7]). As can be seen, \( P^{\text{ds}} \) results in many unwanted clusters, unlike \( P^{\text{se}} \), which is robust to varying noise levels (its Laplacian matrix has only 3 vanishing eigenvalues). We further illustrate this phenomenon on Figure 3 with varying noise levels.

4 **Optimal Transport for Dimension Reduction with SNEkhorn**

In this section, we build upon symmetric entropic affinities to introduce SNEkhorn, a new DR algorithm that fully benefits from the advantages of doubly stochastic affinities.

**SNEkhorn’s objective.** Our proposed method relies on doubly stochastic affinity matrices to capture the dependencies among the samples in both input and latent spaces. The KL divergence, which is the central criterion in most popular DR methods [44], is used to measure the discrepancy between the two affinities. As detailed in sections 2 and 3, \( P^{\text{se}} \) computed using the cost \( \|C_X\|_{ij} = \|X_i - X_j\|_2^2 \), corrects for heterogeneity in the input data density by imposing point-wise entropy constraints. As we do not need such correction for embedding coordinates \( Z \) since they must be optimized, we opt for the standard affinity (DS) built as an OT transport plan with global entropy constraint (EOT). This OT plan can be efficiently computed using Sinkhorn’s algorithm. More precisely, we propose the optimization problem

\[
\min_{Z \in \mathbb{R}^{n \times q}} \text{KL}(P^{\text{se}}|Q_Z^{\text{ds}}) ,
\]

(SNEkhorn)

where \( Q_Z^{\text{ds}} = \exp(f_Z \oplus f_Z - C_Z) \) stands for the (DS) affinity computed with cost \( \|C_Z\|_{ij} = \|Z_i - Z_j\|_2^2 \) and \( f_Z \) is the optimal dual variable found by Sinkhorn’s algorithm. We set the bandwidth to \( \nu = 1 \) in \( Q_Z^{\text{ds}} \) similarly to [46] as the bandwidth in the low dimensional space only affects the scales.

![Figure 3: ARI spectral clustering on the example of three Gaussian clusters with variances: \( \sigma^2 \), \( 2\sigma^2 \) and \( 3\sigma^2 \) (as in Figure 2).](image-url)
of the embeddings and not their shape. Keeping only the terms that depend on \( Z \) and relying on the double stochasticity of \( \text{P}^{\text{se}} \), the objective in SNEkhorn can be expressed as \( \text{KL}(\text{P}^{\text{se}}, \text{C}_{Z}) - 2(\text{f}_{Z}, 1) \).

**Heavy-tailed kernel in latent space.** Since it is well known that heavy-tailed kernels can be beneficial in DR \([21]\), we propose an extension called t-SNEkhorn that simply amounts to computing a doubly stochastic student-t kernel in the low-dimensional space. With our construction, it corresponds to choosing the cost \( \text{C}_{Z_{ij}} = \log(1 + ||Z_{i} - Z_{j}||^{2}) \) instead of \( ||Z_{i} - Z_{j}||^{2} \).

**Inference.** This new DR objective involves computing a doubly stochastic normalization for each update of \( Z \). Interestingly, to compute the optimal dual variable \( f_{Z} \) in \( \text{Q}_{Z}^{d} \), we leverage a well-conditioned Sinkhorn fixed point iteration \([19][13]\), which converges extremely fast in the symmetric setting:

\[
\forall i, [f_{Z}]_{i} \leftarrow \frac{1}{2} \left( [f_{Z}]_{i} - \log \sum_{k} \exp \left( [f_{Z}]_{k} - [\text{C}_{Z}]_{ki} \right) \right).
\]

(Sinkhorn)

On the right side of Figure 4, we plot \( \|\text{Q}_{Z}^{d}1 - 1\|_{\infty} \) as a function of the Sinkhorn iterations for a toy example presented in Section 5. In most practical cases, we found that about 10 iterations were enough to reach a sufficiently small error. \( Z \) is updated through gradient descent with gradients obtained by performing backpropagation through the Sinkhorn iterations. These iterations can be further accelerated with a warm start strategy by plugging the last \( f_{Z} \) to initialize the current one.

**Related work.** Using doubly stochastic affinities for SNE has been proposed in \([29]\), with two key differences from our work. First, they do not consider EAs and resort to \( \text{P}^{\text{ds}} \) (DS). This affinity, unlike \( \text{P}^{\text{se}} \), is not adaptive to the data heterogeneous density (as illustrated in Figure 2). Second, they use the affinity \( \text{Q}_{Z} \) in the low-dimensional space and illustrate empirically that matching the latter with a doubly stochastic matrix (e.g., \( \text{P}^{\text{ds}} \) or \( \text{P}^{\text{se}} \)) can sometimes impose spherical constraints on the embedding \( Z \). This is detrimental for projections onto a 2D flat space (typical use case of DR) where embeddings tend to form circles. This can be verified on the left side of Figure 4. In contrast, in SNEkhorn, the latent affinity is also doubly stochastic so that latent coordinates \( Z \) are not subject to spherical constraints anymore. The corresponding SNEkhorn embedding is shown in Figure 5 (bottom right).

### 5 Numerical experiments

This section aims to illustrate the performances of the proposed affinity matrix \( \text{P}^{\text{se}} \) and DR method SNEkhorn at faithfully representing dependencies and clusters in low dimensions. First, we showcase the relevance of our approach on a simple synthetic dataset with heteroscedastic noise. Then,
we evaluate the spectral clustering performances of symmetric entropic affinities before benchmarking t-SNEkhorn with t-SNE and UMAP [31] on real-world images and genomics datasets.

**Simulated data.** We consider the toy dataset with heteroscedastic noise from [23]. It consists of sampling uniformly two vectors \( p_1 \) and \( p_2 \) in the \( 10^4 \)-dimensional probability simplex. \( n = 10^5 \) samples are then generated as \( x_i = \frac{x_i}{\sum_j x_{ij}} \) where

\[
\mathcal{M}(1000, p_1), \quad 1 \leq i \leq 500 \\
\mathcal{M}(1000, p_2), \quad 501 \leq i \leq 750 \\
\mathcal{M}(2000, p_2), \quad 751 \leq i \leq 1000 .
\]

where \( \mathcal{M} \) stands for the multinomial distribution. The goal of the task is to test the robustness to heteroscedastic noise. Indeed, points generated using \( p_2 \) exhibit different levels of noise due to various numbers of multinomial trials to form an estimation of \( p_2 \). This typically occurs in real-world scenarios when the same entity is measured using different experimental setups thus creating heterogeneous technical noise levels (e.g., in single-cell sequencing [20]). This phenomenon is known as batch effect [43]. In Figure 5, we show that, unlike \( \mathcal{P}^e \) (Symmetric-SNE), \( \mathcal{P}^{se} \) (SEA) manages to properly filter the noise (top row) to discriminate between samples generated by \( p_1 \) and \( p_2 \), and represent these two clusters separately in the embedding space (bottom row). In contrast, \( \mathcal{P}^{se} \) and SNE are misled by the batch effect. This shows that \( \mathcal{P}^{se} \) doesn’t fully benefit from the adaptivity of EAs due to poor normalization and symmetrization. This phenomenon partly explains the superiority of SNEkhorn and t-SNEkhorn over current approaches on real-world datasets as illustrated below.

**Real-world datasets.** We then experiment with various labeled classification datasets including images and genomic data. For images, we use COIL 20 [54], OLIVETTI faces [12], UMNIST [15] and CIFAR 10 [22]. For CIFAR, we experiment with features obtained from the last hidden layer of a pre-trained ResNet [39] while for the other three datasets, we take as input the raw pixel data. Regarding genomics data, we consider the Curated Microarray Database (CuMiDa) [11] made of microarray datasets for various types of cancer, as well as the pre-processed SNAREseq (chromatin accessibility) and scGEM (gene expression) datasets used in [9]. For CuMiDa, we retain the datasets with most samples. For all the datasets, when the dimension exceeds 50 we apply a pre-processing step of PCA in dimension 50, as usually done in practice [46]. In the following experiments, when not specified the hyperparameters are set to the value leading to the best average score on five different seeds with grid-search. For perplexity parameters, we test all multiples of 10 in the interval \([10, \min(n, 300)]\) where \( n \) is the number of samples in the dataset. We use the same grid for the \( k \) of the self-tuning affinity \( \mathcal{P}^{se} \) [55] and for the \( n_{\text{neighbors}} \) parameter of UMAP. For scalar bandwidths, we consider powers of 10 such that the corresponding affinities’ average perplexity belongs to the perplexity range.

**Spectral Clustering.** Building on the strong connections between spectral clustering mechanisms and t-SNE [44, 27] we first consider spectral clustering tasks to evaluate the affinity matrix \( \mathcal{P}^{se} \) (SEA) and compare it against \( \mathcal{P}^{e} \) (Symmetric-SNE). We also consider two versions of the Gaussian affinity with scalar bandwidth \( K = \exp(-C/n) \); the symmetrized row-stochastic \( \mathcal{P}^{rs} = \text{Proj}\_S^f(\mathcal{P}^{rs}) \) where \( \mathcal{P}^{rs} \) is \( K \) normalized by row and \( \mathcal{P}^{ds} \) (DS). We also consider the adaptive Self-Tuning \( \mathcal{P}^{st} \) affinity from [53] which relies on an adaptive bandwidth corresponding to the distance from the \( k \)-th nearest neighbor of each point. We use the spectral clustering implementation of scikit-learn [50] with default parameters which uses the unnormalized graph Laplacian. We measure the quality of clustering using the Adjusted Rand Index (ARI).

In Table 2 and Figure 6 one can notice that, in general, symmetric entropic affinities yield better results than usual entropic affinities with significant improvements in some datasets (e.g., throat microarray and SNAREseq). Overall

\begin{table}[h]
\centering
\caption{ARI (×100) clustering scores on genomics.}
\begin{tabular}{lccccc}
\hline
\textbf{DATA SET} & \textbf{\( \mathcal{P}^{e} \)} & \textbf{\( \mathcal{P}^{se} \)} & \textbf{\( \mathcal{P}^{st} \)} & \textbf{\( \mathcal{P}^{rs} \)} & \textbf{\( \mathcal{P}^{ds} \)} \\
\hline
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 & 14.4 \\
scGEM & 57.3 & 58.5 & 74.8 & 69.9 & 71.6 \\
SNAREseq & 8.89 & 9.95 & 46.3 & 55.4 & 96.6 \\
\hline
\end{tabular}
\end{table}
To evaluate the embeddings’ quality, we make use of the silhouette [40] and trustworthiness [48].

This can be attributed to the fact that symmetric entropic affinities combine the advantages of the relative variation of the loss becomes smaller than $10^{-5}$. For each run, we draw independent $\mathcal{N}(0, 1)$ coordinates and use this same matrix to initialize all the methods that we wish to compare. To evaluate the embeddings’ quality, we make use of the silhouette [40] and trustworthiness [48].

P^∞ outperforms all the other affinities in 8 out of 12 datasets. This shows that the adaptivity of EAs is crucial. Figure 6 also shows that this superiority is verified for the whole range of perplexities. This can be attributed to the fact that symmetric entropic affinities combine the advantages of doubly stochastic normalization in terms of clustering and of EAs in terms of adaptivity. In the next experiment, we show that these advantages translate into better clustering and neighborhood retrieval at the embedding level when running SNEkhorn.

### Dimension Reduction

To guarantee a fair comparison, we implemented not only SNEkhorn, but also t-SNE and UMAP in PyTorch [35]. Note that UMAP also relies on adaptive affinities but sets the degree of each node (related to the hyperparameter n_neighbors which plays a similar role to the perplexity) rather than the entropy. All models were optimized using Adam [19] with default parameters and the same stopping criterion: the algorithm stops whenever the relative variation of the loss becomes smaller than $10^{-5}$. For each run, we draw independent $\mathcal{N}(0, 1)$ coordinates and use this same matrix to initialize all the methods that we wish to compare.

Figure 6: ARI spectral clustering score as a function of the perplexity parameter for image datasets.

Table 3: Scores for the UMAP, t-SNE and t-SNEkhorn embeddings.

<table>
<thead>
<tr>
<th></th>
<th>Silhouette ($\times 100$)</th>
<th>Trustworthiness ($\times 100$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UMAP</td>
<td>t-SNE</td>
</tr>
<tr>
<td>COIL</td>
<td>20.4 ± 3.3</td>
<td>30.7 ± 6.9</td>
</tr>
<tr>
<td>OLIVETTI</td>
<td>6.4 ± 4.2</td>
<td>4.5 ± 3.1</td>
</tr>
<tr>
<td>UMNIST</td>
<td>−1.4 ± 2.7</td>
<td>−0.2 ± 1.5</td>
</tr>
<tr>
<td>CIFAR</td>
<td>13.6 ± 2.4</td>
<td>18.3 ± 0.8</td>
</tr>
<tr>
<td>Liver (14820)</td>
<td>49.7 ± 1.3</td>
<td>50.9 ± 0.7</td>
</tr>
<tr>
<td>Breast (70947)</td>
<td>28.6 ± 0.8</td>
<td>29.0 ± 0.2</td>
</tr>
<tr>
<td>Leukemia (28497)</td>
<td>22.3 ± 0.7</td>
<td>20.6 ± 0.7</td>
</tr>
<tr>
<td>Colorectal (44076)</td>
<td>67.6 ± 2.2</td>
<td>69.5 ± 0.5</td>
</tr>
<tr>
<td>Liver (76427)</td>
<td>39.4 ± 4.3</td>
<td>38.3 ± 0.9</td>
</tr>
<tr>
<td>Breast (45827)</td>
<td>35.4 ± 3.3</td>
<td>39.5 ± 1.9</td>
</tr>
<tr>
<td>Colorectal (21510)</td>
<td>38.0 ± 1.3</td>
<td>42.3 ± 0.6</td>
</tr>
<tr>
<td>Renal (53757)</td>
<td>44.4 ± 1.5</td>
<td>45.9 ± 0.3</td>
</tr>
<tr>
<td>Prostate (6919)</td>
<td>5.4 ± 2.7</td>
<td>8.1 ± 0.2</td>
</tr>
<tr>
<td>Throat (42743)</td>
<td>26.7 ± 2.4</td>
<td>28.0 ± 0.3</td>
</tr>
</tbody>
</table>

scGEM         | 26.9 ± 3.7  | 33.0 ± 1.1  | 39.3 ± 0.7 | 95.0 ± 1.3  | 96.2 ± 0.6  | 96.8 ± 0.3 |
| SNAREseq      | 6.8 ± 6.0   | 35.8 ± 5.2  | 67.9 ± 1.2 | 93.1 ± 2.8  | 99.1 ± 0.1  | 99.2 ± 0.1 |

Figure 7: SNAREseq embeddings produced by t-SNE and t-SNEkhorn with $\xi = 50$.
scores from scikit-learn with default parameters. While the former relies on class labels, the latter measures the agreement between the neighborhoods in input and output spaces, thus giving two complementary metrics to properly evaluate the embeddings. The results, presented in Table 3, demonstrate the notable superiority of t-SNEkhorn compared to the commonly used t-SNE and UMAP algorithms. A sensitivity analysis on perplexity can also be found in Appendix B. Across the 16 datasets examined, t-SNEkhorn almost consistently outperformed the others, achieving the highest silhouette score on 15 datasets and the highest trustworthiness score on 12 datasets. To visually assess the quality of the embeddings, we provide SNAREseq embeddings in Figure 7. Notably, one can notice that the use of t-SNEkhorn results in improved class separation compared to t-SNE.

6 Conclusion

We have introduced a new principled and efficient method for constructing symmetric entropic affinities. Unlike the current formulation that enforces symmetry through an orthogonal projection, our approach allows control over the entropy in each point thus achieving entropic affinities’ primary goal. Additionally, it produces a DS-normalized affinity and thus benefits from the well-known advantages of this normalization. Our affinity takes as input the same perplexity parameter as EAs and can thus be used with little hassle for practitioners. We demonstrate experimentally that both our affinity and DR algorithm (SNEkhorn), leveraging a doubly stochastic kernel in the latent space, achieve substantial improvements over existing approaches.

Note that in the present work, we do not address the issue of large-scale dependencies that are not faithfully represented in the low-dimensional space [44]. The latter shall be treated in future works. Among other promising research directions, one could focus on building multi-scale versions of symmetric entropic affinities [24] as well as fast approximations for SNEkhorn forces by adapting e.g., Barnes-Hut [45] or interpolation-based methods [26] to the doubly stochastic setting. It could also be interesting to use SEAs in order to study the training dynamics of transformers [54].

Acknowledgments

The authors are grateful to Mathurin Massias, Jean Feydy and Aurélien Garivier for insightful discussions. This project was supported in part by the ANR projects AllegroAssai ANR-19-CHIA-0009, SingleStatOmics ANR-18-CE45-0023 and OTTOPIA ANR-20-CHIA-0030. This work was also supported by the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005.

References


A Proofs

A.1 Euclidean Projection onto $S$

For the problem $\arg\min_{P \in S} \|P - K\|_2^2$, the Lagrangian takes the form, with $W \in \mathbb{R}^{n \times n}$,

$$L(P, W) = \|P - K\|_2^2 + \langle W, P - P^T \rangle. \quad (3)$$

Cancelling the gradient of $L$ with respect to $P$ gives $2(P* - K) + W - W^T = 0$. Thus $P* = K + \frac{1}{2}(W^T - W)$. Using the symmetry constraint on $P*$ yields $P* = \frac{1}{2}(K + K^T)$. Hence we have:

$$\arg\min_{P \in S} \|P - K\|_2^2 = \frac{1}{2}(K + K^T). \quad (4)$$

A.2 From Symmetric Entropy-Constrained OT to Sinkhorn Iterations

In this section, we derive Sinkhorn iterations from the problem $\text{EOT}$. Let $C \in D$. We start by making the constraints explicit.

$$\min_{P \in \mathbb{R}^{n \times n}} \langle P, C \rangle \quad (5)$$

$$\text{s.t. } \sum_{i \in [n]} H(P_{i, :}) \geq \eta \quad (6)$$

$$P1 = 1, \quad P = P^T. \quad (7)$$

For the above convex problem the Lagrangian writes, where $\nu \in \mathbb{R}_+, f \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n \times n}$:

$$L(P, f, \nu, \Gamma) = \langle P, C \rangle + \langle \nu, \eta - \sum_{i \in [n]} H(P_{i, :}) \rangle + 2\langle f, 1 - P \rangle + \langle \Gamma, P - P^T \rangle. \quad (8)$$

Strong duality holds and the first order KKT condition gives for the optimal primal $P*$ and dual $(\nu^*, f^*, \Gamma^*)$ variables:

$$\nabla_P L(P^*, f^*, \nu^*, \Gamma^*) = C + \nu^* \log P^* - 2f^*1^T + \Gamma^* - \Gamma^*^T = 0. \quad (9)$$

Since $P^*, C \in S$ we have $\Gamma^* - \Gamma^*^T = f^*1^T - 1f^*^T$. Hence $C + \nu^* \log P^* - f^* \oplus f^* = 0$. Suppose that $\nu^* = 0$ then the previous reasoning implies that $\forall(i, j), C_{ij} = f^*_i + f^*_j$. Using that $C \in D$ we have $C_{ii} = C_{jj} = 0$ thus $\forall i, f^*_i = 0$ and thus this would imply that $C = 0$ which is not allowed by hypothesis. Therefore $\nu^* \neq 0$ and the entropy constraint is saturated at the optimum by complementary slackness. Isolating $P^*$ then yields:

$$P^* = \exp((f^* \oplus f^* - C)/\nu^*). \quad (10)$$

$P^*$ must be primal feasible in particular $P^*1 = 1$. This constraint gives us the Sinkhorn fixed point relation for $f^*$:

$$\forall i \in [n], \quad [f^*]_i = -\nu^* \text{LSE}((f^* - C)_{i, :})/\nu^*, \quad (11)$$

where for a vector $\alpha$, we use the notation $\text{LSE}(\alpha) = \log \sum_k \exp(\alpha_k)$.

A.3 Proof of Proposition 1

We recall the result

**Proposition 1.** Let $C \in \mathbb{R}^{n \times n}$ without constant rows. Then $P^*$ solves the entropic affinity problem (EA) with cost $C$ if and only if $P^*$ is the unique solution of the convex problem

$$\min_{P \in \mathbb{R}_+^n} \langle P, C \rangle. \quad \text{EA as OT} \quad (12)$$

**Proof.** We begin by rewriting the above problem to make the constraints more explicit.

$$\min_{P \in \mathbb{R}_+^{n \times n}} \langle P, C \rangle \quad (13)$$

$$\text{s.t. } \forall i, \ H(P_{i, :}) \geq \log \xi + 1 \quad P1 = 1. \quad (14)$$

$$\forall i, H(P_{i, :}) \geq \log \xi + 1 \quad P1 = 1. \quad (15)$$

We now use the notation $LSE(\alpha) = \log \sum_k \exp(\alpha_k)$.

$$\min_{P \in \mathbb{R}_+^{n \times n}} \langle P, C \rangle \quad (16)$$

$$\text{s.t. } \forall i, \ H(P_{i, :}) \geq \log \xi + 1 \quad P1 = 1. \quad (17)$$

We now use the notation $LSE(\alpha) = \log \sum_k \exp(\alpha_k)$.
By concavity of entropy, one has that the entropy constraint is convex thus the above primal problem is a convex optimization problem. Moreover, the latter is strictly feasible for any \( \xi \in [n-1] \). Therefore Slater’s condition is satisfied and strong duality holds.

Introducing the dual variables \( \lambda \in \mathbb{R}^n \) and \( \varepsilon \in \mathbb{R}_+^n \), the Lagrangian of the above problem writes:

\[
\mathcal{L}(P, \lambda, \varepsilon) = \langle P, C \rangle + \langle \varepsilon, (\log \xi + 1)1 - H_i(P) \rangle + \langle \lambda, 1 - P1 \rangle,
\]

where we recall that \( H_i(P) = (H(P_{ii})) \). Note that we will deal with the constraint \( P \in \mathbb{R}^{n\times n} \) directly, hence there is no associated dual variable. Since strong duality holds, for any solution \( P^* \) to the primal problem and any solution \( (\varepsilon^*, \lambda^*) \) to the dual problem, the pair \( (P^*, (\varepsilon^*, \lambda^*)) \) must satisfy the Karush-Kuhn-Tucker (KKT) conditions. The first-order optimality condition gives:

\[
\nabla_P \mathcal{L}(P^*, \varepsilon^*, \lambda^*) = C + \text{diag}(\varepsilon^*) \log P^* - \lambda^* 1^\top = 0.
\]

(\text{first-order})

Assume that there exists \( \ell \in [n] \) such that \( \varepsilon^*_\ell = 0 \). Then \( (\text{first-order}) \) gives that the \( \ell \)th row of \( C \) is constant which is not allowed by hypothesis. Therefore \( \varepsilon^* > 0 \) (i.e., \( \varepsilon^* \) has positive entries). Thus isolating \( P^* \) in the first order condition results in:

\[
P^* = \text{diag}(u) \exp\left(-\text{diag}(\varepsilon^*)^{-1}C\right)
\]

(13)

where \( u = \exp(\lambda^* \otimes \varepsilon^*) \). This matrix must satisfy the stochasticity constraint \( P1 = 1 \). Hence one has \( u = 1 \otimes (\exp(\text{diag}(\varepsilon^*)^{-1}C)1) \) and \( P^* \) has the form

\[
\forall(i, j) \in [n]^2, \quad P^*_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon^*_i\right)}{\sum_{j} \exp\left(-C_{ij}/\varepsilon^*_i\right)}.
\]

(14)

As a consequence of \( \varepsilon^* > 0 \), complementary slackness in the KKT conditions gives us that for all \( i \), the entropy constraint is saturated i.e., \( H_i(P^*) = \log \xi + 1 \). Therefore \( P^* \) solves the problem \( (SEA) \). Conversely any solution of \( (SEA) \) \( P^*_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon^*_i\right)}{\sum_{j} \exp\left(-C_{ij}/\varepsilon^*_i\right)} \) with \( (\varepsilon^*_i) \) such that \( H_i(P^*) = \log \xi + 1 \) gives an admissible matrix for \( \min_{P \in H_\xi} \langle P, C \rangle \) and the associated variables satisfy the KKT conditions which are sufficient conditions for optimality since the problem is convex.

\[\square\]

### A.4 Proof of Proposition 4 and Proposition 5

The goal of this section is to prove the following results:

**Proposition 4** (Saturation of the entropies). Let \( C \in S \) with zero diagonal, then \( (SEA) \) with cost \( C \) has a unique solution that we denote by \( P^\infty \). If moreover \( C \in D \), then for at least \( n-1 \) indices \( i \in [n] \) the solution satisfies \( H_i(P^\infty_{ii}) = \log \xi + 1 \).

**Proposition 5** (Solving for \( (SEA) \)). Let \( C \in D \), \( \mathcal{L}(P, \gamma, \lambda) = \langle P, C \rangle + \langle \gamma, (\log \xi + 1)1 - H_i(P) \rangle + \langle \lambda, 1 - P1 \rangle \) and \( q(\gamma, \lambda) = \min_{P \in \mathbb{R}^{n\times n}} \mathcal{L}(P, \gamma, \lambda) \). Strong duality holds for \( (SEA) \). Moreover, let \( \gamma^*, \lambda^* \in \arg\min_{\gamma \geq 0, \lambda} q(\gamma, \lambda) \) be the optimal dual variables respectively associated with the entropy and marginal constraints. Then, for at least \( n-1 \) indices \( i \in [n] \), \( \gamma^*_i > 0 \). When \( \forall i \in [n] \), \( \gamma^*_i > 0 \) then \( H_i(P^\infty) = (\log \xi + 1)1 \) and \( P^\infty \) has the form

\[
P^\infty = \exp\left( (\lambda^* \otimes \lambda^* - 2C) \otimes (\gamma^* \otimes \gamma^*) \right).
\]

(2)

The unicity of the solution in Proposition 4 is a consequence of the following lemma

**Lemma 7.** Let \( C \neq 0 \in S \) with zero diagonal. Then the problem \( \min_{P \in H_\xi \cap S} \langle P, C \rangle \) has a unique solution.

**Proof.** Making the constraints explicit, the primal problem of symmetric entropic affinity takes the following form

\[
\min_{P \in \mathbb{R}^{n\times n}} \langle P, C \rangle \\
\text{s.t.} \quad \forall i, \ H(P_{ii}) \geq \log \xi + 1 \quad \text{(SEA)} \\
P1 = 1, \quad P = P^\top.
\]

Suppose that the solution is not unique i.e., there exists a couple of optimal solutions \( (P_1, P_2) \) that satisfy the constraints of \( (SEA) \) and such that \( \langle P_1, C \rangle = \langle P_2, C \rangle \). For \( i \in [n] \), we denote the
function \( f_i : \mathbb{P} \to (\log \xi + 1) - H(P_i) \). Then \( f_i \) is continuous, strictly convex and the entropy conditions of \( \text{(SEA)} \) can be written as \( \forall i \in [n], f_i(P) \leq 0 \).

Now consider \( Q = \frac{1}{2}(P_1 + P_2) \). Then clearly \( Q1 = 1, Q = Q^\top \). Since \( f_i \) is strictly convex we have \( f_i(Q) = f_i(\frac{1}{2}P_1 + \frac{1}{2}P_2) < \frac{1}{2}f_i(P_1) + \frac{1}{2}f(P_2) \leq 0 \). Thus \( f_i(Q) < 0 \) for any \( i \in [n] \).

Take any \( \varepsilon > 0 \) and \( i \in [n] \). By continuity of \( f_i \) there exists \( \delta_i > 0 \) such that, for any \( H \) with \( \|H\|_F \leq \delta_i \), we have \( f_i(Q + H) < f_i(Q) + \varepsilon \). Take \( \varepsilon > 0 \) such that \( \forall i \in [n], 0 < \varepsilon < -\frac{1}{2}f_i(Q) \) (this is possible since for any \( i \in [n] \), \( f_i(Q) < 0 \) and \( H \) with \( \|H\|_F \leq \min_{i \in [n]} \delta_i \). Then for any \( i \in [n] \), \( f_i(Q + H) < 0 \). In other words, we have proven that there exists \( \eta > 0 \) such that for any \( H \) such that \( \|H\|_F \leq \eta \), it holds: \( \forall i \in [n], f_i(Q + H) < 0 \).

Now let us take \( H \) as the Laplacian matrix associated to \( C \) i.e., for any \( (i, j) \in [n]^2 \), \( H_{ij} = -C_{ij} \) if \( i \neq j \) and \( \sum_i C_{ij} \) otherwise. Then we have \( \langle H, C \rangle = -\sum_{i \neq j} C_{ij}^2 + 0 = -\sum_{i \neq j} C_{ij}^2 < 0 \) since \( C \) has zero diagonal (and is nonzero). Moreover, \( H = H^\top \) since \( C \) is symmetric and \( H1 = 0 \) by construction. Consider for \( 0 < \beta < \frac{1}{\|H\|_F} \) the matrix \( H_\beta := \beta H \). Then \( \|H_\beta\|_F = \beta \|H\|_F \leq \eta \).

By the previous reasoning one has: \( \forall i \in [n], f_i(Q + H_{\beta}) < 0 \). Moreover, \( (Q + H_{\beta})^\top = Q + H_{\beta} \) and \( (Q + H_{\beta})1 = 1 \). For \( \beta \) small enough we have \( Q + H_\beta \in \mathbb{R}_+^{n \times n} \) and thus there is a \( \beta \) (that depends on \( P_1 \) and \( P_2 \)) such that \( Q + H_\beta \) is admissible i.e., satisfies the constraints of \( \text{(SEA)} \). Then, for such \( \beta \),

\[
\langle C, Q + H_\beta \rangle - \langle C, P_1 \rangle = \frac{1}{2} \langle C, P_1 + P_2 \rangle + \langle C, H_\beta \rangle = \langle C, H_\beta \rangle = \beta \langle H, C \rangle < 0 .
\]

Thus \( \langle C, Q + H_\beta \rangle < \langle C, P_1 \rangle \) which leads to a contradiction. \( \square \)

We can now prove the rest of the claims of Proposition 4 and Proposition 5.

**Proof.** Let \( C \in \mathcal{D} \). We first prove Proposition 4. The unicity is a consequence of Lemma 7. For the saturation of the entropies we consider the Lagrangian of the problem \( \text{(SEA)} \) that writes

\[
\mathcal{L}(P, \lambda, \gamma, \Gamma) = \langle C, P \rangle + \langle \gamma, (\log \xi + 1)1 - H_r(P) \rangle + \langle \lambda, 1 - P1 \rangle + \langle \Gamma, P - P^\top \rangle
\]

for dual variables \( \gamma \in \mathbb{R}_+^n, \lambda \in \mathbb{R}^n \) and \( \Gamma \in \mathbb{R}^{n \times n} \). Strong duality holds by Slater’s conditions because \( \frac{1}{\xi}1^1 \) is strictly feasible for \( \xi < n - 1 \). Since strong duality holds, for any solution \( P^* \) to the primal problem and any solution \((\gamma^*, \lambda^*, \Gamma^*)\) to the dual problem, the pair \( P^*, (\gamma^*, \lambda^*, \Gamma^*) \) must satisfy the KKT conditions. They can be stated as follows:

\[
P^*1 = 1, H_\beta(P^*) \geq (\log \xi + 1)1, P^* = P^*^\top
\]

\[
\exists \iota, \gamma^*_\iota(H(P^*_\iota) - (\log \xi + 1)) = 0 .
\]

Let us denote \( I = \{ \ell \in [n] \text{ s.t. } \gamma^*_\ell = 0 \} \). For \( \ell \not\in I \), using the first-order condition, one has for \( i \in [n], C_{i\ell} = \lambda^*_i - \Gamma^*_{i\ell} = 0 \). Since \( C \in \mathcal{D} \), we have \( C_{\ell\ell} = 0 \). Thus \( \lambda^*_i = 0 \) and \( C_{\ell\ell} = \Gamma^*_{\ell\ell} - \Gamma^*_{i\ell} \). For \( \ell, \ell' \in I \), one has \( C_{i\ell} = \Gamma^*_{i\ell} - \Gamma^*_{i\ell'} = -\Gamma^*_{\ell\ell'} - \Gamma^*_{\ell'\ell} = -C_{\ell\ell} \). \( C \) is symmetric thus \( C_{\ell\ell} = 0 \). Since \( C \) only has null entries on the diagonal, this shows that \( \ell = \ell' \) and therefore \( I \) has at most one element. By complementary slackness condition (last row of the \( \text{[KKT-SEA]} \) conditions) it holds that \( \forall i \not\in I, H(P^*_i) = \log \xi + 1 \) since the solution of \( \text{(SEA)} \) is unique \( P^* = P^{\text{se}} \) and thus \( \forall i \not\in I, H(P^*_i) = \log \xi + 1 \) which proves Proposition 4 but also that for at least \( n - 1 \) indices \( \gamma^*_i > 0 \). Moreover, from the KKT conditions we have

\[
\forall (i, j) \in [n]^2, \Gamma_{ij} = C_{ij} + \gamma^*_i \log P^*_{ij} - \lambda^*_j .
\]

Now take \( (i, j) \in [n]^2 \) fixed. From the previous equality \( \Gamma_{ij} - \Gamma_{ji} = C_{ij} + \gamma^*_i \log P^*_{ij} - \lambda^*_j \) but also \( \Gamma_{ij} - \Gamma_{ji} = C_{ji} + \gamma^*_j \log P^*_{ij} - \lambda^*_i \). Using that \( P^* = (P^*)^\top \) and \( C \in \mathcal{S} \) we get \( \Gamma_{ij} = \Gamma_{ji} = C_{ij} + \gamma^*_j \log P^*_{ij} - \lambda^*_i \). But \( \Gamma_{ij} - \Gamma_{ji} = -\Gamma_{ji} - \Gamma_{ij} \) which gives

\[
C_{ij} + \gamma^*_j \log P^*_{ij} - \lambda^*_j = -(C_{ij} + \gamma^*_i \log P^*_{ij} - \lambda^*_i) .
\]
This implies
\[
\forall (i, j) \in [n]^2, \quad 2C_{ij} + (\gamma^*_i + \gamma^*_j) \log P^*_{ij} - (\lambda^*_i + \lambda^*_j) = 0.
\] (18)

Consequently, if \( \gamma^* > 0 \) we have the desired form from the above equation and by complementary slackness \( H_i(P^{\pi}) = (\log \xi + 1)1 \) which proves Proposition 5. Note that otherwise, it holds
\[
\forall (i, j) \neq (\ell, \ell), \quad P^*_{ij} = \exp \left( \frac{\lambda^*_i + \lambda^*_j - 2C_{ij}}{\gamma^*_i + \gamma^*_j} \right).
\] (19)

\[\square\]

### A.5 EA and SEA as a KL projection

We prove the characterization as a projection of \([\text{EA}]\) in Lemma 8 and of \([\text{SEA}]\) in Lemma 9.

**Lemma 8.** Let \( C \in \mathcal{D}, \sigma > 0 \) and \( K_\sigma = \exp(-C/\sigma) \). Then for any \( \sigma \leq \min_i \varepsilon_i^* \), it holds \( P^\pi = \text{Proj}_{H_\xi}(K_\sigma) = \arg \min_{P \in H_\xi} \text{KL}(P|K_\sigma) \).

**Proof.** The KL projection of \( K \) onto \( H_\xi \) reads
\[
\min_{P \in \mathbb{R}^{n \times n}} \text{KL}(P|K) \quad \text{s.t.} \quad \forall i, H(P_{ii}) \geq \log \xi + 1 \quad \text{(20)}
\]
\[
P1 = 1.
\] (21)

Introducing the dual variables \( \lambda \in \mathbb{R}^n \) and \( \kappa \in \mathbb{R}^n_+ \), the Lagrangian of this problem reads:
\[
\mathcal{L}(P, \lambda, \kappa) = \text{KL}(P|K) + \langle \kappa, (\log \xi + 1)1 - H(P) \rangle + \langle \lambda, 1 - P1 \rangle
\] (23)

Strong duality holds hence for any solution \( P^* \) to the above primal problem and any solution \( (\kappa^*, \lambda^*) \) to the dual problem, the pair \( P^*, (\kappa^*, \lambda^*) \) must satisfy the KKT conditions. The first-order optimality condition gives:
\[
\nabla_P \mathcal{L}(P^*, \kappa^*, \lambda^*) = \text{diag}(\kappa^*) \log P^* - \lambda^* 1^T = 0.
\] (24)

Solving for \( \lambda^* \) given the stochasticity constraint and isolating \( P^* \) gives
\[
\forall (i, j) \in [n]^2, \quad P^*_{ij} = \exp \left( \frac{\log K_{ij}}{1 + \kappa^*_i} \right) \sum_j \exp \left( \frac{\log K_{ij}}{1 + \kappa^*_j} \right).
\] (25)

We now consider \( P^* \) as a function of \( \kappa \). Plugging this expression back in \( \mathcal{L} \) yields the dual function \( \kappa \mapsto \mathcal{G}(\kappa) \). The latter is concave as any dual function and its gradient reads:
\[
\nabla_\kappa \mathcal{G}(\kappa) = (\log \xi + 1)1 - H(P^*(\kappa^*)).
\] (26)

Denoting by \( \rho = 1 + \kappa \) and taking the dual feasibility constraint \( \kappa \geq 0 \) into account gives the solution: for any \( i, \rho_i^* = \max(\varepsilon_i^*, 1) \) where \( \varepsilon^* \) solves \([\text{EA}]\) with cost \( C = -\log K \). Moreover we have that \( \sigma \leq \min_i \varepsilon_i^* \) where \( \varepsilon^* \in (\mathbb{R}^n_+)^n \) solves \([\text{EA}]\). Therefore for any \( i \in [n] \), one has \( \varepsilon_i^*/\sigma \geq 1 \). Thus there exists \( \kappa_i^* \in \mathbb{R}_+ \) such that \( \sigma(1 + \kappa_i^*) = \varepsilon_i^* \).

This \( \kappa^* \) cancels the above gradient i.e., \((\log \xi + 1)1 = H(P^*(\kappa^*))\) thus solves the dual problem. Therefore given the expression of \( P^* \) we have that \( \text{Proj}_{H_\xi}(K_\sigma) = P^\pi \).

\[\square\]

**Lemma 9.** Let \( C \in \mathcal{D}, \sigma > 0 \) and \( K_\sigma = \exp(-C/\sigma) \). Suppose that the optimal dual variable \( \gamma^* \) associated with the entropy constraint of \([\text{SEA}]\) is positive. Then for any \( \sigma \leq \min_i \gamma_i^* \), it holds \( P^{\pi} = \text{Proj}_{H_\xi \cap \mathcal{S}}(K_\sigma) \).

**Proof.** Let \( \sigma > 0 \). The KL projection of \( K \) onto \( H_\xi \cap \mathcal{S} \) boils down to the following optimization problem.
\[
\min_{P \in \mathbb{R}^{n \times n}} \text{KL}(P|K_\sigma) \quad \text{s.t.} \quad \forall i, H(P_{ii}) \geq \log \xi + 1 \quad \text{(SEA-Proj)}
\]
\[
P1 = 1, \quad P^T = P.
\]
By strong convexity of $P \rightarrow KL(P|K_\sigma)$ and convexity of the constraints the problem (SEA-Proj) admits a unique solution. Moreover, the Lagrangian of this problem takes the following form, where $\omega \in \mathbb{R}_+^n, \mu \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n \times n}$:

$$
\mathcal{L}(P, \mu, \omega, \Gamma) = KL(P|K_\sigma) + \langle \omega, (\log \xi + 1)1 - H_r(P) \rangle + \langle \mu, 1 - P1 \rangle + \langle \beta, P - P^T \rangle.
$$

Strong duality holds by Slater’s conditions thus the KKT conditions are necessary and sufficient. In particular if $P^*$ and $(\omega^*, \mu^*, \beta^*)$ satisfy

$$
\nabla_P \mathcal{L}(P^*, \mu^*, \omega^*, \Gamma^*) = \log (P^* \odot K) + \text{diag}(\omega^*) \log P^* - \mu^*1^T + \beta^* - \beta^*^T = 0
$$

$$
P^*1 = 1, H_r(P^*) \geq (\log \xi + 1)1, P^* = P^{*^T}
$$

$$
\omega^* \geq 0
$$

$$
\forall i, \omega^*_i (H(P^*_i) - (\log \xi + 1)) = 0.
$$

then $P^*$ is a solution to (SEA-Proj) and $(\omega^*, \mu^*, \beta^*)$ are optimal dual variables. The first condition rewrites

$$
\forall (i, j), \log(P^*_{ij}) + \frac{1}{\sigma}C_{ij} + \omega^*_i \log(P^*_{ij}) - \mu^*_i + \beta^*_i - \beta^*_j = 0,
$$

which is equivalent to

$$
\forall (i, j), \sigma(1 + \omega^*_i) \log(P^*_{ij}) + C_{ij} - \sigma \mu^*_i + \sigma(\beta^*_i - \beta^*_j) = 0.
$$

Now take $P^{se}$ the optimal solution of (SEA). As written in the proof Proposition 5 of $P^{se}$ and the optimal dual variables $(\gamma^*, \lambda^*, \Gamma^*)$ satisfy the KKT conditions:

$$
\forall (i, j), C_{ij} + \gamma^*_i \log(P^{se}_{ij}) - \lambda^*_i + \Gamma^*_i - \Gamma^*_j = 0
$$

$$
P^{se}1 = 1, H_r(P^{se}) \geq (\log \xi + 1)1, P^{se} = (P^{se})^T
$$

$$
\gamma^* \geq 0
$$

$$
\forall i, \gamma^*_i (H(P^{se}_i) - (\log \xi + 1)) = 0.
$$

By hypothesis $\gamma^* > 0$ which gives $\forall i, H(P^{se}_i) - (\log \xi + 1) = 0$. Now take $0 < \sigma \leq \min_i \gamma^*_i$ and define $\forall i, \omega^*_i = \frac{\gamma^*_i}{\sigma} - 1$. Using the hypothesis on $\sigma$ we have $\forall i, \omega^*_i \geq 0$ and $\omega^*$ satisfies

$$
\forall i, \sigma(1 + \omega^*_i) = \gamma^*_i.
$$

Moreover for any $i \in [n]$

$$
\omega^*_i (H(P^{se}_i) - (\log \xi + 1)) = 0.
$$

Define also $\forall i, \mu^*_i = \lambda^*_i / \sigma$ and $\forall (i, j), \beta^*_i = \Gamma^*_j / \sigma$. Since $P^{se}, (\gamma^*, \lambda^*, \Gamma^*)$ satisfies the KKT conditions (KKT-SEA) then by the previous reasoning $P^{se}, (\omega^*, \mu^*, \beta^*)$ satisfy the KKT conditions (KKT-Proj) and in particular $P^{se}$ is an optimal solution of (SEA-Proj) since KKT conditions are sufficient. Thus we have proven that $P^{se} \in \arg\min_{P \in H_\xi \cap S} KL(P|K_\sigma)$ and by the uniqueness of the solution this is in fact an equality. 

\[\square\]

## B Sensitivity Analysis for Dimensionality Reduction Experiments

In Figure 8 we extend the sensitivity analysis performed for spectral clustering (Figure 5) to DR scores. One can notice that tSNEkhorn outperforms tSNE on a wide range of perplexity values.

![Figure 8](image_url)

**Figure 8:** Dimensionality reduction scores as a function of the perplexity parameter.