Doubly-Stochastic Normalization of the Gaussian Kernel is Robust to Heteroskedastic Noise

Boris Landa\textsuperscript{1,*} Ronald R. Coifman\textsuperscript{1} Yuval Kluger\textsuperscript{1,2,3}

\textsuperscript{1}Program in Applied Mathematics, Yale University
\textsuperscript{2}Interdepartmental Program in Computational Biology and Bioinformatics, Yale University
\textsuperscript{3}Department of Pathology, Yale University School of Medicine

\textsuperscript{*}Corresponding author. Email: boris.landa@yale.edu

June 2, 2020

Abstract

A fundamental step in many data-analysis techniques is the construction of an affinity matrix describing similarities between data points. When the data points reside in Euclidean space, a widespread approach is to form an affinity matrix by the Gaussian kernel with pairwise distances, and to follow with a certain normalization (e.g. the row-stochastic normalization or its symmetric variant). We demonstrate that the doubly-stochastic normalization of the Gaussian kernel with zero main diagonal (i.e. no self loops) is robust to heteroskedastic noise. That is, the doubly-stochastic normalization is advantageous in that it automatically accounts for observations with different noise variances. Specifically, we prove that in a suitable high-dimensional setting where heteroskedastic noise does not concentrate too much in any particular direction in space, the resulting (doubly-stochastic) noisy affinity matrix converges to its clean counterpart with rate $m^{-1/2}$, where $m$ is the ambient dimension. We demonstrate this result numerically, and show that in contrast, the popular row-stochastic and symmetric normalizations behave unfavourably under heteroskedastic noise. Furthermore, we provide a prototypical example of simulated single-cell RNA sequence data with strong intrinsic heteroskedasticity, where the advantage of the doubly-stochastic normalization for exploratory analysis is evident.

1 Introduction

1.1 Affinity matrix constructions

Given a dataset of points in Euclidean space, a useful approach for encoding the intrinsic geometry of the data is by a weighted graph, where the vertices represent data points, and the edge-weights describe similarities between them. Such a graph can be described by an affinity (or adjacency/similarity) matrix, namely a nonnegative matrix whose $(i,j)$’th entry holds the edge-weight between vertices $i$ and $j$. To measure the similarity between pairs of data points, one can employ the Gaussian kernel with pairwise (Euclidean) distance. In particular, given data points $x_1, \ldots, x_n \in \mathbb{R}^m$, we consider the matrix $K \in \mathbb{R}^{n \times n}$ given by

\[
K_{i,j} = \begin{cases} 
\exp(-\|x_i - x_j\|^2/\varepsilon), & i \neq j, \\
0, & i = j,
\end{cases}
\]
for \(i, j = 1, \ldots, n\), where \(\varepsilon\) is the kernel “width” parameter. For many applications it is a common practice to normalize \(K\), so to equip the resulting affinity matrix with a useful interpretation and favourable properties. Two such normalizations, which are closely related to each other, are the row-stochastic and the symmetric normalizations:

\[(\text{Row-stochastic normalization}) \quad W^{(r)} \overset{\text{def}}{=} \text{diag}(r)K, \quad r_i = \frac{1}{\sum_{j=1}^{n} K_{i,j}}, \quad (2)\]

\[(\text{Symmetric normalization}) \quad W^{(s)} \overset{\text{def}}{=} \sqrt{\text{diag}(r)}K\sqrt{\text{diag}(r)}, \quad (3)\]

where \(r = [r_1, \ldots, r_n]\), and \(\text{diag}(r)\) is a diagonal matrix with \(r\) on its main diagonal.

Notably, the matrix \(W^{(r)}\) is row-stochastic, i.e. the sum of every row of \(W^{(r)}\) is 1, which allows for a useful interpretation of \(W^{(r)}\) as a transition-probability matrix (in the sense of a Markov chain). An important characteristic of the the row-stochastic affinity matrix \(W^{(r)}\) is its relation to the heat kernel and the Laplace-Beltrami operator on a manifold \([4, 12, 24, 45, 51]\). Specifically, under the “manifold assumption” – where the points \(x_1, \ldots, x_n\) are uniformly sampled from a smooth low-dimensional Riemannian manifold embedded in the Euclidean space – \(W^{(r)}\) approximates the heat kernel on the manifold, and the matrix \(L^{(r)} = I - W^{(r)}\) (known as the random-walk graph Laplacian) approximates the Laplace-Beltrami operator. This property of the row-stochastic normalization establishes the relation between \(W^{(r)}\) and the intrinsic local geometry of the data, thereby justifying the use of \(W^{(r)}\) as an affinity matrix.

The affinity matrix \(W^{(s)}\) (obtained by the symmetric normalization) is closely-related to \(W^{(r)}\). Since \(W^{(s)} = [\text{diag}(r)]^{-1/2}W^{(r)}[\text{diag}(r)]^{1/2}\), \(W^{(s)}\) shares the spectrum of \(W^{(r)}\), and their eigenvectors are related through the vector \(r\). Even though \(W^{(s)}\) is not a proper transition-probability matrix, it enjoys symmetry, which is advantageous in various applications.

We also mention that the row stochastic and symmetric normalizations can be used in conjunction with a kernel with variable width, i.e. when a different value of \(\varepsilon\) is taken for each row or column of \(K\) (see for instance \([5]\) and references therein). We further discuss one such variant in the example in Section 3.2.

The matrices \(W^{(r)}\) and \(W^{(s)}\) (or equivalently, their corresponding graph Laplacians \(I - W^{(r)}\) and \(I - W^{(s)}\)) are used extensively in data processing and machine learning, notably in non-linear dimensionality reduction (or manifold learning) \([4, 12, 37, 33]\), community detection and spectral-clustering \([43, 38, 41, 53, 20, 12, 29]\), image denoising \([9, 39, 36, 31, 46]\), and in signal processing and supervised-learning over graph domains \([44, 13, 23, 15, 7]\).

### 1.2 The doubly-stochastic normalization

In this work, we focus on the doubly-stochastic normalization of \(K\):

\[(\text{Doubly-stochastic normalization}) \quad W^{(d)} \overset{\text{def}}{=} \text{diag}(d)K\text{diag}(d), \quad (4)\]

where \(d = [d_1, \ldots, d_n] > 0\) is a vector chosen such that \(W^{(d)}\) is doubly-stochastic, i.e., such that the sum of every row and every column of \(W^{(d)}\) is 1. The problem of finding \(d\) such that \(W^{(d)}\) has prescribed row and column sums is known as a matrix scaling problem, and the entries of \(d\) are often referred to as scaling factors. Matrix scaling problems have a rich history, with a long list of applications and
generalizations \cite{2, 27}. Since the scaling factors are defined implicitly, their existence and uniqueness are not obvious, and depend on the zero-pattern of the matrix to be scaled. For the particular zero-pattern of \( K \), existence and uniqueness are established by the following proposition.

**Proposition 1** (Existence and uniqueness). Suppose that \( A \in \mathbb{R}^{n \times n}, n > 2 \), is symmetric with zero main diagonal and strictly positive off-diagonal entries. Then, there exist scaling factors \( d_1, \ldots, d_n > 0 \) such that \( \sum_{j=1}^{n} d_i A_{i,j} d_j = 1 \) for all \( i = 1, \ldots, n \). Moreover, \( \{ d_i \} \) are unique.

The proof can be found in Appendix A and is based on the simple zero-pattern of \( A \) and on a Lemma by Knight \cite{30}. On the computational side, the scaling factors \( d \) can be obtained by the classical Sinkhorn-Knopp algorithm \cite{47} (known also as the RAS algorithm), or by more recent techniques based on optimization (see \cite{1} and references therein). We detail a lean variant of the Sinkhorn-Knopp algorithm adapted to symmetric matrices in Algorithm 1.

**Algorithm 1** Sinkhorn-Knopp algorithm for symmetric matrices \cite{30}

| Input: Symmetric nonnegative \( n \times n \) matrix \( K \), tolerance \( \delta > 0 \). |
|---|
| 1: Initialize: \( d_i^{(0)} = (\sum_{j=1}^{n} K_{i,j})^{-1}, d_i^{(1)} = (\sum_{j=1}^{n} K_{i,j} d_j^{(0)})^{-1}, d_i^{(2)} = (\sum_{j=1}^{n} K_{i,j} d_j^{(1)})^{-1}, \tau = 2. \) |
| 2: While \( \max_{1 \leq i \leq n} |d_i^{(\tau-2)} / d_i^{(\tau)} - 1| > \delta \), do: |
| • \( d_i^{(\tau+1)} = (\sum_{j=1}^{n} K_{i,j} d_j^{(\tau)})^{-1}, \) for \( i = 1, \ldots, n. \) |
| • Update \( \tau \leftarrow \tau + 1. \) |
| 3: Return \( d_i = \sqrt{d_i^{(\tau)} d_i^{(\tau-1)}}, \) for \( i = 1, \ldots, n. \) |

By definition, \( W^{(d)} \) is a symmetric transition-probability matrix. Hence, it naturally combines the two favorable properties that \( W^{(r)} \) and \( W^{(s)} \) hold separately. It is worthwhile to point-out that \( W^{(d)} \) is in fact the closest symmetric and row-stochastic matrix to \( K \) in KL-divergence \cite{8, 56}, and interestingly, it can also be obtained by iteratively re-applying the symmetric normalization (3) indefinitely (see \cite{55}). Another appealing interpretation of the doubly-stochastic normalization is through the lens of optimal transport with entropy regularization \cite{14}, summarized by the following proposition.

**Proposition 2** (Optimal transport interpretation). \( W^{(d)} \) from (4) is the unique solution to

\[
\text{Minimize} \quad \sum_{i,j=1}^{n} ||x_i - x_j||^2 W_{i,j} + \varepsilon H(W),
\]

Subject to: \( W 1 = 1, \quad W^T 1 = 1, \quad W_{i,i} = 0, \quad i = 1 \ldots, n, \)

where \( 1 \) is a column vector of \( n \) ones, and \( H(W) = \sum_{i,j=1}^{n} W_{i,j} \log W_{i,j} \) is the negative entropy.

The proof of Proposition 2 follows very closely with the proof of Lemma 2 in \cite{14}, with the additional use of Proposition 1 (to account for the constraint \( W_{i,i} = 0 \)), and is omitted for the sake of brevity. In the optimal transport interpretation of the problem (3), each point \( x_i \) holds a unit mass that should be distributed between all the other points \( x_j \neq x_i \), while minimizing the transportation cost between the points (measured by the pair-wise distances \( ||x_i - x_j||^2 \)). The outcome of this process is constrained so that each point ends up holding a unit mass. In this context, the matrix \( W \) describes the distribution
of the masses from all points to all other points, and is therefore required to be doubly-stochastic. The negative entropy regularization term $\varepsilon H(W)$ controls the “fairness” of the mass allocation, such that each mass is distributed more evenly between the points for large values of $\varepsilon$.

The optimization problem (5) can also be interpreted as an optimal graph construction. In this context, the term $\sum_{i,j=1}^{n} \|x_i - x_j\|^2 W_{i,j}$ can be considered as accounting for the regularity of the data (as a multivariate signal) with respect to the weighted graph represented by $W$, while the negative entropy term $\varepsilon H(W)$ controls the approximate sparseness of $W$. Since the solution to (5) is a symmetric matrix, $W(d)$ can be thought of as describing the undirected weighted graph that optimizes the “smoothness” of the dataset, under the constraints of prescribed entropy (or approximate sparseness), no self-loops, and stochasticity (i.e. a transition-probability matrix).

In the context of manifold learning, the relation between the doubly-stochastic normalization and the heat kernel (or the Laplace-Beltrami operator) on a Riemannian manifold has been recently established in [35]. That is, under the manifold assumption (and under certain conditions) $W(d)$ is expected to approximate the heat kernel on the manifold, and therefore to encode the local geometry of the data much like $W(r)$. The doubly-stochastic normalization was also demonstrated to be useful for spectral clustering in [3], where it was shown to achieve the best clustering performance on several datasets.

1.3 Robustness to noise

When considering real-world datasets, it is desirable to construct affinity matrices that are robust to noise. Specifically, suppose that we do not have access to the points $x_1, \ldots, x_n$, but rather to their noisy observations $\tilde{x}_1, \ldots, \tilde{x}_n$, given by

$$\tilde{x}_i = x_i + \eta_i,$$

where $\eta_1, \ldots, \eta_n \in \mathbb{R}^m$ are pairwise independent noise vectors satisfying

$$\mathbb{E}[\eta_i] = 0, \quad \mathbb{E}[\eta_i^T \eta_i] = \Sigma_i^2,$$

for all $i = 1, \ldots, n$, where 0 is the zero (row) vector in $\mathbb{R}^m$, and $\Sigma_i^2$ is the covariance matrix of $\eta_i$. We then define $\tilde{W}(r)$, $\tilde{W}(s)$, $\tilde{W}(d)$, $\tilde{K}$, and $\{\tilde{d}_i\}$ analogously to $W(r)$, $W(s)$, $W(d)$, $K$, and $\{d_i\}$, respectively, when replacing $\{x_i\}$ in (1) with $\{\tilde{x}_i\}$. For the noise model described above, we say that the noise is homoskedastic if $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_n$, and heteroskedastic otherwise.

The influence of homoskedastic noise on kernel matrices (such as $K$) was investigated in [16], and the results therein imply that $\tilde{W}(r)$ and $\tilde{W}(s)$ are robust to high-dimensional homoskedastic noise. Specifically, in the high-dimensional setting considered in [16], $\tilde{K}$ converges to a biased version $\tilde{K}$ where all the off-diagonal entries of $\tilde{K}$ admit the same multiplicative bias. Such bias can therefore be corrected by applying either the row-stochastic or the symmetric normalizations (see [17]). However, this is not the case in the more general setting of heteroskedastic noise.

Heteroskedastic noise is a natural assumption for many real-world applications. For example, het-
eroskedastic noise arises in certain biological, photon-imaging, and Magnetic Resonance Imaging (MRI) applications [10, 40, 22, 19], where observations are modeled as samples from random variables whose variances depend on their means, such as in binomial, negative-binomial, multinomial, Poisson, or Rice distributions. In natural image processing, heteroskedastic noise occurs due to the spatial clipping of values in an image [18]. Additionally, heteroskedastic noise is encountered when the experimental setup varies during the data collection process, such as in spectrophotometry and atmospheric data acquisition [11, 48]. Generally, many modern datasets are inherently heteroskedastic as they are formed by aggregating observations collected at different times and from different sources. Last, we mention that heteroskedastic noise can be considered as a natural relaxation to the popular manifold assumption. In particular, heteroskedastic noise arises whenever data points are sampled from the high-dimensional surroundings of a low-dimensional manifold embedded in the ambient space, where the size of the sampling neighborhood (in the ambient space) around the manifold is determined locally by the manifold itself. See Figure 4 and the corresponding example in Section 3.1.2.

1.4 Contributions

Our main contribution is to establish the robustness of the doubly-stochastic normalization of the Gaussian kernel (with zero main diagonal) to high-dimensional heteroskedastic noise. In particular, we prove that in the high-dimensional setting where the number of points $n$ is fixed, the dimension $m$ is increasing, and the noise does not concentrate too much in specific direction in space, $\widehat{W}^{(d)}$ converges to $W^{(d)}$ with rate $m^{-1/2}$. See Theorem 3 in Section 2. An intuitive justification of the robustness of the doubly-stochastic normalization to heteroskedastic noise, and also why zeroing-out the main diagonal of $K$ is important, can be found in Section 2, equations (9)–(10). The proof of Theorem 3 see Appendix B relies on a perturbation analysis of the doubly stochastic normalization.

We demonstrate the robustness of $W^{(d)}$ to heteroskedastic noise in several simulations (see Section 3). In Section 3.1.1 we corroborate Theorem 3 numerically, and exemplify that $W^{(r)}$ and $W^{(s)}$ suffer from inherent point-wise bias due to heteroskedastic noise (see Figures 1–3). In Section 3.1.2 we demonstrate the robustness of the leading eigenvectors of $W^{(d)}$ to heteroskedastic noise whose characteristics depend locally on the manifold of the clean data (see Figures 4–6). In Section 3.2 we apply the doubly stochastic normalization for analyzing simulated single-cell RNA sequence data with significant heteroskedasticity, showcasing its ability to accurately recover the underlying structure of the data despite the noise (see Figures 7–8).

2 Main result

Theorem 3 (Convergence of $\widehat{W}^{(d)}$ to $W^{(d)}$). Consider the setting where the number of points $n$ is fixed, and the dimension $m$ is increasing. Suppose that $\|x_i\| \leq 1$ and $\|\Sigma_i\|_2 \leq C_\eta m^{-1/2}$, for all $i = 1, \ldots, n$ and all (sufficiently large) dimensions $m$, where $C_\eta$ is a universal constant (independent of $m$). Then,

$$\|\widehat{W}^{(d)} - W^{(d)}\|_F = \mathcal{O}(m^{-1/2}),$$

(8)
with high probability.  

The proof is detailed in Appendix. We mention that the constant 1 in the boundedness condition \(|x_i| \leq 1\) is arbitrary and can be replaced with any other constant. Additionally, note that even though the quantities \(|\Sigma_i|_2^2\) are required to decrease with \(m\), the expected noise magnitudes \(E|\eta_i|^2\) (which are equal to \(\text{Tr}\{\Sigma_i^2\}\)) can remain constant, and can possibly be large compared to the magnitudes of the clean data points \(|x_i|^2\). For example, if we have \(|\Sigma_i|_2^2 = m^{-1/2}\) for all \(i\), then it follows that \(E|\eta_i|^2 = \text{Tr}\{\Sigma_i^2\} = 1\), asserting that the magnitude of the noise is greater or equal to that of the clean data points (assuming \(|x_i| \leq 1\)). In this regime of non-vanishing high-dimensional noise, the condition \(|\Sigma_i|_2^2 \leq C\eta m^{-1/2}\) guarantees that the noise spreads-out in Euclidean space, and does not concentrate too much in any particular direction (note that \(|\Sigma_i|_2^2\) is the largest singular value of \(\Sigma_i\), and is therefore the standard deviation of the noise in the direction with largest variance). Clearly, the setup of Theorem 3 can also accommodate for heteroskedastic noise, as the ratios between the noise magnitudes \(E|\eta_i|^2\) for different data points can be arbitrary. See Remark 1 for further discussion of the setting considered in Theorem 3 and a closely-related high-dimensional setting.

The main reason behind the robustness of the doubly-stochastic normalization to high-dimensional heteroskedastic noise, is that it is invariant to the type of bias introduced by heteroskedastic noise. Specifically, under the conditions of Theorem 3, for all \(i \neq j\)

\[
\|\tilde{x}_i - \tilde{x}_j\|^2 \sim \text{m} \rightarrow \infty \quad E\|\tilde{x}_i - \tilde{x}_j\|^2 = E|\eta_i|^2 + \|x_i - x_j\|^2 + E|\eta_j|^2, \tag{9}
\]

and therefore

\[
\tilde{K}_{i,j} \sim \text{m} \rightarrow \infty \exp(-E|\eta_i|^2/\varepsilon) \cdot K_{i,j} \cdot \exp(-E|\eta_j|^2/\varepsilon), \tag{10}
\]

for all \(i, j\) (since \(\tilde{K}_{i,i} = K_{i,i} = 0\)). Crucially, \(\tilde{K}\) in (10) is biased by symmetric diagonal scaling, which is precisely the type of bias corrected automatically by the doubly-stochastic normalization (11). See remark 2 for an alternative justification and further discussion of the robustness of the doubly-stochastic normalization to heteroskedastic noise.

Equations (9) and (10) also highlight why zeroing-out the main diagonal of the Gaussian kernel (see Eq. (1)) is important. Without it, the entries on the main diagonal of \(\tilde{K}\) would be 1, while the off-diagonal entries of \(\tilde{K}\) would be small due to the bias in the noisy pairwise distances (9). Thus, \(\tilde{K}\) would be close to the identity matrix, which would render any normalization (row-stochastic, symmetric, or doubly-stochastic) ineffective.

Remark 1. Consider the following setting for high-dimensionality, where data coordinates (dimensions) are sampled from some underlying distribution, and the noise is only required to have bounded variance, i.e. \(|\Sigma_i|_2^2 \leq C\eta\) for some universal constant \(C\eta\). Specifically, suppose that each clean observation \(x_i\) is given by

\[
x_i = [F_i(y_1), \ldots, F_i(y_m)], \tag{11}
\]

where \(F_i\) is a bounded function, and \(y_1, \ldots, y_m\) are i.i.d samples from some latent “coordinate” variable

\(^1\)The precise statement of the theorem is that for any probability \(p < 1\), there exists a constant \(C(p, n, C\eta)\) and an integer \(M(p, n, C\eta)\), such that for every \(m \geq M(p, n, C\eta)\) we have \(|\tilde{W}^{(d)} - W^{(d)}|_F \leq C(p, n, C\eta)m^{-1/2}\) with probability at least \(p\).
\[ \|x_i - x_j\|^2 = \sum_{k=1}^{m} (F_i(y_k) - F_j(y_k))^2 = m \left( \mathbb{E}_{y \sim Y} \left[ (F_i(y) - F_j(y))^2 \right] + O(m^{-1/2}) \right), \]

with high probability, where the last equality is due to Hoeffding’s inequality \([25]\) (for sums of independent and bounded random variables). Therefore, the clean pairwise distances \(\|x_i - x_j\|^2\) in this setting grow linearly with \(m\), which suggests that the parameter \(\varepsilon\) of the Gaussian kernel \([1]\) should also grow linearly with \(m\). Ultimately, taking \(\varepsilon \propto m\) when computing \(\tilde{K}\) is equivalent to normalizing the noisy observations \(\tilde{x}_i\) by \(\sqrt{m}\) (while keeping \(\varepsilon\) fixed), which places us in the setting of Theorem \([3]\).

**Remark 2.** According to \([4]\), for \(i \neq j\) we can write

\[ W^{(d)}_{ij} = d_i \exp(-\|x_i\|^2/\varepsilon) \cdot \exp(2\langle x_i, x_j \rangle/\varepsilon) \cdot \exp(-\|x_j\|^2/\varepsilon) d_j = u_i \exp(2\langle x_i, x_j \rangle/\varepsilon) u_j, \]

where we defined \(u_i = d_i \exp(-\|x_i\|^2/\varepsilon) > 0\). Thus, \(W^{(d)}\) can be viewed alternatively as obtained by the doubly-stochastic normalization of the nonnegative matrix \(\exp(2\langle x_i, x_j \rangle/\varepsilon)_{i \neq j}\) (with zero main diagonal) instead of \(K\), where the scaling factors are \(u_1, \ldots, u_n\). This implies that \(W^{(d)}\) depends only on the scalar products \(\{\langle x_i, x_j \rangle\}_{i \neq j}\), which are not biased by heteroskedastic noise in the sense that \(\mathbb{E}[\langle \tilde{x}_i, \tilde{x}_j \rangle] = \langle x_i, x_j \rangle\).

Indeed, this fact is a key ingredient in the proof of Theorem \([3]\).

However, it is important to note that without the doubly-stochastic normalization, the matrix \(\exp(2\langle x_i, x_j \rangle/\varepsilon)_{i,j}\) does not correspond to a local kernel (see \([9]\)). This is because the quantity \(\exp(2\langle x_i, x_j \rangle/\varepsilon)\) is sensitive to the magnitudes \(\|x_i\|, \|x_j\|\), such that it may be large even if \(x_i\) and \(x_j\) are not close (but rather \(\|x_i\|\) or \(\|x_j\|\) are large). Therefore, the row-stochastic and symmetric normalizations of the matrix \(\exp(2\langle x_i, x_j \rangle/\varepsilon)_{i \neq j}\) (with zero main diagonal) are not expected to encode the local geometry of the data (even though they are robust to heteroskedastic noise by the virtue of using only the scalar products \(\{\langle x_i, x_j \rangle\}_{i \neq j}\)). To conclude, the doubly stochastic normalization of \(K\) is special in the following way. On the one hand, it is obtained from a local kernel (the Gaussian kernel), and therefore describes the local neighborhoods of the data points (as can be seen from \([4]\) or Proposition \([2]\) and is established formally in \([35]\)). On the other hand, it only depends on the scalar products \(\{\langle x_i, x_j \rangle\}_{i \neq j}\), which is advantageous for coping with heteroskedastic noise.

### 3 Numerical examples

**3.1 Example 1: The unit circle embedded in high-dimensional space**

In our first example, we sampled \(n = 10^3\) points uniformly from the unit circle in \(\mathbb{R}^2\), and embedded them in \(\mathbb{R}^m\), for \(m \in [10, 10^4]\), using randomly-generated orthogonal transformations. In more details, we first sampled angles \(\theta_1, \ldots, \theta_n\) independently and uniformly from \([0, 2\pi]\). Then, for each embedding dimension \(m\), we generated a random orthogonal matrix \(R_m \in \mathbb{R}^{2 \times m}\) (i.e. such that \(R_mR_m^T = I\)), and computed the data points \(\{x_i\}\) as

\[ x_i = [\cos(\theta_i), \sin(\theta_i)] \cdot R_m, \quad 1 \leq i \leq n. \]
Therefore, the noise magnitudes $E$.

Note that as a result, the magnitude of all points is constant, with $\|x_i\| = 1$ for all $1 \leq i \leq n$ and embedding dimension $m$.

3.1.1 Gaussian noise with arbitrary variances

We begin by demonstrating Theorem 3 numerically. Towards that end, we created the noise as follows. For every embedding dimension $m$, we set $\Sigma_i = \text{diag}([\sigma_{i,1}, \ldots, \sigma_{i,m}])$ (so that the noise is uncorrelated between coordinates), and generated the noise standard-deviations $\sigma_{i,j}$ according to

$$\sigma_{i,j} = \sqrt{\frac{\alpha_i \beta_j}{m}}.$$  \hspace{1cm} (15)

where $\{\alpha_i\}_{i=1}^n$, $\{\beta_j\}_{j=1}^m$ were sampled (independently) from the uniform distribution over $[0.05, 0.5]$. Therefore, the noise magnitudes $E\|\eta_i\|^2$ satisfy

$$\frac{1}{400} \leq E\|\eta_i\|^2 \leq \frac{1}{4},$$  \hspace{1cm} (16)

for all $1 \leq i \leq n$, and can take any values in that range. Importantly, the noise magnitudes can vary substantially between data points, which is key in our setting. Then, $\{\eta_{i,j}\}$ were sampled (independently) according to

$$\eta_{i,j} \sim \mathcal{N}(0, \sigma_{i,j}^2).$$  \hspace{1cm} (17)

Once we generated the noisy data points $\tilde{x}_1, \ldots, \tilde{x}_n$ according to (6), we formed the clean and noisy kernel matrices $K$ and $\tilde{K}$ with $\varepsilon = 0.1$, and computed $W^{(d)}, \tilde{W}^{(d)}$ using Algorithm 1 with $\delta = 10^{-12}$. Last, we also evaluated $W^{(r)}, W^{(s)}$ and $\tilde{W}^{(r)}, \tilde{W}^{(s)}$ using $K$ and $\tilde{K}$, respectively, according to (2) and (3).

The behavior of the errors $\|\tilde{W}^{(d)} - W^{(d)}\|_F^2$, $\|\tilde{W}^{(r)} - W^{(r)}\|_F^2$, $\|\tilde{W}^{(s)} - W^{(s)}\|_F^2$ as a function of $m$ can be seen in Figure 1. It is evident that for $m > 100$ the error for the doubly stochastic normalization is substantially smaller than that for the row-stochastic normalization or for the symmetric normalization. Additionally, the error for the doubly-stochastic normalization decreases linearly in logarithmic scale, while the errors for the row-stochastic and the symmetric normalizations reach saturation and never fall below a certain value. In this experiment, the slope of $\log(\|\tilde{W}^{(d)} - W^{(d)}\|_F^2)$ versus $\log(m$) (between $m = 10^2$ and $m = 10^4$) was $-0.9996$, matching the slope suggested by the upper bound in Theorem 3 (which implies a slope of $-1$ for the squared Frobenius norm).

In Figure 2 we depict the noisy affinities $\tilde{W}_{i,j}^{(d)}, \tilde{W}_{i,j}^{(r)}, \tilde{W}_{i,j}^{(s)}$ versus their corresponding clean affinities $W_{i,j}^{(d)}, W_{i,j}^{(r)}, W_{i,j}^{(s)}$, for $m = 10^4$. It can be observed that the noisy affinities from the doubly-stochastic normalization concentrate near their corresponding clean affinities, while the noisy affinities from the row-stochastic and symmetric normalizations deviate substantially from their clean counterparts, particularly for larger affinity values.

Last, in Figure 3 we visually demonstrate the first row of the clean and noisy affinity matrices $W^{(d)}$, $W^{(r)}$, $W^{(s)}$ and $\tilde{W}^{(d)}$, $\tilde{W}^{(r)}$, $\tilde{W}^{(s)}$, using $m = 10^4$. Note that we only display about a quarter of all the entries, since all the other entries are vanishingly small. It can be seen that the clean row-stochastic, clean symmetric, and clean doubly-stochastic affinities are all very similar, and resemble a Gaussian. This is explained by the fact that both $W^{(d)}$ and $W^{(r)}$ are expected to approximate the heat kernel on the unit circle (see 12 35 and other related references given in the introduction), which is close to...
Figure 1: Squared Frobenius loss (averaged over 10 trials) between clean and noisy affinity matrices from different normalizations, versus the dimension $m$. The dataset is the unit circle embedded in different dimensions (see (14)), with $n = 10^3$ and heteroskedastic noise simulated according to (15)–(17).

(a) Doubly-stochastic normalization (4)
(b) Row-stochastic normalization (2)
(c) Symmetric normalization (3)

Figure 2: Entries of the noisy affinity matrices (y-axis) versus the corresponding entries in the clean affinity matrices (x-axis), using different normalizations. The dataset is the unit circle (see (14)), with $n = 10^3$, $m = 10^4$, and heteroskedastic noise simulated according to (15)–(17).

(a) Doubly-stochastic normalization (4)  (b) Row-stochastic normalization (2)  (c) Symmetric normalization (3)
the Gaussian kernel with geodesic distance (for sufficiently small $\varepsilon$). Additionally, since the sampling density on the circle is uniform, $\text{diag}(r)$ (from (2)) is close to a multiple of the identity, and hence $W(s)$ is expected to be close to $W(r)$ (recall that $W(s) = [\text{diag}(r)]^{-1/2} W(r) [\text{diag}(r)]^{1/2}$). Indeed, we found that $\|W(d) - W(r)\|_F^2 \approx \|W(d) - W(s)\|_F^2 \approx 0.01$.

Importantly, the doubly stochastic normalization recovers the true affinities with high accuracy, with an almost perfect match between the corresponding clean and noisy affinities. On the other hand, there is an evident discrepancy between the corresponding clean and noisy affinities from the row-stochastic normalization and from the symmetric normalization.

### 3.1.2 Noise sampled uniformly from a ball with smoothly varying radius

Next, we proceed by demonstrating the robustness of the leading eigenvectors from the doubly-stochastic normalization under heteroskedastic noise, and in particular, in the presence of noise whose magnitude depends on the local geometry of the clean data. Specifically, we simulated heteroskedastic noise whose magnitude varies smoothly according to the angle $\theta_i$ of each point $x_i$ on the circle (see (14)), according to

$$
\eta_i \sim U(B_\rho(\theta_i)), \quad \rho(\theta) = 0.01 + 0.99 \frac{1 + \cos(2\theta)}{2},
$$

(18)

where $U(B_r)$ stands for the uniform distribution over $B_r$, which is a ball with radius $r$ in $\mathbb{R}^m$ (centered at the origin). That is, every noisy observation $\tilde{x}_i$ is sampled uniformly from a sphere whose center is $x_i$ and its radius is $\rho(\theta_i)$ from (18). Consequently, the maximal noise magnitude varies smoothly between 0.01 (for $\theta = \pi/2, 3\pi/2$) and 1 (for $\theta = 0, \pi$). A typical array of clean and noisy points arising from the noise model (18) for dimension $m = 2$ can be seen in Figure 4.

We generated the noisy data points $\tilde{x}_1, \ldots, \tilde{x}_n$ according to (18) for dimension $m = 500$, and formed the noisy kernel matrix $\tilde{K}$ with $\varepsilon = 0.1$. We next computed $\tilde{W}(d)$ using Algorithm 1 with $\delta = 10^{-12}$, and evaluated $\tilde{W}(r), \tilde{W}(s)$ using $\tilde{K}$ according to (2) and (3).

Figure 5 displays the five leading (right) eigenvectors of $\tilde{W}(d), \tilde{W}(r), \tilde{W}(s)$, denoted by $\{\tilde{\psi}_k^{(d)}\}_{k=1}^5, \{\tilde{\psi}_k^{(r)}\}_{k=1}^5, \{\tilde{\psi}_k^{(s)}\}_{k=1}^5$, respectively. It can be seen that the leading eigenvectors from the doubly-stochastic
Figure 4: Typical array of clean and noisy data points for \( n = 1000, m = 2 \), and additive noise sampled uniformly from a sphere whose radius depends on the angle of the corresponding clean point (on the unit circle) according to (18).

Figure 5: Eigenvectors corresponding to the five largest eigenvalues of affinity matrices obtained from different normalizations. The dataset is the unit circle, with \( n = 10^3, m = 500 \), and heteroskedastic noise generated according to (18).

Normalization are almost unaffected by the noise, which is evident by the fact that they approximate sines and cosines – the eigenfunctions of the Laplace-Beltrami operator on the circle. As sines and cosines are advantageous for expanding periodic functions, it is natural to employ the eigenvectors of \( \tilde{W}^{(d)} \) for the purposes of regression, interpolation, and classification over the dataset. It is important to mention that other useful bases and frames can potentially be constructed from \( \tilde{W}^{(d)} \) (see [13, 23]). On the other hand, the eigenvectors obtained from \( \tilde{W}^{(r)} \) and \( \tilde{W}^{(s)} \) are strongly biased due to the heteroskedastic noise, and exhibit undesired effects such as discontinuities and localization. Specifically, as evident from Figure 5, the leading eigenvectors of \( \tilde{W}^{(r)} \) are discontinuous at \( \theta = 0 \) and \( \theta = \pi \), and the leading eigenvectors of \( \tilde{W}^{(s)} \) are localized around \( \theta = \pi/2 \) and \( \theta = 3\pi/2 \) (their values are close to 0 around \( \theta = 0 \) and \( \theta = \pi \)). Clearly, this behaviour of the leading eigenvectors of \( \tilde{W}^{(r)} \) and \( \tilde{W}^{(s)} \) does not reflect the geometry of the data, but rather the characteristics of the noise (since the noise variance is smallest at \( \theta = \pi/2, 3\pi/2 \) and largest at \( \theta = 0, \pi \)).
Figure 6: Two-dimensional embedding using the second and third eigenvectors (corresponding to the second- and third-largest eigenvalues) of affinity matrices obtained from different normalizations. The dataset is the unit circle, with $n = 10^3$, $m = 500$, and heteroskedastic noise generated according to (18). 

In Figure 6 we illustrate the two-dimensional embedding of the noisy data points $\tilde{x}_1, \ldots, \tilde{x}_n$ using the second and third eigenvectors of $\tilde{W}^{(d)}$, $\tilde{W}^{(r)}$, and $\tilde{W}^{(s)}$ (corresponding to their second- and third-largest eigenvalues). That is, the $x$-axis and $y$-axis values for each embedding are given by the entries of $\tilde{\psi}_2^{(d)}$ and $\tilde{\psi}_3^{(d)}$ for the doubly-stochastic normalization, $\tilde{\psi}_2^{(r)}$ and $\tilde{\psi}_3^{(r)}$ for the row-stochastic normalization, and $\tilde{\psi}_2^{(s)}$ and $\tilde{\psi}_3^{(s)}$ for the symmetric normalization (see also [4, 12]). It is clear that the embedding due to the doubly-stochastic normalization reliably represents the intrinsic structure of the clean dataset – a unit circle with uniform density, whereas the embeddings due to the row-stochastic and the symmetric normalizations are incoherent with the geometry and density of the clean points.

3.2 Example 2: Simulated single-cell RNA sequence data

Single-cell RNA sequencing (scRNA-seq) is a revolutionary technique for measuring target gene expressions of individual cells in large and heterogeneous samples [49, 34]. Due to the method’s high resolution (single-cell level) it allows for the discovery of rare cell populations, which is of paramount importance in immunology and developmental biology. A typical scRNA-seq dataset is an $m \times n$ nonnegative matrix corresponding to $n$ cells and $m$ genes, where its $(i, j)$'th entry is an integer called the read count, describing the expression level of $j$'th gene in the $i$'th cell. Importantly, the total number of read counts (or in short total reads) per cell (i.e. row sums) may vary substantially within a sample [28].

We now exemplify the advantage of using the doubly stochastic normalization for exploratory analysis of scRNA-seq data. Specifically, we provide a simple prototypical example where the gene expression levels of cells are measured in two different batches, such that the number of total reads (per cell) within each batch is constant, but is substantially different between the batches. Therefore, the noise variance (modeled by the variance of the multinomial distribution, to be described shortly) differs between the observations in the two batches, giving rise to heteroskedastic noise. Such a scenario can arise naturally in scRNA-seq, either from the intrinsic read count variability common to such datasets, or when two datasets from two independent experiments are merged for unified analysis.

We consider a simulated dataset which includes only two cell types, denoted by $p_1, p_2 \in \mathbb{R}_+^m$, with $m = 4000$ genes. The prototypes $p_1$ and $p_2$ were created by first sampling their entries uniformly (and
Towards that end, non-linear dimensionality reduction techniques are often employed, among which are fuzzy clustering methods that are loyal to the grouping according to cell types, but rather to batch association. In particular, the other hand, the affinity matrices from the row-stochastic and the symmetric normalizations are not well-behaved in all the other observations. Clearly, the fundamental issue here is the heteroskedasticity of the noise, and consequently, the fact that the noise in the last 250 observations is considerably smaller than the noise in all other observations. Therefore, the dataset consists of 500 (normalized) multinomial observations of 501 p

\[ p_{1,j} = \frac{z_{1,j}}{\sum_{k=1}^{m} z_{1,k}}, \quad p_{2,j} = \frac{z_{2,j}}{\sum_{k=1}^{m} z_{2,k}}, \quad z_{1,j}, z_{2,j} \sim U[0,1]. \]  

Next, each noisy observation \( \tilde{x}_i \) was drawn from a multinomial distribution using either \( p_1 \) or \( p_2 \) as the probability vector, and normalized to sum to 1, as described next. First, we generated a batch containing 500 observations of \( p_1 \) and 250 observations of \( p_2 \), each with 1000 multinomial trials. Second, we added a batch containing 250 observations of \( p_2 \) only, each with \( 10^4 \) multinomial trials. To summarize, the total number of observations is \( n = 1000 \), given explicitly by

\[ \tilde{x}_i = \frac{\tilde{x}_i}{\sum_{j=1}^{m} \tilde{x}_{i,j}}, \quad \tilde{x}_i \sim \begin{cases} \text{Multinomial}(10^3, p_1), & 1 \leq i \leq 500, \\ \text{Multinomial}(10^3, p_2), & 501 \leq i \leq 750, \\ \text{Multinomial}(10^4, p_2), & 751 \leq i \leq 1000. \end{cases} \]

Therefore, the dataset consists of 500 (normalized) multinomial observations of \( p_1 \), followed by 500 (normalized) multinomial observations of \( p_2 \). While all observations of \( p_1 \) are with \( 10^3 \) multinomial trials, the observations of \( p_2 \) are split between 250 observations with \( 10^3 \) multinomial trials, and 250 observations with \( 10^4 \) multinomial trials. Evidently, we can write

\[ \bar{x}_i = \mathbb{E}[\tilde{x}_i] + \eta_i = p_{\ell_i} + \eta_i, \quad \ell_i = \begin{cases} 1, & 1 \leq i \leq 500, \\ 2, & 501 \leq i \leq 1000, \end{cases} \]

where \( \eta_i \) is a zero-mean noise vector (arising from the multinomial sampling) satisfying that \( \mathbb{E}[\|\eta_i\|^2] \) is significantly smaller (by a factor of 10 roughly) for \( 751 \leq i \leq 1000 \) compared to \( 1 \leq i \leq 750 \).

Using the noisy observations \( \bar{x}_1, \ldots, \bar{x}_n \), we formed the noisy kernel matrix \( \tilde{K} \) of (1) with \( \epsilon = 2 \cdot 10^{-5} \), computed the corresponding matrix \( \tilde{W}^{(d)} \) using Algorithm 1 with \( \delta = 10^{-12} \), and evaluated the matrices \( \tilde{W}^{(r)}, \tilde{W}^{(s)} \) according to (2) and (3). Our methodology for choosing \( \epsilon \) was to take it to be the smallest possible such that Algorithm 1 converges within the desired tolerance (specifically, in this experiment we set a maximum of \( 10^6 \) iterations for the algorithm). We note that if \( \epsilon \) is too small, then \( \tilde{K} \) becomes too sparse, and the doubly-stochastic normalization may become numerically ill-posed.

Figure 7 illustrates the values (in logarithmic scale) of the obtained affinity matrices \( \tilde{W}^{(d)}, \tilde{W}^{(r)}, \tilde{W}^{(s)} \). It is evident that the affinity matrix from the doubly-stochastic normalization accurately describes the relationships between the data points. That is, \( \tilde{W}^{(d)} \) indicates the similarities within the two groups of cell types (i.e. \( p_1 \) and \( p_2 \)), but also the dissimilarities between them, regardless of batch association. On the other hand, the affinity matrices from the row-stochastic and the symmetric normalizations are not loyal to the grouping according to cell types, but rather to batch association. In particular, \( \tilde{W}^{(r)} \) and \( \tilde{W}^{(s)} \) highlight the observations from the second batch (observations 751–1000) as being most similar to all other observations. Clearly, the fundamental issue here is the heteroskedasticity of the noise, and specifically, the fact that the noise in the last 250 observations is considerably smaller than the noise in all the other observations.

One of the main goals of exploratory analysis of scRNA-seq data is to identify different cell types. Towards that end, non-linear dimensionality reduction techniques are often employed, among which t-
Figure 7: Entries of the affinity matrices obtained from different normalizations in logarithmic scale (from left to right: $\log_{10}(\tilde{W}^{(d)})$, $\log_{10}(\tilde{W}^{(r)})$, $\log_{10}(\tilde{W}^{(s)})$), for single-cell RNA sequence data simulated according to (19)–(21), with $n = 1000$, $m = 4000$.

distributed stochastic neighbor embedding (t-SNE) is perhaps the most prominent. For its operation, t-SNE employs an affinity matrix which is a close variant of the row-stochastic normalization (2), where the kernel width parameter $\varepsilon$ in (1) is allowed to vary between different rows of $K$, and the resulting row-stochastic matrix is symmetrized by averaging it with its transpose. The different kernel widths are determined by a parameter called the perplexity, which is related to the entropy of each row of the resulting affinity matrix.

Even though the affinity matrix employed by t-SNE is a modification of the standard row-stochastic normalization, and uses a different value of $\varepsilon$ for each row, it is still expected to suffer from the inherent bias observed in Figure 7b. Specifically, note that the order of the entries in each row of $\tilde{W}^{(r)}$ (when sorted by their values) does not depend on $\varepsilon$, and only on the noisy pair-wise distances $\|\tilde{x}_i - \tilde{x}_j\|^2$, which are strongly biased by the magnitudes of the noise, as evident from Figure 7b.

In Figures 8a, 8b, 8c we demonstrate the two-dimensional visualization obtained from t-SNE for the dataset $\tilde{x}_1, \ldots, \tilde{x}_n$, using typical perplexity values of 10, 30, 100. We used MATLAB’s standard implementation of t-SNE, activating the option of forcing the algorithm to be exact (i.e. without approximating the affinity matrix). All other parameters of t-SNE were set to their default values suggested by the code (we also mention that the default suggested perplexity is 30).

In Figure 8d we display the two-dimensional visualization obtained from t-SNE when replacing its default affinity matrix construction with the doubly-stochastic matrix $\tilde{W}^{(d)}$ (obtained using $\varepsilon = 2 \cdot 10^{-5}$), while leaving all other aspects of t-SNE unchanged. Since the optimization procedure in t-SNE is affected by randomness, we ran the experiment several times to verify that the results we exhibit are consistent.

While there are only two types of cell in the data ($p_1$ and $p_2$), no clear evidence of this fact can be found in the visualizations by t-SNE (Figures 8a, 8b, 8c). Furthermore, the visualizations by t-SNE do not provide any noticeable separation between the cell types. On the other hand, the visualization obtained by modifying the t-SNE to employ the doubly-stochastic affinity matrix $\tilde{W}^{(d)}$ (Figure 8d) allows one to easily identify and distinguish between the two cell types.
Figure 8: Two-dimensional visualization from t-SNE with different perplexity values (Figures 8a, 8b, 8c), and from t-SNE modified to use the doubly-stochastic affinity matrix $\tilde{W}^{(d)}$ (Figure 8d). The dataset is a simulated single-cell RNA sequence data (see (19)–(21)) with $n = 1000$, $m = 4000$. 
4 Summary and discussion

In this work, we investigated the robustness of the doubly-stochastic normalization to heteroskedastic noise, both from a theoretical perspective and from a numerical one. Our results imply that the doubly-stochastic normalization is advantageous over the popular row-stochastic and symmetric normalizations, particularly when the data at hand is high-dimensional and suffers from inherent heteroskedasticity. Moreover, our experiments suggest that incorporating the doubly-stochastic normalization into various data analysis, visualization, and processing techniques for real-world datasets can be worthwhile. The doubly stochastic normalization is particularly appealing due to its simplicity, solid theoretical foundation, and resemblance to the row-stochastic/symmetric normalizations—which proved useful in countless applications.

The results reported in this work naturally give rise to several possible future research directions. On the theoretical side, it is of interest to characterize the convergence rate of $\widetilde{W}^{(d)}$ to $W^{(d)}$ also in terms of the number of points $n$ and the covariance matrices $\{\Sigma^2_i\}$ explicitly. As a particular simpler case, one may consider the high-dimensional setting where both $n$ and $m$ tend to infinity, while the quantity $n/m$ is fixed (or tends to a fixed constant). On the practical side, it is of interest to investigate how to best incorporate the affinity matrix from the doubly-stochastic normalization into data analysis and visualization techniques. To that end, it is desirable to derive a method for picking the kernel parameter $\varepsilon$ automatically, or in more generality, to determine how to make use of a variable kernel width (similarly to [57]) while retaining the robustness to heteroskedastic noise.

5 Acknowledgements

We would like to thank Boaz Nadler for his useful comments and suggestions. B.L, R.R.C, and Y.K. acknowledge support by NIH grant R01GM131642. R.R.C and Y.K acknowledge support by NIH grant R01HG008383. Y.K. acknowledges support by NIH grant 2P50CA121974. R.R.C acknowledges support by NIH grant 5R01NS10004903.

Appendix A Proof of Proposition \[1\]

We first recall the definition of a fully indecomposable \[2\] matrix. A matrix $B$ is called fully indecomposable if there are no permutation matrices $P$ and $Q$ such that

$$PBQ = \begin{bmatrix} B_1 & 0 \\ B_2 & B_3 \end{bmatrix},$$

(22)

with $B_1$ square. We now proceed to show that $A$ from Proposition \[1\] is fully-indecomposable. Since the only zeros in $A$ are on its main diagonal, there is only one zero in every row and every column of $A$. Consequently, any permutation of the rows and columns of $A$ would retain this property, namely have a single zero in every row and every column. Therefore, if $n > 2$, it is impossible to find $P$ and $Q$ such that (22) would hold for $B = A$, since there cannot be a block of zeros in $PAQ$ whose number of rows or columns is greater than 1. Hence, $A$ is fully-indecomposable, and the existence and uniqueness of
\( d = [d_1, \ldots, d_n] > 0 \) follows from Lemma 4.1 in [30].

Appendix B  Proof of Theorem 3

Throughout the proof, all quantities should be considered as dependent on the dimension \( m \) (unless stated otherwise), while the number of points \( n \) is fixed.

Let us define
\[
\tilde{u}_i = d_i \exp(-\|x_i\|^2/\varepsilon), \quad \tilde{H}_{i,j} = \begin{cases} \exp(2\langle x_i, x_j \rangle/\varepsilon), & i \neq j, \\ 0, & i = j, \end{cases}
\]
for \( i, j = 1, \ldots, n \). By the definition of \( W^{(d)} \) in (4), for \( i \neq j \) we can write
\[
W_{i,j}^{(d)} = d_i \exp(-\|x_i - x_j\|^2/\varepsilon) d_j = d_i e^{-\|x_i\|^2/\varepsilon} e^{2\langle x_i, x_j \rangle/\varepsilon} e^{-\|x_j\|^2/\varepsilon} d_j = u_i H_{i,j} u_j.
\]

Analogously, we define \( \tilde{H}_{i,j} \) and \( \tilde{u}_i \) by replacing \( \{x_i\} \) and \( \{d_i\} \) in (23) with \( \{\tilde{x}_i\} \) and \( \{\tilde{d}_i\} \), respectively, and we have that \( \tilde{W}_{i,j} = \tilde{u}_i \tilde{H}_{i,j} \tilde{u}_j \).

Let \( \odot \) denote the Hadamard (element-wise) product, \( u = [u_1, \ldots, u_n]^T \), and \( \tilde{u} = [\tilde{u}_1, \ldots, \tilde{u}_n]^T \). We can write
\[
\|\tilde{W}^{(d)} - W^{(d)}\|_F = \| \text{diag}(\tilde{u}) \tilde{H} \text{diag}(\tilde{u}) - \text{diag}(u) H \text{diag}(u) \|_F
\]
\[
= \| (uu^T) \odot (\tilde{H} - H) + \tilde{H} \odot (\tilde{u}u^T - uu^T) \|_F
\]
\[
\leq \max_{i,j} \{u_i u_j\} \cdot \|\tilde{H} - H\|_F + \max_{i,j} \{H_{i,j}\} \cdot \|\tilde{u}u^T - uu^T\|_F.
\]

We begin by bounding the quantity \( \|\tilde{H} - H\|_F \), which is the subject of the following Lemma.

Lemma 4. For all \( i \neq j \),
\[
|\tilde{H}_{i,j} - H_{i,j}| = \mathcal{O}(m^{-1/2}),
\]
with high probability.

Proof. Let us write
\[
\langle \tilde{x}_i, \tilde{x}_j \rangle = \langle x_i, x_j \rangle + \langle x_i, \eta_j \rangle + \langle \eta_i, x_j \rangle + \langle \eta_i, \eta_j \rangle.
\]

According to (7) and the conditions in Theorem 3 for \( i \neq j \) we have
\[
\mathbb{E} \{\langle x_i, \eta_j \rangle + \langle \eta_i, x_j \rangle + \langle \eta_i, \eta_j \rangle\} = 0,
\]
\[
\text{Var} \{\langle x_i, \eta_j \rangle\} = \mathbb{E}[x_i \eta_j^T \eta_i x_i^T] = x_i \Sigma_j x_i^T \leq \|x_i\|^2 \|\Sigma_j\|_2 \leq C_m^2 m^{-1},
\]
\[
\text{Var} \{\langle x_j, \eta_i \rangle\} = \mathbb{E}[x_j \eta_i^T \eta_i x_j^T] = x_j \Sigma_i x_j^T \leq \|x_j\|^2 \|\Sigma_i\|_2 \leq C_m^2 m^{-1},
\]
\[
\text{Var} \{\langle \eta_i, \eta_j \rangle\} = \mathbb{E}[\eta_i \eta_j^T \eta_i \eta_j^T] = \sum_{k=1}^{m} \sum_{\ell=1}^{n} \mathbb{E}[\eta_i \eta_k \eta_j \eta_j] \mathbb{E}[\eta_i \eta_k \eta_j \eta_j]
\]
\[
= \text{Tr}\{\Sigma_i^2 \Sigma_j^2\} \leq m \|\Sigma_i^2 \Sigma_j^2\|_2 \leq m \|\Sigma_i\|_2^2 \|\Sigma_j\|_2^2 \leq C_m^4 m^{-1}.
\]

Therefore,
\[
\text{Var} \{\langle x_i, \eta_j \rangle + \langle \eta_i, x_j \rangle + \langle \eta_i, \eta_j \rangle\} \leq \frac{6C_m^2 + 3C_m^4}{m},
\]
where we used the inequality \((a+b+c)^2 \leq 3(a^2+b^2+c^2)\). Consequently, for any \(0 < p < 1\), Chebyshev’s inequality yields that with probability at least \(p\)
\[
|\langle x_i, \eta_j \rangle + \langle \eta_i, x_j \rangle + \langle \eta_i, \eta_j \rangle| \leq \sqrt{\frac{6C_n^2 + 3C_n^4}{m(1-p)}} = O\left(m^{-1/2}\right). \tag{33}
\]
Using the above for \(i \neq j\), a first-order Taylor expansion of \(\exp(y)\) around \(y = 0\) gives
\[
\exp\left\{\frac{2(\langle x_i, \eta_j \rangle + \langle \eta_i, x_j \rangle + \langle \eta_i, \eta_j \rangle)}{\varepsilon}\right\} = 1 + O(m^{-1/2}), \tag{34}
\]
and by (27) we have
\[
\tilde{H}_{i,j} = e^{2\langle \tilde{x}_i, \tilde{x}_j \rangle/\varepsilon} = e^{2\langle x_i, x_j \rangle/\varepsilon}(1 + O(m^{-1/2})) = H_{i,j} + O(m^{-1/2}), \tag{35}
\]
where we used \(H_{i,j} = e^{\langle x_i, x_j \rangle/\varepsilon} \leq e\|x_i\|\|x_j\|/\varepsilon \leq e^{1/\varepsilon} = O(1)\) in the last equality. \(\square\)

Using Lemma 4 and applying the union bound on the off-diagonal entries of \(\tilde{H} - H\), we obtain
\[
\|\tilde{H} - H\|_F = O(m^{-1/2}), \tag{36}
\]
with high probability.

Continuing, we bound the quantities \(\max_{i,j}\{u_i u_j\}\) and \(\max_{i,j}\{\tilde{H}_{i,j}\}\) from (25). Towards that end, we have the following result.

**Proposition 5.** Under the conditions of Theorem 3, \(\{H_{i,j}\}_{i \neq j}\) and \(\{u_i\}_{i=1}^n\) are upper- and lower-bounded by positive constants independent of \(m\).

**Proof.** Observe that for \(i \neq j\) and for all \(m\),
\[
0 < e^{-1/\varepsilon} \leq H_{i,j} = e^{\langle x_i, x_j \rangle/\varepsilon} \leq e^{1/\varepsilon}. \tag{37}
\]
Since the set of \(n \times n\) matrices satisfying the above is compact, and using the fact that \(\{u_i\} > 0\) can be uniquely determined by \(H\) (from Proposition 1 applied to \(H\)), there must exist constants \(c_u, C_u\) independent of \(m\) such that
\[
0 < c_u \leq u_i \leq C_u, \tag{38}
\]
for all \(i\) and \(m\). \(\square\)

Consequently, Proposition 5 together with Lemma 4 guarantee that
\[
\max_{i,j}\{u_i u_j\} = O(1), \quad \max_{i,j}\{\tilde{H}_{i,j}\} = \max_{i,j}\{H_{i,j}\} + O(m^{-1/2}) = O(1), \tag{39}
\]
with high probability.

Next, we turn to bound the quantity \(\|\tilde{u}\tilde{u}^T - uu^T\|_F\) from (25). From Proposition 1 applied to \(H\) and \(\tilde{H}\), it follows that \(u\) and \(\tilde{u}\) are unique. Additionally, by Lemma 4 it is clear that \(\tilde{H} \overset{m \to \infty}{\to} H\) almost surely. Therefore, we also have that \(\tilde{u} \overset{m \to \infty}{\to} u\) almost surely (as otherwise we have a contradiction to the
Towards that end, proposition 5 can be used to verify that the second-order partial derivatives of 
\[ f \] 
around \((H, u)\) are bounded by constants independent of \(m\). In particular,

\[ \sum_{j=1}^{n} \tilde{W}_{i,j}^{(d)} = \sum_{j=1}^{n} \tilde{u}_{i,j} = 1. \tag{40} \]

Let us define the multivariate functions \( \{f_i(A, v)\}_{i=1}^{n} \), where \( A \in \mathbb{R}^{n \times n} \), \( v \in \mathbb{R}^{n} \), as

\[ f_i(A, v) = \sum_{j=1}^{m} v_i A_{i,j} v_j. \tag{41} \]

To bound the error \( \|\tilde{u} - u\| \), we expand \( f_i(A, v) \) around \((H, u)\) using a first-order Taylor expansion. Towards that end, proposition 5 can be used to verify that the second-order partial derivatives of \( f_i \) in the vicinity of \((H, u)\) are bounded by constants independent of \(m\). In particular,

\[ \max_{(A, v) \in \mathcal{B}_1(H, u)} \left| \frac{\partial^2 f_i}{\partial v_k \partial v_j} \right| = O(1), \quad \max_{(A, v) \in \mathcal{B}_1(H, u)} \left| \frac{\partial^2 f_i}{\partial v_k A_{m,j}} \right| = O(1), \quad \frac{\partial^2 f_i}{\partial A_{k,j} A_{m,\ell}} = 0, \tag{42} \]

for all \(i, j, k, m, \ell\), where \( \mathcal{B}_1(H, u) \) is a ball of radius 1 in Euclidean space around \((H, u)\):

\[ \mathcal{B}_1(H, u) = \{(A, v) : \|A - H\|_F^2 + \|v - u\|^2 \leq 1\}. \tag{43} \]

The choice of the radius of the ball \( \mathcal{B}_1(H, u) \) is arbitrary, and is only required to guarantee that the point \((\tilde{H}, \tilde{u})\) is included in \( \mathcal{B}_1(H, u) \) for sufficiently large \(m\). Therefore, by (40) and (42), the first-order Taylor expansion of \( f_i(A, v) \) around \((H, u)\) gives

\[ 1 = f_i(\tilde{H}, \tilde{u}) = f_i(H, u) + \sum_{j=1}^{n} \left. \frac{\partial f_i}{\partial v_j} \right|_{(H, u)} (\tilde{u}_j - u_j) + \sum_{k,j=1}^{n} \left. \frac{\partial f_i}{\partial A_{k,j}} \right|_{(H, u)} (\tilde{H}_{k,j} - H_{k,j}) \]

\[ + O(\|\tilde{u} - u\|^2) + O(\|\tilde{H} - H\|_F^2). \tag{44} \]

where

\[ \left. \frac{\partial f_i}{\partial v_j} \right|_{(H, u)} = \begin{cases} u_i, & j = i, \\ u_i H_{i,j}, & j \neq i, \end{cases} \quad \left. \frac{\partial f_i}{\partial A_{k,j}} \right|_{(H, u)} = \begin{cases} u_i u_j, & k = i, \\ 0, & k \neq i, \end{cases} \tag{45} \]

and we used the fact that \( \sum_{j=1}^{n} u_i H_{i,j} u_j = 1 \) (\( W^{(d)} \) is doubly-stochastic). Next, using that \( f_i(H, u) = 1 \), denoting \( \tilde{u}_j - u_j := e_j \), and multiplying both hand sides of (44) by \( u_i \) \((u_i \) is bounded according to Proposition 5), we can write

\[ e_i = -\sum_{j=1}^{n} u_i^2 H_{i,j} e_j - \sum_{j \neq i}^{n} u_i^2 u_j (\tilde{H}_{i,j} - H_{i,j}) + O(\|e\|^2) + O(\|\tilde{H} - H\|_F^2), \tag{46} \]

where \( e = [e_1, \ldots, e_n]^T \). Consequently, since \( H_{i,i} = 0 \), writing (46) in matrix form gives

\[ (I_n + [\text{diag}(u)]^2 H) e = -[\text{diag}(u)]^2 (\tilde{H} - H) u + O(\|e\|^2) + O(\|\tilde{H} - H\|_F^2), \tag{47} \]
where \( I_n \) is the \( n \times n \) identity matrix. In order to bound the vector \( e \), we must be able to invert the matrix

\[
G := I_n + \text{diag}(u)^2 H, \tag{48}
\]

which is the subject of the following Lemma.

**Lemma 6.** Under the conditions of Theorem 3, \( G \) from (48) is invertible for all \( m \), and \( \|G^{-1}\|_2 \leq C_G \) for some constant \( C_G \) independent of \( m \).

**Proof.** Notice that \( G \) is similar to the matrix

\[
[\text{diag}(u)]^{-1} G \text{diag}(u) = I_n + \text{diag}(u) H \text{diag}(u) = I_n + W(d). \tag{49}
\]

Therefore, \( G \) is invertible if \( I_n + W(d) \) is invertible. Since \( W(d) \) is symmetric and doubly-stochastic, its largest eigenvalue is exactly 1, and \( \lambda_{\min}\{W(d)\} \geq -1 \). Moreover, since \( W_{i,j} > 0 \) for all \( i \neq j \), we have that \( \{(W(d)^2)_{i,j} \geq 0 \) for all \( i,j \). Therefore, by Lemma 8.4.3 in [26] \( W(d) \) has only one eigenvalue with maximal absolute-value (which is 1). Hence, \( \lambda_{\min}\{W(d)\} > -1 \), and we obtain that

\[
\lambda_{\min}\{G\} = \lambda_{\min}\{I_n + W(d)\} = 1 + \lambda_{\min}\{W(d)\} > 0. \tag{50}
\]

The fact that \( \|G^{-1}\|_2 \) is bounded by some constant independent of \( m \) is established by Proposition 5 (since the set of all possible matrices \( G \) that satisfy (37) and (38) is compact).

Using (47) together with Lemma 6 and Proposition 5, we have that

\[
\|e\| \leq C_G \|\text{diag}(u)^2 (\tilde{H} - H)u\| + \mathcal{O}(\|e\|^2) + \mathcal{O}(\|\tilde{H} - H\|_F^2) \leq C_G \|\text{diag}(u)\|^2 \cdot \|\tilde{H} - H\|_2 \cdot \|u\|_2 + \mathcal{O}(\|e\|^2) + \mathcal{O}(\|\tilde{H} - H\|_F^2), \tag{51}
\]

where we used the inequality \( \|\tilde{H} - H\|_2 \leq \|\tilde{H} - H\|_F \). From (36), (51), and the fact that \( \tilde{H} \rightarrow H \), \( \tilde{u} \rightarrow u \) almost surely, it follows that

\[
\|e\| = \|\tilde{u} - u\| = \mathcal{O}(m^{-1/2}), \tag{52}
\]

with high probability. Consequently,

\[
\|\tilde{u}^T u - uu^T\|_F = \|(u + e)(u + e)^T - uu^T\|_F = \|ue^T + ee^T + uu^T\|_F \leq 2\|u\| \cdot \|e\| + \|e\|^2 = \mathcal{O}(m^{-1/2}), \tag{53}
\]

with high probability, where we used Proposition 5 to bound \( \|u\| \). Overall, substituting (53), (36), and (39) into (25), we arrive at the required result

\[
\|\tilde{W}^{(d)} - W^{(d)}\|_F = \mathcal{O}(m^{-1/2}), \tag{54}
\]

with high probability.
References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Much faster algorithms for matrix scaling. In 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pages 890–901. IEEE, 2017.

[2] Ravi B Bapat, Ravindra B Bapat, TES Raghavan, et al. Nonnegative matrices and applications, volume 64. Cambridge University Press, 1997.

[3] Mario Beauchemin. On affinity matrix normalization for graph cuts and spectral clustering. Pattern Recognition Letters, 68:90–96, 2015.

[4] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6):1373–1396, 2003.

[5] Tyrus Berry and John Harlim. Variable bandwidth diffusion kernels. Applied and Computational Harmonic Analysis, 40(1):68–96, 2016.

[6] Tyrus Berry and Timothy Sauer. Local kernels and the geometric structure of data. Applied and Computational Harmonic Analysis, 40(3):439–469, 2016.

[7] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4):18–42, 2017.

[8] Jack B Brown, Phillip J Chase, and Arthur O Pittenger. Order independence and factor convergence in iterative scaling. Linear algebra and its applications, 190:1–38, 1993.

[9] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), volume 2, pages 60–65. IEEE, 2005.

[10] Yuanpei Cao, Anru Zhang, and Hongzhe Li. Multi-sample estimation of bacterial composition matrix in metagenomics data. arXiv preprint arXiv:1706.02380, 2017.

[11] Robert N Cochran and Frederick H Horne. Statistically weighted principal component analysis of rapid scanning wavelength kinetics experiments. Analytical Chemistry, 49(6):846–853, 1977.

[12] Ronald R Coifman and Stéphane Lafon. Diffusion maps. Applied and computational harmonic analysis, 21(1):5–30, 2006.

[13] Ronald R Coifman and Mauro Maggioni. Diffusion wavelets. Applied and Computational Harmonic Analysis, 21(1):53–94, 2006.

[14] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in neural information processing systems, pages 2292–2300, 2013.

[15] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In Advances in neural information processing systems, pages 3844–3852, 2016.
[16] Noureddine El Karoui et al. On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216, 2010.

[17] Noureddine El Karoui, Hau-Tieng Wu, et al. Graph connection laplacian methods can be made robust to noise. *The Annals of Statistics*, 44(1):346–372, 2016.

[18] Alessandro Foi. Clipped noisy images: Heteroskedastic modeling and practical denoising. *Signal Processing*, 89(12):2609–2629, 2009.

[19] Alessandro Foi. Noise estimation and removal in mr imaging: The variance-stabilization approach. In *2011 IEEE International symposium on biomedical imaging: from nano to macro*, pages 1809–1814. IEEE, 2011.

[20] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[21] Naomi Habib, Yinqing Li, Matthias Heidenreich, Lukasz Swiech, Inbal Avraham-David, John J Trombetta, Cynthia Hession, Feng Zhang, and Aviv Regev. Div-seq: Single-nucleus rna-seq reveals dynamics of rare adult newborn neurons. *Science*, 353(6302):925–928, 2016.

[22] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):1–15, 2019.

[23] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[24] Matthias Hein, Jean-Yves Audibert, and Ulrike Von Luxburg. From graphs to manifolds—weak and strong pointwise consistency of graph laplacians. In *International Conference on Computational Learning Theory*, pages 470–485. Springer, 2005.

[25] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[26] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[27] Martin Idel. A review of matrix scaling and sinkhorn’s normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

[28] Tae Kim, Xiang Zhou, and Mengjie Chen. Demystifying” drop-outs” in single cell umi data. *bioRxiv*, 2020.

[29] Yuval Kluger, Ronen Basri, Joseph T Chang, and Mark Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research*, 13(4):703–716, 2003.

[30] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[31] Boris Landa and Yoel Shkolnisky. The steerable graph laplacian and its application to filtering image datasets. *SIAM Journal on Imaging Sciences*, 11(4):2254–2304, 2018.
[32] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nature methods*, 16(3):243–245, 2019.

[33] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[34] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[35] Nicholas F Marshall and Ronald R Coifman. Manifold learning with bi-stochastic kernels. *IMA Journal of Applied Mathematics*, 84(3):455–482, 2019.

[36] François G Meyer and Xilin Shen. Perturbation of the eigenvectors of the graph laplacian: Application to image denoising. *Applied and Computational Harmonic Analysis*, 36(2):326–334, 2014.

[37] Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[38] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[39] Jiahao Pang and Gene Cheung. Graph laplacian regularization for image denoising: Analysis in the continuous domain. *IEEE Transactions on Image Processing*, 26(4):1770–1785, 2017.

[40] Joseph Salmon, Zachary Harmany, Charles-Alban Deledalle, and Rebecca Willett. Poisson noise reduction with non-local pca. *Journal of mathematical imaging and vision*, 48(2):279–294, 2014.

[41] Purnamrita Sarkar, Peter J Bickel, et al. Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962–990, 2015.

[42] Uri Shaham, Kelly Stanton, Henry Li, Ronen Basri, Boaz Nadler, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *International Conference on Learning Representations*, 2018.

[43] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[44] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

[45] Amit Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.
[46] Amit Singer, Yoel Shkolnisky, and Boaz Nadler. Diffusion interpretation of nonlocal neighborhood filters for signal denoising. *SIAM Journal on Imaging Sciences*, 2(1):118–139, 2009.

[47] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[48] Omer Tamuz, Tsevi Mazeh, and Shay Zucker. Correcting systematic effects in a large set of photometric light curves. *Monthly Notices of the Royal Astronomical Society*, 356(4):1466–1470, 2005.

[49] Fuchou Tang, Catalin Barbaciou, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377, 2009.

[50] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

[51] Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error estimates for spectral convergence of the graph laplacian on random geometric graphs toward the laplace–beltrami operator. *Foundations of Computational Mathematics*, pages 1–61, 2019.

[52] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, et al. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335):eaah4573, 2017.

[53] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[54] Fei Wang, Ping Li, Arnd Christian König, and Muting Wan. Improving clustering by learning a bi-stochastic data similarity matrix. *Knowledge and information systems*, 32(2):351–382, 2012.

[55] Ron Zass and Amnon Shashua. A unifying approach to hard and probabilistic clustering. In *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, volume 1, pages 294–301. IEEE, 2005.

[56] Ron Zass and Amnon Shashua. Doubly stochastic normalization for spectral clustering. In *Advances in neural information processing systems*, pages 1569–1576, 2007.

[57] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2005.