Electron. Commun. Probab. **27** (2022), article no. 61, 1–13.
https://doi.org/10.1214/22-ECP502
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

Scaling positive random matrices: concentration and asymptotic convergence*

Boris Landa†

Abstract

It is well known that any positive matrix can be scaled to have prescribed row and column sums by multiplying its rows and columns by certain positive scaling factors. This procedure is known as matrix scaling, and has found numerous applications in operations research, economics, image processing, and machine learning. In this work, we establish the stability of matrix scaling to random bounded perturbations. Specifically, letting $\widetilde{A} \in \mathbb{R}^{M\times N}$ be a positive and bounded random matrix whose entries assume a certain type of independence, we provide a concentration inequality for the scaling factors of $\widetilde{A}$ around those of $A = \mathbb{E}[\widetilde{A}]$. This result is employed to study the convergence rate of the scaling factors of $\widetilde{A}$ to those of A, as well as the concentration of the scaled version of $\widetilde{A}$ around the scaled version of A in operator norm, as $M, N \to \infty$. We demonstrate our results in several simulations.

Keywords: matrix scaling; concentration inequality; matrix balancing; doubly stochastic matrix.
MSC2020 subject classifications: 60B20; 60F10.
Submitted to ECP on November 22, 2021, final version accepted on November 20, 2022.

1 Introduction

Let $A \in \mathbb{R}^{M\times N}$ be a nonnegative matrix. It was established in a series of classical papers [28, 29, 30, 4, 3, 21] that under certain conditions one can find a positive vector $\mathbf{x} = [x_1, \ldots, x_M]$ and a positive vector $\mathbf{y} = [y_1, \ldots, y_N]$, such that the matrix $P = D(\mathbf{x})AD(\mathbf{y})$ has prescribed row sums $\mathbf{r} = [r_1, \ldots, r_M]$ and column sums $\mathbf{c} = [c_1, \ldots, c_N]$, where $D(\mathbf{v})$ is a diagonal matrix with $\mathbf{v}$ on its main diagonal. The problem of finding $\mathbf{x}$ and $\mathbf{y}$ is known as *matrix scaling* or *matrix balancing*; see [14] for a comprehensive review. Formally, we say that a pair of positive vectors $(\mathbf{x}, \mathbf{y})$ *scales* A to row sums $\mathbf{r}$ and column sums $\mathbf{c}$, if

$$r_i = \sum_{j=1}^{N} P_{i,j} = \sum_{j=1}^{N} x_i A_{i,j} y_j, \qquad \text{and} \qquad c_j = \sum_{i=1}^{M} P_{i,j} = \sum_{i=1}^{M} x_i A_{i,j} y_j, \tag{1.1}$$

for all $i \in [M]$ and $j \in [N]$. We refer to $\mathbf{x}$ and $\mathbf{y}$ from (1.1) (or their entries) as *scaling factors* of A. In the special case of $M = N$ and $r_i = c_j = 1$ for all $i \in [M]$ and $j \in [N]$, the problem of matrix scaling becomes that of finding a doubly stochastic normalization of A,

*This research was supported by the National Institute of Health, grant numbers R01GM131642 and UM1DA051410.

†Department of Mathematics, Yale University. E-mail: boris.landa@yale.edu

originally studied by Sinkhorn [28] with the motivation of estimating doubly stochastic transition probability matrices.

It is important to mention that (1.1) is a system of nonlinear equations in $\mathbf{x}$ and $\mathbf{y}$ with no closed-form solution. Nevertheless, if the scaling factors $\mathbf{x}$ and $\mathbf{y}$ exist, they can be found by the Sinkhorn–Knopp algorithm [30] (also known as the RAS algorithm), which is a simple iterative procedure that alternates between computing $\mathbf{x}$ via (1.1) using $\mathbf{y}$ from the previous iteration, and vice versa.

Given a nonnegative matrix A, existence and uniqueness of the scaling factors depend primarily on the particular zero-pattern of A; see [3] and references therein. In this work, we focus on the simpler case that A is strictly positive, in which case existence and uniqueness of the scaling factors are guaranteed by the following theorem (see [29]).

Theorem 1.1 (Existence and uniqueness [29])**.** *Suppose that A, $\mathbf{r}$, and $\mathbf{c}$ are positive, and $\|\mathbf{r}\|_1 = \|\mathbf{c}\|_1$. Then, there exists a pair of positive vectors $(\mathbf{x}, \mathbf{y})$ that scales A to row sums $\mathbf{r}$ and column sums $\mathbf{c}$. Furthermore, the resulting scaled matrix $P = D(\mathbf{x})AD(\mathbf{y})$ is unique, and the pair $(\mathbf{x}, \mathbf{y})$ can be replaced only with $(\alpha\mathbf{x}, \alpha^{-1}\mathbf{y})$, for any $\alpha > 0$.*

Note that the condition $\|\mathbf{r}\|_1 = \|\mathbf{c}\|_1$ in Theorem 1.1 is necessary for existence of the scaling factors, as each of the quantities $\|\mathbf{r}\|_1$ and $\|\mathbf{c}\|_1$ must be the sum of all entries in the scaled matrix P. From this point onward we will always assume that $\mathbf{r}$ and $\mathbf{c}$ are positive and $\|\mathbf{r}\|_1 = \|\mathbf{c}\|_1$.

Over the years, matrix scaling and the Sinkhorn–Knopp algorithm have found a wide array of applications in science and engineering. In economy and operations research, classical applications of matrix scaling include transportation planning [17], analyzing migration fields [31], and estimating social accounting matrices [27]. In image processing and computer vision, matrix scaling was employed for image denoising [24] and graph matching [7]. Recently, matrix scaling has been attracting a growing interest from the machine learning community, with applications in manifold learning [22, 33], clustering [34, 20], and classification [10]. See also [26, 8] for applications of matrix scaling in data science through the machinery of optimal transport.

In many practical situations, matrix scaling is applied to a random matrix that represents a perturbation, or a random observation, of an underlying deterministic population matrix; see for example [19, 33, 24, 6]. Arguably, this is the case in all of the previously-mentioned applications of matrix scaling whenever real data is involved. In particular, applications of matrix scaling in machine learning and data science often involve large data matrices that suffer from corruptions and measurement errors. Consequently, it is important to understand the influence of random perturbations of A on the required scaling factors and on the resulting scaled matrix, particularly in the setting where A is large and the entrywise perturbations are not necessarily small.

The existing literature on stability of matrix scaling under random perturbations is mostly concerned with special instances of the problem. In manifold learning, [33, 19, 18] investigated the doubly stochastic scaling of kernel matrices constructed from random point clouds. The perturbation in this case, which stems from the point cloud sampling, exhibits a special form that depends on the particular kernel (see [9]). The kernel matrix also admits properties such as symmetry and positive definiteness that are utilized in the analysis. Another line of work is on entropically-regularized optimal transport between discrete distributions; see [2, 25, 11, 5, 1, 23, 15] and references therein. In this case, the matrix to be scaled takes the form $(\exp\{-C_{i,j}/\varepsilon\})_{i,j}$, where C is a matrix of ground distances (e.g., Euclidean distances) between two sets of point, and ε is a parameter of the entropic regularization. The prescribed row and column sums $\mathbf{r}$ and $\mathbf{c}$ represent the desired marginals of the transport map, namely the source and target distributions in the transport. The analysis is focused on the stability and convergence of the scaling

scheme when the points are sampled from certain populations, or under perturbations of the marginals. In contrast to the aforementioned lines of work, in this work we focus on scaling unstructured rectangular matrices corrupted by centered entrywise perturbations, describing, e.g., noisy tabular data obtained experimentally.

Let $\widetilde{A} \in \mathbb{R}^{M\times N}$ be a positive random matrix and define $A = \mathbb{E}[\widetilde{A}]$, where $\widetilde{A}$ represents a random bounded perturbation of A. Theorem 1.1 establishes the existence and uniqueness of a set of scaling factors $\{(\alpha\mathbf{x}, \alpha^{-1}\mathbf{y})\}_{\alpha>0}$ of A, together with the existence and uniqueness of the corresponding scaled matrix $P = D(\mathbf{x})AD(\mathbf{y})$. Theorem 1.1 can also be applied analogously to $\widetilde{A}$, establishing the existence and uniqueness of a set of random scaling factors $\{(\alpha\widetilde{\mathbf{x}}, \alpha^{-1}\widetilde{\mathbf{y}})\}_{\alpha>0}$ of $\widetilde{A}$, as well as the existence and uniqueness of the corresponding scaled random matrix $\widetilde{P} = D(\widetilde{\mathbf{x}})\widetilde{A}D(\widetilde{\mathbf{y}})$. The main purpose of this work is to establish that under suitable conditions on $\widetilde{A}$, $\mathbf{r}$, and $\mathbf{c}$, there is a pair of scaling factors $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ of $\widetilde{A}$ that concentrates entrywise around a pair of scaling factors $(\mathbf{x}, \mathbf{y})$ of A, and furthermore, the resulting scaled random matrix $\widetilde{P}$ concentrates around P in operator norm. Our results demonstrate that matrix scaling is robust to random perturbations of A when the dimensions of the matrix are sufficiently large.

2 Main results

2.1 Concentration of matrix scaling factors

Denote $s = \|\mathbf{r}\|_1 = \|\mathbf{c}\|_1$. Out main result is the following theorem, which provides a concentration inequality for a certain pair of scaling factors of $\widetilde{A}$ around any pair of scaling factors of A.

Theorem 2.1 (Concentration of scaling factors)**.** *Let $\widetilde{A} \in \mathbb{R}^{M\times N}$ be a positive random matrix, $A = \mathbb{E}[\widetilde{A}]$, and $(\mathbf{x}, \mathbf{y})$ be a pair of positive vectors that scales A to row sums $\mathbf{r}$ and column sums $\mathbf{c}$. Suppose that $\widetilde{A}_{i,j} \in [a_{i,j}, b_{i,j}]$ a.s. for all $i \in [M]$ and $j \in [N]$, and denote $a = \min_{i,j} a_{i,j}$, $b = \max_{i,j} b_{i,j}$, and $d = \max_{i,j}\{b_{i,j} - a_{i,j}\}$. Suppose further that $\{\widetilde{A}_{i,j}\}_{j=1}^N$ are independent for each $i \in [M]$, and $\{\widetilde{A}_{i,j}\}_{i=1}^M$ are independent for each $j \in [N]$. Then, there exists a pair of positive random vectors $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ that scales $\widetilde{A}$ to row sums $\mathbf{r}$ and column $\mathbf{c}$, such that for any $\delta \in (0, 1]$, with probability at least*

$$1 - 2M\exp\left(-\frac{\delta^2 s^2}{C_p^2\|\mathbf{c}\|_2^2}\right) - 2N\exp\left(-\frac{\delta^2 s^2}{C_p^2\|\mathbf{r}\|_2^2}\right), \tag{2.1}$$

we have for all $i \in [M]$ and $j \in [N]$,

$$\frac{|\widetilde{x}_i - x_i|}{x_i} \le \frac{C_e\delta s}{M\min_i r_i}, \qquad \frac{|\widetilde{y}_j - y_j|}{y_j} \le \frac{C_e\delta s}{N\min_j c_j}, \tag{2.2}$$

where $C_p = \sqrt{2}\left(bd/a^2\right)$ and $C_e = 1 + 2\left(b/a\right)^{7/2}$.

The proof of Theorem 2.1 can be found in Section 4.1. Note that Theorem 2.1 requires the entries of $\widetilde{A}$ to be independent in each of its rows and each of its columns separately. Clearly, this condition is less restrictive than requiring all entries of $\widetilde{A}$ to be independent. For instance, consider the matrix $\widetilde{A}_{i,j} = g_{i,j}(u_i v_j)$, where $\{u_i\}_{i=1}^M$, $\{v_j\}_{j=1}^N$ are independent Rademacher variables, and $g_{i,j} : \{-1, 1\} \to \{a_{i,j}, b_{i,j}\}$ are deterministic functions with $0 < a_{i,j} < b_{i,j}$, for all $i \in [M]$, $j \in [N]$. Evidently, each row and column of $\widetilde{A}$ contains independent entries, yet the entries of $\widetilde{A}$ are strongly dependent since knowing any single row (column) of $\widetilde{A}$ substantially restricts the distribution of any other row (column).

We next apply Theorem 2.1 to study the convergence rate of the scaling factors of $\widetilde{A}$ to those of A in relative error, as the dimensions M and N grow. Let $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ be a pair of

scaling factors of a random positive matrix $\widetilde{A} \in \mathbb{R}^{M\times N}$, and let $(\mathbf{x}, \mathbf{y})$ be a pair of scaling factors of $A = \mathbb{E}[\widetilde{A}]$. We then define the error

$$\mathcal{E} = \min_{\alpha>0} \max \left\{ \max_{i\in[M]} \frac{|\alpha\widetilde{x}_i - x_i|}{x_i}, \max_{j\in[N]} \frac{|\alpha^{-1}\widetilde{y}_j - y_j|}{y_j} \right\}. \tag{2.3}$$

Note that the scaling factors $(\mathbf{x}, \mathbf{y})$ may converge to 0 or grow unbounded as $M, N \to \infty$, depending on the behavior of the prescribed row sums $\mathbf{r}$ and column sums $\mathbf{c}$. Consequently, the normalization by x_i and y_j in the error (2.3) is important for making $\mathcal{E}$ meaningful in the asymptotic regime of $M, N \to \infty$. In what follows, we use the notation $X \lesssim f(M, N)$ to mean $X \le Cf(M, N)$ for a universal constant $C > 0$ independent of M and N. We now have the following corollary of Theorem 2.1.

Corollary 2.2 (Convergence rate of scaling factors)**.** *Suppose that $\widetilde{A}$ satisfies the conditions in Theorem 2.1 with universal positive constants a, b (independent of M and N). Then, with probability at least $1 - 4/\min\{M, N\}$,*

$$\mathcal{E} \lesssim \rho_1\rho_2\sqrt{\log(\max\{M, N\})}, \tag{2.4}$$

where

$$\rho_1 = \max\left\{\frac{\|\mathbf{r}\|_2}{s}, \frac{\|\mathbf{c}\|_2}{s}\right\}, \qquad \rho_2 = \max\left\{\frac{s}{M\min_i r_i}, \frac{s}{N\min_j c_j}\right\}. \tag{2.5}$$

The proof can be found in Section 4.2. To exemplify Corollary 2.2, let us consider the setting of doubly stochastic matrix scaling, namely $M = N$, and $r_i = c_j = 1$ for all $i \in [M]$, $j \in [N]$. According to (2.5), we have $\rho_1 = 1/\sqrt{N}$ and $\rho_2 = 1$. Hence, Corollary 2.2 asserts that $\mathcal{E} \lesssim \sqrt{\log N/N}$ with probability approaching 1 as $N \to \infty$. Similarly, it is easy to verify that the same rate $\sqrt{\log N/N}$ holds whenever M grows proportionally to N and $\max_i r_i/\min_i r_i \le c$, $\max_j c_j/\min_j c_j \le c$, for some universal constant $c > 0$. Under the same conditions on $\mathbf{r}$ and $\mathbf{c}$ but for general M and N, we have that

$$\mathcal{E} \lesssim \sqrt{\frac{\log(\max\{M, N\})}{\min\{M, N\}}}, \tag{2.6}$$

with probability at least $1 - 4/\min\{M, N\}$, which follows immediately from the fact that $\rho_1 \le c\max\{M^{-1/2}, N^{-1/2}\}$ and $\rho_2 \le c$ (if $\max_i r_i/\min_i r_i \le c$ and $\max_j c_j/\min_j c_j \le c$).

Aside from the setting where the r_i's and c_j's have the same orders of magnitude, Corollary 2.2 provides guarantees on the convergence rate of the scaling factors even if some of the r_i's or c_j's grow unbounded with M or N relative to others. For instance, considering again the setting of doubly stochastic matrix scaling, we can set a fixed number of the r_i's or c_j's to be $\sqrt{N}$ instead of 1, without affecting the behavior of $\|\mathbf{r}\|_2$, $\|\mathbf{c}\|_2$, and s asymptotically as N grows. Consequently, the convergence rate of $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ to $(\mathbf{x}, \mathbf{y})$ (in the sense of (2.3)) would remain $\sqrt{\log N/N}$ in this case.

2.2 Concentration of $\widetilde{P}$ around P in operator norm

Let $\widetilde{P}$ and P be the matrices obtained from $\widetilde{A}$ and A, respectively, after scaling them to row sums $\mathbf{r}$ and column sums $\mathbf{c}$, i.e., $\widetilde{P} = D(\widetilde{\mathbf{x}})\widetilde{A}D(\widetilde{\mathbf{y}})$, $P = D(\mathbf{x})AD(\mathbf{y})$. We now have the following result.

Corollary 2.3. *(Concentration of $\widetilde{P}$ around P) Suppose that the entries of $\widetilde{A}$ are independent, and $\widetilde{A}_{i,j} \in [a, b]$ a.s. for all $i \in [M]$, $j \in [N]$ and universal positive constants a, b (independent of M and N). Then, with probability at least $1 - 6/\min\{M, N\}$,*

$$\|\widetilde{P} - P\|_2 \lesssim \rho_1\rho_2\rho_3\sqrt{\log(\max\{M, N\})}, \tag{2.7}$$

where ρ_1 and ρ_2 are from (2.5)*, and*

$$\rho_3 = \frac{\sqrt{M}\max_i r_i \cdot \sqrt{N}\max_j c_j}{s}. \tag{2.8}$$

The proof can be found in Section 4.3. To exemplify Corollary 2.3, we consider again the case of doubly stochastic matrix scaling, where $\rho_1 = 1/\sqrt{N}$, $\rho_2 = \rho_3 = 1$. In this case, $\|\widetilde{P} - P\|_2 \lesssim \sqrt{\log N/N}$ with probability approaching 1 as $N \to \infty$. Note that since P is doubly stochastic, it follows that $\|P\|_2 = 1$ (see [13]). In the more general case where the prescribed row and column sums are not 1, $\|P\|_2$ can converge to zero or grow unbounded with M and N, depending on the asymptotic behavior of $\mathbf{r}$ and $\mathbf{c}$. Therefore, it is worthwhile to consider the normalized error $\|\widetilde{P} - P\|_2/\|P\|_2$. If the conditions in Corollary 2.3 hold and in addition $\max_i r_i/\min_i r_i \le c$, $\max_j c_j/\min_j c_j \le c$, for some universal constant $c > 0$, then with probability at least $1 - 6/\min\{M, N\}$,

$$\frac{\|\widetilde{P} - P\|_2}{\|P\|_2} \lesssim \sqrt{\frac{\log(\max\{M, N\})}{\min\{M, N\}}}. \tag{2.9}$$

Equation (2.9) follows from the fact that

$$\|P\|_2 \ge \max\left\{\left\|\frac{P\mathbf{1}_N}{\sqrt{N}}\right\|_2, \left\|\frac{\mathbf{1}_M^T P}{\sqrt{M}}\right\|_2\right\} \ge \sqrt{\min_i r_i \min_j c_j}, \tag{2.10}$$

where $\mathbf{1}_N$ is the column vector of N ones, and we used $\rho_1 \le c\max\{M^{-1/2}, N^{-1/2}\}$, $\rho_2 \le c$, and

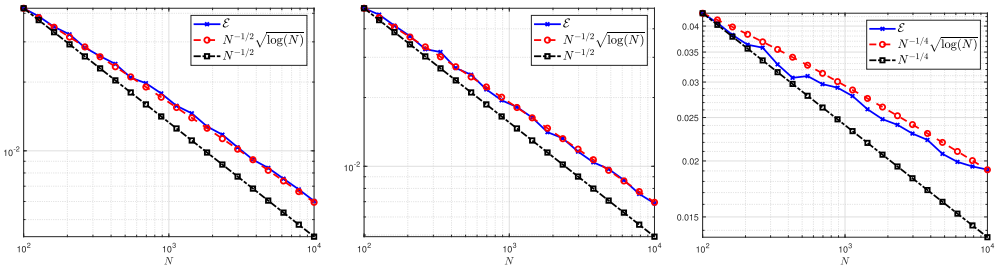
$$\frac{\rho_3}{\|P\|_2} \le \frac{\max_i r_i \max_j c_j}{\|P\|_2\sqrt{\min_i r_i \min_j c_j}} \le c^2, \tag{2.11}$$

since $s = \sqrt{s}\sqrt{s} \ge \sqrt{M\min_i r_i}\sqrt{N\min_j c_j}$.

3 Numerical experiments

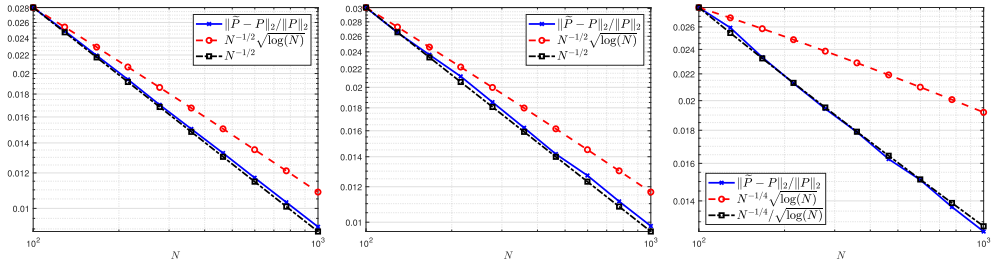
We now exemplify our results in several experiments. In all of our experiments, the matrix A was generated by sampling its entries independently and uniformly from $[1.5, 2.5]$, and $\widetilde{A}_{i,j}$ were sampled independently and uniformly from $[A_{i,j} - 0.5, A_{i,j} + 0.5]$ for all $i \in [M]$ and $j \in [N]$. Then, the Sinkhorn–Knopp algorithm [30, 16] was applied to both A and $\widetilde{A}$, where the algorithm iterations were terminated once the row and column sums of the scaled matrices reached their targets up to an error of 10^{-12}. The resulting pairs of scaling factors were normalized so that $\|\widetilde{\mathbf{x}}\|_1 = \|\widetilde{\mathbf{y}}\|_1$ and $\|\mathbf{x}\|_1 = \|\mathbf{y}\|_1$. This process was repeated 20 times (each time for a different realization of A and $\widetilde{A}$) and the error measures appearing in the left-hand sides of (2.4) and (2.9) were computed and averaged over the 20 randomized trials.

Figure 1 depicts the behavior of the empirical error $\mathcal{E}$ (see (2.3)) as a function of N in several scenarios. Specifically, Figure 1a exemplifies the scenario of doubly stochastic matrix scaling, i.e., $M = N$ and $r_i = c_j = 1$ for all $i \in [M]$ and $j \in [N]$, in which case (2.6) guarantees that $\mathcal{E} \lesssim N^{-1/2}\sqrt{\log N}$ with probability approaching 1 as $N \to \infty$. Figure 1b illustrates the case of a rectangular matrix with $M = 3N$, where the prescribed row and column sums were sampled independently and uniformly from $[0.1, 1]$ and normalized to sum to 1. In this case, since M is proportional to N, and in addition $\max_i r_i/\min_i r_i \le 10$, $\max_j c_j/\min_j c_j \le 10$, (2.6) again dictates that $\mathcal{E} \lesssim N^{-1/2}\sqrt{\log N}$ (as for the doubly stochastic case). Figure 1c illustrates the scenario of a rectangular matrix with $M = 10\sqrt{N}$ and $r_i = N$, $c_j = M$, for all $i \in [M]$ and $j \in [N]$. In this case it follows from (2.6) that $\mathcal{E} \lesssim N^{-1/4}\sqrt{\log N}$. It is evident from Figures 1a, 1b, 1c that the probabilistic bound in Corollary 2.2 agrees very well with the experimental results,



(a) $M = N$, $r_i = c_j = 1$ (b) $M = 3N$, r_i, c_j are random (c) $M = 10\sqrt{N}$, $r_i = N$, $c_j = M$

Figure 1: Empirical values of $\mathcal{E}$ as function of N in log-log scale, compared to the corresponding bounds from (2.6). Figure 1a corresponds to $M = N$ and $r_i = c_j = 1$ for all $i \in [M]$ and $j \in [N]$. Figure 1b corresponds to $M = 3N$, and $\{r_i\}, \{c_j\}$ sampled independently and uniformly from $[0.1, 1]$ and normalized to sum to 1. Figure 1c corresponds to $M = 10\sqrt{N}$ and $r_i = N,\ c_j = M$ for all $i \in [M]$ and $j \in [N]$.



(a) $M = N$, $r_i = c_j = 1$ (b) $M = 3N$, r_i, c_j are random (c) $M = 10\sqrt{N}$, $r_i = N$, $c_j = M$

Figure 2: Empirical values of $\|\widetilde{P} - P\|_2/\|P\|_2$ as a function of N in log-log scale for the same settings as in Figure 1, compared to the corresponding bounds from (2.9).

suggesting that this bound is tight for the considered scenarios, and in particular that the factor $\sqrt{\log(N)}$ in the corresponding bounds is necessary.

Figure 2 shows the behavior of the empirical error $\|\widetilde{P} - P\|_2/\|P\|_2$ as a function of N for the same scenarios as in Figure 1. For these scenarios, the rates that govern the bounds on $\|\widetilde{P} - P\|_2/\|P\|_2$ according to (2.9) are the same as those for $\mathcal{E}$ from (2.6) (described previously in the context of Figure 1). In the scenario where M grows proportionally to N, it is evident from Figures 2a and 2b that the bound in (2.9) is tight up to the factor $\sqrt{\log(N)}$, suggesting that this factor is probably not required in the bound (in contrast to the bound on $\mathcal{E}$ depicted in Figure 1). In the scenario where M grows disproportionately to N, Figure 1c empirically suggests that the bound in (2.9) can be improved by a factor of $\log(N)$, which would bring the rate in this case to be $N^{-1/4}/\sqrt{\log N}$. Overall, these experiments suggest that the bound in (2.9) describes the correct behavior of $\|\widetilde{P} - P\|_2/\|P\|_2$ up to poly-logarithmic factors in the tested scenarios.

4 Proofs

4.1 Proof of Theorem 2.1

Let us define

$$\overline{r}_i = \frac{r_i}{\sqrt{\|\mathbf{r}\|_1}}, \qquad \overline{c}_j = \frac{c_j}{\sqrt{\|\mathbf{c}\|_1}}, \tag{4.1}$$

for $i \in [M]$ and $j \in [N]$. We begin with the following lemma, which describes a useful normalization of the scaling factors and the resulting bounds on their entries.

Lemma 4.1 (Boundedness of scaling factors)**.** *Let A be a positive matrix, and denote $a = \min_{i,j} A_{i,j}$ and $b = \max_{i,j} A_{i,j}$. There exists a unique pair of positive vectors $(\mathbf{x}, \mathbf{y})$ that satisfies $\|\mathbf{x}\|_1 = \|\mathbf{y}\|_1$ and scales A to row sums $\mathbf{r}$ and column sums $\mathbf{c}$, and moreover, for all $i \in [M]$ and $j \in [N]$,*

$$\frac{\sqrt{a}}{b} \le \frac{x_i}{\overline{r}_i} \le \frac{\sqrt{b}}{a}, \qquad \frac{\sqrt{a}}{b} \le \frac{y_j}{\overline{c}_j} \le \frac{\sqrt{b}}{a}. \tag{4.2}$$

Proof. By Theorem 1.1, we have a pair $(\mathbf{x}', \mathbf{y}')$ of scaling factors of A, and denoting $(\mathbf{x}, \mathbf{y}) = (\alpha \mathbf{x}', \alpha^{-1}\mathbf{y}')$ with $\|\alpha \mathbf{x}'\|_1 = \|\alpha^{-1}\mathbf{y}'\|_1$, determines α uniquely. According to (1.1), we have $x_i = r_i / \sum_{j=1}^N A_{i,j} y_j$ and $y_j = c_j / \sum_{i=1}^M A_{i,j} x_i$. Therefore, since $a \le A_{i,j} \le b$,

$$\frac{r_i}{b\|\mathbf{y}\|_1} \le x_i \le \frac{r_i}{a\|\mathbf{y}\|_1}, \qquad \frac{c_j}{b\|\mathbf{x}\|_1} \le y_j \le \frac{c_j}{a\|\mathbf{x}\|_1}, \tag{4.3}$$

for all $i \in [M]$, $j \in [N]$. Summing the inequalities in (4.3) for x_i over $i = 1, \dots, M$, and using $\|\mathbf{y}\|_1 = \|\mathbf{x}\|_1$ together with $\|\mathbf{r}\|_1 = s$, gives $\sqrt{s/b} \le \|\mathbf{y}\|_1 = \|\mathbf{x}\|_1 \le \sqrt{s/a}$. Plugging this last inequality back into (4.3) gives the required result. □

We now proceed with the proof of Theorem 2.1. Without loss of generality, we can always assume that the pair of scaling factors $(\mathbf{x}, \mathbf{y})$ satisfies $\|\mathbf{x}\|_1 = \|\mathbf{y}\|_1$, as otherwise we can replace $(\mathbf{x}, \mathbf{y})$ and $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ with $(\alpha \mathbf{x}, \alpha^{-1}\mathbf{y})$ and $(\alpha\widetilde{\mathbf{x}}, \alpha^{-1}\widetilde{\mathbf{y}})$, respectively, for an appropriate $\alpha > 0$, and clearly we are not changing the normalized errors in (2.2). Since $\widetilde{A}_{i,j} \in [a_{i,j}, b_{i,j}]$ a.s., then $A_{i,j} = \mathbb{E}[\widetilde{A}_{i,j}] \in [a_{i,j}, b_{i,j}]$, and Lemma 4.1 dictates that $x_i y_j \le \overline{r}_i \overline{c}_j (b/a^2) = C r_i c_j / s$ for all $i \in [M]$ and $j \in [N]$, where $C = b/a^2$. Therefore, using the fact that $\sum_j x_i A_{i,j} y_j = r_i$, Hoeffding's inequality [12] implies that

$$\Pr\left\{ \left| \frac{1}{r_i} \sum_{j=1}^N x_i \widetilde{A}_{i,j} y_j - 1 \right| > \varepsilon \right\} \le 2 \exp\left(\frac{-2\varepsilon^2 s^2}{C^2 \sum_{j=1}^N c_j^2 (b_{i,j} - a_{i,j})^2} \right), \tag{4.4}$$

for all $i \in [M]$. Similarly, by a symmetry argument (considering (4.4) when replacing $\widetilde{A}$ with $\widetilde{A}^T$, interchanging the roles of $\mathbf{r}$ and $\mathbf{c}$), one can verify that

$$\Pr\left\{ \left| \frac{1}{c_j} \sum_{i=1}^M x_i \widetilde{A}_{i,j} y_j - 1 \right| > \varepsilon \right\} \le 2 \exp\left(\frac{-2\varepsilon^2 s^2}{C^2 \sum_{i=1}^M r_i^2 (b_{i,j} - a_{i,j})^2} \right), \tag{4.5}$$

for all $j \in [N]$. Consequently, applying the union bound over all $i \in [M]$ and $j \in [N]$ asserts that with probability at least

$$1 - 2\sum_{i=1}^M \exp\left(-\frac{2\varepsilon^2 s^2}{C^2 \sum_{j=1}^N c_j^2 (b_{i,j} - a_{i,j})^2} \right) - 2\sum_{j=1}^N \exp\left(-\frac{2\varepsilon^2 s^2}{C^2 \sum_{i=1}^M r_i^2 (b_{i,j} - a_{i,j})^2} \right), \tag{4.6}$$

we have

$$\max_{j \in [N]} \left| \frac{1}{M} \sum_{i=1}^M x_i \widetilde{A}_{i,j} y_j - 1 \right| \le \varepsilon, \qquad \text{and} \qquad \max_{i \in [M]} \left| \frac{1}{N} \sum_{j=1}^N x_i \widetilde{A}_{i,j} y_j - 1 \right| \le \varepsilon. \tag{4.7}$$

In what follows we assume that the event (4.7) holds, and extend Sinkhorn's original proof of uniqueness of the scaling factors (see [29]) to describe their stability under

approximate scaling (as in (4.7)). Let $(\overline{\mathbf{x}}, \overline{\mathbf{y}})$ be the unique pair of scaling factors of $\widetilde{A}$ with $\|\overline{\mathbf{x}}\|_1 = \|\overline{\mathbf{y}}\|_1$ (applying Lemma 4.1 to $\widetilde{A}$), and define

$$\hat{P}_{i,j} = x_i \widetilde{A}_{i,j} y_j, \qquad \widetilde{P}_{i,j} = \overline{x}_i \widetilde{A}_{i,j} \overline{y}_j = u_i \hat{P}_{i,j} v_j, \qquad u_i = \frac{\overline{x}_i}{x_i}, \qquad v_j = \frac{\overline{y}_j}{y_j}, \tag{4.8}$$

for all $i \in [M]$ and $j \in [N]$. Observe that $\sum_i \widetilde{P}_{i,j} = c_j$, $\sum_j \widetilde{P}_{i,j} = r_i$, and

$$\left| \frac{1}{c_j} \sum_{i=1}^{M} \hat{P}_{i,j} - 1 \right| \le \varepsilon, \qquad \left| \frac{1}{r_i} \sum_{j=1}^{N} \hat{P}_{i,j} - 1 \right| \le \varepsilon, \tag{4.9}$$

for all $i \in [M]$, $j \in [N]$. Using the first inequality in (4.9) for $j = \operatorname{argmin}_k v_k$, we have

$$1 = \frac{1}{c_j} \sum_{i=1}^{M} \widetilde{P}_{i,j} = \frac{1}{c_j} \sum_{i=1}^{M} u_i \hat{P}_{i,j} v_j \le \min_j v_j \max_i u_i \frac{1}{c_j} \sum_{i=1}^{M} \hat{P}_{i,j} \le (1+\varepsilon) \min_j v_j \max_i u_i. \tag{4.10}$$

Similarly, using the second inequality in (4.9) for $i = \operatorname{argmax}_\ell v_\ell$ gives

$$1 = \frac{1}{r_i} \sum_{j=1}^{N} \widetilde{P}_{i,j} = \frac{1}{r_i} \sum_{j=1}^{N} u_i \hat{P}_{i,j} v_j \ge \max_i u_i \min_j v_j \frac{1}{r_i} \sum_{j=1}^{N} \hat{P}_{i,j} \ge (1-\varepsilon) \max_i u_i \min_j v_j, \tag{4.11}$$

and by combining (4.10) and (4.11) we obtain

$$\frac{1}{1+\varepsilon} \le \max_i u_i \min_j v_j \le \frac{1}{1-\varepsilon}. \tag{4.12}$$

By a symmetry argument (considering (4.12) in the setting where $\widetilde{A}$ is replaced with $\widetilde{A}^T$, interchanging the roles of $\mathbf{u}$ and $\mathbf{v}$), we also have

$$\frac{1}{1+\varepsilon} \le \min_i u_i \max_j v_j \le \frac{1}{1-\varepsilon}. \tag{4.13}$$

Let us denote $\ell = \operatorname{argmax}_i u_i$. By the second inequality in (4.9) together with (4.12), we can write

$$1 = \frac{1}{r_\ell} \sum_{j=1}^{N} \widetilde{P}_{\ell,j} = \frac{1}{r_\ell} \sum_{j=1}^{N} u_\ell \hat{P}_{\ell,j} v_j \ge \frac{1}{(1+\varepsilon) r_\ell} \sum_{j=1}^{N} \hat{P}_{\ell,j} \frac{v_j}{\min_i v_i}, \tag{4.14}$$

implying that

$$\frac{1}{r_\ell} \sum_{j=1}^{N} \hat{P}_{\ell,j} \left(\frac{v_j}{\min_i v_i} - 1 \right) \le 1 + \varepsilon - \frac{1}{r_\ell} \sum_{j=1}^{N} \hat{P}_{\ell,j} \le 2\varepsilon. \tag{4.15}$$

Multiplying (4.15) by $\min_j v_j / \min_k \hat{P}_{\ell,k}$, it follows that

$$\frac{1}{r_\ell} \sum_{j=1}^{N} (v_j - \min_i v_i) \le \frac{1}{r_\ell} \sum_{j=1}^{N} \frac{\hat{P}_{\ell,j}}{\min_k \hat{P}_{\ell,k}} (v_j - \min_i v_i) \le 2\varepsilon \frac{\min_j v_j}{\min_k \hat{P}_{\ell,k}} \le \frac{2\varepsilon \min_j v_j}{a C_1 C_2 \overline{r}_\ell \min_j \overline{c}_j}, \tag{4.16}$$

where $C_1 = \min_i \{x_i / \overline{r}_i\}$ and $C_2 = \min_k \{y_k / \overline{c}_k\}$. Multiplying (4.16) by r_ℓ / N and employing the definitions of $\overline{r}_i$ and $\overline{c}_j$ (see (4.1)) gives

$$\frac{1}{N} \sum_{j=1}^{N} (v_j - \min_i v_i) \le \frac{2\varepsilon s \min_j v_j}{a C_1 C_2 N \min_j c_j} \le \frac{2\varepsilon s \max_j v_j}{a C_1 C_2 N \min_j c_j}. \tag{4.17}$$

Similarly, using the second inequality in (4.9) together with (4.13), one can verify by a derivation analogous to (4.14)–(4.17) that

$$\frac{1}{N}\sum_{j=1}^{N}(\max_j v_j - v_j) \le \frac{2\varepsilon s \max_j v_j}{aC_1C_2N\min_j c_j}. \tag{4.18}$$

Summing (4.17) and (4.18) gives

$$\max_j v_j - \min_j v_j \le \frac{4\varepsilon s \max_j v_j}{aC_1C_2N\min_j c_j}, \tag{4.19}$$

and by a symmetry argument (considering (4.19) in the setting where $\widetilde{A}$ is replaced with its transpose, so N is replaced with M, $\mathbf{c}$ is replaced with $\mathbf{r}$, and $\mathbf{v}$ is replaced with $\mathbf{u}$) we also have

$$\max_i u_i - \min_i u_i \le \frac{4\varepsilon s \max_i u_i}{aC_1C_2M\min_i r_i}. \tag{4.20}$$

Observe that $|\tau - v_j| \le \max_j v_j - \min_j v_j$ for any $\tau \in [\min_j v_j, \max_j v_j]$ and all $j \in [N]$. Taking τ as the geometric mean of $\max_j v_j$ and $\min_j v_j$, together with (4.19) gives

$$\left|\sqrt{\max_j v_j \min_j v_j} - v_j\right| \le \frac{4\varepsilon s \max_j v_j}{aC_1C_2N\min_j c_j}, \tag{4.21}$$

for all $j \in [N]$. Multiplying both hand sides of (4.21) by $\alpha^{-1} = \sqrt{\max_i u_i / \max_j v_j}$ we get

$$\left|\sqrt{\max_i u_i \min_j v_j} - \alpha^{-1}v_j\right| \le \frac{4\varepsilon s\sqrt{\max_j v_j \max_i u_i}}{aC_1C_2N\min_j c_j} \le \frac{4\varepsilon s\sqrt{b}}{a^2C_1^{3/2}C_2^{3/2}N\min_j c_j}, \tag{4.22}$$

where we also used $u_i \le \sqrt{b}/(aC_1)$ and $v_j \le \sqrt{b}/(aC_2)$ (by Lemma 4.1). According to (4.12), we have for all $\varepsilon \in (0,1)$ that

$$1 - \frac{\varepsilon}{1-\varepsilon} \le \frac{1}{1+\varepsilon} \le \sqrt{\frac{1}{1+\varepsilon}} \le \sqrt{\max_i u_i \min_j v_j} \le \sqrt{\frac{1}{1-\varepsilon}} \le \frac{1}{1-\varepsilon} = 1 + \frac{\varepsilon}{1-\varepsilon}, \tag{4.23}$$

which together with (4.21) implies that

$$\left|1 - \alpha^{-1}v_j\right| \le \frac{\varepsilon}{1-\varepsilon} + \frac{4\varepsilon s\sqrt{b}}{a^2C_1^{3/2}C_2^{3/2}N\min_j c_j} \le \varepsilon\left(2 + \frac{4s}{N\min_j c_j}\left(\frac{b}{a}\right)^{7/2}\right), \tag{4.24}$$

for all $i \in [M]$, $j \in [N]$, and $\varepsilon \in (0, 1/2]$, where we also used $C_1 = \min_i\{x_i/\overline{r}_i\} \ge \sqrt{a}/b$ and $C_2 = \min_j\{y_j/\overline{c}_j\} \ge \sqrt{a}/b$ (see (4.2)). By a derivation analogous to (4.21)– (4.24), one can verify that

$$|1 - \alpha u_i| \le \varepsilon\left(2 + \frac{4s}{M\min_i r_i}\left(\frac{b}{a}\right)^{7/2}\right), \tag{4.25}$$

for all $i \in [M]$, $j \in [N]$, and $\varepsilon \in (0, 1/2]$.

Overall, taking $\varepsilon = \delta/2$ and using the fact that $s = \|\mathbf{r}\|_1 \ge M\min_i r_i$ and $s = \|\mathbf{c}\|_1 \ge N\min_j c_j$, it follows that with probability at least (4.6), we have

$$\frac{|\alpha\overline{x}_i - x_i|}{x_i} \le \frac{\delta s}{M\min_i r_i}\left(1 + 2\left(\frac{b}{a}\right)^{7/2}\right), \qquad \frac{|\alpha^{-1}\overline{y}_j - y_j|}{y_j} \le \frac{\delta s}{N\min_j c_j}\left(1 + 2\left(\frac{b}{a}\right)^{7/2}\right), \tag{4.26}$$

for all $i \in [M]$, $j \in [N]$, and $\delta \in (0, 1]$. Denoting $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = (\alpha\overline{\mathbf{x}}, \alpha^{-1}\overline{\mathbf{y}})$ (which is a pair of scaling factors of $\widetilde{A}$), and using the fact that $b_{i,j} - a_{i,j} \le d$, we proved Theorem 2.1 with $C_p = \sqrt{2}\left(bd/a^2\right)$ and $C_e = 1 + 2\left(b/a\right)^{7/2}$.

4.2 Proof of Corollary 2.2

We begin by considering the case of $\rho_1\rho_2\sqrt{\log(\max\{M,N\})} \le 1/\sqrt{2C_p}$, where C_p is from Theorem 2.1. In this case, we apply Theorem 2.1 using $\delta = \rho_1\sqrt{2C_p\log(\max\{M,N\})}$, noting that $\delta \le 1$ as required since $\rho_2 \ge 1$. Therefore, there exists a pair of scaling factors $(\overline{\mathbf{x}},\overline{\mathbf{y}})$ of $\widetilde{A}$, such that with probability at least

$$\begin{aligned}&1 - 2M\exp\left(-\frac{\delta^2 s^2}{C_p\|\mathbf{c}\|_2^2}\right) - 2N\exp\left(-\frac{\delta^2 s^2}{C_p\|\mathbf{r}\|_2^2}\right)\\ &\ge 1 - 2M\exp\left(-2\log(\max\{M,N\})\right) - 2N\exp\left(-2\log(\max\{M,N\})\right) \ge 1 - \frac{4}{\min\{M,N\}},\end{aligned} \tag{4.27}$$

we have for all $i \in [M]$ and $j \in [N]$,

$$\begin{aligned}\frac{|\overline{x}_i - x_i|}{x_i} &\le \frac{C_e\delta s}{M\min_i r_i} \le C_c\rho_1\rho_2\sqrt{2C_p\log(\max\{M,N\})},\\ \frac{|\overline{y}_j - y_j|}{y_j} &\le \frac{C_e\delta s}{N\min_j c_j} \le C_c\rho_1\rho_2\sqrt{2C_p\log(\max\{M,N\})}.\end{aligned} \tag{4.28}$$

Next, we consider the case of $\rho_1\rho_2\sqrt{\log(\max\{M,N\})} > 1/\sqrt{2C_p}$. In this case, we apply Lemma 4.1 to $\widetilde{A}$, which states there exists a pair of scaling factors $(\overline{\mathbf{x}},\overline{\mathbf{y}})$ of $\widetilde{A}$, such that $\sqrt{a}/b \le \overline{x}_i/\overline{r}_i \le \sqrt{b}/a$ and $\sqrt{a}/b \le \overline{y}_j/\overline{c}_j \le \sqrt{b}/a$ for all $i \in [M]$ and $j \in [N]$. Using these inequalities with (4.2) asserts that $\overline{x}_i/x_i \le (b/a)^{3/2}$ and $\overline{y}_j/y_j \le (b/a)^{3/2}$, for all $i \in [M]$ and $j \in [N]$. Therefore,

$$\mathcal{E} \le \left(\frac{b}{a}\right)^{3/2} < \sqrt{2C_p}\left(\frac{b}{a}\right)^{3/2}\rho_1\rho_2\sqrt{\log(\max\{M,N\})}. \tag{4.29}$$

Combining (4.28) and (4.29) proves the required result, where we used the fact that $(\overline{\mathbf{x}},\overline{\mathbf{y}}) = (\alpha\widetilde{\mathbf{x}},\alpha^{-1}\widetilde{\mathbf{y}})$ for some $\alpha > 0$.

4.3 Proof of Corollary 2.3

Let us define $\eta = \widetilde{\mathbf{x}} - \mathbf{x}$, and $\xi = \widetilde{\mathbf{y}} - \mathbf{y}$. We can write

$$\begin{aligned}&\left\|D(\widetilde{\mathbf{x}})\widetilde{A}D(\widetilde{\mathbf{y}}) - D(\mathbf{x})AD(\mathbf{y})\right\|_2 \le \|D(x)(\widetilde{A}-A)D(y)\|_2\\ &+ \|D(\eta)\widetilde{A}D(y)\|_2 + \|D(x)\widetilde{A}D(\xi)\|_2 + \|D(\eta)\widetilde{A}D(\xi)\|_2.\end{aligned} \tag{4.30}$$

We now bound the summands in the right-hand side of (4.30) one by one. For the first summand in (4.30), applying Lemma 4.1 to A we have

$$\begin{aligned}&\|D(\mathbf{x})(\widetilde{A}-A)D(\mathbf{y})\|_2 \le \|D(\mathbf{x})\|_2\cdot\|\widetilde{A}-A\|_2\cdot\|D(\mathbf{y})\|_2\\ &\le \frac{b}{a^2}\frac{\max_i r_i}{\sqrt{s}}\|\widetilde{A}-A\|_2\frac{\max_j c_j}{\sqrt{s}} = \frac{b\rho_3}{a^2\sqrt{MN}}\|\widetilde{A}-A\|_2.\end{aligned} \tag{4.31}$$

Since $a \le \widetilde{A}_{i,j} \le b$, then also $a \le A_{i,j} \le b$, which implies that $a - b \le \widetilde{A}_{i,j} - A_{i,j} \le b - a$. Hence, $\{\widetilde{A}_{i,j} - A_{i,j}\}_{i,j}$ are independent, have mean zero, and are bounded (and therefore sub-Gaussian). Applying Theorem 4.4.5 in [32] with $t = \sqrt{\log(\max\{M,N\})}$ gives that

$$\|\widetilde{A}-A\|_2 \lesssim \sqrt{N} + \sqrt{M} + \sqrt{\log(\max\{M,N\})}, \tag{4.32}$$

with probability at least $1 - 2/\max\{M,N\}$. Combining (4.32) with (4.31) asserts that

$$\|D(\mathbf{x})(\widetilde{A}-A)D(\mathbf{y})\|_2 \lesssim \frac{\rho_3}{\sqrt{MN}}(\sqrt{N} + \sqrt{M} + \sqrt{\log(\max\{M,N\})}), \tag{4.33}$$

with probability at least $1-2/\max\{M,N\}$. Observe that $s=\|\mathbf{r}\|_1 \le \sqrt{M}\|\mathbf{r}\|_2$, and $s=\|\mathbf{c}\|_1 \le \sqrt{N}\|\mathbf{c}\|_2$. Therefore, $\rho_1 = \max\{\|\mathbf{r}\|_2/s, \|\mathbf{c}\|_2/s\} \ge \max\{1/\sqrt{M}, 1/\sqrt{N}\}$. Using this fact together with $\rho_2 \ge 1$, it follows that

$$\begin{aligned}
&\rho_1\rho_2\rho_3\sqrt{\log(\max\{M,N\})} \ge \rho_3 \max\{\frac{1}{\sqrt{M}}, \frac{1}{\sqrt{N}}\}\sqrt{\log(\max\{M,N\})} \\
&= \frac{\rho_3}{\sqrt{MN}}\max\{\sqrt{M},\sqrt{N}\}\sqrt{\log(\max\{M,N\})} \\
&\ge \frac{\rho_3}{8\sqrt{MN}}(\sqrt{M}+\sqrt{N}+\sqrt{\log(\max\{M,N\})}). \qquad (4.34)
\end{aligned}$$

Applying the above inequality to (4.33) we obtain that

$$\|D(\mathbf{x})(\widetilde{A}-A)D(\mathbf{y})\|_2 \lesssim \rho_1\rho_2\rho_3\sqrt{\log(\max\{M,N\})}, \qquad (4.35)$$

with probability at least $1-2/\max\{M,N\}$. Continuing, for the second summand in (4.30),

$$\begin{aligned}
&\|D(\eta)\widetilde{A}D(y)\|_2 \le \|D(\eta)\|_2\cdot\|\widetilde{A}\|_2\cdot\|D(y)\|_2 \le \|D(\eta)\|_2\cdot\|\widetilde{A}\|_F\frac{\max_j c_j}{\sqrt{s}} \\
&\le \max_i|\eta_i|\frac{b\sqrt{MN}\max_j c_j}{\sqrt{s}} \le \max_i\frac{|\widetilde{x}_i-x_i|}{x_i}\cdot\max_i x_i\cdot\frac{b\sqrt{MN}\max_j c_j}{\sqrt{s}} \\
&\le \max_i\frac{|\widetilde{x}_i-x_i|}{x_i}\cdot\frac{b^{3/2}\sqrt{MN}\max_i r_i\cdot\max_j c_j}{as} \lesssim \rho_1\rho_2\rho_3\sqrt{\log(\max\{M,N\})}, \qquad (4.36)
\end{aligned}$$

with probability at least $1-4/\min\{M,N\}$, where we used Lemma 4.1 and Corollary 2.2. Analogously, it is easy to verify that the third summand in (4.30) admits the same probabilistic bound as the second summand. For the fourth summand in (4.30),

$$\begin{aligned}
&\|D(\eta)\widetilde{A}D(\xi)\|_2 \le \|D(\eta)\|_2\cdot\|\widetilde{A}\|_2\cdot\|D(\xi)\|_2 \le \max_i|\eta_i|\cdot\|\widetilde{A}\|_F\cdot\max_j|\xi_j| \\
&\le \max_i\frac{|\widetilde{x}_i-x_i|}{x_i}\cdot\max_i x_i\cdot\sqrt{MN}\cdot\max_j\frac{|\widetilde{y}_j-y_j|}{y_j}\cdot\max_j y_j \\
&\le \max_i\frac{|\widetilde{x}_i-x_i|}{x_i}\cdot\max_j\frac{|\widetilde{y}_j-y_j|}{y_j}\cdot\frac{b\sqrt{M}\max_i r_i\cdot\sqrt{N}\max_j c_j}{a^2 s} \\
&\lesssim \left(\rho_1\rho_2\sqrt{\log(\max\{M,N\})}\right)^2\rho_3, \qquad (4.37)
\end{aligned}$$

with probability at least $1-4/\min\{M,N\}$, where we again used Lemma 4.1 and Corollary 2.2. We now consider the case where $\rho_1\rho_2\sqrt{\log(\max\{M,N\})} \le 1$. In this case, using all of the above and applying the union bound (on the events (4.32) and (2.4)) gives that

$$\left\|D(\widetilde{\mathbf{x}})\widetilde{A}D(\widetilde{\mathbf{y}}) - D(\mathbf{x})AD(\mathbf{y})\right\|_2 \lesssim \rho_1\rho_2\rho_3\sqrt{\log(\max\{M,N\})}, \qquad (4.38)$$

with probability at least $1-2/\max\{M,N\}-4/\min\{M,N\} \ge 1-6/\min\{M,N\}$. Lastly, we consider the case where $\rho_1\rho_2\sqrt{\log(\max\{M,N\})} > 1$. In this case, we write

$$\left\|D(\widetilde{\mathbf{x}})\widetilde{A}D(\widetilde{\mathbf{y}}) - D(\mathbf{x})AD(\mathbf{y})\right\|_2 \le \left\|D(\widetilde{\mathbf{x}})\widetilde{A}D(\widetilde{\mathbf{y}})\right\|_F + \|D(\mathbf{x})AD(\mathbf{y})\|_F. \qquad (4.39)$$

By applying Lemma 4.1 to A and $\widetilde{A}$, we have that

$$\widetilde{x}_i\widetilde{A}\widetilde{y}_j \le \left(\frac{b}{a}\right)^2\frac{r_i c_j}{s}, \qquad \text{and} \qquad x_i A y_j \le \left(\frac{b}{a}\right)^2\frac{r_i c_j}{s}, \qquad (4.40)$$

for all $i \in [M]$ and $j \in [N]$. Therefore, Combining (4.40) with (4.39) implies

$$\begin{aligned}\left\|D(\widetilde{\mathbf{x}})\widetilde{A}D(\widetilde{\mathbf{y}}) - D(\mathbf{x})AD(\mathbf{y})\right\|_2 &\leq 2\left(\frac{b}{a}\right)^2 \frac{\|\mathbf{r}\|_2 \cdot \|\mathbf{c}\|_2}{s} \\ &\leq 2\left(\frac{b}{a}\right)^2 \frac{\sqrt{M}\max_i r_i \cdot \sqrt{N}\max_j c_j}{s} < 2\left(\frac{b}{a}\right)^2 \rho_1\rho_2\rho_3\sqrt{\log(\max\{M,N\})}, \qquad (4.41)\end{aligned}$$

where we used $1 < \rho_1\rho_2\sqrt{\log(\max\{M,N\})}$ in the last inequality.

References

[1] Jason M Altschuler, Jonathan Niles-Weed, and Austin J Stromme, *Asymptotics for semidiscrete entropic optimal transport*, SIAM Journal on Mathematical Analysis **54** (2022), no. 2, 1718–1741. MR4393198

[2] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis, *Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications*, Electronic Journal of Statistics **13** (2019), no. 2, 5120–5150. MR4041704

[3] Richard A Brualdi, *The dad theorem for arbitrary row sums*, Proceedings of the American Mathematical Society **45** (1974), no. 2, 189–194. MR0354737

[4] Richard A Brualdi, Seymour V Parter, and Hans Schneider, *The diagonal equivalence of a nonnegative matrix to a stochastic matrix*, Journal of Mathematical Analysis and Applications **16** (1966), no. 1, 31–50. MR0206019

[5] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer, *Convergence of entropic schemes for optimal transport and gradient flows*, SIAM Journal on Mathematical Analysis **49** (2017), no. 2, 1385–1418. MR3635459

[6] Minsu Cho, Jungmin Lee, and Kyoung Mu Lee, *Reweighted random walks for graph matching*, European conference on Computer vision, Springer, 2010, pp. 492–505.

[7] Timothee Cour, Praveen Srinivasan, and Jianbo Shi, *Balanced graph matching*, Advances in Neural Information Processing Systems **19** (2006), 313–320.

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

[9] Noureddine El Karoui, *On information plus noise kernel random matrices*, The Annals of Statistics **38** (2010), no. 5, 3191–3216. MR2722468

[10] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio, *Learning with a wasserstein loss*, Advances in neural information processing systems, 2015, pp. 2053–2061.

[11] Promit Ghosal, Marcel Nutz, and Espen Bernton, *Stability of entropic optimal transport and schrödinger bridges*, Journal of Functional Analysis **283** (2022), no. 9, 109622. MR4460341

[12] Wassily Hoeffding, *Probability inequalities for sums of bounded random variables*, The Collected Works of Wassily Hoeffding, Springer, 1994, pp. 409–426. MR1307621

[13] Roger A Horn and Charles R Johnson, *Matrix analysis*, Cambridge university press, 2012. MR2978290

[14] Martin Idel, *A review of matrix scaling and sinkhorn's normal form for matrices and positive maps*, arXiv preprint arXiv:1609.06349 (2016). MR3314325

[15] Marcel Klatt, Carla Tameling, and Axel Munk, *Empirical regularized optimal transport: Statistical theory and applications*, SIAM Journal on Mathematics of Data Science **2** (2020), no. 2, 419–443. MR4105566

[16] Philip A Knight, *The sinkhorn–knopp algorithm: convergence and applications*, SIAM Journal on Matrix Analysis and Applications **30** (2008), no. 1, 261–275. MR2399579

[17] B Lamond and Neil F Stewart, *Bregman's balancing method*, Transportation Research Part B: Methodological **15** (1981), no. 4, 239–248. MR0624430

[18] Boris Landa and Xiuyuan Cheng, *Robust inference of manifold density and geometry by doubly stochastic scaling*, arXiv preprint arXiv:2209.08004 (2022).

[19] Boris Landa, Ronald R Coifman, and Yuval Kluger, *Doubly stochastic normalization of the gaussian kernel is robust to heteroskedastic noise*, SIAM journal on mathematics of data science **3** (2021), no. 1, 388–413. MR4234155

[20] Derek Lim, René Vidal, and Benjamin D Haeffele, *Doubly stochastic subspace clustering*, arXiv preprint arXiv:2011.14859 (2020).

[21] Albert W Marshall and Ingram Olkin, *Scaling of matrices to achieve specified row and column sums*, Numerische Mathematik **12** (1968), no. 1, 83–90. MR0238875

[22] Nicholas F Marshall and Ronald R Coifman, *Manifold learning with bi-stochastic kernels*, IMA Journal of Applied Mathematics **84** (2019), no. 3, 455–482. MR3954914

[23] Gonzalo Mena and Jonathan Niles-Weed, *Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem*, Advances in Neural Information Processing Systems **32** (2019).

[24] Peyman Milanfar, *Symmetrizing smoothing filters*, SIAM Journal on Imaging Sciences **6** (2013), no. 1, 263–284. MR3032954

[25] Marcel Nutz and Johannes Wiesel, *Entropic optimal transport: Convergence of potentials*, Probability Theory and Related Fields (2021), 1–24. MR4498514

[26] Gabriel Peyré and Marco Cuturi, *Computational optimal transport: With applications to data science*, Foundations and Trends® in Machine Learning **11** (2019), no. 5-6, 355–607.

[27] Michael H Schneider and Stavros A Zenios, *A comparative study of algorithms for matrix balancing*, Operations research **38** (1990), no. 3, 439–455.

[28] Richard Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, The annals of mathematical statistics **35** (1964), no. 2, 876–879. MR0161868

[29] Richard Sinkhorn, *Diagonal equivalence to matrices with prescribed row and column sums*, The American Mathematical Monthly **74** (1967), no. 4, 402–405. MR0210730

[30] Richard Sinkhorn and Paul Knopp, *Concerning nonnegative matrices and doubly stochastic matrices*, Pacific Journal of Mathematics **21** (1967), no. 2, 343–348. MR0210731

[31] Paul B Slater, *Measuring migration fields of us counties*, Geographical Analysis **16** (1984), no. 1, 65–73.

[32] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018. MR3837109

[33] Caroline L Wormell and Sebastian Reich, *Spectral convergence of diffusion maps: Improved error bounds and an alternative normalization*, SIAM Journal on Numerical Analysis **59** (2021), no. 3, 1687–1734. MR4273695

[34] Ron Zass and Amnon Shashua, *Doubly stochastic normalization for spectral clustering*, Advances in neural information processing systems **19** (2006), 1569–1576.

Acknowledgments. The author would like to thank Thomas Zhang, Yuval Kluger, and Dan Kluger for their useful comments and suggestions.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4], ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[2]EJMS: Electronic Journal Management System: `https://vtex.lt/services/ejms-peer-review/`
[3]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[4]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[5]Project Euclid: `https://projecteuclid.org/`
[6]IMS Open Access Fund: `https://imstat.org/shop/donation/`