Mathematical Programming (2021) 188:395–407
https://doi.org/10.1007/s10107-020-01503-3

SHORT COMMUNICATION

Series A

Better and simpler error analysis of the Sinkhorn–Knopp algorithm for matrix scaling

Deeparnab Chakrabarty[1] · Sanjeev Khanna[2]

Received: 17 February 2018 / Accepted: 2 April 2020 / Published online: 15 April 2020

Abstract

Given a non-negative $n \times m$ real matrix A, the *matrix scaling* problem is to determine if it is possible to scale the rows and columns so that each row and each column sums to a specified positive target values. The Sinkhorn–Knopp algorithm is a simple and classic procedure which alternately scales all rows and all columns to meet these targets. The focus of this paper is the worst-case theoretical analysis of this algorithm. We present an elementary convergence analysis for this algorithm that improves upon the previous best bound. In a nutshell, our approach is to show (i) a simple bound on the number of iterations needed so that the KL-divergence between the current row-sums and the target row-sums drops below a specified threshold δ, and (ii) then show that for a suitable choice of δ, whenever KL-divergence is below δ, then the ℓ_1-error or the ℓ_2-error is below ε. The well-known Pinsker's inequality immediately allows us to translate a bound on the KL divergence to a bound on ℓ_1-error. To bound the ℓ_2-error in terms of the KL-divergence, we establish a new inequality, referred to as (**KL vs** ℓ_1/ℓ_2). This inequality is a strengthening of Pinsker's inequality and may be of independent interest.

Keywords Matrix scaling · Alternate minimization · KL divergence · Matchings

Mathematics Subject Classification 68Q25 · 68W40

The second author's work was supported in part by the National Science Foundation Grants CCF-1552909 and CCF-1617851. A preliminary version of this work was presented in the 1st Symposium on Simplicity in Algorithms, January 10th, 2018, co-located with SODA 2018 in New Orleans.

✉ Deeparnab Chakrabarty
deeparnab@dartmouth.edu

Sanjeev Khanna
sanjeev@cis.upenn.edu

[1] Department of Computer Science, Dartmouth College, Hanover, USA

[2] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA

1 Introduction

In the matrix scaling problem one is given an $n \times m$ non-negative, non-zero real matrix A, and positive vectors $\mathbf{r} \in \mathbb{R}^n_{>0}$ and $\mathbf{c} \in \mathbb{R}^m_{>0}$ with the same ℓ_1 norm $\sum_{i=1}^n \mathbf{r}_i = \sum_{j=1}^m \mathbf{c}_j = h$. The objective is to determine if there exist diagonal matrices $R \in \mathbb{R}^{n\times n}$ and $S \in \mathbb{R}^{m\times m}$ such that the ith row of the matrix RAS sums to $\mathbf{r}_i$ for all $1 \le i \le n$ *and* the jth column of RAS sums to $\mathbf{c}_j$ for all $1 \le j \le m$. Of special importance is the case when $n = m$ and $\mathbf{r} \equiv \mathbf{c} \equiv \mathbf{1}_n$, the n-dimensional all-ones vector—the $(\mathbf{1}, \mathbf{1})$-matrix scaling problem wishes to scale the rows and columns of A to make it doubly stochastic. This problem arises in many different areas ranging from transportation planning [13,27] to quantum mechanics [1,35]; we refer the reader to a recent comprehensive survey by Idel [16] for more examples.

One of the most natural algorithms for the matrix scaling problem is the following Sinkhorn–Knopp algorithm [36,37], which is known by many names including the RAS method [5] and the Iterative Proportional Fitting Procedure [33]. The algorithm starts off by multiplicatively scaling all the columns by the columns-sum times $\mathbf{c}_j$ to get a matrix $A^{(0)}$ with column-sums $\mathbf{c}$. Subsequently, for $t \ge 0$, it obtains the $B^{(t)}$ by scaling each row of $A^{(t)}$ by the respective row-sum times $\mathbf{r}_i$, and obtain $A^{(t+1)}$ by scaling each column of $B^{(t)}$ by the respective column sums time $\mathbf{c}_j$. More precisely,

$$A^{(0)}_{ij} := \frac{A_{ij}}{\sum_{i=1}^n A_{ij}} \cdot \mathbf{c}_j; \quad \forall t \ge 0, \quad B^{(t)}_{ij} := \frac{A^{(t)}_{ij}}{\sum_{j=1}^m A^{(t)}_{ij}} \cdot \mathbf{r}_i, \quad A^{(t+1)}_{ij} := \frac{B^{(t)}_{ij}}{\sum_{i=1}^n B^{(t)}_{ij}} \cdot \mathbf{c}_j \tag{SK}$$

The above algorithm is simple and easy to implement and each iteration takes $O(\mathsf{nnz}(A))$, the number of non-zero entries of A. Furthermore, it has been known for almost five decades [14,36–38] that if A is $(\mathbf{r}, \mathbf{c})$-scalable then the above algorithm asymptotically[1] converges to a right solution. More precisely, given $\varepsilon > 0$, there is some finite t by which one obtains a matrix which is "ε-close to having row- and column-sums $\mathbf{r}$ and $\mathbf{c}$". Since the rate depends on how we measure "ε-closeness", we look at two natural error definitions. For any t, let $\mathbf{r}^{(t)} := A^{(t)}\mathbf{1}_m$ denote the vector of row-sums of $A^{(t)}$. Similarly, we define $\mathbf{c}^{(t)} := {B^{(t)}}^\top \mathbf{1}_n$ to be the vector of the column-sums of $B^{(t)}$. Note that $\sum_{i=1}^n \mathbf{r}^{(t)}_i = \sum_{j=1}^m \mathbf{c}^{(t)}_j = h$ for all t. The error of the matrix A_t (the error of matrix B_t similarly defined) is

$$\ell_1\text{-error}: \ \mathsf{error}_1(A_t) := ||\mathbf{r}^{(t)} - \mathbf{r}||_1 \quad \ell_2\text{-error}: \ \mathsf{error}_2(A_t) := ||\mathbf{r}^{(t)} - \mathbf{r}||_2$$

[1] Computationally, this asymptotic viewpoint is unavoidable in the sense that there are simple examples for which the unique matrix scaling matrices need to have irrational entries. For instance, consider the following example from Rothblum and Schneider [32]. The matrix is $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ with $\mathbf{r} \equiv \mathbf{c} \equiv [1, 1]^\top$. The unique (up to scaling) R and S matrices are $\begin{bmatrix} (\sqrt{2}+1)^{-1} & 0 \\ 0 & (\sqrt{2}+2)^{-1} \end{bmatrix}$ and $\begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{bmatrix}$, respectively, giving $RAS = \begin{bmatrix} 2-\sqrt{2} & \sqrt{2}-1 \\ \sqrt{2}-1 & 2-\sqrt{2} \end{bmatrix}$.

The main objective of this note is to understand the number of iterations this simple algorithm takes to attain an error of $\leq \varepsilon$ in the *worst case*. We give simple convergence analysis for both error norms which improves upon previously known results.

Theorem 1 *Given a matrix $A \in \mathbb{R}_{\geq 0}^{n\times m}$ which is $(\mathbf{r}, \mathbf{c})$-scalable, and any $\varepsilon > 0$, the Sinkhorn–Knopp algorithm*

1. *in time $t = O\left(\frac{h^2 \ln(\Delta/\nu)}{\varepsilon^2}\right)$ returns a matrix A_t or B_t with ℓ_1-error $\leq \varepsilon$.*
2. *in time $t = O\left(h \ln(\Delta/\nu) \cdot \left(\frac{1}{\varepsilon} + \frac{\rho}{\varepsilon^2}\right)\right)$ returns a matrix A_t or B_t with ℓ_2-error $\leq \varepsilon$.*

Here $h = \sum_{i=1}^{n} \mathbf{r}_i = \sum_{j=1}^{m} \mathbf{c}_j$, $\rho = \max(\max_i \mathbf{r}_i, \max_j \mathbf{c}_j)$, $\nu = \frac{\min_{i,j:A_{ij}>0} A_{ij}}{\max_{i,j} A_{ij}}$, and $\Delta = \max_j |\{i : A_{ij} > 0\}|$ is the maximum number of non-zeros in any column of A.

For the special case of $n = m$ and $\mathbf{r} \equiv \mathbf{c} \equiv \mathbf{1}_n$, we get the following as a corollary.

Corollary 1 *Given a matrix $A \in \mathbb{Z}_{\geq 0}^{n\times n}$ which is $(\mathbf{1}, \mathbf{1})$-scalable, and any $\varepsilon > 0$, the Sinkhorn–Knopp algorithm*

1. *in time $t = O\left(\frac{n^2 \ln(\Delta/\nu)}{\varepsilon^2}\right)$ returns a matrix A_t or B_t with ℓ_1-error $\leq \varepsilon$.*
2. *in time $t = O\left(n \ln(\Delta/\nu) \cdot \left(\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2}\right)\right)$ returns a matrix A_t or B_t with ℓ_2-error $\leq \varepsilon$.*

Here $\Delta = \max_j |\{i : A_{ij} > 0\}|$ is the maximum number of non-zeros in any column of A.

Remark 1 To our knowledge, the ℓ_1-error hasn't been explicitly studied[2] in the literature, although for small $\varepsilon \in (0, 1)$ the same can be deduced from previous papers on matrix scaling [15,17,20,22]. One of our main motivations to look at ℓ_1-error arose from the connections to perfect matchings in bipartite graphs as observed by Linial, Samorodnitsky and Wigderson [22]. For the ℓ_2 error, which is the better studied notion in the matrix scaling literature, the best analysis is due to Kalantari et al. [19,20]. They give a $\tilde{O}(\rho h^2/\varepsilon^2)$ upper bound on the number of iterations for the general problem, and for the special case when $m = n$ and the square matrix has positive permanent (see [19]), they give a $\tilde{O}(\rho(h^2 - nh + n)/\varepsilon^2)$ upper bound. Thus, for $(\mathbf{1}, \mathbf{1})$-scaling, they get the same result as in Corollary 1. We get a quadratic improvement on h in the general case, and we think our proof is more explicit and simpler.

Remark 2 Both parts of Theorem 1 and Corollary 1 are interesting in certain regimes of error. When the error ε is "small" (say, ≤ 1) so that $1/\varepsilon^2 \geq 1/\varepsilon$, then statement 2 of Corollary 1 implies statement 1 by Cauchy-Schwarz.

However, this breaks down when ε is "large" (say $\varepsilon = \delta n$ for some constant $\delta > 0$). In that case, statement 1 implies that in $O(\ln n/\delta^2)$ iterations, the ℓ_1-error is $\leq \delta n$, but Statement 2 implies that in $O(\ln n/\delta^2)$ iterations, only the ℓ_2-error is $\leq \delta n$ (the ℓ_1-error could be large). This "large ℓ_1-error regime" is of particular interest for an application to approximate matchings in bipartite graphs discussed below.

[2] After the first version of this paper was made public, we were pointed to concurrent work by Altschuler, Weed and Rigollet [4] studying the ℓ_1-error and obtaining the same result as part 1 of Theorem 1.

Applications to Parallel Algorithms for Bipartite Perfect Matching. As a corollary, we get the following application, first pointed by Linial et al. [22], to the existence of perfect matchings in bipartite graphs. Let A be the adjacency matrix of a bipartite graph $G = (L \cup R, E)$ with $A_{ij} = 1$ iff $(i, j) \in E$. If G has a perfect matching, then clearly there is a doubly stochastic matrix X in the support of A. This suggests the algorithm of running the Sinkhorn–Knopp algorithm to A, and the following claim suggests when to stop. Note that each iteration can be run in $O(1)$ parallel time with m-processors where m is the number of edges.

Lemma 1 *If we find a column (or row) stochastic matrix Y in the support of A such that $\mathsf{error}_1(Y) \le n\varepsilon$, then G has a matching of size at least $n(1-\varepsilon)$.*

Proof Suppose Y is column stochastic. Given $S \subseteq L$, consider $\sum_{i\in S, j\in \Gamma S} Y_{ij} = |S| + \sum_{i\in S}\left(\sum_{j=1}^{n} Y_{ij} - 1\right) \ge |S| - \sum_{i=1}^{n}\left|\sum_{j=1}^{n} Y_{ij} - 1\right| \ge |S| - \mathsf{error}_1(Y) \ge |S| - n\varepsilon$. On the other hand, $\sum_{i\in S, j\in\Gamma S} Y_{ij} \le \sum_{j\in\Gamma S}\sum_{i=1}^{n} Y_{ij} = |\Gamma S|$. Therefore, for every $S \subseteq L$, $|\Gamma S| \ge |S| - n\varepsilon$. The claim follows by approximate Hall's theorem. □

Corollary 2 (Fast Parallel Approximate Matchings) *Given a bipartite graph G of max-degree Δ and an $\varepsilon \in (0, 1)$, $O(\ln \Delta/\varepsilon^2)$-iterations of Sinkhorn–Knopp algorithm suffice to distinguish between the case when G has a perfect matching and the case when the largest matching in G has size at most $n(1-\varepsilon)$.*

Thus the approximate perfect matching problem in bipartite graphs is in NC for ε as small as polylogarithmic in n. This is not a new result and can indeed be obtained from the works on parallel algorithms for packing-covering LPs [3,23,25,40], but the Sinkhorn–Knopp algorithm is arguably simpler.

Remark 3 At present, we do not know if the dependence on h and ε in Theorem 1 is tight. The best lower bound examples we know take $\Omega(h/\varepsilon)$-rounds to achieve ε ℓ_1-error. We think that the result asserted in our theorem is tight, but we currently cannot prove this. Given that Sinkhorn–Knopp is such a fundamental algorithm, closing this gap is an important and interesting question.

1.1 Perspective

As mentioned above, the matrix scaling problem and in particular the Sinkhorn–Knopp algorithm has been extensively studied over the past 50 years. We refer the reader to Idel's survey [16] and the references within for a broader perspective. Below, we mention works which we feel are most relevant to our paper.

We have already discussed the previously best known, in their dependence on h, analysis for the Sinkhorn–Knopp algorithm in Remark 1. For the special case of *strictly positive* matrices, better rates are known. Kalantari and Khachiyan [17] showed that for positive matrices and the $(\mathbf{1}, \mathbf{1})$-scaling problem, the Sinkhorn–Knopp algorithm obtains ℓ_2 error $\le \varepsilon$ in $O(\sqrt{n}\ln(1/\nu)/\varepsilon)$-iterations; this result was extended to the general matrix scaling problem by Kalantari et al. [20]. In a different track, Franklin and Lorenz [14] show that in fact the dependence on ε can be made logarithmic, and thus the algorithm has "linear convergence", however their analysis[3] has a polynomial

[3] [14] never make the base of the logarithm explicit, but their proof shows it can be as large as $1 - 1/\nu^2$.

dependence of $(1/\nu)$. These results were improved in subsequent works of Soules [38] and Knight [21]; we refer the reader to Idel's survey [16] for more details. All these results use the positivity crucially.

As mentioned in the last remark in the previous subsection, the Sinkhorn–Knopp algorithm has polynomial dependence on the error parameter and therefore is not a polynomial time approximation in the complexity theoretic sense. We conclude by briefly describing bounds obtained by *other* algorithms for the matrix scaling problem whose dependence on ε is logarithmic rather than polynomial. Kalantari and Khachiyan [18] describe a method based on the ellipsoid algorithm which runs in time $O(n^4 \ln(n/\varepsilon)\ln(1/\nu))$. The first strongly polynomial time approximation scheme (with no dependence on ν) was due to Linial, Samoridnitsky, and Wigderson [22] who gave a $\tilde{O}(n^7 \ln(h/\varepsilon))$ time algorithm. Rote and Zachariasen [31] improved this to a $O(n^4 \ln(h/\varepsilon))$ time algorithm via reduction to network flow prolems. Much more recently, two independent works obtain vastly improved running times for matrix scaling. Cohen et al. [10] give $\tilde{O}(\mathsf{nnz}(A)^{3/2})$ time algorithm, while Allen-Zhu et al. [2] give a $\tilde{O}(n^{7/3}+\mathsf{nnz}(A)\cdot(n+n^{1/3}h^{1/2}))$ time algorithm; the tildes in both the above running times hide the logarithmic dependence on ε and ν. To compare, recall that Theorem 1 shows that the Sinkhorn–Knopp algorithm runs in time $O(\mathsf{nnz}(A)h^2/\varepsilon^2)$ time (for ℓ_1-error).

2 Entropy minimization viewpoint of the Sinkhorn–Knopp algorithm

There have been many approaches (see Idel [16], Section 3 for a discussion) towards analyzing the Sinkhorn–Knopp algorithm including convex optimization and log-barrier methods [6,17,20,24], non-linear Perron-Frobenius theory [9,14,17,26,38], topological methods [7,29], connections to the permanent [19,22], and the entropy minimization method [8,11,12,15] which is what we use for our analysis.

We briefly describe the entropy minimization viewpoint. Given two non-negative matrices M and N let us define the *Kullback-Leibler* divergence[4] between M and N as follows

$$\mathbf{D}(M, N) := \frac{1}{h} \sum_{1\le i\le n} \sum_{1\le j\le m} M_{ij} \ln\left(\frac{M_{ij}}{N_{ij}}\right) \tag{1}$$

with the convention that the summand is zero if both M_{ij} and N_{ij} are 0, and is ∞ if $M_{ij} > 0$ and $N_{ij} = 0$. Let Φ_r be the set of $n \times m$ matrices whose row-sums are $\mathbf{r}$ and let Φ_c be the set of $n \times m$ matrices whose column sums are $\mathbf{c}$. Given matrix A suppose we wish to find the matrix $A^* = \arg\min_{B\in\Phi_r\cap\Phi_c} \mathbf{D}(B, A)$. One algorithm for this is to use the method of alternate projections with respect to the KL-divergence [8] (also known as I-projections [11]) which alternately finds the matrices in Φ_r and Φ_c closest in the KL-divergence sense to the current matrix at hand, and then sets the minimizer to be the current matrix. It is not too hard to see (see Idel [16], Observation 3.17 for a proof) that the above alternate projection algorithm is precisely the Sinkhorn–Knopp algorithm. Therefore, at least in this sense, the right metric to measure the distance

[4] The KL-divergence is normally stated between two distributions and doesn't have the $1/h$ factor. Also the logarithms are usually base 2.

to optimality is not the ℓ_1 or the ℓ_2 error as described in the previous section, but the rather the KL-divergence between the normalized vectors as described below.

Let $\pi_{\mathbf{r}}^{(t)} := \mathbf{r}^{(t)}/h$ be the n-dimensional probability vector whose ith entry is $\mathbf{r}_i^{(t)}/h$; similarly define the m-dimensional vector $\pi_{\mathbf{c}}^{(t)}$. Let $\pi_{\mathbf{r}}$ denote the n-dimensional probability vector with the ith entry being $\mathbf{r}_i/h$; similarly define $\pi_{\mathbf{c}}$. Recall that the KL-divergence between two probability distributions p, q is defined as $\mathbf{D}_{KL}(p||q) := \sum_{i=1}^{n} p_i \ln(q_i/p_i)$. The following theorem gives the convergence time for the KL-divergence.

Theorem 2 *If the matrix $A \in \mathbb{R}_{\geq 0}^{n\times m}$ is $(\mathbf{r}, \mathbf{c})$-scalable, then for any $\delta > 0$ there is a $t \leq T = \lceil \left(\frac{\ln(1+2\Delta/\nu)}{\delta}\right) \rceil$ with either $\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) \leq \delta$ or $\mathbf{D}_{KL}(\pi_{\mathbf{c}}||\pi_{\mathbf{c}}^{(t)}) \leq \delta$. Recall, $\rho = \max(\max_i \mathbf{r}_i, \max_j \mathbf{c}_j)$, $\nu = \frac{\min_{i,j:A_{ij}>0} A_{ij}}{\max_{i,j} A_{ij}}$, and $\Delta = \max_j |\{i : A_{ij} > 0\}|$ is the maximum number of non-zeros in any column of A.*

Proof Let $Z := RAS$ be a matrix with row-sums $\mathbf{r}$ and column-sums $\mathbf{c}$ for diagonal matrices R, S. Recall $A^{(0)}$ is the matrix obtained by column-scaling A [see (SK)]. Note that the minimum non-zero entry in column j of $A_j^{(0)}$ is $\geq \frac{\nu \mathbf{c}_j}{\Delta}$.

Lemma 2 *$\mathbf{D}(Z, A^{(0)}) \leq \ln(1 + 2\Delta/\nu)$ and $\mathbf{D}(Z, A^t) \geq 0$ for all t.*

Proof By definition,

$$\mathbf{D}(Z, A^{(t)}) = \frac{1}{h}\sum_{j=1}^{m}\sum_{i=1}^{n} Z_{ij} \ln\left(\frac{Z_{ij}}{A_{ij}^{(t)}}\right) = \frac{1}{h}\sum_{j=1}^{m} \mathbf{c}_j \sum_{i=1}^{n} \frac{Z_{ij}}{\mathbf{c}_j} \ln\left(\frac{Z_{ij}}{A_{ij}^{(t)}}\right)$$

For a fixed j, the vectors $\left(\frac{Z_{1j}}{\mathbf{c}_j}, \frac{Z_{2j}}{\mathbf{c}_j}, \ldots, \frac{Z_{nj}}{\mathbf{c}_j}\right)$ and $\left(\frac{A_{1j}^{(t)}}{\mathbf{c}_j}, \frac{A_{2j}^{(t)}}{\mathbf{c}_j}, \ldots, \frac{A_{nj}^{(t)}}{\mathbf{c}_j}\right)$ are probability vectors, and therefore the above is a sum of $\mathbf{c}_j$-weighted KL-divergences which is always non-negative. For the upper bound, one can use the fact (Inequality 27, [34]) that for any two distributions p and q, $D(p||q) \leq \ln(1+\frac{||p-q||_2^2}{q_{\min}}) \leq \ln(1+\frac{2}{q_{\min}})$ where $q_{\min}$ is the smallest non-zero entry of q. For our purpose, we note that the minimum non-zero probability, $q_{\min}$, of the $\frac{A_j^{(0)}}{\mathbf{c}_j}$ distribution is $\geq \nu/\Delta$. Therefore, the second summand is at most $\ln(1+2\Delta/\nu)$ giving us $D(Z, A^{(0)}) \leq \frac{1}{h}\sum_{j=1}^{m} \mathbf{c}_j \cdot \ln(1+2\Delta/\nu) = \ln(1+2\Delta/\nu)$. This ends the proof of Lemma 2. □.

Lemma 3

$$\mathbf{D}(Z, A^{(t)}) - \mathbf{D}(Z, B^{(t)}) = \mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) \quad \text{and} \quad \mathbf{D}(Z, B^{(t)}) - \mathbf{D}(Z, A^{(t+1)}) = \mathbf{D}_{KL}(\pi_{\mathbf{c}}||\pi_{\mathbf{c}}^{(t)})$$

Proof The LHS of the first equality is simply

$$\begin{aligned}\frac{1}{h}\sum_{j=1}^{m}\sum_{i=1}^{n} Z_{ij}\ln\left(\frac{B_{ij}^{(t)}}{A_{ij}^{(t)}}\right) &= \frac{1}{h}\sum_{j=1}^{m}\sum_{i=1}^{n} Z_{ij}\ln\left(\frac{\mathbf{r}_i}{\mathbf{r}_i^{(t)}}\right)\\ &= \frac{1}{h}\sum_{i=1}^{n}\ln\left(\frac{\mathbf{r}_i}{\mathbf{r}_i^{(t)}}\right)\sum_{j=1}^{m} Z_{ij}\\ &= \sum_{i=1}^{n}\left(\frac{\mathbf{r}_i}{h}\right)\cdot\ln\left(\frac{\mathbf{r}_i/h}{\mathbf{r}_i^{(t)}/h}\right)\end{aligned}$$

since $\sum_{j=1}^{m} Z_{ij} = \mathbf{r}_i$. The last summand is precisely $\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)})$. The other equation follows analogously. This ends the proof of Lemma 3. □.

The above two lemmas easily imply the theorem. If for all $0 \leq t \leq T$, both $\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) > \delta$ and $\mathbf{D}_{KL}(\pi_{\mathbf{c}}||\pi_{\mathbf{c}}^{(t)}) > \delta$, then substituting in Lemma 3 and summing we get $\mathbf{D}(Z, A^{(0)}) - \mathbf{D}(Z, A^{(T+1)}) > T\delta > \ln(1 + 2\Delta/\nu)$ contradicting Lemma 2. This concludes the proof of Theorem 2. □

Theorem 1 follows from Theorem 2 using connections between the KL-divergence and the ℓ_1 and ℓ_2 norms. One is the following famous Pinsker's inequality which allows us to easily prove part 1 of Theorem 1. Given any two probability distributions p, q,

$$\mathbf{D}_{KL}(p||q) \geq \frac{1}{2}\cdot||p-q||_1^2 \qquad \textbf{(Pinsker)}$$

Proof (Theorem 1, Part 1) Apply (**Pinsker**) on the vectors $\pi_{\mathbf{r}}$ and $\pi_{\mathbf{r}}^{(t)}$ to get

$$\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) \geq \frac{1}{2h^2}||\mathbf{r}^{(t)} - \mathbf{r}||_1^2$$

Set $\delta := \frac{\varepsilon^2}{2h^2}$ and apply Theorem 2. In $O\left(\frac{h^2\ln(\Delta/\nu)}{\varepsilon^2}\right)$ time we would get a matrix with $\delta > \mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)})$ which from the above inequality would imply $||\mathbf{r}^{(t)} - \mathbf{r}||_1 \leq \varepsilon$. This proves part 1 of Theorem 1. □

To prove Part 2, we need a way to relate the ℓ_2 norm and the KL-divergence. In order to do so, we prove a different lower bound which implies Pinsker's inequality (with a worse constant), but is significantly stronger in certain regimes. This may be of independent interest in other domains. Below we state the version which we need for the proof of Theorem 1, part 2. This is an instantiation of the general inequality Lemma 5 which we prove in Sect. 3.

Lemma 4 *Given any pair of probability distributions p, q over a finite domain, define $\mathcal{A} := \{i : q_i > 2p_i\}$ and $\mathcal{B} := \{i : q_i \leq 2p_i\}$. Then,*

$$\mathbf{D}_{KL}(p||q) \geq (1 - \ln 2) \cdot \left(\sum_{i \in \mathcal{A}} |q_i - p_i| + \sum_{i \in \mathcal{B}} \frac{(q_i - p_i)^2}{p_i} \right) \qquad (\mathbf{KL\ vs}\ \ell_1/\ell_2)$$

Proof (Theorem 1, Part 2) We apply Lemma 4 on the vectors $\pi_{\mathbf{r}}$ and $\pi_{\mathbf{r}}^{(t)}$.

Lemma 4 gives us

$$\begin{aligned}\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) &\geq C \cdot \left(\frac{1}{h} \sum_{i \in A} |\mathbf{r}_i^{(t)} - \mathbf{r}_i| \ + \ \frac{1}{h} \sum_{i \in B} \frac{(\mathbf{r}_i^{(t)} - \mathbf{r}_i)^2}{\mathbf{r}_i} \right) \\ &\geq \frac{C}{h} \left(\sum_{i \in A} |\mathbf{r}_i^{(t)} - \mathbf{r}_i| \ + \ \frac{1}{\rho} \sum_{i \in B} (\mathbf{r}_i^{(t)} - \mathbf{r}_i)^2 \right)\end{aligned}$$

where $C = 1 - \ln 2$. If the second summand in the parenthesis of the RHS is $\geq \frac{1}{2}||\mathbf{r}^{(t)} - \mathbf{r}||_2^2$, then we get $\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) \geq \frac{C}{2\rho h}||\mathbf{r}^{(t)} - \mathbf{r}||_2^2$. Otherwise, we have $\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) \geq \frac{C}{\sqrt{2}h}||\mathbf{r}^{(t)} - \mathbf{r}||_2$, where we used the weak fact that the sum of some positive numbers is at least the square-root of the sum of their squares. In any case, we get the following

$$\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) \geq \min\left(\frac{C}{2\rho h}||\mathbf{r}^{(t)} - \mathbf{r}||_2^2, \frac{C}{\sqrt{2}h}||\mathbf{r}^{(t)} - \mathbf{r}||_2 \right) \qquad (2)$$

To complete the proof of part 2 of Theorem 1, we set $\delta := \frac{C}{2h\left(\frac{1}{\varepsilon} + \frac{\rho}{\varepsilon^2}\right)}$. In particular, this implies that $\delta < \min\left(\frac{\varepsilon C}{2h}, \frac{C\varepsilon^2}{2h\rho}\right)$. Applying Theorem 2, we get that in $O\left(h \ln(\Delta/\nu) \cdot \left(\frac{1}{\varepsilon} + \frac{\rho}{\varepsilon^2}\right)\right)$ iterations, we would get a matrix with $\mathbf{D}_{KL}(\pi_{\mathbf{r}}||\pi_{\mathbf{r}}^{(t)}) \leq \delta$.

If the minimum of the RHS of (2) is the first term, then we get $||\mathbf{r}^{(t)} - \mathbf{r}||_2^2 \leq \frac{2\rho h}{C} \cdot \delta < \frac{2\rho h}{C} \cdot \frac{C\varepsilon^2}{2h\rho} = \varepsilon^2$, implying the ℓ_2-error is $\leq \varepsilon$. If the minimum is the second term, then we get $||\mathbf{r}^{(t)} - \mathbf{r}||_2 \leq \frac{\sqrt{2}h}{C} \cdot \delta < \frac{\sqrt{2}h}{C} \cdot \frac{\varepsilon C}{2h} < \varepsilon$. This completes the proof of part 2 of Theorem 1. □

3 New lower bound on the KL-divergence

We now establish a new lower bound on KL-divergence which yields ($\mathbf{KL\ vs}\ \ell_1/\ell_2$) as a corollary.

Lemma 5 *Let p and q be two distributions over a finite n-element universe. For any fixed $\theta > 0$, define the sets $\mathcal{A}_\theta := \{i \in [n] : q_i > (1 + \theta)p_i\}$ and $\mathcal{B}_\theta = [n] \setminus \mathcal{A}_\theta =$*

$\{i \in [n] : q_i \le (1+\theta)p_i\}$. *Then we have the following inequality*

$$\mathbf{D}_{KL}(p||q) \ge \left(1 - \frac{\ln(1+\theta)}{\theta}\right) \cdot \left(\sum_{i\in\mathcal{A}_\theta} |q_i - p_i| + \frac{1}{\theta}\sum_{i\in\mathcal{B}_\theta} p_i \left(\frac{q_i - p_i}{p_i}\right)^2\right) \quad (3)$$

When $\theta = 1$, we get (**KL vs** ℓ_1/ℓ_2).

Proof We need the following fact which follows from calculus; we provide a proof later for completeness. □

Lemma 6 *Given any* $\theta > 0$, *define* $a_\theta := \frac{\ln(1+\theta)}{\theta}$ *and* $b_\theta := \frac{1}{\theta}\left(1 - \frac{\ln(1+\theta)}{\theta}\right)$. *Then,*

- *For* $t \ge \theta$, $(1+t) \le e^{a_\theta t}$
- *For* $t \le \theta$, $(1+t) \le e^{t - b_\theta t^2}$

Define $\eta_i := \frac{q_i - p_i}{p_i}$. Note that $\mathcal{A}_\theta = \{i : \eta_i > \theta\}$ and $\mathcal{B}_\theta$ is the rest. We can write the KL-divergence as follows

$$\mathbf{D}_{KL}(p||q) := \sum_{i=1}^{n} p_i \ln(p_i/q_i) = -\sum_{i=1}^{n} p_i \ln(1+\eta_i)$$

For $i \in \mathcal{A}_\theta$, since $\eta_i > \theta$, we upper bound $(1+\eta_i) \le e^{a_\theta \eta_i}$ using Lemma 6. For $i \in \mathcal{B}_\theta$, that is $\eta_i \le \theta$, we have the upper bound $(1+\eta_i) \le e^{\eta_i - b_\theta \eta_i^2}$ using Lemma 6. Lastly, we note $\sum_i p_i \eta_i = 0$ since p, q both sum to 1, implying $\sum_{i\in\mathcal{B}_\theta} p_i\eta_i = -\sum_{i\in\mathcal{A}_\theta} p_i\eta_i$. Putting all this in the definition above we get

$$\mathbf{D}_{KL}(p||q) \ge -a_\theta \cdot \sum_{i\in\mathcal{A}_\theta} p_i\eta_i - \sum_{i\in\mathcal{B}_\theta} p_i\eta_i + b_\theta \sum_{i\in\mathcal{B}_\theta} p_i\eta_i^2$$
$$= (1 - a_\theta)\sum_{i\in\mathcal{A}_\theta} p_i\eta_i + b_\theta \sum_{i\in\mathcal{B}_\theta} p_i\eta_i^2$$

The proof of inequality (3) follows by noting that $b_\theta = \frac{1-a_\theta}{\theta}$. This completes the proof of Lemma 5. □

Proof of Lemma 6 The proof of both facts follow by proving non-negativity of the relevant function in the relevant interval. Recall $a_\theta = \ln(1+\theta)/\theta$ and $b_\theta = \frac{1}{\theta}(1-a_\theta)$. We start with the following three inequalities about the log-function.

$$\text{For all } z > 0, \quad z + z^2/2 > (1+z)\ln(1+z) > z \quad \text{and} \quad \ln(1+z) > z - z^2/2 \quad (4)$$

The third inequality in (4) implies $a_\theta > 1 - \theta/2$ and thus, $b_\theta < 1/2$. The first inequality in (4) implies $a_\theta < \frac{1+\frac{\theta}{2}}{1+\theta}$ which in turn implies $b_\theta > 1/2(1+\theta)$. For brevity, henceforth let us lose the subscript on a_θ and b_θ.

Consider the function $f(t) = e^{at} - (1+t)$. Note that $f'(t) = ae^{at} - 1$ which is increasing in t since $a > 0$. So, for any $t \ge \theta$, we have $f'(t) \ge ae^{a\theta} - 1 =$

$\frac{(1+\theta)\ln(1+\theta)}{\theta} - 1 \geq 0$, by the second inequality in (4). Therefore, f is increasing when $t \geq \theta$. The first part of Lemma 6 follows since $f(\theta) = 0$ by definition of a.

Consider the function $g(t) = e^{t(1-bt)} - (1+t)$. Note that $g(0) = g(\theta) = 0$. We break the argument in two parts: we argue that $g(t)$ is strictly positive for all $t \leq 0$, and that $g(t)$ is strictly positive for $t \in (0, \theta)$. This will prove the second part of Lemma 6.

The first derivative is $g'(t) = (1-2bt)e^{t(1-bt)} - 1$ and the second derivative is $g''(t) = e^{t(1-bt)} \cdot \big((1-2bt)^2 - 2b\big)$. Since $b < 1/2$, we have $2b < 1$, and thus for $t \leq 0$, $g''(t) > 0$. Therefore, g' is strictly increasing for $t \leq 0$. However, $g'(0) = 0$, and so $g'(t) < 0$ for all $t < 0$. This implies g is strictly decreasing in the interval $t < 0$. Noting $g(0) = 0$, we get $g(t) > 0$ for all $t < 0$. This completes the first part of the argument.

For the second part, we first note that $g'(\theta) < 0$ since $b > \frac{1}{2(1+\theta)}$. That is, g is strictly decreasing at θ. On the other hand g is increasing at θ. To see this, looking at g' is not enough since $g'(0) = 0$. However, $g''(0) > 0$ since $b < 1/2$. This means that 0 is a strict (local) minimum for g implying g is increasing at 0. In sum, g vanishes at 0 and θ, and is increasing at 0 and decreasing at θ. This means that if g does vanish at some $r \in (0, \theta)$, then it must vanish once again in $[r, \theta)$ for the it to be decreasing at θ. In particular, g' must vanish three times in $(0, \theta)$ and thus four times in $[0, \theta)$ since $g'(0) = 0$. This in turn implies g'' vanishes three times in $[0, \theta)$ which is a contradiction since g'' is a quadratic in t multiplied by a positive term.

We end by proving (4). This also follows the same general methodology. Define $p(z) := (1+z)\ln(1+z) - z$ and $q(z) := p(z) - z^2/2$. Differentiating, we get $p'(z) = \ln(1+z) > 0$ for all $z > 0$, and $q'(z) = \ln(1+z) - z < 0$ for all $z > 0$. Thus, p is increasing, and q is decreasing, in $(0, \infty)$. The first two inequalities of (4) follow since $p(0) = q(0) = 0$. To see the third inequality, define $r(z) = \ln(1+z) - z + z^2/2$ and observe $r'(z) = \frac{1}{1+z} - 1 + z = \frac{z^2}{1+z}$ which is > 0 if $z > 0$. Thus r is strictly increasing, and the third inequality of (4) follows since $r(0) = 0$. □

3.1 Comparison with other well-known inequalities

We connect (**KL vs** ℓ_1/ℓ_2) with two well known lower bounds on the KL-Divergence. First we compare with Pinsker's inequality (**Pinsker**). To see that (**KL vs** ℓ_1/ℓ_2) generalizes (**Pinsker**) with a weaker constant, note that

$$||p-q||_1^2 = \left(\sum_{i\in\mathcal{A}} |q_i - p_i| + \sum_{i\in\mathcal{B}} |q_i - p_i|\right)^2 \leq 2\left(\sum_{i\in\mathcal{A}} |q_i - p_i|\right)^2$$
$$+2\left(\sum_{i\in\mathcal{B}} p_i \frac{|q_i - p_i|}{p_i}\right)^2$$

The first parenthetical term above, since it is ≤ 1, is at most the first summation in the parenthesis of (**KL vs** ℓ_1/ℓ_2). The second parenthetical term above, by Cauchy-Schwarz, is at most the second summation in the parenthesis of (**KL vs** ℓ_1/ℓ_2). Thus (**KL vs** ℓ_1/ℓ_2) implies

$$\mathbf{D}_{KL}(p||q) \geq \frac{(1-\ln 2)}{2}||p-q||_1^2$$

On the other hand, the RHS of (**KL vs** ℓ_1/ℓ_2) can be much larger than that of (**Pinsker**). For instance, suppose $p_i = 1/n$ for all i, $q_1 = 1/n + 1/\sqrt{n}$, and for $i \neq 1$, $q_i = 1/n - \frac{1}{(n-1)\sqrt{n}}$. The RHS of (**Pinsker**) is $\Theta(1/n)$ while that of (**KL vs** ℓ_1/ℓ_2) is $\Theta(1/\sqrt{n})$ which is the correct order of magnitude for $\mathbf{D}_{KL}(p||q)$.

The KL-divergence between two distributions is also at least the *Hellinger distance* between them. Before proceeding, let us define this distance.

$$\text{Given two distributions } p, q \text{ over } [n], \quad \mathbf{D}_{\mathsf{Hellinger}}(p,q) := \left(\sum_{i=1}^{n}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2\right)^{1/2}$$

The following inequality is known (see Reiss [30] p 99, Pollard [28] Chap 3.3, or the webpage [39] for a proof).

$$\text{For any two distributions } p, q, \quad \mathbf{D}_{KL}(p||q) \geq \mathbf{D}^2_{\mathsf{Hellinger}}(p,q) \quad \textbf{(KL-vs-Hellinger)}$$

It seems natural to compare the RHS of (**KL vs** ℓ_1/ℓ_2) and (**KL-vs-Hellinger**) (we thank Daniel Dadush for bringing this to our attention). As the subsequent calculation shows, the RHS of (**KL vs** ℓ_1/ℓ_2) is in fact $\Theta(\mathbf{D}^2_{\mathsf{Hellinger}}(p,q))$. In particular, this implies one can obtain (by reverse engineering the argument below) part 2 of Theorem 2 via the application of (**KL-vs-Hellinger**) as well.

For the set $\mathcal{A} = \{i : q_i > 2p_i\}$, we know $\sqrt{q_i} + \sqrt{p_i} = \Theta(\sqrt{q_i} - \sqrt{p_i})$. Therefore,

$$\sum_{i\in\mathcal{A}}(q_i - p_i) = \sum_{i\in A}\left(\sqrt{q_i}+\sqrt{p_i}\right)\left(\sqrt{q_i}-\sqrt{p_i}\right) = \Theta\left(\sum_{i\in\mathcal{A}}\left(\sqrt{q_i}-\sqrt{p_i}\right)^2\right)$$

For any $i \in \mathcal{B} = \{i : q_i \leq 2p_i\}$, let $q_i = (1+\eta_i)p_i$ where $-1 \leq \eta_i \leq 1$. Via a Taylor series expansion it is not hard to check $\left(1 + \frac{\eta_i}{2} - \sqrt{1+\eta_i}\right) = \Theta(\eta_i^2)$ in this range of η_i. Observing that

$$p_i\left(\frac{q_i-p_i}{p_i}\right)^2 = \eta_i^2 p_i \quad \text{and} \quad \left(\sqrt{p_i}-\sqrt{q_i}\right)^2 = 2p_i\left(1+\frac{\eta_i}{2}-\sqrt{1+\eta_i}\right)$$

we get that the RHS of (**KL vs** ℓ_1/ℓ_2) is $\Theta(\mathbf{D}^2_{\mathsf{Hellinger}}(p,q))$.

Acknowledgements We thank Daniel Dadush for asking the connection between our inequality and Hellinger distance, and Jonathan Weed for letting us know of [4]. We would also like to thank the anonymous reviewers for insightful suggestions.

References

1. Aaronson, S.: Quantum computing and hidden variables. Phys. Rev. A **71**, 032325 (2005)
2. Allen Zhu, Z., Li, Y., Oliveira, R., Wigderson, A.: Much faster algorithms for matrix scaling. In: 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS (2017)
3. Allen Zhu, Z., Orecchia, L.: Using optimization to break the epsilon barrier: a faster and simpler width-independent algorithm for solving positive linear programs in parallel. In: Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, pp. 1439–1456, San Diego, CA, USA (2015)
4. Altschuler, J., Weed, J., Rigollet, P.: Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, pp. 1961–1971, Long Beach, CA, USA (2017)
5. Bacharach, M.: Estimating nonnegative matrices from marginal data. Int. Econ. Rev. **6**(3), 294–310 (1965)
6. Balakrishnan, H., Hwang, I., Tomlin, C.J.: Polynomial approximation algorithms for belief matrix maintenance in identity management. In: 2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601), vol. 5, pp. 4874–4879 (2004)
7. Bapat, R., Raghavan, T.: An extension of a theorem of Darroch and Ratcliff in loglinear models and its application to scaling multidimensional matrices. Linear Algebra Appl. **114**, 705–715 (1989). Special Issue Dedicated to Alan J. Hoffman
8. Bregman, L.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
9. Brualdi, R.A., Parter, S.V., Schneider, H.: The diagonal equivalence of a nonnegative matrix to a stochastic matrix. J. Math. Anal. Appl. **16**(1), 31–50 (1966)
10. Cohen, M.B., Madry, A., Tsipras, D., Vladu, A.: Matrix scaling and balancing via box constrained newton's method and interior point methods. In: 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS (2017)
11. Csiszar, I.: I-divergence geometry of probability distributions and minimization problems. Ann. Probab. **3**(1), 146–158 (1975)
12. Csiszar, I.: A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. Ann. Stat. **17**(3), 1409–1413 (1989)
13. Deming, W.E., Stephan, F.F.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Stat. **11**(4), 427–444 (1940)
14. Franklin, J., Lorenz, J.: On the scaling of multidimensional matrices. Linear Algebra Appl. **114**, 717–735 (1989). Special Issue Dedicated to Alan J. Hoffman
15. Gurvits, L., Yianilos, P.N.: The deflation–inflation method for certain semidefinite programming and maximum determinant completion problems. Technical report, NEC Research Institute, 4 Independence Way, Princeton, NJ 08540 (1998)
16. Idel, M.: A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps (2016). ArXiv e-prints. arXiv:1609.06349
17. Kalantari, B., Khachiyan, L.: On the rate of convergence of deterministic and randomized RAS matrix scaling algorithms. Oper. Res. Lett. **14**(5), 237–244 (1993)
18. Kalantari, B., Khachiyan, L.: On the complexity of nonnegative-matrix scaling. Linear Algebra Appl. **240**, 87–103 (1996)
19. Kalantari, B., Lari, I., Ricca, F., Simeone, B.: On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. Technical Report, no. 24. Department of Statistics and Applied probability, La Sapienza University, Rome (2002)
20. Kalantari, B., Lari, I., Ricca, F., Simeone, B.: On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. Math. Program. **112**(2), 371–401 (2008)
21. Knight, P.A.: The Sinkhorn–Knopp algorithm: convergence and applications. SIAM J. Matrix Anal. Appl. **30**(1), 261–275 (2008)
22. Linial, N., Samorodnitsky, A., Wigderson, A.: A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. Combinatorica **20**(4), 545–568 (2000)

23. Luby, M., Nisan, N.: A parallel approximation algorithm for positive linear programming. In: Proceedings of the Twenty-fifth Annual ACM Symposium on Theory of Computing, STOC ’93, pp. 448–457. ACM, New York, NY, USA (1993)
24. Macgill, S.M.: Theoretical properties of biproportional matrix adjustments. Environ. Plan. A **9**(6), 687–701 (1977)
25. Mahoney, M.W., Rao, S., Wang, D., Zhang, P.: Approximating the solution to mixed packing and covering LPs in parallel $O(\varepsilon^{-3})$ time. In: 43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11–15, 2016, pp. 52:1–52:14, Rome, Italy (2016)
26. Menon, M.: Reduction of a matrix with positive elements to a doubly stochastic matrix. Proc. Am. Math. Soc. **18**(2), 244–247 (1967)
27. Ortúzar, J d D, Willumsen, L .G.: Modelling Transport. Wiley, New York (2011)
28. Pollard, D.: A User’s Guide to Measure Theoretic Probability. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2001)
29. Raghavan, T.: On pairs of multidimensional matrices. Linear Algebra Appl. **62**, 263–268 (1984)
30. Reiss, R.-D.: Approximate Distributions of Order Statistics. Springer, Berlin (1989)
31. Rote, G., Zachariasen, M.: Matrix scaling by network flow. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA ’07, pp. 848–854. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2007)
32. Rothblum, U., Schneider, H.: Scalings of matrices which have prespecified row sums and column sums via optimization. Linear Algebra Appl. **114**, 737–764 (1989). **(Special Issue Dedicated to Alan J. Hoffman)**
33. Ruschendorf, L.: Convergence of the iterative proportional fitting procedure. Ann. Stat. **23**(4), 1160–1174 (1995)
34. Sason, I., Verdú, S.: Upper bounds on the relative entropy and Rényi divergence as a function of total variation distance for finite alphabets. In: 2015 IEEE Information Theory Workshop—Fall (ITW), Jeju Island, South Korea, October 11–15, 2015, pp. 214–218. IEEE (2015)
35. Schrödinger, E.: Über die umkehrung der naturgesetze. Preuss. Akad. Wiss., Phys.-Math. Kl, pp. 412–422 (1931)
36. Sinkhorn, R.: Diagonal equivalence to matrices with prescribed row and column sums. Am. Math. Mon. **74**(4), 402–405 (1967)
37. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pac. J. Math. **21**(2), 343–348 (1967)
38. Soules, G.W.: The rate of convergence of Sinkhorn balancing. Linear Algebra Appl. **150**, 3–40 (1991)
39. Xi’an: Statistics stack exchange. https://stats.stackexchange.com/questions/130432/differences-between-bhattacharyya-distance-and-kl-divergence (2014). Accessed 15 Jan 2018
40. Young, N.E.: Sequential and parallel algorithms for mixed packing and covering. In: 42nd Annual Symposium on Foundations of Computer Science, FOCS, pp. 538–546, Las Vegas, Nevada, USA (2001)