SPECTRAL ANALYSIS OF MATRIX SCALING AND OPERATOR SCALING[*]

TSZ CHIU KWOK[†], LAP CHI LAU[‡], AND AKSHAY RAMACHANDRAN[‡]

Abstract. We present a spectral analysis of a continuous scaling algorithm for matrix scaling and operator scaling. The main result is that if the input matrix or operator has a spectral gap, then a natural gradient flow has linear convergence. This implies that a simple gradient descent algorithm also has linear convergence under the same assumption. The spectral gap condition for operator scaling is closely related to the notion of quantum expander studied in quantum information theory. The spectral analysis also provides bounds on some important quantities of the scaling problems, such as the condition number of the scaling solution and the capacity of the matrix and operator. These results can be used in various applications of scaling problems, including matrix scaling on expander graphs, permanent lower bounds on random matrices, the Paulsen problem on random frames, and Brascamp–Lieb constants on random operators. In some applications, the inputs of interest satisfy the spectral condition and we prove significantly stronger bounds than the worst case bounds.

Key words. operator scaling, spectral analysis, frame theory

AMS subject classification. 68Q25

DOI. 10.1137/20M1315981

1. Introduction. In the matrix scaling problem, we are given a nonnegative matrix $B \in \mathbb{R}^{n \times n}$, and the goal is to find a left diagonal scaling matrix $L \in \mathbb{R}_+^{n \times n}$ and a right diagonal scaling matrix $R \in \mathbb{R}_+^{n \times n}$ such that LBR is doubly stochastic (every row sum and every column sum is one), or report that such scaling matrices do not exist. This problem has been extensively studied in different communities; see [41] for a detailed survey.

The operator scaling problem is a significant generalization of the matrix scaling problem. Given a tuple of matrices $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$, a linear operator $\Phi_{\mathcal{A}} : \mathbb{R}^{n \times n} \to \mathbb{R}^{m \times m}$ is defined as

$$\Phi_{\mathcal{A}}(X) = \sum_{i=1}^{k} A_i X A_i^*,$$

where A_i^* denotes the conjugate transpose of A_i which is just the transpose when A_i is real. We will simply refer to $\mathcal{A}$ as an operator. The size of an operator $\mathcal{A}$ is defined as $s(\mathcal{A}) := \sum_{i=1}^{k} \|A_i\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. An operator $\mathcal{A}$ is called ϵ-nearly doubly balanced if

$$(1-\epsilon)\frac{s(\mathcal{A})}{m}I_m \preceq \sum_{i=1}^{k} A_i A_i^* \preceq (1+\epsilon)\frac{s(\mathcal{A})}{m}I_m \text{ and } (1-\epsilon)\frac{s(\mathcal{A})}{n}I_n \preceq \sum_{i=1}^{k} A_i^* A_i \preceq (1+\epsilon)\frac{s(\mathcal{A})}{n}I_n$$

[†]Institute for Theoretical Computer Science, Shanghai University of Finance and Economics, Shanghai, China (kwok@mail.shufe.edu.cn).

[‡]School of Computer Science, University of Waterloo, Waterloo ON N2L 3G1, Canada (lapchi@uwaterloo.ca, a5ramach@uwaterloo.ca).

and is called doubly balanced when $\epsilon = 0$. The operator scaling problem is defined by Gurvits [31]. The objective is to scale the input operator so that it becomes doubly balanced with size one.

DEFINITION 1.1 (operator scaling problem).
Input: *An operator $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$ for $1 \leq i \leq k$.*
Output: *A left scaling matrix $L \in \mathbb{R}^{m \times m}$ and a right scaling matrix $R \in \mathbb{R}^{n \times n}$ such that*

$$\sum_{i=1}^{k} (LA_iR)(LA_iR)^* = \frac{I_m}{m} \quad \text{and} \quad \sum_{i=1}^{k} (LA_iR)^*(LA_iR) = \frac{I_n}{n},$$

or report that such scaling matrices L, R do not exist.

There is a simple reduction from the matrix scaling problem to the operator scaling problem, by having one matrix $A_{ij} \in \mathbb{R}^{n \times n}$ for each entry B_{ij} with the (i, j)-entry of A_{ij} being $\sqrt{B_{ij}}$ and all other entries zero; see section 4.1 for details.

The operator scaling problem generalizes matrix scaling and frame scaling, defined in section 1.4.2, and has many applications; see sections 1.4 and 4. Much work has been done in analyzing algorithms for these scaling problems and understanding the scaling solutions and related quantities.

1.1. Previous algorithms. For matrix scaling, the best-known algorithm is Sinkhorn's algorithm [56], which is a simple iterative algorithm that alternatively rescales the rows and rescales the columns. This algorithm is analyzed in [19] and it is shown that the alternating algorithm finds an η-nearly doubly stochastic scaling in time polynomial in n and $1/\eta$.

The alternating scaling algorithm is generalized in [31] for the operator scaling problem. In this algorithm, we alternately find a left scaling matrix $L = (\sum_i A_i A_i^*)^{-1/2}$ and set $A_i \leftarrow LA_i$ so that the first condition of doubly balanced is satisfied, and we find a right scaling matrix $R = (\sum_i A_i^* A_i)^{-1/2}$ and set $A_i \leftarrow A_i R$ so that the second condition of doubly balanced is satisfied, and repeat.

This alternating algorithm was proposed in [31] as a solution to the operator scaling problem, and its polynomial time convergence was verified in [22, 21].

THEOREM 1.2 ([56, 19, 22, 21]). *The alternating scaling algorithm returns an η-nearly doubly balanced scaling in $O(\text{poly}(n, m, k, 1/\eta))$ iterations if such a scaling exists.*

This theorem is used in [22, 21] to give the first polynomial time algorithm for computing the noncommutative rank of a symbolic matrix, as it is sufficient to set η to be inverse polynomial in n to solve that problem exactly. For some applications, however, faster convergence in terms of η is required.

For matrix scaling, there are several algorithms with dependency on η being $\log(1/\eta)$, including the ellipsoid method in [42], the interior point method in [53], and a strongly polynomial time combinatorial algorithm in [49]. The dependency on n in these algorithms is at least $\Omega(n^{7/2})$ even for sparse matrices. Recently, two independent groups [14, 2] developed a fast second order method for matrix scaling, and this method is extended to geodesic convex optimization in [1] for the operator scaling problem.

THEOREM 1.3 ([14, 2, 1]). *There is a second order method to return an η-nearly doubly balanced scaling in time $O(\text{poly}(n, m, k, \log(1/\eta)))$ for operator scaling and in*

time $O(\|B\|_0 \log \kappa \log^2(1/\eta))$ for matrix scaling where $\|B\|_0$ denotes the number of nonzero entries in B and κ denotes the condition number of the scaling solution.

For matrix scaling, this theorem can be used to obtain a fast deterministic e^{-n} approximation algorithm for the permanent of a matrix [49]. For operator scaling, this theorem is used to obtain a polynomial time algorithm for an orbit intersection problem in invariant theory [1].

1.2. Gradient flow. An important quantity in [31, 22, 1] to measure the progress of the algorithms is the ℓ_2-error of the current solution. Given an operator $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$ for $1 \le i \le k$, define

$$\Delta(\mathcal{A}) = \frac{1}{m} \left\| s(\mathcal{A}) \cdot I_m - m \sum_{i=1}^k A_i A_i^* \right\|_F^2 + \frac{1}{n} \left\| s(\mathcal{A}) \cdot I_n - n \sum_{i=1}^k A_i^* A_i \right\|_F^2 .$$

Note that $\Delta(\mathcal{A}) = 0$ if and only if $\mathcal{A}$ is doubly balanced. In the matrix scaling problem for general $m \times n$ matrix where the objective is to scale the input matrix B such that every row sum is the same and every column sum is the same, this definition simplifies to

$$\Delta(B) = \frac{1}{m} \sum_{i=1}^m (s - mr_i)^2 + \frac{1}{n} \sum_{j=1}^n (s - nc_j)^2,$$

where r_i and c_j are the ith row sum and the jth column sum of the matrix B, and $s = \sum_{i=1}^m \sum_{j=1}^n B_{ij}$ is the size of the matrix B.

A continuous version of the alternating algorithm for operator scaling is studied in [47], where both operations are done simultaneously and continuously. The following differential equation describes how $\mathcal{A}$ changes over time:

$$\frac{d}{dt} A_i := \left(s(\mathcal{A}) \cdot I_m - m \sum_{j=1}^k A_j A_j^* \right) A_i + A_i \left(s(\mathcal{A}) \cdot I_n - n \sum_{j=1}^k A_j^* A_j \right) \quad \text{for } 1 \le i \le k.$$

In the matrix case, this continuous scaling algorithm simplifies to

$$\frac{d}{dt} B_{ij} = 2\big((s - mr_i) + (s - nc_j)\big) \cdot B_{ij}.$$

The continuous operator scaling algorithm is developed to bound the "total movement" of the operator in order to solve the Paulsen problem in [47]. Its convergence rate is shown to be similar to that of the alternating scaling algorithm, with dependency on η being $1/\eta$.

The continuous operator scaling algorithm can be understood as a natural first order method for the operator scaling problem. As we will show in Lemma A.1 in Appendix A, the dynamical system in continuous operator scaling is equivalent to the gradient flow (or continuous gradient descent) that always moves in the direction of minimizing $\Delta(\mathcal{A})$ at each time. This shows a close connection between gradient descent and the alternating algorithm.

This gradient flow was studied in much greater generality (see [43, 29]) and in particular was used by Kirwan to prove a beautiful convexity result in the setting of symplectic geometry [3, 27, 28, 45, 44]. An important lemma in that work considered the topological properties of the set of fixed points of the gradient flow. The current result, as well as the result in [47], can be seen as quantitative convergence analyses in the neighborhoods of fixed points of this natural gradient flow in the operator scaling setting.

1.3. Contributions. In this paper, we analyze this gradient flow for the operator scaling problem. We identify a natural spectral condition under which the gradient flow converges in time $t = O(\log(1/\eta))$ (corresponding to the number of iterations in the alternating algorithm) where η is the output accuracy. The spectral condition is closely related to the notion of "quantum expander" and is satisfied in many random instances. A key feature of our approach is that it also provides bounds on some important mathematical quantities such as the condition number of the scaling solution and the capacity of the matrix and operator. These bounds can be used in various applications of the operator scaling problem to show significantly stronger results for inputs that satisfy the spectral condition such as random matrices and random frames. We emphasize that the new results in various applications cannot be obtained through previous work (e.g., the fast algorithm for operator scaling in [1]), as the analyses of previous algorithms do not provide mathematical bounds for the condition number of the scaling solution and the operator capacity.

Spectral condition. We first state the spectral condition in the general operator setting.

DEFINITION 1.4 (spectral gap condition). *Given an operator $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$ for $1 \le i \le k$, define the $m^2 \times n^2$ matrix*

$$M_{\mathcal{A}} := \sum_{i=1}^{k} A_i \otimes A_i,$$

where $\otimes$ denotes the tensor product. The operator $\mathcal{A}$ is said to have a λ-spectral gap if

$$\sigma_2(M_{\mathcal{A}}) \le (1 - \lambda) \frac{s(\mathcal{A})}{\sqrt{mn}},$$

where $\sigma_2(M_{\mathcal{A}})$ is the second largest singular value of $M_{\mathcal{A}}$.

Note that the spectral condition can be checked in polynomial time through standard eigenvalue computation.

The matrix $M_{\mathcal{A}}$ associated with $\mathcal{A}$ is studied in the quantum information theory literature (e.g., [63]), as the natural matrix representation of the completely positive map $\Phi(X) := \sum_i A_i X A_i^*$ defined by $\mathcal{A}$. It can be shown that the largest singular value of $M_{\mathcal{A}}$ satisfies

$$\frac{s(\mathcal{A})}{\sqrt{mn}} \le \sigma_1(M_{\mathcal{A}}) \le (1 + \epsilon) \frac{s(\mathcal{A})}{\sqrt{mn}},$$

when $\mathcal{A}$ is ϵ-nearly doubly balanced (Lemma 3.6). The spectral gap condition is also studied under the name of "quantum expander" in [7, 37]. We will discuss more about this spectral gap condition in section 2.1 after some background on quantum information theory is reviewed.

For matrix scaling, given the input matrix $B \in \mathbb{R}^{m \times n}$, the spectral gap condition is simply

$$\sigma_2(B) \le (1 - \lambda) \frac{s(B)}{\sqrt{mn}}.$$

If we interpret the input matrix B as a weighted undirected bipartite graph, then the spectral gap condition is closely related to the expansion/conductance of the graph. We will explain more about these in sections 1.4.1 and 4.1.

Linear convergence. The main technical result is that the gradient flow has linear convergence when the input satisfies the spectral gap condition.

THEOREM 1.5 (linear convergence). *Given an operator $\mathcal{A} = (A_1, \ldots, A_k)$ where each $A_i \in \mathbb{R}^{m \times n}$ with $m \leq n$, if $\mathcal{A}$ is ϵ-nearly doubly balanced and $\mathcal{A}$ satisfies the λ-spectral gap condition in Definition 1.4 with $\lambda^2 \geq C\epsilon \log m$ for a sufficiently large constant C, then in the gradient flow,*

$$\Delta^{(t)} \leq \Delta^{(0)} e^{-\lambda s^{(0)} t} \quad \text{for any } t \geq 0.$$

In particular, the gradient flow converges to an η-nearly doubly balanced scaling in time $t = O(\frac{1}{\lambda} \log(\frac{m}{\eta}))$, and such a scaling always exists under our assumptions.

By discretizing the gradient flow with step size $\Theta((m + n)^{-2})$, it follows that a natural gradient descent algorithm returns an η-nearly doubly stochastic scaling in polynomial time in the input size and logarithmic in $1/\eta$, when the input satisfies the spectral gap condition.

COROLLARY 1.6 (gradient descent). *Under the assumptions in Theorem 1.5, there is a gradient descent algorithm to return an η-nearly doubly balanced scaling in $O(\frac{(n+m)^2}{\lambda} \log(\frac{m+n}{\eta}))$ iterations.*

It is an interesting open question whether the alternating algorithm also has the same convergence rate as the gradient flow under the same assumptions. Subsequent to our work, the linear convergence of the alternating algorithm was proved by Franks and Moitra [20]. A more detailed discussion of their techniques is given in Remark 3.24.

Condition number. The condition number of the scaling solutions L, R is defined as $\kappa(L) := \sigma_{\max}(L)/\sigma_{\min}(L)$ where $\sigma_{\max}(L)$ and $\sigma_{\min}(L)$ denote the largest and smallest singular values of L, respectively. For matrix scaling, $\kappa(L)$ is simply the ratio between the largest entry and the smallest entry in the diagonal matrix L.

In general, the condition numbers could be exponential in the input size. It is of interest to identify instances with small condition numbers as these are closely related to the performance of matrix/operator scaling algorithms (e.g., Theorem 1.3), but not much is known even in the simpler matrix scaling setting. Kalantari and Khachiyan [42] proved a bound for strictly positive matrices in terms of the ratio of the sum of the entries and the minimum entry. We show that the condition numbers are bounded by a small constant when the input satisfies the spectral gap condition (not necessarily strictly positive).

THEOREM 1.7 (condition number). *Under the assumptions in Theorem 1.5, the condition number of the scaling solutions $L \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{n \times n}$ satisfies*

$$\kappa(L) \leq 1 + O\left(\frac{\epsilon \log m}{\lambda}\right) \quad \text{and} \quad \kappa(R) \leq 1 + O\left(\frac{\epsilon \log m}{\lambda}\right).$$

The condition number of the scaling solutions is used in bounding the time complexity of the scaling algorithms using the second order method [1, 14], in analyzing an approximation algorithm for permanent [55], and in bounding the optimal transport cost [15, 54]. We will discuss the implications of Theorem 1.7 to these applications in section 4.

Operator capacity. The capacity of an operator $\mathcal{A}$ is defined by Gurvits [31] as

$$\operatorname{cap}(\mathcal{A}) := \inf_{X \succ 0} \frac{m \det\left(\sum_{i=1}^{k} A_i X A_i^*\right)^{1/m}}{\det(X)^{1/n}}.$$

The capacity of a matrix $B \in \mathbb{R}^{m \times n}$ has a simpler form (section 4.1.6) where

$$\operatorname{cap}(B) := \inf_{x \in \mathbb{R}^n : x > 0} \frac{m\left(\prod_{i=1}^{m}(Bx)_i\right)^{1/m}}{\left(\prod_{j=1}^{n} x_j\right)^{1/n}}.$$

Optimization problems of this form are also studied in functional analysis [5] and in approximation algorithms [52].

In general, when $\mathcal{A}$ is ϵ-nearly doubly balanced, it is proved that [31, 22, 47]

$$s(\mathcal{A}) \geq \operatorname{cap}(\mathcal{A}) \geq (1 - mn\epsilon)s(\mathcal{A}).$$

Using a connection between the convergence rate of the gradient flow and the operator capacity developed in [47], we show a much stronger bound for operators that also satisfy the spectral gap condition.

THEOREM 1.8 (capacity). *Under the assumptions in Theorem* 1.5,

$$s(\mathcal{A}) \geq \operatorname{cap}(\mathcal{A}) \geq \left(1 - \frac{4\epsilon^2}{\lambda}\right) s(\mathcal{A}).$$

The capacity of an operator is used in bounding the permanent of a matrix [49], the Brascamp–Lieb constant of an operator [23], and the total movement to a nearby doubly balanced operator [47]. We will discuss the implications of Theorem 1.8 to these applications in section 1.4.

1.4. Applications of matrix scaling and operator scaling. The matrix scaling and operator scaling problems have many applications and we will discuss some implications of our results in this section.

1.4.1. Matrix scaling. In the matrix scaling problem, we are given a nonnegative matrix $B \in \mathbb{R}^{m \times n}$, and the goal is to find a left diagonal scaling matrix $L \in \mathbb{R}_+^{m \times m}$ and a right diagonal scaling matrix $R \in \mathbb{R}_+^{n \times n}$ such that LBR is doubly balanced (i.e., every row sum is the same and every column sum is the same; see section 4.1 for a definition) or to report that such scaling matrices do not exist.

The matrix scaling problem is a special case of the operator scaling problem (section 4.1.1) and so the spectral analysis also applies. In the case of matrix scaling, the spectral condition in Definition 1.4 is simply $\sigma_2(B) \leq (1 - \lambda)s(B)/\sqrt{mn}$ (section 4.1.2). Using Cheeger's inequality, we show that this spectral gap condition is closely related to the conductance of the weighted bipartite graph associated to B (section 4.1.3). These imply that many random matrices will satisfy the condition in Theorem 1.5 (section 4.1.4).

Our results have implications for the matrix scaling problem, e.g., to obtain stronger results for random matrices. For bipartite matching, we show that the gradient flow converges quickly to a fractional perfect matching in an almost regular bipartite expander graph (section 4.1.5).

COROLLARY 1.9. *Suppose $G = (X, Y; E)$ is a bipartite graph with $|X| = |Y|$ where each vertex v satisfies $(1 - \epsilon)|E|/|X| \leq \deg(v) \leq (1 + \epsilon)|E|/|X|$ for some ϵ. If the graph conductance $\phi(G)$ satisfies $\phi(G)^4 \geq C\epsilon \log |X|$ for some sufficiently large constant C, then the gradient flow converges to an η-nearly perfect fractional matching in time $t = O(\frac{1}{\phi^2(G)} \log(\frac{|X|}{\eta}))$.*

For permanent, Van der Waerden's conjecture states that the permanent of a doubly stochastic $n \times n$ matrix is at least $n!/n^n \geq e^{-n}$, which is proven in [16, 17, 30]. The capacity lower bound in Theorem 1.8 can be used to prove a Van der Waerden's type lower bound on the permanent of matrices satisfying the spectral gap condition (not necessarily doubly stochastic).

COROLLARY 1.10. *If a nonnegative matrix $B \in \mathbb{R}^{n \times n}$ is ϵ-nearly doubly balanced with $s(B) = n$, and $\sigma_2(B) \leq 1 - \lambda$ with $\lambda^2 \geq C\epsilon \log n$ for some sufficiently large constant C, then*

$$\text{per}(B) \geq \exp\left(-n\left(1 + \Theta\left(\frac{\epsilon^2}{\lambda}\right)\right)\right).$$

For example, consider a random matrix A with each entry an independent random variable $A_{ij} = g_{ij}^2$ where g_{ij} is sampled from the Gaussian distribution $N(0, \frac{1}{n})$. The corollary implies that $\text{per}(A) \geq e^{-n}/\text{poly}(n)$ with high probability. See section 4.1.6 for details.

For optimal transportation distance, we can use the condition number result in Theorem 4.1.7 to bound the Sinkhorn distance [15, 54], which is receiving increasing attention in computer vision and machine learning (section 4.1.7).

The condition number result in Theorem 4.1.7 can also be used to show that the second order method for matrix scaling [14, 2] as stated in Theorem 1.3 is near linear time in the instances satisfying the spectral gap assumption.

1.4.2. Frame scaling. In the frame scaling problem, we are given n vectors $u_1, \ldots, u_n \in \mathbb{R}^d$, and the goal is to find a matrix (a linear transformation) $M \in \mathbb{R}^{d \times d}$ such that if we set $v_i = Mu_i / \|Mu_i\|_2$, then $\sum_{i=1}^{n} v_i v_i^* = I_d$. This problem was studied in communication complexity [18], machine learning [35], and frame theory [47, 34].

The frame scaling problem is a special case of the operator scaling problem (section 4.2.1) and so the spectral analysis also applies. In the case of frame scaling, the spectral condition in Definition 1.4 has a nice form (section 4.2.2): Let $G \in \mathbb{R}^{n \times n}$ be the squared Gram matrix where $G_{ij} = \langle u_i, u_j \rangle^2$. Then the spectral condition is equivalent to $\lambda_2(G) \leq (1 - \lambda)^2 s^2/(dn)$ where $\lambda_2(G)$ is the second largest eigenvalue of G and s is the size of the frame defined as $\sum_{i=1}^{n} \|u_i\|^2$. We will prove in section 5 that this condition is satisfied for random frames with high probability.

THEOREM 1.11. *If we generate n random unit vectors $u_1, \ldots, u_n \in \mathbb{R}^d$ with $n = \Omega(d^{4/3})$, then the resulting frame is ϵ-nearly doubly balanced for $\epsilon \ll 1/\log d$ and satisfies the spectral gap condition with constant λ with probability at least 0.99.*

For intuition, suppose each u_i is a random unit vector; then the expected value of $G_{ij} = \langle u_i, u_j \rangle^2$ for $i \neq j$ is $1/d$ and so the expected matrix G is $J_n/d + (d-1)I_n/d$ where J_n is the $n \times n$ all-one matrix. The matrix J_n has large spectral gap, and we expect that a random frame will have its squared Gram matrix G close to $J_n/d + (d-1)I_n/d$ and thus a large spectral gap. The proof is by a low moment analysis of the trace method commonly used in random matrix theory (section 5). Subsequent to our work, Franks and Moitra [20] show that $n = \Omega(d \text{ polylog } d)$ random unit vectors suffice to satisfy the conclusions of Theorem 1.11. A more detailed discussion of their results and techniques is given in Remark 5.2.

One significant implication of our result is the Paulsen problem on random frames. Given a frame $U = (u_1, \ldots, u_n)$ where each $u_i \in \mathbb{R}^d$ satisfying

$$(1 - \epsilon)I_d \preceq \sum_{i=1}^n u_i u_i^* \preceq (1 + \epsilon)I_d \quad \text{and} \quad (1 - \epsilon)\frac{d}{n} \leq \|u_i\|_2^2 \leq (1 + \epsilon)\frac{d}{n} \quad \text{for } 1 \leq i \leq n,$$

the Paulsen problem asks whether there always exists a frame $V = (v_1, \ldots, v_n)$ where each $v_i \in \mathbb{R}^d$ satisfying $\sum_{i=1}^n v_i v_i^* = I_d$, $\|v_i\|_2^2 = d/n$ for $1 \leq i \leq n$, and $\text{dist}^2(U, V) := \sum_{i=1}^n \|u_i - v_i\|_2^2$ small. It was an open problem whether $\text{dist}^2(U, V)$ can be bounded by a function independent of the number of vectors n. Recently, this question was answered positively in [47], showing that $\text{dist}^2(U, V) \leq O(d^{13/2}\epsilon)$. This bound is improved to $O(d^2\epsilon)$ by Hamilton and Moitra [34] with a much simpler proof. There are examples showing that $\text{dist}^2(U, V) \geq \Omega(d\epsilon)$, so the upper bound and the lower bound almost match in the worst case.

The Paulsen problem was asked [38] because it is difficult to generate V that satisfies the conditions exactly but easier to generate U that almost satisfies the conditions. But actually not many ways are known to generate U that almost satisfies the conditions with small ϵ, and almost all known constructions are random frames [38, 61]. Even for the few constructions that are deterministic (such as equiangular lines), it is likely that they satisfy the spectral gap assumption. So, for the Paulsen problem, the inputs of interest satisfy the spectral gap assumption, and we can prove a much stronger bound $O(d\epsilon^2)$ that goes beyond the worst case lower bound.

THEOREM 1.12. *Let $U = (u_1, \ldots, u_n)$ be a random frame with $n = \Omega(d^{4/3})$, where each $u_i \in \mathbb{R}^d$ is an independent random vector with $\|u_i\|_2^2 = d/n$. Suppose $(1 - \epsilon)I_d \preceq \sum_{i=1}^n u_i u_i^* \preceq (1 + \epsilon)I_d$. Then, with probability at least 0.99, there exists a frame $V = (v_1, \ldots, v_n)$ with $\sum_{i=1}^n v_i v_i^* = I_d$, $\|v_i\|_2^2 = d/n$ for $1 \leq i \leq n$, and $\text{dist}^2(U, V) \leq O(d\epsilon^2)$.*

We also demonstrate how the results in spectral analysis can be used to construct V with the additional property that $|\langle v_i, v_j \rangle|$ is small for $1 \leq i \neq j \leq n$, which is an original motivation for the Paulsen problem (section 4.2.4).

THEOREM 1.13. *For $n = d^2$, there exists a doubly balanced frame $V = (v_1, \ldots, v_n)$ where each $v_i \in \mathbb{R}^d$ with $\|v_i\| = 1$ and*

$$\max_{i \neq j} \langle v_i, v_j \rangle^2 \leq O\left(\frac{\log^3 d}{d}\right).$$

Note that this is nearly tight (up to log factors), as any doubly balanced frame with $\|v_i\| = 1$ satisfies $\max_{i \neq j}\langle v_i, v_j \rangle^2 \geq \frac{1}{d} \cdot \frac{n-d}{n-1}$. Using the improvement in [20], the conclusion of this theorem can be tightened by a $\log d$ factor. See Remark 4.22 for a more detailed discussion.

1.4.3. Operator scaling. The operator scaling problem was used to compute the Brascamp–Lieb constant [23]. A Brascamp–Lieb datum is specified by an m-tuple $\mathbf{B} = \{B_j : \mathbb{R}^n \to \mathbb{R}^{n_j} \mid 1 \leq j \leq m\}$ of linear transformations and an m-tuple of exponents $\mathbf{p} = \{p_1, \ldots, p_m\}$. The Brascamp–Lieb constant $\text{BL}(\mathbf{B}, \mathbf{p})$ of this datum is defined as the smallest C such that for every m-tuple $\{f_j : \mathbb{R}^{n_j} \to \mathbb{R}_{\geq 0} \mid 1 \leq j \leq m\}$ of nonnegative functions which are integrable, we have

$$\int_{x \in \mathbb{R}^n} \prod_{j=1}^m \left(f_j(B_j x)\right)^{p_j} dx \leq C \prod_{j=1}^m \left(\int_{x_j \in \mathbb{R}^{n_j}} f_j(x_j) dx_j\right)^{p_j}.$$

This is a common generalization of many useful inequalities; see [8, 23]. It turns out that the functions f_i for which the inequality is tight are density functions of Gaussians [48], and this implies the Brascamp–Lieb constant can be written in a form very similar to the capacity of an operator (see section 4.3.1). This is used in [23] to compute the Brascamp–Lieb constant through operator scaling.

Using this connection, we can derive upper bounds on the Brascamp–Lieb constant using the capacity lower bound in Theorem 1.8.

COROLLARY 1.14. *Given a datum* $(\mathbf{B}, \mathbf{p})$ *with* $B_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$ *for* $1 \le j \le m$ *and* $\sum_{j=1}^m p_j n_j = n$, *if* $(\mathbf{B}, \mathbf{p})$ *is* ϵ-*nearly geometric and satisfies the* λ-*spectral gap condition with* $\lambda^2 \ge C\epsilon \log n$ *for some sufficiently large constant* C *and* $\sum_{j=1}^m p_j \|B_j\|_F^2 = n$, *then*

$$1 \le \mathrm{BL}(\mathbf{B}, \mathbf{p}) \le \left(1 - \frac{4\epsilon^2}{\lambda}\right)^{-n/2} \le \exp\left(\Theta\left(\frac{n\epsilon^2}{\lambda}\right)\right).$$

An interesting special case of the Brascamp–Lieb inequality is the rank one case $B_j = u_j^*$ where $u_j \in \mathbb{R}^d$ and $n_j = 1$ and $p_j = d/m$ for $1 \le j \le m$ which was studied in [5]. In this case, the capacity of the operator $\mathcal{A}$ from the reduction (section 4.3.1) is

$$\mathrm{cap}(\mathcal{A}) = \inf_{x \in \mathbb{R}^n : x > 0} \frac{d \left(\det\left(\sum_{j=1}^m x_j u_j u_j^*\right)\right)^{1/d}}{\left(\prod_{j=1}^m x_j\right)^{1/m}},$$

which is a form that is also studied in approximation algorithms [52]. If $\{u_1, \ldots, u_m\}$ are independent random unit vectors where each $u_i \in \mathbb{R}^d$, the worst case capacity lower bound in [22] and [23] is only in terms of ϵ and gives $\mathrm{cap}(\mathcal{A}) \ge m(1 - md\epsilon)$, which is only nontrivial for $\epsilon < \frac{1}{md}$, whereas the random frame $\mathcal{A}$ will be $\epsilon \approx \sqrt{d/m}$-doubly balanced with constant probability. Using the results in section 5 and the above corollary, we can show that if $m \ge \Omega(d^{4/3})$, then $m \ge \mathrm{cap}(\mathcal{A}) \ge m\left(1 - 4d \log d/m\right)$ and $1 \le \mathrm{BL}(\mathbf{B}, \mathbf{p}) \le d^{\Theta(d)}$; see Example 4.29. Note that this is independent of the number of vectors.

The operator scaling algorithm is used in [22, 21] to compute the noncommutative rank of a symbolic matrix. We show in section 4.3.2 that an operator satisfying the spectral gap condition has full noncommutative rank.

In solving the orbit intersection problem [1], the result of a generalization of the Paulsen problem to the operator setting in [47] was used. As in Theorem 1.12, we prove a much stronger bound in section 4.3.3 on the squared distance when the operator satisfies the spectral gap condition.

1.5. Techniques. We are not aware of previous work on spectral analysis of matrix scaling and operator scaling. The closest work in this direction that we are aware of is a recent work by Rudelson, Samorodnitsky, and Zeitouni [55], who analyze the condition number of the matrix scaling solution when the matrix satisfies some strong (vertex) expansion property using a combinatorial argument.

In the following, we discuss the previous techniques used in analyzing the continuous operator scaling algorithm, and then discuss the techniques used in this paper.

1.5.1. Comparisons with previous techniques. The operator capacity defined by Gurvits [31] was used crucially as a potential function to analyze the discrete operator scaling algorithms in [31, 22] as well as the continuous operator scaling algorithm in [47].

A smoothed analysis of matrix scaling was presented in [47] for solving the Paulsen problem, which also gave a condition for linear convergence of gradient flow. It was shown that if most of the entries of an $m \times n$ matrix with $m \leq n$ are at least σ^2 for a large enough σ, then the continuous matrix scaling algorithm has linear convergence with rate at least $\sigma^2 n$. This combinatorial assumption is restrictive and only applies in the matrix scaling setting. Note that the convergence result in this work is strictly stronger, as the combinatorial assumption implies the spectral gap assumption in Definition 1.4 with $\lambda \geq \Omega(\sigma^2)$ but not vice versa. Another advantage is that the natural spectral condition defined in this work guarantees linear convergence directly in the general operator setting, whereas the combinatorial condition in [47] could only be lifted to the frame setting with great difficulty. As a consequence, we do not need to go through the operator capacity to analyze the convergence rate of the operator scaling algorithm, which is different from previous analyses. Nonetheless, we can use the linear convergence to prove a lower bound on the operator capacity as was done in [47]. We leave it as an open question whether the spectral condition can be used to replace the combinatorial assumption in [47] to simplify and extend the main results there.

1.5.2. Outline of spectral analysis. We illustrate the main ideas of the spectral analysis in the simpler matrix scaling setting and mention how these ideas can be generalized to the operator setting. For gradient descent, a common approach to prove linear convergence is to show that the Hessian matrix has small condition number. Instead, our approach is to directly analyze the change of Δ. In the matrix scaling setting, it follows from Lemma 4.2.9 in [47] that

$$-\frac{1}{4}\frac{d}{dt}\Delta = \sum_{i=1}^{m}(s - mr_i)^2 r_i + \sum_{j=1}^{n}(s - nc_j)^2 c_j + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(s - mr_i)(s - nc_j)B_{ij},$$

where $B \in \mathbb{R}^{m \times n}$ is the current nonnegative matrix, and s, r_i, c_j are the size, the ith row sum, and the jth column sum of B, respectively. We call the first two terms in the right-hand side the quadratic terms and the last term the cross term. Our goal is to lower bound their sum by $\lambda s \Delta$. To do so, we will prove a lower bound on the sum of the quadratic terms and an upper bound on the absolute value of the cross term.

First, we prove a structural result that the maximum violation of a row and a column will not increase much throughout the continuous matrix scaling algorithm, and then we use this to show that the sum of the quadratic terms is at least $(1 - \epsilon)s\Delta$ for an ϵ-nearly doubly balanced matrix B. Then, we write the cross term as a quadratic form of the matrix B as $\vec{r}B\vec{c}$, where $\vec{r} \in \mathbb{R}^m$ is the vector with the ith entry being $s - mr_i$ and $\vec{c} \in \mathbb{R}^n$ is the vector with the jth entry being $s - nc_j$. The observation is that $\vec{r} \perp \vec{1_m}$ and $\vec{c} \perp \vec{1_n}$ while $\vec{1_m}, \vec{1_n}$ are close to the first singular vectors of B, so the cross term would be small if there is a spectral gap of the matrix B. By a spectral argument, we can show that the absolute value of the cross term is at most $(1 + \epsilon - \lambda)s\Delta$. Combining these two bounds, we can lower bound the convergence rate to be at least $4(\lambda - 4\epsilon)s\Delta$ initially.

To prove that the convergence rate is at least $\lambda s \Delta$ for all time, we need to prove that the spectral gap condition is maintained throughout the continuous matrix scaling algorithm. To do so, we argue through the condition number of the scaling solutions. We use the structural result and the linear convergence to show that the condition number of the scaling solution is small, and then we show that the singular values of the matrix would not change much if we scale the matrix B by diagonal matrices of small condition numbers. Finally, we use an inductive argument to prove that

the linear convergence is maintained for all time. The results for condition numbers and capacity follow from the arguments developed and the linear convergence.

The proof for the general operator setting has the same structure, with more involved technical details in some steps. To prove the structural result that the operator norm of the error matrices would not increase much throughout the continuous operator scaling algorithm, we need to use the envelope theorem to bound the maximum eigenvalue and the minimum eigenvalue. To bound the condition number of the scaling solutions, we need to use results from the theory of product integration to analyze the scaling solutions. For readers who are more interested in matrix scaling and/or who would like to understand the spectral analysis in a simpler setting first, we include a self-contained proof for the matrix scaling case in Appendix B even though the matrix scaling result is completely generalized by the operator scaling result.

1.6. Subsequent work. After the submission of this article, we learned of applications and improvements of the results in this work. Franks and Moitra [20] use Theorem 1.5 to prove nearly optimal sample guarantees for an important estimation problem in statistics. In the process, they improve the result of Theorem 5.1 to require only $n \geq \Omega(d \log^2 d)$ vectors (instead of our bound $n \geq \Omega(d^{4/3})$). The improved theorem is discussed in more detail in Remark 5.2. In order to give a rigorous analysis of the natural scaling algorithm for finding the estimator in their application, they also extend our analysis of discrete gradient flow to show that Sinkhorn's alternating algorithm also has linear convergence for operator scaling inputs satisfying the spectral condition. This result is discussed in more detail in Remark 3.24.

1.7. Organization. We first review some background about completely positive linear operators and the continuous operator scaling algorithm in section 2. We then prove the main technical results in section 3 and show various applications in section 4. We provide a proof in section 5 that a random frame satisfies the spectral condition with high probability. In Appendix B, we provide a self-contained proof of Theorem 1.5 in the special case of matrix scaling.

2. Preliminaries. We first review in section 2.1 some background in quantum information theory about completely positive maps and discuss the spectral gap condition stated in Definition 1.4. Then, we review the known results about the continuous operator scaling algorithm in section 2.2.

2.1. Positive linear maps, matrix representations, quantum expanders. First, we define completely positive linear maps and their natural matrix representation in section 2.1.1. Then, in section 2.1.2, we present the spectral gap condition in Definition 1.4 using this language, and compare to the notion of quantum expanders studied in the literature. Finally, we introduce the Choi matrix in section 2.1.3 and state some facts about tensors and completely positive maps that we will use in our proof.

2.1.1. Completely positive linear map. Given $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$ for $1 \leq i \leq k$, it can be used to define a linear map $\Phi : \mathbb{R}^{n \times n} \to \mathbb{R}^{m \times m}$ as

$$(2.1) \qquad \Phi_{\mathcal{A}}(Y) = \sum_{i=1}^{k} A_i Y A_i^* \quad \text{and} \quad \Phi_{\mathcal{A}}^*(X) = \sum_{i=1}^{k} A_i^* X A_i,$$

where $\Phi^* : \mathbb{R}^{m \times m} \to \mathbb{R}^{n \times n}$ is the adjoint map so that $\langle X, \Phi(Y) \rangle = \langle \Phi^*(X), Y \rangle$ for any $X \in \mathbb{R}^{m \times m}$ and $Y \in \mathbb{R}^{n \times n}$, where $\langle P, Q \rangle := \operatorname{tr}(P^* Q) = \sum_{i,j} P_{ij}^* Q_{ij}$ is the Hilbert–Schmidt inner product.

DEFINITION 2.1 (completely positive map). *A linear map Φ is positive if $\Phi(Y) \succeq 0$ for every $Y \succeq 0$, where $Y \succeq 0$ denotes that Y is a positive semidefinite matrix. A linear map Φ is completely positive if $\Phi \otimes \mathcal{I}_l$ is positive for every natural number $l \geq 1$, where $\mathcal{I}_l$ is the identity mapping on $l \times l$ matrices (see [63] for more details).*

THEOREM 2.2 (Choi [13]). *A linear map Φ is completely positive if and only if it can be written as the form described in* (2.1).

The matrices $A_1, \ldots, A_k$ are called the Kraus operators of Φ. Note that the Kraus operators are not uniquely defined for a linear map Φ.

DEFINITION 2.3 (doubly balanced map). *A linear map Φ is called unital if $\Phi(I_n) = I_m$. A linear map Φ is called trace preserving if $\Phi^*(I_m) = I_n$ (which implies that $\mathrm{tr}(\Phi(Y)) = \mathrm{tr}(Y)$ for any $Y \in \mathbb{R}^{n \times n}$). A linear map Φ is called doubly balanced if there exists $c > 0$ such that $c\sqrt{n}\Phi$ is unital and $c\sqrt{m}\Phi$ is trace preserving.*

Using this terminology, the operator scaling problem can be rephrased as given the Kraus operators $(A_1, \ldots, A_k)$ of a completely positive map, find a left scaling matrix L and a right scaling matrix R so that the completely positive map defined by the Kraus operators $(LA_1R, \ldots, LA_kR)$ is nonzero doubly balanced.

For each completely positive linear map Φ, we can associate a matrix representation describing the same linear transformation.

DEFINITION 2.4 (natural matrix representation of linear map). *Given a linear map $\Phi : \mathbb{R}^{n \times n} \to \mathbb{R}^{m \times m}$, we can interpret it as a matrix $M_\Phi : \mathbb{R}^{n^2} \to \mathbb{R}^{m^2}$ by vectorizing the input and output matrices such that*

$$M_{\mathcal{A}} \cdot \mathrm{vec}(Y) = \mathrm{vec}(\Phi(Y)),$$

where $\mathrm{vec} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n^2}$ is the linear map satisfying $\mathrm{vec}(E_{i,j}) = e_i \otimes e_j$ for all $1 \leq i, j \leq n$, where $E_{i,j}$ is the $n \times n$ matrix with one in the (i,j)th entry and zero otherwise and $e_i \in \mathbb{R}^n$ is the vector with one in the ith entry and zero otherwise.

There is a one-to-one correspondence between the matrix representations and the linear maps. Given a matrix $M : \mathbb{R}^{n^2} \to \mathbb{R}^{m^2}$, we can also interpret it as a map $\Phi_M : \mathbb{R}^{n \times n} \to \mathbb{R}^{m \times m}$ by matrixizing the input and output vectors such that

$$\Phi_M(\mathrm{mat}(y)) = \mathrm{mat}(M_{\mathcal{A}} \cdot y),$$

where $\mathrm{mat} : \mathbb{R}^{n^2} \to \mathbb{R}^{n \times n}$ is the linear map satisfying $\mathrm{mat}(e_i \otimes e_j) = E_{i,j}$.

The matrix representation of a completely positive map has a nice form in terms of its Kraus operators.

FACT 2.5 (Proposition 2.20 in [63]). *Given a completely positive map $\Phi_{\mathcal{A}}$ with Kraus operators $\mathcal{A}$, the matrix representation $M_{\mathcal{A}}$ can be written in the form described in Definition* 1.4 *such that*

$$M_{\mathcal{A}} = \sum_{i=1}^{k} A_i \otimes A_i.$$

2.1.2. Spectral gap condition and quantum expanders. Given the correspondence between the completely positive linear map $\Phi_{\mathcal{A}}$ and the natural matrix representation $M_{\mathcal{A}}$, the spectral gap condition in Definition 1.4 can be presented as follows.

DEFINITION 2.6 (operator). *An operator $\mathcal{A}$ is defined by a tuple of matrices $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$ for $1 \leq i \leq k$.*

DEFINITION 2.7 (spectral gap condition of Φ). *Given an operator $\mathcal{A}$ let*

$$\sigma_1(\Phi_{\mathcal{A}}) := \max_{Y \in \mathbb{R}^{n \times n}} \frac{\|\Phi(Y)\|_F}{\|Y\|_F} = \max_{y \in R^{n^2}} \frac{\|M_{\mathcal{A}} \cdot y\|_2}{\|y\|_2} = \sigma_1(M_{\mathcal{A}})$$

and Y_1, y_1 as maximizers to the optimization problems with $y_1 = \text{vec}(Y_1)$. Let

$$\sigma_2(\Phi_{\mathcal{A}}) := \max_{Y \in \mathbb{R}^{n \times n}, \ \langle Y, Y_1 \rangle = 0} \frac{\|\Phi(Y)\|_F}{\|Y\|_F} = \max_{y \in R^{n^2}, \ y \perp y_1} \frac{\|M_{\mathcal{A}} \cdot y\|_2}{\|y\|_2} = \sigma_2(M_{\mathcal{A}}).$$

The spectral gap condition in Definition 1.4 is equivalent to $\sigma_2(\Phi_{\mathcal{A}}) \leq (1-\lambda)s(\mathcal{A})/\sqrt{mn}$. Here $\{\sigma_1(M_{\mathcal{A}}) \geq \sigma_2(M_{\mathcal{A}}) \geq \ldots\}$ are the singular values repeated according to multiplicity.

The concept of quantum expander was studied by Hastings [37] and Ben-Aroya, Schwartz, and Ta-Shma [7], which was stated using the above language with $m = n$.

DEFINITION 2.8 (quantum expander [37, 7]). *An operator $\mathcal{A} = (A_1, \ldots, A_k)$ where each $A_i \in \mathbb{R}^{n \times n}$ is called a $(1-\lambda)$-quantum expander if*
1. *the largest singular value is $s(\mathcal{A})/n$ and the identity matrix I_n is the largest left and right singular vector, i.e.,*

$$\sigma_1(\Phi_{\mathcal{A}}) = \frac{\|\Phi(I_n)\|_F}{\|I_n\|_F} = \frac{s(\mathcal{A})}{n};$$

2. *for any Y orthogonal to I_n, it holds that*

$$\sigma_2(\Phi_{\mathcal{A}}) = \max_{Y: \langle Y, I_n \rangle = 0} \frac{\|\Phi(Y)\|_F}{\|Y\|_F} \leq (1-\lambda) \frac{\|\Phi(I_n)\|_F}{\|I_n\|_F} = \frac{(1-\lambda)s(\mathcal{A})}{n}.$$

In [7, 37], the map Φ is defined as $\frac{1}{k} \sum_{i=1}^{k} U_i Y U_i^*$, where $U_i \in \mathbb{R}^{n \times n}$ is a unitary matrix. Then, the size of this operator is equal to n, and the largest singular value is 1 achieved at the identity matrix.

When the operator $\mathcal{A}$ is ϵ-nearly doubly balanced, we will show in Lemma 3.6 that $\sigma_1(\Phi_{\mathcal{A}}) \leq (1+\epsilon)s(\mathcal{A})/\sqrt{mn}$ and I_n is an approximate optimizer. Therefore, in the case $m = n$, the spectral gap condition in Definition 1.4 is a more relaxed version of the quantum expander definition in [7], where we do not require I_n to be the optimizer (but only an approximate optimizer).

From random matrix theory [60], almost all random nonnegative matrices (from reasonable distributions) have a constant spectral gap, i.e., λ is a constant. For random operators, Hastings [37] proved that the operator $\mathcal{A}$ has spectral gap approaching the optimal value of $\lambda = 1 - 2\sqrt{k-1}/k$ if each A_i is a random unitary matrix. This result has been extended recently by González-Guilén, Junge, and Nechita to more general distributions [26]. It is reasonable to expect that most random operators have a constant spectral gap. There are also deterministic constructions of quantum expanders [7]. See [7, 37] for some applications of quantum expanders.

2.1.3. Choi matrix and useful facts. There is another matrix representation that is useful in studying completely positive linear maps.

DEFINITION 2.9 (Choi matrix). *Given a completely positive linear map $\Phi_{\mathcal{A}} : \mathbb{R}^{n \times n} \to \mathbb{R}^{m \times m}$, the Choi matrix $Q_{\mathcal{A}} \in \mathbb{R}^{mn \times mn}$ is defined as*

$$Q_{\mathcal{A}} := \sum_{i=1}^{n} \sum_{j=1}^{n} \Phi_{\mathcal{A}}(E_{i,j}) \otimes E_{i,j}.$$

Using the Choi matrix, we can rephrase the operator scaling problem as finding left scaling matrix $L \in \mathbb{R}^{m \times m}$ and right scaling matrix $R \in \mathbb{R}^{n \times n}$ so that the scaled Choi matrix $P := (L \otimes R)Q(L \otimes R)^*$ satisfies

$$\mathrm{tr}_n(P) = \frac{s}{m}I_m \quad \text{and} \quad \mathrm{tr}_m(P) = \frac{s}{n}I_n,$$

where the partial trace operations tr_n and tr_m are linear functions that satisfy $\mathrm{tr}_n(X \otimes Y) := \mathrm{tr}(Y) \cdot X$ and $\mathrm{tr}_m(X \otimes Y) = \mathrm{tr}(X) \cdot Y$ for $X \in \mathbb{R}^{m \times m}$ and $Y \in \mathbb{R}^{n \times n}$. This phrasing of the operator scaling problem is in line with the more general quantum marginal problem [11].

The following facts will be useful in our proofs.

FACT 2.10. *In the following, $\Phi_{\mathcal{A}}$ is the completely positive map with Kraus operators $\mathcal{A} = (A_1, \ldots, A_k)$ where each $A_i \in \mathbb{R}^{m \times n}$.*
   1. *For any matrices $A, X \in \mathbb{R}^{m \times m}$ and $B, Y \in \mathbb{R}^{n \times n}$,*

$$(A \otimes B)(X \otimes Y) = AX \otimes BY \quad \text{and} \quad \langle A \otimes B, X \otimes Y \rangle = \langle A, X \rangle \langle B, Y \rangle.$$

   2. $\Phi_{\mathcal{A}}(Y) \succeq 0$ *for any $Y \succeq 0$.*
   3. *For any symmetric matrices $X \in \mathbb{R}^{m \times m}$ and $Y \in \mathbb{R}^{n \times n}$,*

$$\langle Q_{\mathcal{A}}, X \otimes Y \rangle = \langle X, \Phi_{\mathcal{A}}(Y) \rangle = \langle \Phi_{\mathcal{A}}^*(X), Y \rangle.$$

   4. *Let $L \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{n \times n}$ and define the scaled operator $L\mathcal{A}R :=\{LA_1R, \ldots, LA_kR\}$. Then,*

$$\Phi_{L\mathcal{A}R}(I_n) = L \cdot \Phi_{\mathcal{A}}(RR^*) \cdot L^* \quad \text{and} \quad \Phi_{L\mathcal{A}R}^*(I_m) = R^* \cdot \Phi_{\mathcal{A}}^*(L^*L) \cdot R.$$

2.2. Continuous operator scaling. The continuous operator scaling algorithm was studied in [47]. We collect the definitions and the results that we will use in this subsection. We start with some definitions about operator scaling that we have already stated in the introduction.

2.2.1. Operator scaling.

DEFINITION 2.11 (size of an operator). *The size of an operator $\mathcal{A}$ is defined as*

$$s(\mathcal{A}) = \sum_{i=1}^{k} \|A_i\|_F^2 = \sum_{i=1}^{k} \mathrm{tr}(A_i A_i^*) = \mathrm{tr}(\Phi_{\mathcal{A}}(I_n)).$$

DEFINITION 2.12 (ϵ-nearly doubly balanced operator). *An operator $\mathcal{A}$ is called ϵ-nearly doubly balanced if*

$$(1 - \epsilon)\frac{s}{m}I_m \preceq \sum_{i=1}^{k} A_i A_i^* = \Phi_{\mathcal{A}}(I_n) \preceq (1 + \epsilon)\frac{s}{m}I_m \quad \text{and}$$

$$(1 - \epsilon)\frac{s}{n}I_n \preceq \sum_{i=1}^{k} A_i^* A_i = \Phi_{\mathcal{A}}^*(I_m) \preceq (1 + \epsilon)\frac{s}{n}I_n.$$

$\mathcal{A}$ *is called doubly balanced when $\epsilon = 0$.*

DEFINITION 2.13 (ℓ_2-error). *Given an operator $\mathcal{A}$, define*

$$\Delta(\mathcal{A}) = \frac{1}{m} \left\| sI_m - m\sum_{i=1}^{k} A_i A_i^* \right\|_F^2 + \frac{1}{n} \left\| sI_n - n\sum_{i=1}^{k} A_i^* A_i \right\|_F^2$$

$$= \frac{1}{m} \operatorname{tr}\left(sI_m - m\sum_{i=1}^{k} A_i A_i^* \right)^2 + \frac{1}{n} \operatorname{tr}\left(sI_n - n\sum_{i=1}^{k} A_i^* A_i \right)^2.$$

DEFINITION 2.14 (error matrices). *We define the error matrices as*

$$E := sI_m - m\sum_{i=1}^{k} A_i A_i^* \quad \text{and} \quad F := sI_n - n\sum_{i=1}^{k} A_i^* A_i.$$

Note that $\operatorname{tr}(E) = \operatorname{tr}(F) = 0$, *as*

$$\operatorname{tr}(E) = \operatorname{tr}\left(sI_m - m\sum_{i=1}^{k} A_i A_i^* \right) = sm - m\sum_{i=1}^{k} \operatorname{tr}(A_i A_i^*) = 0,$$

where the last equality is by Definition 2.11. *Also, we write*

$$\Delta_E := \frac{1}{m} \|E\|_F^2 = \frac{1}{m} \operatorname{tr}(E^2) \quad \text{and} \quad \Delta_F := \frac{1}{n} \|F\|_F^2 = \frac{1}{n} \operatorname{tr}(F^2)$$

so that $\Delta = \Delta_E + \Delta_F$.

The ℓ_2-error is bounded for an ϵ-nearly doubly balanced operator.

LEMMA 2.15 (Lemma 3.6.1 in [47]). *For an ϵ-nearly doubly balanced operator $\mathcal{A}$,*

$$\Delta(\mathcal{A}) \leq 2\epsilon^2 s(\mathcal{A})^2.$$

2.2.2. Dynamical system.

DEFINITION 2.16 (dynamical system). *The following dynamical system describes how $\mathcal{A}$ changes over time in the continuous operator scaling algorithm:*

$$\frac{d}{dt} A_i := \left(sI_m - m\sum_{j=1}^{k} A_j A_j^* \right) A_i + A_i \left(sI_n - n\sum_{j=1}^{k} A_j^* A_j \right) = EA_i + A_i F \quad \text{for } 1 \leq i \leq k.$$

We show in Lemma A.1 in Appendix A that the dynamical system is equivalent to the gradient flow with potential function $\Delta(\mathcal{A})$.

It is shown in [47] that the dynamical system will converge to a solution $\mathcal{A}^{(\infty)}$ with $\Delta(\mathcal{A}^{(\infty)}) = 0$. The following lemmas describe how the different quantities evolve in the dynamical system. We use the superscript $^{(t)}$ to represent the quantity of interest at time t in the dynamical system and omit it when the time t is clear from context.

LEMMA 2.17 (Lemma 3.4.2 in [47]). *The change of the size of the operator $\mathcal{A}^{(t)}$ at time t is*

$$\frac{d}{dt} s^{(t)} = -2\Delta^{(t)}.$$

The following lemma was proved directly in [47]. It can also be seen as a consequence that the dynamical system is the gradient flow on Δ.

LEMMA 2.18 (Lemma 3.4.3 in [47]). *The change of $\Delta^{(t)}$ at time t is*

$$\frac{d}{dt}\Delta^{(t)} = -4\left(\sum_{i=1}^{k}\left\|\frac{d}{dt}A_i^{(t)}\right\|_F^2\right).$$

The following result was used in [47] for the smoothed analysis when the dynamical system has linear convergence.

LEMMA 2.19 (Proposition 4.3.1 in [47]). *Suppose there exists $\mu > 0$ such that for all $0 \le t \le T$,*

$$-\frac{d}{dt}\Delta^{(t)} \ge \mu\Delta^{(t)}.$$

Then

$$\Delta^{(T)} \le \Delta^{(0)}e^{-\mu T} \quad \text{and} \quad s^{(0)} - s^{(T)} \le \frac{2\Delta^{(0)}}{\mu}.$$

2.2.3. Operator capacity.

DEFINITION 2.20 (capacity). *The capacity of an operator $\mathcal{A}$ is defined as*

$$\text{cap}(\mathcal{A}) := \inf_{X \succ 0} \frac{m\det\left(\sum_{i=1}^{k}A_iXA_i^*\right)^{1/m}}{\det(X)^{1/n}}.$$

It was shown in [47] that the convergence rate of Δ can be used to derive a lower bound on operator capacity.

PROPOSITION 2.21 (Proposition 4.3.1 in [47]). *Suppose there exists $\mu > 0$ such that for all $t \ge 0$, it holds that*

$$-\frac{d}{dt}\Delta^{(t)} \ge \mu\Delta^{(t)}.$$

Then, it follows that

$$\text{cap}^{(0)} \ge s^{(0)} - \frac{2\Delta^{(0)}}{\mu}.$$

3. Spectral analysis of operator scaling. We prove the main technical results in this section.

3.1. Overview. The main goal is to show that the dynamical system in Definition 2.16 has linear convergence. Let $\mathcal{A}$ be an ϵ-nearly doubly balanced operator with λ-spectral gap. Assuming $\lambda^2 \ge C\epsilon \ln m$ for a sufficiently large constant C, we will prove that for all time $t \ge 0$,

$$-\frac{d}{dt}\Delta^{(t)} \ge \lambda s^{(0)}\Delta^{(t)}.$$

We start by looking more closely at the expression for the change of Δ.

LEMMA 3.1. *The change of Δ is*

$$-\frac{1}{4}\frac{d}{dt}\Delta = \langle E^2, \Phi(I_n)\rangle + \langle F^2, \Phi^*(I_m)\rangle + 2\langle Q, E \otimes F\rangle.$$

Proof. By Lemma 2.18 and Definition 2.16,

$$-\frac{1}{4}\frac{d}{dt}\Delta = \sum_{i=1}^{k}\left\|\frac{d}{dt}A_i\right\|_F^2 = \sum_{i=1}^{k}\langle EA_i + A_iF, EA_i + A_iF\rangle$$

$$= \left\langle E^2, \sum_{i=1}^{k}A_iA_i^*\right\rangle + \left\langle \sum_{i=1}^{k}A_i^*A_i, F^2\right\rangle + 2\left\langle E, \sum_{i=1}^{k}A_iFA_i^*\right\rangle$$

$$= \langle E^2, \Phi(I_n)\rangle + \langle F^2, \Phi^*(I_m)\rangle + 2\langle E, \Phi(F)\rangle,$$

and the lemma follows from Fact 2.10.3 that $\langle E, \Phi(F)\rangle = \langle Q, E \otimes F\rangle$. □

We call the terms $\langle E^2, \Phi(I_m)\rangle$ and $\langle F^2, \Phi^*(I_n)\rangle$ the quadratic terms as they are always nonnegative, and we call the term $2\langle Q, E \otimes F\rangle$ the cross term. The proof outline is the following:

1. In section 3.2, we prove a structural result that bounds the operator norms of $E^{(t)}$ and $F^{(t)}$ throughout the dynamical system using the envelope theorem. This implies a bound on the operator norm of $\Phi^{(t)}(I_n)$ and $\Phi^{(t)^*}(I_m)$, which is used to show that the sum of the quadratic terms is at least $(1 - \epsilon)s\Delta$.
2. In section 3.3, we bound the largest singular value of the matrix $M_{\mathcal{A}}$ and show that I is an approximate largest singular vector, and then we use a spectral argument to upper bound the absolute value of the cross term to be at most $(1 + \epsilon - \lambda)s\Delta$.
3. These two parts combine to show that $-\Delta' \geq \lambda s\Delta$ when the spectral gap condition holds. To prove the linear convergence for all time $t \geq 0$, we need to prove that the spectral gap condition is maintained throughout the dynamical system. To do this, we bound the condition number of the scaling solutions in section 3.5 and use it to conclude that the spectral gap condition and the linear convergence hold throughout in section 3.6.

In sections 3.7 and 3.8, we use the results to prove Theorems 1.7 and 1.8 about condition number and operator capacity, respectively.

Finally, in section 3.9, we explain how to discretize the gradient flow to obtain a discrete algorithm with linear convergence under the spectral assumption.

3.2. Lower bounding the quadratic terms. First, we prove a structural result bounding the operator norm of the error matrices $E^{(t)}$ and $F^{(t)}$ for all $t \geq 0$ in Proposition 3.2, which will also be useful in bounding the condition number of the scaling solution in section 3.5. Then we will use this to prove Proposition 3.4, which will allow us to lower bound the quadratic terms in Lemma 3.5.

PROPOSITION 3.2. *If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced, then for any $t \geq 0$,*

$$\|E^{(t)}\|_{\mathrm{op}} \leq (1 + \epsilon)s^{(0)} - s^{(t)} \quad \text{and} \quad \|F^{(t)}\|_{\mathrm{op}} \leq (1 + \epsilon)s^{(0)} - s^{(t)}.$$

Proof. The main idea is to show that the change of the quadratic form $\frac{d}{dt}u^*E^{(t)}u$ in the direction u achieving $\|E^{(t)}\|_{\mathrm{op}}$ is at most $2\Delta^{(t)}$, and then to use it to conclude that $\|E^{(t)}\|_{\mathrm{op}} \leq \|E^{(0)}\|_{\mathrm{op}} + \int_0^t 2\Delta^{(\tau)}d\tau$ to complete the proof using Lemma 2.17. Note that the direction u achieving $\|E^{(t)}\|_{\mathrm{op}}$ varies over time t. To turn this idea into a formal proof, we use the generalized envelope theorem proven by Milgrom and Segal [51].

THEOREM 3.3 (Corollary 4 in Milgrom and Segal [51]). *Suppose that X is a nonempty compact space, $f(x,t)$ is continuous in x, and $\frac{\partial}{\partial t}f(x,t)$ is continuous in*

(x, t). *Then the function $g(t) = \max_{x \in X} f(x, t)$ is differentiable almost everywhere and satisfies*

$$g(\tau) = g(0) + \int_0^\tau \frac{\partial}{\partial t} f(x, t)|_{x = x^*(t)} dt,$$

where $x^*(t)$ is any optimizer at time t satisfying $g(t) = f(x^*(t), t)$.

To apply the theorem, we define the space X to be $\{0, 1\} \times \{0, 1\} \times \mathbb{S}^{m-1} \times \mathbb{S}^{n-1}$, which is clearly nonempty and compact. The first coordinate indicates whether we are considering the error matrix E or F. The second coordinate indicates whether we are considering the largest or smallest eigenvalue of the error matrix. The third and fourth coordinates indicate the unit test vectors we are applying to E and F. The function f is defined as follows:

$$f(0, 0, u, v, t) = u^* E^{(t)} u,$$
$$f(0, 1, u, v, t) = -u^* E^{(t)} u,$$
$$f(1, 0, u, v, t) = v^* F^{(t)} v,$$
$$f(1, 1, u, v, t) = -v^* F^{(t)} v.$$

It is clear that $f(x, t)$ is continuous in $x \in X$ and $\frac{\partial}{\partial t} f(x, t)$ is continuous in (x, t). Hence, by Theorem 3.3, the function $g(t) = \max_{x \in X} f(x, t)$ satisfies

$$g(\tau) = g(0) + \int_0^\tau \frac{\partial}{\partial t} f(x, t)|_{x = x^*(t)} dt.$$

Since $E^{(t)}$ and $F^{(t)}$ are Hermitian matrices,

$$g(t) = \max \left\{ \lambda_{\max}(E^{(t)}), -\lambda_{\min}(E^{(t)}), \lambda_{\max}(F^{(t)}), -\lambda_{\min}(F^{(t)}) \right\}$$
$$= \max \left\{ \|E^{(t)}\|_{\mathrm{op}}, \|F^{(t)}\|_{\mathrm{op}} \right\},$$

and so $g(0) \leq \epsilon s^{(0)}$ by the assumption that $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced. To compute the partial derivative, we consider the four cases of the optimizer $x^*(t)$ at time t one by one.

1. $x^*(t) = (0, 0, u, v)$. As $E^{(t)}$ and $F^{(t)}$ are Hermitian matrices, the optimizer u of $\|E^{(t)}\|_{\mathrm{op}}$ is a maximum eigenvector of $E^{(t)}$ satisfying $E^{(t)} u = g(t) \cdot u$, and $F^{(t)} \succeq -g(t) \cdot I_n$ as $\|E^{(t)}\|_{\mathrm{op}} \geq \|F^{(t)}\|_{\mathrm{op}}$ in this case. Then, by the definition of $\frac{d}{dt} A_i^{(t)}$ in Definition 2.16 and $\frac{d}{dt} s^{(t)} = -2\Delta^{(t)}$ from Lemma 2.17, it follows that

$$\frac{\partial}{\partial t} f(x, t)|_{x = (0, 0, u, v)} = u^* \frac{d}{dt} \left(s I_m - m \sum_{i=1}^k A_i A_i^* \right) u$$

$$= -2\Delta - 2m \sum_{i=1}^k u^* (E A_i + A_i F) A_i^* u$$

$$= -2\Delta - 2m u^* E \left(\sum_{i=1}^k A_i A_i^* \right) u - 2m u^* \left(\sum_{i=1}^k A_i F A_i^* \right) u$$

$$\leq -2\Delta - 2m g(t) u^* \left(\sum_{i=1}^k A_i A_i^* \right) u + 2m g(t) u^* \left(\sum_{i=1}^k A_i A_i^* \right) u$$

$$= -2\Delta,$$

where the inequality follows from $\sum_{i=1}^{k} A_i F A_i^* = \Phi_{\mathcal{A}}(F) \succeq \Phi_{\mathcal{A}}(-g(t) \cdot I_n) = -g(t) \sum_{i=1}^{k} A_i A_i^*$ by Fact 2.10.2.

2. $x^*(t) = (0, 1, u, v)$. In this case, $E^{(t)} u = -g(t) \cdot u$, $F^{(t)} \preceq g(t) \cdot I_n$ and by similar calculations to the first case, we have

$$\frac{\partial}{\partial t} f(x, t)|_{x=(0,1,u,v)} = -u^* \frac{d}{dt}\left(s I_m - m \sum_{i=1}^{k} A_i A_i^*\right) u \le 2\Delta.$$

3. $x^*(t) = (1, 0, u, v)$. By symmetry of $E^{(t)}$ and $F^{(t)}$, we get the same bound as the first case:

$$\frac{\partial}{\partial t} f(x, t)|_{x=(1,0,u,v)} \le -2\Delta.$$

4. $x^*(t) = (1, 1, u, v)$. By symmetry of $E^{(t)}$ and $F^{(t)}$, we get the same bound as the second case:

$$\frac{\partial}{\partial t} f(x, t)|_{x=(1,1,u,v)} \le 2\Delta.$$

Therefore, in any case we have $f_t(x^*(t), t) \le 2\Delta(t)$, and we conclude that

$$g(t) \le \epsilon s^{(0)} + \int_0^t 2\Delta^{(\tau)} d\tau = \epsilon s^{(0)} - \int_0^t \frac{d}{d\tau} s^{(\tau)} d\tau = \epsilon s^{(0)} + s^{(0)} - s^{(t)},$$

where the first equality is by Lemma 2.17 that $\frac{d}{dt} s^{(t)} = -2\Delta^{(t)}$. $\qquad\square$

We have the following corollary by rewriting the conclusions of Proposition 3.2 using the definitions that $E^{(t)} = s I_m - m \Phi^{(t)}(I_n)$ and $F^{(t)} = s I_n - n \Phi^{(t)^*}(I_m)$.

PROPOSITION 3.4. *If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced, then for any $t \ge 0$,*

$$\frac{2 s^{(t)} - (1+\epsilon) s^{(0)}}{m} I_m \preceq \Phi^{(t)}(I_n) \preceq \frac{(1+\epsilon) s^{(0)}}{m} I_m$$

and

$$\frac{2 s^{(t)} - (1+\epsilon) s^{(0)}}{n} I_n \preceq \Phi^{(t)^*}(I_m) \preceq \frac{(1+\epsilon) s^{(0)}}{n} I_n.$$

We can use Proposition 3.4 to lower bound the quadratic terms in Lemma 3.1.

LEMMA 3.5. *If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced, then for any $t \ge 0$,*

$$\langle (E^{(t)})^2, \Phi^{(t)}(I_n) \rangle + \langle (F^{(t)})^2, \Phi^{(t)^*}(I_m) \rangle \ge \left(2 s^{(t)} - (1+\epsilon) s^{(0)}\right) \Delta^{(t)}.$$

Proof. By Proposition 3.4 and the fact that $\langle X, Y \rangle \ge 0$ for positive semidefinite matrices X, Y,

$$\langle E^2, \Phi(I_n) \rangle + \langle F^2, \Phi^*(I_m) \rangle \ge \frac{2 s^{(t)} - (1+\epsilon) s^{(0)}}{m} \langle E^2, I_m \rangle + \frac{2 s^{(t)} - (1+\epsilon) s^{(0)}}{n} \langle F^2, I_n \rangle$$

$$= \left(2 s^{(t)} - (1+\epsilon) s^{(0)}\right) \left[\frac{1}{m} \|E\|_F^2 + \frac{1}{n} \|F\|_F^2\right]$$

$$= \left(2 s^{(t)} - (1+\epsilon) s^{(0)}\right) \Delta. \qquad\square$$

3.3. Upper bounding the cross term. We will first bound the largest singular value of the matrix $M_{\mathcal{A}}$ for any ϵ-nearly doubly balanced operator $\mathcal{A}$. Then, we will use a spectral argument to upper bound the absolute value of the cross term in Lemma 3.1.

Given a nonnegative matrix, it is known that the square of the largest singular value is bounded by the product of the maximum row sum and the maximum column sum (see [40]). The proof of this bound is generalized to prove the following lemma.

John Watrous provided a different proof of Lemma 3.6 by generalizing the proof of Theorem 4.27 in his book [63]. We include his proof in Lemma A.2 in Appendix A.

LEMMA 3.6. *If $\mathcal{A}$ is an ϵ-nearly doubly balanced operator, then the largest singular value of its matrix representation $M_{\mathcal{A}}$ in Definition 1.4 is*

$$\sigma_1(M_{\mathcal{A}}) \leq (1+\epsilon)\frac{s(\mathcal{A})}{\sqrt{mn}}.$$

Proof. Given a vector norm $\|\cdot\|$, we can define an induced matrix norm $\||M\|| := \sup_x \|Mx\| / \|x\|$. To prove the lemma, we define the vector norm for vectors in $\mathbb{R}^{n^2}$ for any n and its induced matrix norm for matrices in $\mathbb{R}^{m^2 \times n^2}$ for any m as

$$\|x\|_{\mathrm{op}} := \|\mathrm{mat}(x)\|_{\mathrm{op}} \quad \text{and} \quad \||M\||_{\mathrm{op}} := \sup_x \frac{\|Mx\|_{\mathrm{op}}}{\|x\|_{\mathrm{op}}},$$

where $\mathrm{mat}(\cdot)$ is the matrixizing operation in Definition 2.4 and $\|x\|_{\mathrm{op}}$ is the standard operator norm of a matrix.

For a positive semidefinite matrix $H \in \mathbb{R}^{m^2 \times m^2}$ for some m, we can bound its largest eigenvalue by this matrix norm, i.e., $\lambda_1(H) \leq \||H\||_{\mathrm{op}}$. To see this, let $v \in \mathbb{R}^{m^2}$ be an eigenvector with $Hv = \lambda_1 v$, then

$$\lambda_1 \|v\|_{\mathrm{op}} = \|\lambda_1 v\|_{\mathrm{op}} = \|Hv\|_{\mathrm{op}} \leq \||H\||_{\mathrm{op}} \|v\|_{\mathrm{op}} \quad \Longrightarrow \quad \lambda_1 \leq \||H\||_{\mathrm{op}}.$$

We apply this inequality to bound the largest singular value of $M_{\mathcal{A}}$, by considering the square matrix $M_{\mathcal{A}} M_{\mathcal{A}}^*$ and its largest eigenvalue:

$$\sigma_1(M_{\mathcal{A}})^2 = \lambda_1(M_{\mathcal{A}} M_{\mathcal{A}}^*) \leq \||M_{\mathcal{A}} M_{\mathcal{A}}^*\||_{\mathrm{op}} \leq \||M_{\mathcal{A}}\||_{\mathrm{op}} \||M_{\mathcal{A}}^*\||_{\mathrm{op}}.$$

As $M_{\mathcal{A}} \in \mathbb{R}^{m^2 \times n^2}$ is the natural matrix representation of the completely positive map $\Phi_{\mathcal{A}}$ defined by the operator $\mathcal{A}$,

$$\||M_{\mathcal{A}}\||_{\mathrm{op}} = \sup_{y \in \mathbb{R}^{n^2}} \frac{\|M_{\mathcal{A}} \cdot y\|_{\mathrm{op}}}{\|y\|_{\mathrm{op}}} = \sup_{Y \in \mathbb{R}^{n \times n}} \frac{\|\Phi_{\mathcal{A}}(Y)\|_{\mathrm{op}}}{\|Y\|_{\mathrm{op}}} = \|\Phi_{\mathcal{A}}(I_n)\|_{\mathrm{op}},$$

where the second equality is from Definition 2.4 and the last equality is by the theorem [9] that

$$\sup_{Y \in \mathbb{R}^{n \times n}} \frac{\|\Phi_{\mathcal{A}}(Y)\|_{\mathrm{op}}}{\|Y\|_{\mathrm{op}}} = \|\Phi_{\mathcal{A}}(I_n)\|_{\mathrm{op}} \quad \text{and} \quad \sup_{X \in \mathbb{R}^{m \times m}} \frac{\|\Phi_{\mathcal{A}}^*(X)\|_{\mathrm{op}}}{\|X\|_{\mathrm{op}}} = \|\Phi_{\mathcal{A}}^*(I_m)\|_{\mathrm{op}}.$$

By a similar argument, $\||M_{\mathcal{A}}^*\||_{\mathrm{op}} = \|\Phi_{\mathcal{A}}^*(I_m)\|_{\mathrm{op}}$. Therefore,

$$\sigma_1(M_{\mathcal{A}})^2 \leq \||M_{\mathcal{A}}\||_{\mathrm{op}} \||M_{\mathcal{A}}^*\||_{\mathrm{op}} = \|\Phi_{\mathcal{A}}(I_n)\|_{\mathrm{op}} \|\Phi_{\mathcal{A}}^*(I_m)\|_{\mathrm{op}} \leq (1+\epsilon)\frac{s(\mathcal{A})}{m} \cdot (1+\epsilon)\frac{s(\mathcal{A})}{n},$$

where the last inequality follows from the assumption that $\mathcal{A}$ is ϵ-nearly balanced in Definition 2.12. Taking the square root on both sides gives the lemma. $\square$

Lemma 3.6 implies that $\text{vec}(I_n)$ is an "approximate" first singular vector of $M_{\mathcal{A}}$. By the spectral gap condition in Definition 1.4, it will follow that any vector perpendicular to $\text{vec}(I_n)$ has a "small" quadratic form of $M_{\mathcal{A}}$, and this can be used to bound the cross term in Lemma 3.1. The following lemma summarizes the spectral argument, which will be used to bound the cross term in the next lemma.

LEMMA 3.7. *Let $A \in \mathbb{R}^{m \times n}$. Let $p \in \mathbb{R}^m$ and $q \in \mathbb{R}^n$ be unit vectors. Suppose the following assumptions hold:*

$$\sigma_1(A)^2 \leq 1 + \delta_1 \quad \text{and} \quad \sigma_2(A)^2 \leq 1 - \delta_2 \quad \text{and} \quad p^* A q = 1.$$

Then, for any unit vectors $x \perp p$ and $y \perp q$, it holds that $|x^* A y| \leq 1 + \delta_1 - \frac{\delta_2}{2}$.

Proof. First, we show that p and q are highly correlated with the first singular vectors of A. Let $A = \sum_i \sigma_i u_i v_i^*$ be its singular value decomposition with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ and $\{u_i\}$ and $\{v_i\}$ are orthonormal bases. Write p and q as linear combinations of singular vectors as $p = \sum_i c_i u_i$ and $q = \sum_i d_i v_i$. We will show that c_1 and d_1 are large. Observe that, since $I_m \succeq pp^*$,

$$\|Aq\|_2^2 = q^* A^* I_m A q \geq q^* A^* pp^* A q = 1,$$

and similarly $\|A^* p\|_2^2 \geq 1$. So we have

$$1 \leq \|A^* p\|_2^2 = \left\| \sum_i \sigma_i c_i v_i \right\|_2^2 = \sum_i \sigma_i^2 c_i^2 \leq \sigma_1^2 c_1^2 + \sigma_2^2 (1 - c_1^2),$$

where the last inequality is because $\sum_i c_i^2 = \|p\|_2^2 = 1$ and $\sigma_2^2 \geq \sigma_j^2$ for $j \geq 2$. Using our assumptions about σ_1 and σ_2, it follows that

$$1 \leq (1 + \delta_1) c_1^2 + (1 - \delta_2)(1 - c_1^2) = 1 + \delta_1 c_1^2 - \delta_2 (1 - c_1^2),$$

which implies that

$$\delta_2 (1 - c_1^2) \leq \delta_1 c_1^2 \quad \implies \quad \delta_2 \leq (\delta_1 + \delta_2) c_1^2 \quad \implies \quad c_1^2 \geq \frac{\delta_2}{\delta_1 + \delta_2}.$$

By the same calculation, we have $d_1^2 \geq \delta_2 / (\delta_1 + \delta_2)$.

Next, we show that $x \perp p$ and $y \perp q$ are not highly correlated with the first singular vectors. Write $x = \sum_i \alpha_i u_i$ and $y = \sum_i \beta_i v_i$ with $\sum_i \alpha_i^2 = \|x\|_2^2 = 1$ and $\sum_i \beta_i^2 = \|y\|_2^2 = 1$. We will show that α_1 and β_1 are small. Since $\langle x, p \rangle = 0$ by our assumption,

$$\sum_i \alpha_i c_i = 0 \implies |\alpha_1 c_1| = \left| \sum_{i \geq 2} \alpha_i c_i \right| \leq \sqrt{\sum_{i \geq 2} \alpha_i^2} \sqrt{\sum_{i \geq 2} c_i^2}$$

$$\implies \alpha_1^2 c_1^2 \leq (1 - \alpha_1^2)(1 - c_1^2) = 1 - \alpha_1^2 - c_1^2 + \alpha_1^2 c_1^2$$

$$\implies \alpha_1^2 \leq 1 - c_1^2 \leq 1 - \frac{\delta_2}{\delta_1 + \delta_2} = \frac{\delta_1}{\delta_1 + \delta_2}.$$

By the same calculation, we have $\beta_1^2 \leq \delta_1 / (\delta_1 + \delta_2)$.

Finally, we bound the absolute value of the quadratic form

$$|x^* A y| = \left| \left(\sum_i \alpha_i u_i \right) A \left(\sum_j \beta_j v_j \right) \right| = \left| \left(\sum_i \alpha_i u_i \right) \left(\sum_j \beta_j \sigma_j u_j \right) \right| = \left| \sum_i \alpha_i \beta_i \sigma_i \right|$$

$$\leq |\alpha_1 \beta_1 \sigma_1| + \sigma_2 \sum_{i \geq 2} |\alpha_i \beta_i|.$$

Using our assumptions on σ_1 and σ_2 and the Cauchy–Schwarz inequality,

$$|x^* A y| \leq \sqrt{1 + \delta_1} \cdot |\alpha_1 \beta_1| + \sqrt{1 - \delta_2} \cdot \sqrt{\sum_{i \geq 2} \alpha_i^2} \sqrt{\sum_{i \geq 2} \beta_i^2}$$
$$= \sqrt{1 + \delta_1} |\alpha_1 \beta_1| + \sqrt{1 - \delta_2} \sqrt{1 - \alpha_1^2} \sqrt{1 - \beta_1^2}.$$

Finally, using the Taylor approximation for $\sqrt{1 + \delta}$ and putting in the upper bounds on α_1^2 and β_1^2 derived above, we conclude

$$|x^* A y| \leq (1 + \delta_1) \frac{\delta_1}{\delta_1 + \delta_2} + \left(1 - \frac{\delta_2}{2} \right) \frac{\delta_2}{\delta_1 + \delta_2} = 1 + \frac{2\delta_1^2 - \delta_2^2}{2(\delta_1 + \delta_2)} \leq 1 + \frac{2\delta_1 - \delta_2}{2}. \qquad \square$$

We use Lemma 3.7 to bound the cross term in Lemma 3.1.

LEMMA 3.8. *If $\mathcal{A}$ satisfies the spectral gap condition in Definition* 1.4 *with the additional assumption that $\sigma_1(M_{\mathcal{A}}) \leq (1 + \delta)s(\mathcal{A})/\sqrt{mn}$ for $\delta \leq 1$, then*

$$2|\langle Q_{\mathcal{A}}, E \otimes F \rangle| \leq \left(1 + 3\delta - \frac{\lambda}{2} \right) s\Delta.$$

Proof. Note that the cross term

$$\langle Q_{\mathcal{A}}, E \otimes F \rangle = \langle E, \Phi_{\mathcal{A}}(F) \rangle = \text{vec}(E) \cdot M_{\mathcal{A}} \cdot \text{vec}(F),$$

where the first equality is by Fact 2.10.3 and the second equality is by the definition of matrix representation in Definition 2.4.

To prove the lemma, we apply Lemma 3.7 with

$$A := \frac{\sqrt{mn}}{s} M_{\mathcal{A}} \in \mathbb{R}^{m^2 \times n^2}, \quad p := \frac{1}{\sqrt{m}} \text{vec}(I_m) \in \mathbb{R}^{m^2}, \quad q := \frac{1}{\sqrt{n}} \text{vec}(I_n) \in \mathbb{R}^{n^2},$$

and

$$x := \frac{1}{\sqrt{m\Delta_E}} \text{vec}(E) \in \mathbb{R}^{m^2}, \quad y := \frac{1}{\sqrt{n\Delta_F}} \text{vec}(F) \in \mathbb{R}^{n^2}.$$

Clearly, p, q are unit vectors, and x, y are also unit vectors as $\|x\|_2 = \|E\|_F / \sqrt{m\Delta_E} = 1$ by Definition 2.14 and similarly $\|y\|_2 = 1$. Note that $x \perp p$ as $\langle x, p \rangle = \langle E, I_m \rangle / (m\sqrt{\Delta_E})$ and $\langle E, I_m \rangle = \text{tr}(E) = 0$ from Definition 2.14, and similarly $y \perp q$.

We check the assumptions of Lemma 3.7. By the additional assumption,

$$\sigma_1(A)^2 = \sigma_1(M_{\mathcal{A}})^2 \cdot \frac{mn}{s^2} \leq (1 + \delta)^2 \cdot \frac{s^2}{mn} \cdot \frac{mn}{s^2} = 1 + 2\delta + \delta^2,$$

and so we can set $\delta_1 := 2\delta + \delta^2$. By the spectral gap condition in Definition 1.4,

$$\sigma_2(A)^2 = \sigma_2(M_{\mathcal{A}})^2 \cdot \frac{mn}{s^2} \leq (1 - \lambda)^2 \cdot \frac{s^2}{mn} \cdot \frac{mn}{s^2} = 1 - 2\lambda + \lambda^2,$$

and so we can set $\delta_2 := 2\lambda - \lambda^2$. Also, we check that

$$p^* A q = \frac{1}{s} \text{vec}(I_m) \cdot M_{\mathcal{A}} \cdot \text{vec}(I_n) = \frac{1}{s} \text{tr}(\Phi_{\mathcal{A}}(I_n)) = 1,$$

where the second equality is from Definition 2.4 and the last equality is from Definition 2.11.

Therefore, we can conclude from Lemma 3.7 that

$$1 + 2\delta + \delta^2 - \frac{2\lambda - \lambda^2}{2} \geq |x^* A y| = \frac{1}{s\sqrt{\Delta_E \Delta_F}} |\langle \text{vec}(E), M_{\mathcal{A}} \cdot \text{vec}(F) \rangle|$$

$$= \frac{1}{s\sqrt{\Delta_E \Delta_F}} |\langle Q_{\mathcal{A}}, E \otimes F \rangle|.$$

Finally, we complete the proof using the inequality $\sqrt{\Delta_E \Delta_F} \leq (\Delta_E + \Delta_F)/2 = \Delta/2$, and $\delta \leq 1$ by our assumption, and $\lambda \leq 1$ by definition. □

3.4. Lower bounding the convergence rate. Putting the bounds in Lemmas 3.5 and 3.8 into Lemma 3.1, we obtain the following lower bound on the convergence rate of Δ at any time t.

PROPOSITION 3.9. *If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced and the matrix representation $M_{\mathcal{A}^{(t)}}$ of $\mathcal{A}^{(t)}$ satisfies the spectral conditions that*

$$\sigma_1(M_{\mathcal{A}^{(t)}}) = (1 + \delta^{(t)}) \frac{s^{(t)}}{\sqrt{mn}} \quad \text{and} \quad \sigma_2(M_{\mathcal{A}^{(t)}}) = (1 - \lambda^{(t)}) \frac{s^{(t)}}{\sqrt{mn}},$$

for continuous functions $\lambda^{(t)}, \delta^{(t)} \leq 1$, then

$$-\frac{1}{4} \frac{d}{dt} \Delta^{(t)} \geq \left(\left(1 + \frac{\lambda^{(t)}}{2} - 3\delta^{(t)} \right) s^{(t)} - (1 + \epsilon) s^{(0)} \right) \Delta^{(t)}.$$

Note that Proposition 3.9 implies that the dynamical system has linear convergence at time $t = 0$. To see this, note that $\delta^{(0)} \leq \epsilon$ by Lemma 3.6, and $\lambda^{(0)} = \lambda$ from Definition 1.4, and therefore

$$-\frac{d}{dt} \Delta^{(0)} \geq 2(\lambda - 8\epsilon) s^{(0)} \Delta^{(0)}.$$

Under our assumption that $\lambda \gg \epsilon$, the dynamical system has linear convergence at time $t = 0$ with rate $\lambda s^{(0)}$.

To prove that the dynamical system has linear convergence with rate $\lambda s^{(0)}$ for all time $t \geq 0$, we will prove that the quantities in Proposition 3.9 do not change much when we move from $\mathcal{A}^{(0)}$ to $\mathcal{A}^{(t)}$, i.e., $s^{(t)} \approx s^{(0)}$, $\delta^{(t)} \approx \delta^{(0)}$, and $\lambda^{(t)} \approx \lambda$.

To bound the change of the singular values of $M_{\mathcal{A}^{(t)}}$, we will bound the condition number of the scaling solutions in the dynamical system in section 3.5, and then use these bounds to argue about the change of the singular values and establish Theorem 1.5 in section 3.6.

3.5. Scaling solutions and condition numbers. We first present the results in product integration in Slavík's book [57] in section 3.5.1, and then use these results to bound the condition number of the scaling solutions in section 3.5.2.

3.5.1. Scaling solutions. The dynamical system in Definition 2.16 describes the change of $\mathcal{A}$ by a differential equation. The solution to the differential equation can be analyzed using the theory of product integration in [57].

DEFINITION 3.10. *Let $A : [a, b] \to \mathbb{R}^{n \times n}$ be a matrix valued function. A partition t of the interval $[a, b]$ is a sequence of numbers $a = t_0 < t_1 < t_2 < \cdots < t_m = b$. Let $\Delta t_i = t_i - t_{i-1}$ for $i = 1, \ldots, n$ and $\Delta t = \max_{i=1,\cdots n} \Delta t_i$. When the limits over all partitions with $\Delta t \to 0$ exist, the left product integral is defined as*

$$\prod_a^b (I + A(x)dx) := \lim_{\Delta t \to 0} (I + A(t_{m-1})\Delta t_m) \cdots (I + A(t_1)\Delta t_2)(I + A(t_0)\Delta t_1),$$

and the right product integral is defined as

$$(I + A(x)dx) \prod_a^b := \lim_{\Delta t \to 0} (I + A(t_0)\Delta t_1)(I + A(t_1)\Delta t_2) \cdots (I + A(t_{m-1})\Delta t_m).$$

THEOREM 3.11 (Theorem 2.5.1 in [57]). *If $P, Q : [a, b] \to \mathbb{R}^{n \times n}$ are continuous matrix functions, then the product integrals*

$$Y(x) = \prod_a^x (I + P(t)dt) \quad \text{and} \quad Z(x) = (I + Q(t)dt) \prod_a^x$$

exist and satisfy the equations

$$\frac{d}{dx}Y(x) = P(x)Y(x) \quad \text{and} \quad \frac{d}{dx}Z(x) = Z(x)Q(x)$$

for every $x \in [a, b]$.

Applying Theorem 3.11 with $P(t) = E^{(t)}$, $Q(t) = F^{(t)}$, $Y(x) = L^{(T)}$, and $Z(x) = R^{(T)}$, we can explicitly describe the scaling matrices of the dynamical system.

COROLLARY 3.12. *The solution to the dynamical system in Definition* 2.16 *is $A_i^{(T)} = L^{(T)} A_i^{(0)} R^{(T)}$ where*

$$L^{(T)} := \prod_{t=0}^T (I + E^{(t)}dt) \quad \text{and} \quad R^{(T)} := (I + F^{(t)}dt) \prod_{t=0}^T.$$

We are interested in bounding the condition number of $L^{(T)}$ and $R^{(T)}$.

DEFINITION 3.13 (condition number). *The condition number of a matrix A is defined as*

$$\kappa(A) := \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)},$$

where $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ are the maximum singular value and the minimum singular value of A, respectively.

The following theorem in Slavík [57] will be used to bound $\kappa(L^{(T)})$ and $\kappa(R^{(T)})$.

THEOREM 3.14 (Corollary 3.4.3 in [57]). *If $P, Q : [a, b] \to \mathbb{R}^{n \times n}$ are Riemann integrable functions, then*

$$\left\| (I + Q(x)dx) \prod_a^b - (I + P(x)dx) \prod_a^b \right\|_{\mathrm{op}}$$

$$\leq \exp \left(\int_a^b \|P(x)\|_{\mathrm{op}} \, dx \right) \left(\exp \left(\int_a^b \|Q(x) - P(x)\|_{\mathrm{op}} \, dx \right) - 1 \right).$$

Applying Theorem 3.14 with $Q(x) = E^{(t)}$ and $P(x) = 0$, we have the following bound of the maximum and minimum eigenvalues of $L^{(T)}$.

COROLLARY 3.15. *For any $T \geq 0$,*

$$\left\| L^{(T)} - I \right\|_{\mathrm{op}} = \left\| \prod_{t=0}^T (I + E^{(t)}dt) - I \right\|_{\mathrm{op}} \leq \exp \left(\int_0^T \|E^{(t)}\|_{\mathrm{op}} \, dt \right) - 1.$$

This corollary will be used to bound the condition number of $L^{(T)}$ in Lemma 3.16, which will then be used to bound the condition number of $R^{(T)}$ in Lemma 3.18.

3.5.2. Bounding the condition number. To bound the condition number, we use Corollary 3.15 and bound the integral in the exponent. To bound the integral, we divide the time into two phases. In the first phase, we use Proposition 3.2 to argue that $\left\|E^{(t)}\right\|_{\mathrm{op}} \approx \left\|E^{(0)}\right\|_{\mathrm{op}}$. In the second phase, we use that $\Delta^{(t)}$ is converging linearly to argue that $\left\|E^{(t)}\right\|_{\mathrm{op}} \leq \left\|E^{(t)}\right\|_{F} \leq \sqrt{m\Delta^{(t)}}$ is converging linearly. In the following lemma, we should think of g as the spectral gap parameter in Definition 1.4.

LEMMA 3.16. *Suppose there exists $g > 0$ such that for all $0 \leq t \leq T$, it holds that*

$$-\frac{d}{dt}\Delta^{(t)} \geq gs^{(0)}\Delta^{(t)}.$$

If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced for $\epsilon \leq g$, then

$$\left\|L^{(T)} - I\right\|_{\mathrm{op}} \leq \exp\left(O\left(\frac{\epsilon \ln m}{g}\right)\right) - 1.$$

Proof. To bound the condition number, we use Corollary 3.15 and bound the integral

$$\int_0^T \|E^{(t)}\|_{\mathrm{op}} = \int_0^\tau \|E^{(t)}\|_{\mathrm{op}} + \int_\tau^T \|E^{(t)}\|_{\mathrm{op}}.$$

We split the integral into two terms. For the first term, we use Proposition 3.2 to bound

$$\int_0^\tau \|E^{(t)}\|_{\mathrm{op}}\, dt \leq \int_0^\tau \left((1+\epsilon)s^{(0)} - s^{(t)}\right) dt \leq \tau\left(s^{(0)} - s^{(T)} + \epsilon s^{(0)}\right),$$

where the second inequality is by the fact that $s^{(t)}$ is nonincreasing from Lemma 2.17. Applying Lemma 2.19 with our assumption that $\mu = gs^{(0)}$, it follows that

$$\int_0^\tau \|E^{(t)}\|_{\mathrm{op}}\, dt \leq \tau\left(\frac{2\Delta^{(0)}}{gs^{(0)}} + \epsilon s^{(0)}\right) \leq \tau\left(\frac{4\epsilon^2 s^{(0)}}{g} + \epsilon s^{(0)}\right) \leq 5\tau\epsilon s^{(0)},$$

where the second inequality is by Lemma 2.15, and the last inequality is by our assumption that $g \geq \epsilon$.

For the second term,

$$\int_\tau^T \|E^{(t)}\|_{\mathrm{op}}\, dt \leq \int_\tau^T \|E^{(t)}\|_F\, dt \leq \int_\tau^T \sqrt{m\Delta^{(t)}} dt \leq \sqrt{m\Delta^{(\tau)}} \int_\tau^T e^{-gs^{(0)}(t-\tau)/2} dt$$
$$\leq \frac{2\sqrt{m\Delta^{(\tau)}}}{gs^{(0)}},$$

where the second inequality is from the inequality that $\|E^{(t)}\|_F^2 \leq m\Delta^{(t)}$ from Definition 2.14, and the third inequality follows from Lemma 2.19 using the assumption that Δ is converging linearly with $\mu = gs^{(0)}$.

We choose

$$\tau = \frac{\ln m}{gs^{(0)}} \quad \Longrightarrow \quad e^{-gs^{(0)}\tau} \leq \frac{1}{m}.$$

This implies that

$$\Delta^{(\tau)} \leq \Delta^{(0)} e^{-gs^{(0)}\tau} \leq \frac{\Delta^{(0)}}{m} \leq \frac{2\epsilon^2(s^{(0)})^2}{m} \quad \Longrightarrow \quad \frac{2\sqrt{m\Delta^{(\tau)}}}{gs^{(0)}} \leq \frac{3\epsilon}{g},$$

and so the second term is at most $3\epsilon/g$. The first term is at most $5\tau\epsilon s^{(0)} \leq 5\epsilon \ln m/g$, and so Corollary 3.15 implies that

$$\left\|L^{(T)} - I\right\|_{op} \leq \exp\left(\int_0^T \|E(t)\|_{op} \, dt\right) - 1 \leq \exp\left(\frac{8\epsilon \ln m}{g}\right) - 1. \qquad \square$$

Remark 3.17. We leave it as an open question whether the $\log m$ term in the condition number is necessary.

We cannot use the same argument to bound $\|R^{(T)} - I\|_{op}$, as it will only give us a bound with dependency on n (where we assumed $m \leq n$). Instead, we use the bound on $\|L^{(T)} - I\|_{op}$ to derive a similar bound on $\|R^{(T)} - I\|_{op}$.

LEMMA 3.18. *Suppose there exists $g > 0$ such that for all $0 \leq t \leq T$, it holds that*

$$-\frac{d}{dt}\Delta^{(t)} \geq gs^{(0)}\Delta^{(t)}.$$

If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced for $\epsilon \leq g$, and also $\epsilon, \ell \leq 1/2$, then

$$\left\|L^{(T)} - I\right\|_{op} \leq \ell \quad \Longrightarrow \quad \left\|R^{(T)} - I\right\|_{op} \leq O(\ell + \epsilon).$$

Proof. We would like to bound

$$r_{\max} := \max_{\|u\|_2 \leq 1} \left\|R^{(T)}u\right\|_2 \quad \text{and} \quad r_{\min} := \min_{\|u\|_2 \leq 1} \left\|R^{(T)}u\right\|_2.$$

First, we bound $r_{\max}$. Let $u \in \mathbb{R}^n$ be a maximizer such that $\left\|R^{(T)}u\right\|_2 = r_{\max}$ and $\|u\|_2 = 1$.

Consider $\langle \Phi^{(T)^*}(I_m), uu^* \rangle$. On one hand, we use Proposition 3.4 to upper bound

$$\left\langle \Phi^{(T)^*}(I_m), uu^* \right\rangle \leq \left\langle \frac{(1+\epsilon)s^{(0)}}{n}I_n, uu^* \right\rangle = \frac{(1+\epsilon)s^{(0)}}{n}.$$

On the other hand, by Fact 2.10.4,

$$\left\langle \Phi^{(T)^*}(I_m), uu^* \right\rangle = \left\langle R^{(T)^*} \cdot \Phi^{(0)^*}\left(L^{(T)^*}L^{(T)}\right) \cdot R^{(T)}, \ uu^* \right\rangle$$
$$= \left\langle \Phi^{(0)^*}\left(L^{(T)^*}L^{(T)}\right), \ (R^{(T)}u)(R^{(T)}u)^* \right\rangle.$$

Since $\left\|L^{(T)} - I_m\right\|_{op} \leq \ell$, all singular values of $L^{(T)}$ are at least $1 - \ell$, and thus all eigenvalues of $L^{(T)}L^{(T)^*}$ are at least $(1-\ell)^2$, i.e., $L^{(T)}L^{(T)^*} \succeq (1-\ell)^2 I_m$. It follows from Fact 2.10.2 that $\Phi^{(0)^*}(L^{(T)^*}L^{(T)}) \succeq \Phi^{(0)^*}((1-\ell)^2 I_m)$, and using it in the above equation gives

$$\left\langle \Phi^{(T)^*}(I_m), uu^* \right\rangle \geq \left\langle \Phi^{(0)^*}\left((1-\ell)^2 I_m\right), \ (R^{(T)}u)(R^{(T)}u)^* \right\rangle$$
$$\geq \left\langle (1-\ell)^2(1-\epsilon)\frac{s^{(0)}}{n}I_n, \ (R^{(T)}u)(R^{(T)}u)^* \right\rangle$$
$$= r_{\max}^2(1-\ell)^2(1-\epsilon)\frac{s^{(0)}}{n},$$

where the second inequality uses that $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced. Combining the upper bound and lower bound gives

$$r_{\max}^2 \leq \frac{1+\epsilon}{(1-\epsilon)(1-\ell)^2} \leq 1 + O(\epsilon + \ell) \quad \implies \quad r_{\max} \leq 1 + O(\epsilon + \ell),$$

where we use the assumptions that $\epsilon, \ell \leq 1/2$.

Next, we bound $r_{\min}$ using a similar argument. Let $v \in \mathbb{R}^n$ be a minimizer such that $\left\| R^{(T)} v \right\|_2 = r_{\min}$ and $\|v\|_2 = 1$. Consider $\langle \Phi^{(T)^*}(I_m), vv^* \rangle$. On one hand, we use Proposition 3.4 to lower bound

$$\left\langle \Phi^{(T)^*}(I_m), vv^* \right\rangle \geq \left\langle \frac{2s^{(T)} - (1+\epsilon)s^{(0)}}{n} I_n, \ vv^* \right\rangle = \frac{2s^{(T)} - (1+\epsilon)s^{(0)}}{n} \geq \frac{(1-9\epsilon)s^{(0)}}{n},$$

where the second inequality uses the assumption that $\Delta^{(t)}$ is converging linearly for $0 \leq t \leq T$ to apply Lemma 2.19 with $\mu = gs^{(0)}$ to obtain

$$s^{(0)} - s^{(T)} \leq \frac{2\Delta^{(0)}}{gs^{(0)}} \leq \frac{4\epsilon^2 s^{(0)}}{g} \leq 4\epsilon s^{(0)} \quad \implies \quad s^{(T)} \geq (1 - 4\epsilon) \cdot s^{(0)},$$

where the second inequality is by Lemma 2.15 and the last inequality is from the assumption that $\epsilon \leq g$.

On the other hand, by a similar calculation as above with $L^{(T)} L^{(T)^*} \leq (1+\ell)^2 I_m$, we obtain

$$\left\langle \Phi^{(T)^*}(I_m), vv^* \right\rangle \leq r_{\min}^2 (1+\ell)^2 (1+\epsilon) \frac{s^{(0)}}{n}.$$

Combining the upper bound and lower bound gives

$$r_{\min}^2 \geq \frac{(1-9\epsilon)}{(1+\ell)^2(1+\epsilon)} \geq 1 - O(\epsilon + \ell) \quad \implies \quad r_{\min} \geq 1 - O(\epsilon + \ell),$$

where we used the assumptions that ϵ and ℓ are sufficiently small. Therefore, we conclude that

$$\left\| R^{(T)} - I \right\|_{\text{op}} = \max\{r_{\max} - 1, 1 - r_{\min}\} \leq O(\epsilon + \ell). \qquad \square$$

3.6. Invariance of linear convergence. We will first use Lemma 3.16 and Lemma 3.18 to bound the change of the singular values of $M_{\mathcal{A}^{(t)}}$. Then, we will combine the previous results to prove Theorem 1.5 that $\Delta^{(t)}$ is converging linearly for all $t \geq 0$.

To bound the change of the singular values, we use the following inequality.

LEMMA 3.19 (Theorem 3.3.16 in [39]). *Let A and B be two $m \times n$ matrices. For any $1 \leq k \leq m$,*

$$|\sigma_k(A) - \sigma_k(B)| \leq \sigma_1(A - B) = \|A - B\|_{\text{op}}.$$

The following lemma bounds the change of the singular values after scaling the operator.

LEMMA 3.20. *Given operator $\mathcal{A} := (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$, if we produce scaling $\mathcal{B} := (LA_1R, \ldots LA_kR)$, with left/right scalings $\|L - I_m\|_{\text{op}} \leq \zeta$ and $\|R - I_n\|_{\text{op}} \leq \zeta$ for some $\zeta \leq 1$, then for all k,*

$$|\sigma_k(M_{\mathcal{B}}) - \sigma_k(M_{\mathcal{A}})| \leq O(\zeta) \cdot \|M_{\mathcal{A}}\|_{\text{op}}.$$

Proof. By Fact 2.5, the matrix representation of the new operator is

$$M_{\mathcal{B}} = \sum_{i=1}^{k} (LA_i R) \otimes (LA_i R)$$

$$= (L \otimes L) \left(\sum_{i=1}^{k} A_i \otimes A_i \right) (R \otimes R)$$

$$= (L \otimes L) \cdot M_{\mathcal{A}} \cdot (R \otimes R),$$

where the second equality is by Fact 2.10.1. By Lemma 3.19,

$$|\sigma_k(M_{\mathcal{B}}) - \sigma_k(M_{\mathcal{A}})| \le \|M_{\mathcal{B}} - M_{\mathcal{A}}\|_{\mathrm{op}} = \|(L \otimes L) \cdot M_{\mathcal{A}} \cdot (R \otimes R) - M_{\mathcal{A}}\|_{\mathrm{op}}.$$

To bound the right-hand side, we expand $L \otimes L$ as $(L - I) \otimes (L - I) + (L - I) \otimes I + I \otimes (L - I) + I \otimes I$ and expand $R \otimes R$ similarly. Then $(L \otimes L) \cdot M_{\mathcal{A}} \cdot (R \otimes R) - M_{\mathcal{A}}$ can be written as the sum of 15 terms, where $M_{\mathcal{A}}$ cancels with $(I \otimes I) M_{\mathcal{A}} (I \otimes I)$. To bound the operator norm, we use the triangle inequality and bound the sum of the 15 operator norms. For each term, we use the facts that $\|A \otimes B\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}}$ and $\|ABC\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}} \|C\|_{\mathrm{op}}$ to bound its norm. For example,

$$\left\|\big((L - I_m) \otimes (L - I_m)\big) \cdot M_{\mathcal{A}} \cdot \big((R - I_n) \otimes I_n\big)\right\|_{\mathrm{op}}$$

$$\le \|(L - I_m) \otimes (L - I_m)\|_{\mathrm{op}} \|M_{\mathcal{A}}\|_{\mathrm{op}} \|(R - I_n) \otimes I_n\|_{\mathrm{op}}$$

$$\le \|(L - I_m)\|_{\mathrm{op}}^2 \|(R - I_n)\|_{\mathrm{op}} \|M_{\mathcal{A}}\|_{\mathrm{op}}.$$

Since we assumed that $\|L - I_m\|_{\mathrm{op}} \le \zeta$ and $\|R - I_n\|_{\mathrm{op}} \le \zeta$ for some $\zeta \le 1$, each of these terms is at most $\zeta \|M_{\mathcal{A}}\|_{\mathrm{op}}$ and thus we conclude that $\|M_{\mathcal{B}} - M_{\mathcal{A}}\|_{\mathrm{op}} \le 15\zeta \cdot \|M_{\mathcal{A}}\|_{\mathrm{op}}$. $\square$

We are ready to put together the results to prove the following theorem which implies Theorem 1.5.

THEOREM 3.21. *If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced and $\mathcal{A}^{(0)}$ satisfies the λ-spectral gap condition in Definition 1.4 with $\lambda^2 \ge C\epsilon \ln m$ for a sufficiently large universal constant C, then for all $t \ge 0$ it holds that*

$$-\frac{d}{dt} \Delta^{(t)} \ge \lambda s^{(0)} \Delta^{(t)}.$$

Proof. Let $\delta(t)$ and $\lambda(t)$ be the functions satisfying the hypothesis of Proposition 3.9. Note that $\delta^{(0)} \le \epsilon$ by Lemma 3.6 and $\lambda^{(0)} = \lambda$ from Definition 1.4. Let T be the supremum over all t such that $s^{(t)} \ge (1 - \epsilon) s^{(0)}$ and $\lambda^{(t)} - 6\delta^{(t)} \ge \frac{1}{2}(1-\epsilon)^{-1}(\lambda^{(0)} + 8\epsilon)$. Our goal is to prove that $\Delta^{(t)}$ is converging linearly for $0 \le t \le T$ and T is unbounded.

First, we show that $\Delta^{(t)}$ is converging linearly for $0 \le t \le T$. By Proposition 3.9,

$$-\frac{d}{dt} \Delta^{(t)} \ge 4 \left(\left(1 + \frac{1}{2} \left(\lambda^{(t)} - 6\delta^{(t)} \right) \right) s^{(t)} - (1 + \epsilon) s^{(0)} \right) \Delta^{(t)}$$

$$\ge 4 \left((1 - \epsilon) \left(1 + \frac{1}{4}(1 - \epsilon)^{-1} \left(\lambda^{(0)} + 8\epsilon \right) \right) - (1 + \epsilon) \right) s^{(0)} \Delta^{(t)}$$

$$= \left((1 - \epsilon)(1 - \epsilon)^{-1} \left(\lambda^{(0)} + 8\epsilon \right) - 8\epsilon \right) s^{(0)} \Delta^{(t)} = \lambda s^{(0)} \Delta^{(t)},$$

where in the second inequality we used that $s^{(t)} \geq (1 - \epsilon)s^{(0)}$ and $\lambda^{(t)} - 6\delta^{(t)} \geq \frac{1}{2}(1 - \epsilon)^{-1}(\lambda^{(0)} + 8\epsilon)$ for $0 \leq t \leq T$. Since $\lambda^{(0)} = \lambda$, we have the required convergence up to time T.

Next, we argue that the size condition $(s^{(t)} \geq (1 - \epsilon)s^{(0)})$ and the spectral gap condition $(\lambda^{(t)} - 6\delta^{(t)} \geq \frac{1}{2}(1 - \epsilon)^{-1}(\lambda^{(0)} + 8\epsilon))$ will still be maintained beyond time T. For the size change, by Lemma 2.19 with $\mu = \lambda s^{(0)}$,

$$s^{(0)} - s^{(T)} \leq \frac{2\Delta^{(0)}}{\lambda s^{(0)}} \leq \frac{4\epsilon^2 s^{(0)}}{\lambda} \ll \epsilon s^{(0)},$$

where the second inequality is by Lemma 2.15 and the last inequality is by $\lambda \geq C\epsilon$ for a sufficiently large constant C.

For the change of the second largest singular value, by definition,

$$
\begin{aligned}
\sigma_2(M_{\mathcal{A}^{(T)}}) - \sigma_2(M_{\mathcal{A}^{(0)}}) &= \frac{(1 - \lambda^{(T)})s^{(T)}}{\sqrt{mn}} - \frac{(1 - \lambda^{(0)})s^{(0)}}{\sqrt{mn}} \\
&\geq \frac{(1 - \lambda^{(T)})(1 - \epsilon)s^{(0)}}{\sqrt{mn}} - \frac{(1 - \lambda^{(0)})s^{(0)}}{\sqrt{mn}} \\
&= \frac{s^{(0)}}{\sqrt{mn}}(\lambda^{(0)} - (1 - \epsilon)\lambda^{(T)} - \epsilon).
\end{aligned}
$$

On the other hand, we can upper bound $\sigma_2(M_{\mathcal{A}^{(T)}}) - \sigma_2(M_{\mathcal{A}^{(0)}})$ using condition numbers. Using Lemma 3.16 with $g = \lambda$, $\left\|L^{(T)} - I\right\|_{\mathrm{op}} \leq \exp\left(O(\epsilon \ln m/\lambda)\right) - 1$. Note that our assumption implies that

$$O\left(\frac{\epsilon \ln m}{\lambda}\right) \leq O\left(\frac{\lambda}{C}\right) \ll 1 \quad \Longrightarrow \quad \left\|L^{(T)} - I\right\|_{\mathrm{op}} \leq O\left(\frac{\lambda}{C}\right) \ll 1,$$

where the implication is by the inequality $e^x - 1 = O(x)$ for x close to zero. Note that we can take C large enough so that $\epsilon \leq O(\lambda/C) \leq 1/2$ so they satisfy the conditions of Lemma 3.18. Then, by Lemma 3.18, we also have $\left\|R^{(T)} - I\right\|_{\mathrm{op}} \leq O(\lambda/C)$. Therefore we can apply Lemma 3.20, with $\mathcal{A} := \mathcal{A}^{(0)}$ and bounded scalings $L := L^{(T)}, R := R^{(T)}$, to obtain

$$\sigma_2(M_{\mathcal{A}^{(t)}}) - \sigma_2(M_{\mathcal{A}^{(0)}}) \leq O\left(\frac{\lambda}{C}\right) \cdot \|M_{\mathcal{A}^{(0)}}\|_{\mathrm{op}} \leq O\left(\frac{\lambda}{C}\right) \frac{(1 + \delta_1^{(0)})s^{(0)}}{\sqrt{mn}}.$$

Combining the upper bound and lower bound and using $\delta_1^{(0)} \leq \epsilon$ from Lemma 3.6, it follows that

$$\lambda^{(T)} \geq \frac{\lambda - \epsilon - (1 + \epsilon) \cdot O(\lambda/C)}{1 - \epsilon} \geq \lambda - O\left(\frac{\lambda}{C}\right),$$

where the last inequality is by the assumption that $\lambda \geq C\epsilon$.

For the change of the largest singular value, by Proposition 3.4,

$$\frac{(1 - 3\epsilon)s^{(T)}}{m}I_m \preceq \frac{2s^{(T)} - (1 + \epsilon)s^{(0)}}{m}I_m \preceq \Phi^{(T)}(I_n) \preceq \frac{(1 + \epsilon)s^{(0)}}{m}I_m \preceq \frac{(1 + 3\epsilon)s^{(T)}}{m}I_m,$$

where the first and last inequalities use that $s^{(T)} \geq (1 - \epsilon)s^{(0)}$. The same holds for $\Phi^{(T)*}$ and these imply that $\mathcal{A}^{(T)}$ is 3ϵ-nearly doubly balanced. By Lemma 3.6, this implies that $\delta^{(T)} \leq 3\epsilon$. Therefore,

$$\lambda^{(T)} - 6\delta^{(T)} \geq \lambda - O\left(\frac{\lambda}{C}\right) - 9\epsilon \geq \lambda - O\left(\frac{\lambda}{C}\right) \geq \frac{2}{3}\lambda > \frac{1}{2}(1-\epsilon)^{-1}(\lambda + 8\epsilon),$$

where the second to last inequality uses that C is a sufficiently large constant.

Since our dynamical system is continuous and the functions $\delta^{(t)}$ and $\lambda^{(t)}$ depend continuously on the operator, we still have both conditions satisfied at time $T + \eta$ for some finite $\eta > 0$, which contradicts that T is the supremum that both conditions are satisfied. Therefore, T is unbounded and the linear convergence of Δ is maintained throughout the execution of the dynamical system. □

3.7. Condition number. With the invariance of the linear convergence, we can apply Lemmas 3.16 and 3.18 to bound the condition number of the scaling solutions and prove Theorem 1.7

THEOREM 3.22. *If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced and $\mathcal{A}^{(0)}$ satisfies the λ-spectral gap condition in Definition* 1.4 *with $\lambda^2 \geq C\epsilon \log m$ for a sufficiently large constant C, then for any $t \geq 0$,*

$$\kappa\left(L^{(t)}\right) \leq 1 + O\left(\frac{\epsilon \log m}{\lambda}\right) \quad \text{and} \quad \kappa\left(R^{(t)}\right) \leq 1 + O\left(\frac{\epsilon \log m}{\lambda}\right).$$

In particular, these bounds hold for the final scaling solutions $L^{(\infty)}$ and $R^{(\infty)}$.

Proof. By Theorem 3.21, $\Delta^{(t)}$ is linearly converging for all time t with rate at least $\lambda s^{(0)}$. By Lemma 3.16, this implies that

$$\left\|L^{(t)} - I_m\right\|_{\text{op}} \leq \exp\left(O\left(\frac{\epsilon \ln m}{\lambda}\right)\right) - 1 \leq O\left(\frac{\epsilon \log m}{\lambda}\right) \ll 1,$$

where we used the assumption that $\lambda^2 \geq C\epsilon \ln m$ and $e^x - 1 \leq O(x)$ for x close to zero. By Lemma 3.18, this implies the same bound on

$$\left\|R^{(T)} - I\right\|_{\text{op}} \leq O\left(\frac{\epsilon \log m}{\lambda}\right).$$

Therefore, $\lambda_{\min}(L^{(t)}) \geq 1 - O(\epsilon \log m/\lambda)$ and $\lambda_{\max}(L^{(t)}) \leq 1 + O(\epsilon \log m/\lambda)$, and hence

$$\kappa(L^{(t)}) \leq \frac{\lambda_{\max}(L^{(t)})}{\lambda_{\min}(L^{(t)})} \leq \frac{1 + O(\epsilon \log m/\lambda)}{1 - O(\epsilon \log m/\lambda)} \leq 1 + O\left(\frac{\epsilon \log m}{\lambda}\right),$$

where we used that $\epsilon \log m/\lambda \ll 1$. The same argument applies to give the same bound for $\kappa(R^{(t)})$. □

3.8. Operator capacity. We restate Theorem 1.8 below and show that it follows easily from Theorem 3.21.

THEOREM 3.23. *If $\mathcal{A}^{(0)}$ is ϵ-nearly doubly balanced and $\mathcal{A}^{(0)}$ satisfies the λ-spectral gap condition in Definition* 1.4 *with $\lambda^2 \geq C\epsilon \ln m$ for a sufficiently large constant C, then*

$$\text{cap}^{(0)} \geq \left(1 - \frac{4\epsilon^2}{\lambda}\right) s^{(0)}.$$

Proof. By Theorem 3.21, $\Delta^{(t)}$ is linearly converging for all time t with rate $\lambda s^{(0)}$. Apply Proposition 2.21 with $\mu = \lambda s^{(0)}$,

$$\text{cap}^{(0)} \geq s^{(0)} - \frac{2\Delta^{(0)}}{\lambda s^{(0)}} \geq s^{(0)} - \frac{4\epsilon^2 s^{(0)}}{\lambda} = \left(1 - \frac{4\epsilon^2}{\lambda}\right) s^{(0)},$$

where the second inequality is by Lemma 2.15. □

3.9. Discrete gradient flow. The gradient flow can be discretized to give a polynomial time algorithm with linear convergence when the input has a spectral gap. The analysis follows the continuous case closely, so we will just provide a sketch.

Recall that the gradient flow is defined as

$$\frac{d}{dt}A_i := EA_i + A_iF,$$

where E and F are the error matrices (Definition 2.14) of the current operator $\mathcal{A}$.

In the discrete case, a natural update step is

$$\widetilde{A}_i \leftarrow A_i + \alpha(EA_i + A_iF)$$

for some small step size α, but the problem of this update step is that $\widetilde{\mathcal{A}}$ may not be a scaling of $\mathcal{A}$. So we modified the discrete algorithm slightly as follows. In each step, we update

$$\widetilde{A}_i \leftarrow (I_m + \alpha E)A_i(I_n + \alpha F),$$

where α is the step size. This update is to maintain that the current operator is a scaling of the original operator.

We assume that $s = 1$ and $\Delta \leq 1$ initially. We will set the step size to be $\alpha = O((m+n)^{-2})$ for the same analysis in the continuous case to go through. With this choice of the step size, we can show that

$$s(\mathcal{A}) - s(\widetilde{\mathcal{A}}) \leq 4\alpha\Delta(\mathcal{A})$$

by expanding the change of the size s and use the small step size α to argue that the higher order terms are negligible. By a similar but more tedious calculation (since the degree is higher), we can also show that

$$\left|\Delta(\widetilde{\mathcal{A}}) - \left(\Delta(\mathcal{A}) - \alpha\frac{d}{dt}\Delta\right)\right| \leq O(\alpha s^2\Delta(\mathcal{A})),$$

where $\frac{d}{dt}\Delta$ is the change of Δ in the continuous case. This is also the step that we need $\alpha = O((m+n)^{-2})$ to hold. Since we know $-\frac{d}{dt}\Delta \geq \lambda s\Delta$, this implies that

$$\Delta(\widetilde{\mathcal{A}}) \leq \left(1 - \frac{1}{2}\alpha\lambda s\right)\Delta(\mathcal{A}),$$

that Δ is decreasing geometrically with rate λs, when the current operator $\mathcal{A}$ satisfies the spectral condition.

As in the continuous case, we use an inductive argument to prove that the spectral gap condition is maintained to establish that the convergence rate λs is maintained throughout the algorithm. Again, we go through the condition number of the error matrices and use the arguments in Lemma 3.20 to show that the change of the singular value is

$$|\sigma_k(M_{\widetilde{\mathcal{A}}}) - \sigma_k(M_{\mathcal{A}})| \leq O(\alpha\epsilon s),$$

and it follows that the λ-spectral gap condition holds throughout as

$$|\lambda^{(\infty)} - \lambda^{(0)}| \leq O\left(\frac{\epsilon\log(m+n)}{\lambda^{(0)}}\right)$$

which is negligible when the spectral assumption $(\lambda^{(0)})^2 \gg \epsilon\log(m+n)$ holds initially.

In the discrete algorithm, we will set the step size to be $\alpha = \Theta((m+n)^{-2})$. If the continuous algorithm converges to an η-approximate solution in time T, the discrete algorithm will converge to an η-approximate solution in $T \cdot \Theta((m+n)^2)$ number of iterations, and the dependency on η is $\log(1/\eta)$ by Theorem 1.5.

Remark 3.24. The step size $\alpha = O((m+n)^{-2})$ is chosen for the same analysis as in the continuous to hold. An interesting open question is whether the analysis can be extended to constant step size, in particular whether Sinkhorn's alternating algorithm has the same convergence rate as in the gradient flow. After the submission of this article, both of these questions were resolved positively in [20]. They were able to use the perspective of *geodesic convex optimization* to view these algorithms as natural descent methods on the space of scalings. Then they were able to leverage the conditions and result of Theorem 1.5 to show both of these descent methods stay within a zone around the optimum where the function is sufficiently strongly convex to produce linear convergence.

4. Applications of matrix scaling and operator scaling. In this section, we show some implications of our results in various applications of the operator scaling problem.

4.1. Matrix scaling. Given a nonnegative matrix $B \in \mathbb{R}^{m \times n}$, let $s(B) := \sum_{i=1}^m \sum_{j=1}^n B_{i,j}$ be the size of the matrix, $r_i(B) := \sum_{j=1}^n B_{i,j}$ be the ith row sum of B, and $c_j(B) := \sum_{i=1}^m B_{i,j}$ be the jth column sum of B. A nonnegative matrix is called ϵ-nearly doubly balanced if for every $1 \le i \le m$ and for every $1 \le j \le n$,

$$(1-\epsilon)\frac{s(B)}{m} \le r_i(B) \le (1+\epsilon)\frac{s(B)}{m} \quad \text{and} \quad (1-\epsilon)\frac{s(B)}{n} \le c_j(B) \le (1+\epsilon)\frac{s(B)}{n},$$

and is called doubly balanced when $\epsilon = 0$. A common setting is when B is an $n \times n$ matrix when the average row sum is equal to one, in which case $s(B) = n$ and the matrix is called "doubly stochastic" when every row sum and every column sum are equal to one.

DEFINITION 4.1 (matrix scaling problem). *We are given a nonnegative matrix $B \in \mathbb{R}^{m \times n}$, and the goal is to find a left diagonal scaling matrix $L \in \mathbb{R}^{m \times m}$ and a right diagonal scaling matrix $R \in \mathbb{R}^{n \times n}$ such that LBR is doubly balanced, or report that such scaling matrices do not exist.*

Outline. In the following, we will show that the matrix scaling problem can be reduced to the operator scaling problem in section 4.1.1. Then, we will see that the spectral condition has a simple form in section 4.1.2, and there is a natural combinatorial condition that implies the spectral condition in section 4.1.3. We then argue that many random matrices will satisfy our condition in section 4.1.4. Finally, we see the implications of our results in several applications of matrix scaling, including bipartite matching in section 4.1.5, permanent lower bound in section 4.1.6, and optimal transportation in section 4.1.7.

4.1.1. Reduction to operator scaling. The matrix scaling problem is a special case of the operator scaling problem.

LEMMA 4.2. *Given a nonnegative matrix $B \in \mathbb{R}^{m \times n}$, let $\mathcal{A} = (A_{11}, \ldots, A_{mn})$ be the operator where each $A_{ij} \in \mathbb{R}^{m \times n}$ for $1 \le i \le m$ and $1 \le j \le n$ is the matrix with the (i,j)th entry equal to $\sqrt{B_{i,j}}$ and all other entries equal to zero. Then, B is ϵ-nearly doubly balanced if and only if $\mathcal{A}$ is ϵ-nearly doubly balanced. Furthermore,*

there is a solution to the matrix scaling problem for B if and only if there is a solution to the operator scaling problem for $\mathcal{A}$.

Proof. By construction, $A_{ij}A_{ij}^*$ is the $m \times m$ matrix with B_{ij} in the (i,i)th entry and zero otherwise, and $A_{ij}^*A_{ij}$ is the $n \times n$ matrix with B_{ij} in the (j,j)th entry and zero otherwise. So, $\sum_{i=1}^m \sum_{j=1}^n A_{ij}A_{ij}^*$ is the $m \times m$ diagonal matrix where the ith diagonal entry is the ith row sum of B, and $\sum_{i=1}^m \sum_{j=1}^n A_{ij}^*A_{ij}$ is the $n \times n$ diagonal matrix where the jth diagonal entry is the jth column sum of B. Therefore, $\mathcal{A}$ is ϵ-nearly doubly balanced if and only if B is ϵ-nearly doubly balanced. It should be clear that the square root of a scaling solution L, R to B is also a (diagonal) scaling solution to $\mathcal{A}$.

Because of the special structure that each A_{ij} has only one nonzero entry, there is always a scaling solution with L, R being diagonal matrices if a scaling solution exists. To see this, let L, R be a scaling solution to $\mathcal{A}$ with $\sum_{i,j} LA_{ij}RR^*A_{ij}^*L^* = \sum_{i,j} LA_{ij}RR^*A_{ij}^*L^* = sI_m/m$ and $\sum_{i,j}(LA_{ij}R)^*(LA_{ij}R) = \sum_{i,j} R^*A_{ij}^*L^*LA_{ij}R = sI_n/n$. Define $D_L = (L^*L)^{1/2}$. We claim that D_L, R is also a scaling solution to $\mathcal{A}$ and D_L is a diagonal matrix. First, $\sum_{i,j}(D_LA_{ij}R)^*(D_LA_{ij}R) = \sum_{i,j} R^*A_{ij}^*D_L^*D_LA_{ij}R = \sum i, j R^*A_{ij}^*L^*LA_{ij}R = sI_n/n$. Next, it follows from $\sum_{i,j} LA_{ij}RR^*A_{ij}^*L^* = sI_m/m$ that $(s/m)(L^*L)^{-1} = \sum_{i,j} A_{ij}RR^*A_{ij}$, and this implies that L^*L is a diagonal matrix as $\sum_{i,j} A_{ij}RR^*A_{ij}$ is a diagonal matrix because each A_{ij} has only one nonzero entry. Finally, we check that $\sum_{i,j}(D_LA_{ij}R)(D_LA_{ij}R)^* = D_L(\sum_{i,j} A_{ij}RR^*A_{ij}^*)D_L^* = sD_L(L^*L)^{-1}D_L^*/m = sI_m/m$. By the same argument, we can define $D_R = (RR^*)^{1/2}$ so that D_L, D_R is also a scaling solution to $\mathcal{A}$ and both D_L and D_R are diagonal matrices. Therefore, we conclude that the matrix scaling problem can be reduced to the operator scaling problem. □

4.1.2. Spectral condition. The spectral condition for operator scaling has a simple form for matrix scaling.

LEMMA 4.3. *Using the reduction from Lemma* 4.2, *the spectral condition for operator scaling in Definition* 1.4 *becomes*

$$\sigma_2(B) \leq (1-\lambda)\frac{s(B)}{\sqrt{mn}}.$$

Proof. Note that each $A_l \otimes A_l \in \mathbb{R}^{m^2 \times n^2}$ has only one nonzero entry B_{ij}, and $M_{\mathcal{A}} = \sum_l A_l \otimes A_l$ in Definition 1.4 has only an $m \times n$ submatrix with nonzero entries and this submatrix is exactly B. So, the condition that $\sigma_2(M_{\mathcal{A}}) \leq (1-\lambda)s(B)/\sqrt{mn}$ becomes $\sigma_2(B) \leq (1-\lambda)s(B)/\sqrt{mn}$. □

4.1.3. Combinatorial condition. To better understand the spectral gap condition in the matrix case, we present a natural combinatorial condition that implies the spectral condition.

DEFINITION 4.4 (edge-weighted bipartite graph and conductance). *Given a nonnegative matrix $B \in \mathbb{R}^{m \times n}$, we define its edge-weighted bipartite graph G_B as follows. In G_B, there is one vertex u_i for each row i, one vertex v_j for each column j, and an edge ij with weight $w_{ij} = B_{ij}$ between u_i and v_j.*

The conductance of an edge-weighted graph $G = (V, E)$ with $w : E \to \mathbb{R}_{\geq 0}$ is defined as

$$\phi(G) := \min_{S \subseteq V : \text{vol}(S) \leq \text{vol}(V)/2} \phi(S), \text{ where } \phi(S) := \frac{\sum_{i \in S} \sum_{j \notin S} w_{ij}}{\text{vol}(S)} \text{ and } \text{vol}(S) := \sum_{i \in S} \sum_{j \in V} w_{ij}.$$

Using Cheeger's inequality from spectral graph theory, we can show that B satisfies the spectral gap condition if its edge-weighted bipartite graph has large conductance.

LEMMA 4.5. *If $B \in \mathbb{R}^{m \times n}$ is ϵ-nearly doubly balanced for $\epsilon \leq 1/2$, then*

$$\sigma_2(B) \leq \left(1 - \frac{1}{2}\phi^2(G_B) + 3\epsilon\right) \cdot \frac{s(B)}{\sqrt{mn}},$$

where G_B is the edge-weighted bipartite graph of B.

Proof. The adjacency matrix A_G of the edge-weighted bipartite graph G_B is $\begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}$. Note that if $\sum_i \sigma_i x_i y_i^*$ is the singular value decomposition of B, then A_G has eigenvalues $\{\pm\sigma_i\}$ and eigenvectors $\{(x_i, \pm y_i)\}$. Therefore, $\sigma_2(B) = \lambda_2(A_G)$ where $\lambda_2(A_G)$ is the second largest eigenvalue of A_G.

To relate $\sigma_2(B)$ to the conductance $\phi(G_B)$, we will consider the normalized adjacency matrix of A_G and apply Cheeger's inequality. The normalized adjacency matrix $\mathbb{A}$ of a matrix A is defined as $\mathbb{A} := D^{-1/2}AD^{-1/2}$ where D is the diagonal degree matrix with $D_{i,i} := \sum_j A_{i,j}$. For A_G, note that $D_G := \begin{bmatrix} R & 0 \\ 0 & C \end{bmatrix}$, where $R \in \mathbb{R}^{m \times m}$ is the diagonal matrix with the (i,i)th entry being the ith row sum $r_i(B)$ of B and $C \in \mathbb{R}^{n \times n}$ is the diagonal matrix with the (j,j)th entry being the jth column sum $c_j(B)$ of B. Then,

$$\mathbb{A}_G = \begin{bmatrix} R^{-1/2} & 0 \\ 0 & C^{-1/2} \end{bmatrix} \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix} \begin{bmatrix} R^{-1/2} & 0 \\ 0 & C^{-1/2} \end{bmatrix} = \begin{bmatrix} 0 & R^{-1/2}BC^{-1/2} \\ C^{-1/2}B^*R^{-1/2} & 0 \end{bmatrix}.$$

Let $\mathbb{B} = R^{-1/2}BC^{-1/2}$. Note that $\sigma_2(\mathbb{B}) = \lambda_2(\mathbb{A}_G)$ by the argument in the first paragraph. Each entry of $\mathbb{B}$ is

$$(1 - 2\epsilon)\frac{\sqrt{mn}}{s}B_{ij} \leq \sqrt{\frac{m}{s(1+\epsilon)}}\sqrt{\frac{n}{s(1+\epsilon)}}B_{ij} \leq r_i^{-1/2}B_{ij}c_j^{-1/2}$$

$$\leq \sqrt{\frac{m}{s(1-\epsilon)}}\sqrt{\frac{n}{s(1-\epsilon)}}B_{ij} \leq (1 + 2\epsilon)\frac{\sqrt{mn}}{s}B_{ij},$$

where we used the assumptions that B is ϵ-nearly doubly balanced and $\epsilon \leq 1/2$. Hence, we can write $\mathbb{B} = (\sqrt{mn}/s)B + \mathcal{E}$, where $\mathcal{E}$ is the "error" matrix with $|\mathcal{E}_{ij}| \leq 2\epsilon\sqrt{mn}B_{ij}/s$ for all i, j. By Lemma 3.19, $(\sqrt{mn}/s) \cdot \sigma_2(B) \leq \sigma_2(\mathbb{B}) + \|\mathcal{E}\|_{\mathrm{op}}$. By the fact that the square of the largest singular value is at most the maximum row sum times the maximum column sum,

$$\|\mathcal{E}\|_{\mathrm{op}} \leq \sqrt{\max_i \sum_j |\mathcal{E}_{ij}|} \cdot \sqrt{\max_j \sum_i |\mathcal{E}_{ij}|} \leq \frac{2\epsilon\sqrt{mn}}{s}\sqrt{\max_j \sum_i B_{ij}}\sqrt{\max_i \sum_j B_{ij}} \leq 2\epsilon(1 + \epsilon),$$

where the last inequality uses that $r_i(B) \leq (1+\epsilon)s/m$ for $1 \leq i \leq m$ and $c_j(B) \leq (1+\epsilon)s/n$ for $1 \leq j \leq n$. Finally, Cheeger's inequality states that $\phi(G) \leq \sqrt{2(1 - \lambda_2(\mathbb{A}_G))}$. Therefore, we conclude that

$$\frac{\sqrt{mn}}{s} \cdot \sigma_2(B) \leq \sigma_2(\mathbb{B}) + \|\mathcal{E}\|_{\mathrm{op}} \leq \lambda_2(\mathbb{A}_G) + 2\epsilon(1 + \epsilon) \leq 1 - \frac{1}{2}\phi^2(G_B) + 2\epsilon(1 + \epsilon).$$

$\square$

4.1.4. Random matrices. One source of matrices satisfying the spectral condition is random matrices. If we generate $B \in \mathbb{R}_{\geq 0}^{m \times n}$ as a random bipartite graph (e.g., each entry is one with probability p independently), then the resulting graph has $\phi(G_B) = \Omega(1)$ with high probability by a standard probabilistic method. Also, B is ϵ-nearly doubly balanced for small ϵ by standard concentration inequality (e.g., $\epsilon = O(\sqrt{\log m/(pm)})$ in the above example). So, by Lemma 4.5, the λ in Lemma 4.3 is $\Omega(1)$, which implies that the assumption $\lambda^2 \geq C\epsilon \ln m$ in Theorem 1.5 is satisfied with high probability. We can then apply our results to conclude that for those matrices,

1. the continuous operator scaling algorithm converges to a η-nearly doubly balanced solution in time $t = O(\log(m/\eta))$,
2. the condition number of the scaling solution is $O(1)$ from Theorem 1.7,
3. the capacity of the matrix is close to s from Theorem 1.8.

Indeed, the assumption $\lambda^2 \geq C\epsilon \ln m$ in Theorem 1.5 should hold for a large class of random nonnegative matrices where each entry is an independent random variable with reasonable distribution such as the chi-squared distribution [60], and even for some limited dependent random matrices such as k-wise independent random graphs. One can either verify the assumption by using the combinatorial condition in Lemma 4.5 or bound the second largest singular value directly using the trace method as in section 5.

4.1.5. Bipartite matching. It is known that a matrix $B \in \mathbb{R}^{n \times n}$ can be scaled to arbitrarily close to doubly stochastic if and only if the underlying bipartite graph has a perfect matching [49], and so the decision version of the bipartite perfect matching problem can be reduced to the matrix scaling problem. Moreover, the doubly stochastic scaling solution provides a fractional solution to the perfect matching problem, which can be converted to an integral solution to the perfect matching problem very efficiently using the random walks technique in [25] (see also [50]).

Our results imply that the continuous operator scaling algorithm can be used to find a fractional perfect matching in an almost regular bipartite expander graph.

COROLLARY 4.6. *Suppose $G = (X, Y; E)$ is a bipartite graph with $|X| = |Y|$ where each vertex v satisfies $(1 - \epsilon)|E|/|X| \leq \deg(v) \leq (1 + \epsilon)|E|/|X|$ for some ϵ. If $\phi(G)^4 \geq C\epsilon \ln |X|$ for some sufficiently large constant C, then the gradient flow converges to an η-nearly doubly balanced scaling (i.e., η-nearly perfect fractional matching) in time $t = O(\log |X| \log(1/\eta)/\phi^2(G))$.*

We remark that our results also imply that the second order methods for matrix scaling in [14, 2] are near linear time algorithms for the instances in Corollary 4.6. This is because the condition number κ of the scaling solution for those instances is a constant by Theorem 1.7 and the algorithms in [14, 2] have time complexity $\widetilde{O}(|E| \log \kappa)$. We also note that classical combinatorial algorithms can also achieve a similar running time in the instances in Corollary 4.6.

4.1.6. Permanent lower bound. Given a matrix $A \in \mathbb{R}^{n \times n}$, the permanent is defined as

$$\text{per}(A) = \sum_{\pi \in S_n} \prod_{i=1}^{n} a_{i,\pi(i)},$$

where S_n is the set of all permutations of n elements. Linial, Samorodnitsky, and Wigderson [49] used the matrix scaling algorithm to design a deterministic e^n-approximation algorithm for computing the permanent of a nonnegative $n \times n$ matrix. The

algorithm works by scaling the input matrix to a doubly stochastic matrix and keeping track of the change of the permanent, and then using the results in Van der Waerden's conjecture that any doubly stochastic matrix has permanent at least $n!/n^n$ and at most one to conclude the e^n-approximation.

For matrices satisfying the spectral gap condition in Lemma 4.3 (e.g., random matrices in section 4.1.4), we can use the capacity lower bound in Theorem 1.7 to argue that the continuous operator scaling algorithm doesn't do much, and thus to establish a permanent lower bound for those matrices similar to that of Van der Waerden's.

To see the proof, we first define the capacity of a matrix.

DEFINITION 4.7 (matrix capacity). *Given a matrix $B \in \mathbb{R}^{m \times n}$, define*

$$\mathrm{cap}(B) := \inf_{x \in \mathbb{R}^n, x > 0} \frac{m\big(\prod_{i=1}^m (Bx)_i\big)^{1/m}}{\big(\prod_{j=1}^n x_j\big)^{1/n}}.$$

The following lemma is probably known but it was not stated in the literature.

LEMMA 4.8. *Following the reduction in Lemma* 4.2 *from matrix scaling of B to operator scaling of $\mathcal{A}$, we have that $\mathrm{cap}(B)$ in Definition* 4.7 *is equivalent to $\mathrm{cap}(\mathcal{A})$ in Definition* 2.20.

Proof. Recall that the capacity of an operator $\mathcal{A}$ is defined as

$$\mathrm{cap}(\mathcal{A}) := \inf_{X \succ 0} \frac{m \det \left(\sum_{i=1}^k A_i X A_i^* \right)^{1/m}}{\det(X)^{1/n}}.$$

Using the reduction from Lemma 4.2, given a nonnegative matrix $B \in \mathbb{R}^{m \times n}$, we define $\mathcal{A} = (A_{11}, \ldots, A_{mn})$ where each A_{ij} is the matrix with the (i,j)th entry equal to $\sqrt{B_{i,j}}$ and all other entries zero. Then, $\sum_{i=1}^m \sum_{j=1}^n A_{ij} X A_{ij}$ is the $m \times m$ diagonal matrix with the (i,i)th entry equal to $\sum_{j=1}^n B_{i,j} X_{j,j}$. If we let $x \in \mathbb{R}^n$ be the vector of the diagonal entries of X, then the (i,i)th entry of $\sum_{i=1}^m \sum_{j=1}^n A_{ij} X A_{ij}$ is simply $(Bx)_i$. Then, the determinant of $\sum_{i=1}^m \sum_{j=1}^n A_{ij} X A_{ij}$ is simply $\prod_{i=1}^m (Bx)_i$. Finally, by Hadamard's inequality, $\det(X) \leq \prod_{j=1}^n X_{j,j}$ for any positive definite matrix X, and so we can assume the optimizer to $\mathrm{cap}(\mathcal{A})$ is a diagonal matrix, and thus $\mathrm{cap}(\mathcal{A})$ simplifies to $\mathrm{cap}(B)$ in Definition 4.7. $\square$

We are ready to prove the main result in this subsubsection.

COROLLARY 4.9. *If a nonnegative matrix $B \in \mathbb{R}^{n \times n}$ is ϵ-nearly doubly balanced with $s(B) = n$ and it satisfies the λ-spectral gap condition in Definition* B.1 *with $\lambda^2 \geq C\epsilon \log n$ for some sufficiently large constant C, then*

$$1 \geq \mathrm{per}(B) \geq \exp \left(-n \left(1 + \Theta \left(\frac{\epsilon^2}{\lambda} \right) \right) \right).$$

Proof. Let $B \in R^{n \times n}$ be the input nonnegative matrix with $s(B) = n$. Find the scaling solution L, R such that LBR is doubly stochastic (i.e., every row sum and every column sum equal to one), which is guaranteed to exist under our assumptions. Gurvits [33, 31] defined the (unnormalized) capacity of $B \in \mathbb{R}^{n \times n}$ as

$$\overline{\mathrm{cap}}(B) = \inf_{x \in \mathbb{R}^n, x > 0} \frac{\prod_{i=1}^n (Bx)_i}{\prod_{j=1}^n x_j}.$$

Note that $\overline{\mathrm{cap}}(LBR) = \det(L) \cdot \det(R) \cdot \overline{\mathrm{cap}}(B)$ and also $\mathrm{per}(LBR) = \det(L) \cdot \det(R) \cdot \mathrm{per}(B)$. Using the fact that $\overline{\mathrm{cap}}(A) = 1$ for a doubly stochastic matrix A [31, 22],

$$\overline{\mathrm{cap}}(B) = \frac{\overline{\mathrm{cap}}(B)}{\overline{\mathrm{cap}}(LBR)} = \frac{\mathrm{per}(B)}{\mathrm{per}(LBR)}.$$

Note that $\overline{\mathrm{cap}}(B) = (\mathrm{cap}(B)/n)^n$, and so the results on Van der Waerden's conjecture imply that

$$\mathrm{per}(B) = \left(\frac{\mathrm{cap}(B)}{n}\right)^n \cdot \mathrm{per}(LBR) \geq \left(\frac{\mathrm{cap}(B)}{n}\right)^n \cdot e^{-n}.$$

If B is ϵ-nearly doubly balanced with $s(B) = n$ and B satisfies the spectral gap condition in Definition B.1, then Theorem 1.8 and Lemma 4.8 imply that

$$\mathrm{cap}(B) = \mathrm{cap}(\mathcal{A}) \geq \left(1 - \frac{4\epsilon^2}{\lambda}\right) s(\mathcal{A}) = \left(1 - \frac{4\epsilon^2}{\lambda}\right) s(B) = \left(1 - \frac{4\epsilon^2}{\lambda}\right) n,$$

where $\mathcal{A}$ is the operator in the reduction from Lemma 4.2. Therefore, we conclude that

$$\mathrm{per}(B) \geq \left(1 - \frac{4\epsilon^2}{\lambda}\right)^n \cdot e^{-n} = \exp\left(-n\left(1 + \Theta\left(\frac{\epsilon^2}{\lambda}\right)\right)\right). \qquad \square$$

Example 4.10. If B is a random matrix where each entry B_{ij} is an independent random variable g_{ij}^2, where g_{ij} is sampled from the normal distribution $N(0, 1/n)$, then $\lambda = \Omega(1)$ and $\epsilon = \sqrt{\log n/n}$ with high probability. Hence, the conditions in Corollary 4.9 are satisfied and it follows that

$$\mathrm{per}(B) \geq \exp(-n - O(\log n)) = e^{-n}/\mathrm{poly}(n).$$

So, the permanent of a random matrix from this distribution has a Van der Waerden's type lower bound even though it is not doubly stochastic.

4.1.7. Optimal transport distance. Given two probability distributions and a cost function C, the optimal transport distance is the earth mover distance to move from one distribution to another distribution under the cost function. When the two probability distributions are discrete, the cost function can be represented as a cost matrix C, and the problem of computing the optimal transport distance can be formulated as the assignment problem (i.e., a generalization of the minimum cost perfect matching). So the problem can be solved in polynomial time and there is a linear programming formulation for the problem. In large scale data analysis, however, the polynomial time algorithms are not fast enough.

Using the maximum entropy principle, Cuturi [15] proposed to add an entropic regularizer to the linear program and showed that the optimal solution is the matrix scaling solution to a matrix K associated to C (more precisely $K_{i,j} = \exp(-C_{i,j}/\beta)$ where β is a parameter in the regularizer). Cuturi showed that Sinkhorn's algorithm for matrix scaling is very efficient in computing the optimal solution to the regularized linear program, and he even mentioned that Sinkhorn's algorithm exhibits linear convergence in practice [15]. Since then the "Sinkhorn distance" becomes a popular alternative/approximation to the earth mover distance and is used in computer vision and machine learning research; see the book [54] and the references therein. Theorem 1.5 provides a condition to establish the linear convergence observed, which is satisfied in many random matrices as discussed in section 4.1.4.

Also, it is of interest to bound the Sinkhorn distance, which is shown in [15, 54] to be at most

$$\langle e^{f^*/\beta}, (K \circ C) \cdot e^{g^*/\beta} \rangle,$$

where f^* and g^* are the scaling solutions to K and β is the regularizer parameter. This result states that the distance is small if the condition number of the scaling solution is small. Theorem 1.7 provides a strong bound on the condition number for Sinkhorn distance when the spectral assumption is satisfied.

4.2. Frame scaling. A frame is a collection of vectors $U = (u_1, \dots, u_n)$ where each $u_i \in \mathbb{R}^d$ for $1 \le i \le n$. The size of a frame U is defined as $s(U) := \sum_{i=1}^n \|u_i\|_2^2$. A frame U is called ϵ-nearly doubly balanced if

$$(1 - \epsilon)\frac{s(U)}{d}I_d \preceq \sum_{i=1}^n u_i u_i^* \preceq (1 + \epsilon)\frac{s(U)}{d}I_d \quad \text{and}$$

$$(1 - \epsilon)\frac{s(U)}{n}I_n \preceq \text{diag}\left(\left\{\|u_i\|_2^2\right\}_{i=1}^n\right) \preceq (1 + \epsilon)\frac{s(U)}{n}I_n$$

and is called doubly balanced when $\epsilon = 0$.

DEFINITION 4.11 (frame scaling problem). *Given a frame $U = (u_1, \dots, u_n)$ where each $u_i \in \mathbb{R}^d$, the goal is to find a matrix $M \in \mathbb{R}^{d \times d}$ such that $v_i = Mu_i / \|Mu_i\|$ satisfies $\sum_{i=1}^n v_i v_i^* = I_d$.*

Outline. In the following, we will show that the frame scaling problem can be reduced to the operator scaling problem in section 4.2.1. Then, we will see that the spectral condition has a nice form in section 4.2.2 and explain that random frames will satisfy our condition in section 4.2.3. Finally, we show a significant implication of our results to the Paulsen problem in section 4.2.4 and a construction of doubly stochastic frame with small inner products in section 4.2.5.

4.2.1. Reduction to operator scaling. The frame scaling problem is a special case of the operator scaling problem.

LEMMA 4.12. *Given a frame $U = (u_1, \dots, u_n)$ where each $u_i \in \mathbb{R}^d$, let $\mathcal{A} = (A_1, \dots, A_n)$ where each $A_i \in \mathbb{R}^{d \times n}$ for $1 \le i \le n$ is the matrix with the ith column being u_i and all other columns equal to zero. Then, U is ϵ-nearly doubly stochastic if and only if $\mathcal{A}$ is ϵ-nearly doubly stochastic. Furthermore, there is a solution to the frame scaling problem for U if and only if there is a solution to the operator scaling problem for $\mathcal{A}$.*

Proof. By construction, $\sum_{i=1}^n A_i A_i^* = \sum_{i=1}^n u_i u_i^* \in \mathbb{R}^{d \times d}$ and $\sum_{i=1}^n A_i^* A_i = \text{diag}(\{\|u_i\|_2^2\}_{i=1}^n) \in \mathbb{R}^{n \times n}$, and so U is ϵ-nearly doubly stochastic if and only if $\mathcal{A}$ is ϵ-nearly doubly stochastic. If $M \in \mathbb{R}^{d \times d}$ is a solution to the frame scaling problem for U, then we can set $L := M$ and $R := \text{diag}(\{\|Mu_i\|_2^{-1}\}_{i=1}^n)$ and see that it is a solution to the operator scaling problem for $\mathcal{A}$.

If L and R is a solution to the operator scaling problem for $\mathcal{A}$, then we can use a similar argument as in Lemma 4.2 to show that L and $(RR^*)^{1/2}$ is also a solution and $(RR^*)^{1/2}$ is a diagonal matrix as $\mathcal{A}$ has the special structure that each A_i has only one nonzero column. This is also proved in Lemma 3.7.4 in [47] so we omit the details. Since R is diagonal, the (i, i)th entry must necessarily be $\|Lu_i\|_2^{-1}$ for the doubly stochastic conditions to be satisfied, and so $M := L$ is a solution to the frame scaling problem for U. $\square$

4.2.2. Spectral condition. The spectral condition for operator scaling is related to the following Hermitian matrix.

DEFINITION 4.13 (entrywise squared Gram matrix). *Given a frame $U = (u_1, \ldots, u_n)$ where each $u_i \in \mathbb{R}^d$, the squared Gram matrix $G \in \mathbb{R}^{n \times n}$ is defined as $G_{i,j} = \langle u_i, u_j \rangle^2$ for $1 \leq i, j \leq n$.*

Note that G is a positive semidefinite matrix. To see this, let V be the $d \times n$ matrix with the ith column being u_i. Then, we can write $G = (V^*V) \circ (V^*V)$ where $\circ$ denotes the Hadamard (or entrywise) product of two matrices. As V^*V is a positive semidefinite matrix, G is a positive semidefinite matrix by the Schur product theorem. The spectral condition in Definition 1.4 translates to the following spectral condition for the squared Gram matrix in the frame scaling case.

LEMMA 4.14. *Using the reduction from Lemma* 4.12*, the spectral condition for operator scaling for $\mathcal{A}$ in Definition* 1.4 *becomes*

$$\lambda_2(G) \leq (1 - \lambda)^2 \cdot \frac{s(U)^2}{dn},$$

where $\lambda_2(G)$ is the second largest eigenvalue of G.

Proof. Since each A_i has only one nonzero column, each $A_i \otimes A_i$ has only one nonzero column which is $u_i \otimes u_i \in \mathbb{R}^d$. The matrix $M_{\mathcal{A}} \in \mathbb{R}^{d^2 \times n^2}$ has only n nonzero columns $(u_1 \otimes u_1, \ldots, u_n \otimes u_n)$. Hence, $M_{\mathcal{A}}^* M_{\mathcal{A}}$ has only a $n \times n$ nonzero submatrix, where the (i, j)th entry is $\langle u_i \otimes u_i, u_j \otimes u_j \rangle = \langle u_i, u_j \rangle^2$. So, the $n \times n$ nonzero submatrix of $M_{\mathcal{A}}$ is exactly G. Therefore, $\lambda_2(G) = \lambda_2(M_{\mathcal{A}}^* M_{\mathcal{A}}) = \sigma_2(M_{\mathcal{A}})^2$ and the spectral condition $\sigma_2(M_{\mathcal{A}}) \leq (1 - \lambda)s(\mathcal{A})/\sqrt{mn}$ is equivalent to $\lambda_2(G) \leq (1 - \lambda)^2 s(U)^2/(dn)$ as $s(\mathcal{A}) = s(U)$ and $m = d$ in the reduction from Lemma 4.12. $\square$

4.2.3. Random frames. In section 5, we will prove that if we generate $\Omega(d^{4/3})$ random unit vectors, then the resulting frame is ϵ-nearly doubly balanced for $\epsilon = O(1/\operatorname{poly}(d))$ and the λ in Lemma 4.14 satisfies $\lambda = \Omega(1)$ with high probability. Hence, a random frame generated in this way will satisfy the condition $\lambda^2 \geq C\epsilon \ln d$ and our results apply to these random frames. The proof is by a trace method. We believe that the trace method can be improved to prove that generating $\Omega(d \operatorname{polylog} d)$ random unit vectors will satisfy our condition.

Remark 4.15. After the submission of this article, it was shown in [20] that in fact generating $\Omega(d \operatorname{polylog} d)$ random unit vectors produces a frame satisfying the required spectral condition with high probability. The authors of [20] did not improve the trace method, but instead used a version of Cheeger's inequality to better use concentration properties of random unit vectors. More details are given in section 5.

4.2.4. The Paulsen problem in random frames. Given an ϵ-nearly doubly balanced frame $U = (u_1, \ldots, u_n)$ with size $s(U) = d$ where each $u_i \in \mathbb{R}^d$, the Paulsen problem asks to find a doubly balanced frame $V = (v_1, \ldots, v_n)$ that is "close" to U. Given two frames U, V, the squared distance between them is defined as $\operatorname{dist}^2(U, V) = \sum_{i=1}^{n} \|u_i - v_i\|_2^2$. It was an open question whether for every ϵ-nearly doubly balanced frame U with $s(U) = d$, there is always a doubly balanced frame V with $\operatorname{dist}^2(U, V)$ bounded by a function only dependent on d and ϵ but independent of n. Recently, this question was answered affirmatively in [47], showing that for any ϵ-nearly doubly balanced frame U with $s(U) = d$, there is always a doubly balanced frame V with $\operatorname{dist}^2(U, V) = O(d^{13/2}\epsilon)$. Very recently, Hamilton and Moitra [34] proved a stronger bound $O(d^2\epsilon)$ with a much simpler proof. On the other hand, there are examples

showing that the best bound is at least $\Omega(d\epsilon)$, so the upper bound and the lower bound are within a factor of d.

The Paulsen problem was asked because it is difficult to generate doubly balanced frames and easier to generate nearly doubly balanced frames, but actually not many ways are known to even generate ϵ-nearly doubly balanced frames for small ϵ. Most nearly doubly balanced frames that we know are random frames (e.g., random Gaussian vectors, random unit vectors), which can be shown to be ϵ-nearly doubly balanced for small ϵ by matrix concentration inequalities (see section 5.1). So, for the Paulsen problem, the inputs of interest are random frames.

We will prove that for a random frame U with $s(U) = d$ that is ϵ-nearly doubly balanced, there is a doubly balanced frame V with $\mathrm{dist}^2(U, V) = O(d\epsilon^2)$ with high probability, which is much smaller than the worst case $\Omega(d\epsilon)$ bound. We will also show how this result can be used to generate a frame in which every pair of vectors has small inner product in the next subsubsection.

The proof has two steps. The first step is to show that if we generate $n = \Omega(d^{4/3})$ random unit vectors, then the resulting frame U is ϵ-nearly doubly balanced for $\epsilon \leq O(1/\operatorname{poly}(d))$ and also satisfies the spectral gap condition in Lemma 4.14 with $\lambda = \Omega(1)$. Therefore, the assumption in Theorem 1.5 is satisfied and the continuous operator scaling algorithm has linear convergence. The second step is to show that if the continuous operator scaling algorithm has linear convergence, then the "total movement" to a doubly balanced frame is $O(d\epsilon^2)$.

The first step will be proved in section 5. We will prove the second step here. The following lemma states the result in [47] that we will use.

LEMMA 4.16 (Theorem 3.3.5, Lemma 3.3.1, Lemma 3.4.3 in [47]). *The dynamical system in Definition* 2.16 *will move the input operator $\mathcal{A}^{(0)}$ to a doubly balanced operator $\mathcal{A}^{(\infty)}$. For any time $T \geq 0$,*

$$\mathrm{dist}^2(\mathcal{A}^{(T)}, \mathcal{A}^{(0)}) \leq \left(\int_0^T \sqrt{\sum_{i=1}^k \left\| \frac{d}{dt} A_i^{(t)} \right\|_F^2} \, dt \right)^2 = \frac{1}{4} \left(\int_0^T \sqrt{-\frac{d}{dt}\Delta^{(t)}} \, dt \right)^2.$$

The second step actually holds in the more general operator setting, not just in the frame setting.

LEMMA 4.17. *Given an operator $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$ with $m \leq n$ for $1 \leq i \leq k$, if $\mathcal{A}$ is ϵ-nearly doubly balanced and $\mathcal{A}$ satisfies the λ-spectral gap condition in Definition 1.4 with $\lambda^2 \geq C\epsilon \ln m$ for a sufficiently large constant C, then*

$$\mathrm{dist}^2(\mathcal{A}^{(0)}, \mathcal{A}^{(\infty)}) \leq \frac{s^{(0)}\epsilon^2}{\lambda}.$$

Proof. Given the assumptions, Theorem 3.21 implies that

$$-\frac{d}{dt}\Delta^{(t)} \geq \lambda s^{(0)}\Delta^{(t)} \implies \frac{-\frac{d}{dt}\Delta^{(t)}}{\sqrt{\lambda s^{(0)}\Delta^{(t)}}} \geq \sqrt{-\frac{d}{dt}\Delta^{(t)}} \implies -\frac{2}{\sqrt{\lambda s^{(0)}}}\frac{d}{dt}\sqrt{\Delta^{(t)}} \geq \sqrt{-\frac{d}{dt}\Delta^{(t)}}.$$

By Lemma 4.16 and the above inequality, for any $T \geq 0$,

$$\mathrm{dist}^2\left(\mathcal{A}^{(T)}, \mathcal{A}^{(0)}\right) \leq \frac{1}{4}\left(\int_0^T \sqrt{-\frac{d}{dt}\Delta^{(t)}} \, dt \right)^2 \leq \frac{1}{\lambda s^{(0)}}\left(\int_0^T \frac{d}{dt}\sqrt{\Delta^{(t)}} \, dt \right)^2$$

$$\leq \frac{\Delta^{(0)}}{\lambda s^{(0)}} \leq \frac{s^{(0)}\epsilon^2}{\lambda},$$

where the last inequality is by Lemma 2.15.   □

Combining the two steps gives the following theorem.

THEOREM 4.18. *Let* $U = (u_1, \ldots, u_n)$ *be a random frame with* $n = \Omega(d^{4/3})$, *where each* $u_i \in \mathbb{R}^d$ *is an independent random vector with* $\|u_i\|_2^2 = d/n$. *Then, with probability at least* 0.99, *there is a doubly balanced frame* V *with* $\mathrm{dist}^2(U, V) \leq O(d\epsilon^2)$ *if* U *is* ϵ-*nearly doubly balanced.*

Proof. By Theorem 5.1, the random frame U satisfies the spectral gap condition in Lemma 4.14 with constant λ and $\epsilon \ll 1/\ln d$ with probability at least 0.99. Note that Theorem 5.1 is stated when each $\|u_i\|_2^2 = 1$ but it is easy to see that the nearly doubly balanced condition and the spectral gap condition are unchanged upon scaling the vectors to $\|u_i\|_2^2 = d/n$ for $1 \leq i \leq n$. By the reduction in Lemma 4.12 and the spectral gap condition in Lemma 4.14, this implies that the condition $\lambda^2 \geq C\epsilon \ln d$ for operator scaling is satisfied and also $s(U) = d$. Therefore, by Lemma 4.17, the continuous operator scaling algorithm will move U to a doubly balanced frame V with $\mathrm{dist}^2(U, V) \leq O(d\epsilon^2)$. □

4.2.5. Constructing frames with small inner products. The original motivation for the Paulsen problem was to construct doubly balanced frames with some additional structure.

DEFINITION 4.19. *A frame* $V = \{v_1, \ldots, v_n\}$ *is equiangular if* $\langle v_i, v_j \rangle^2$ *is the same for all* $i \neq j$.

For $n = \Theta(d^2)$, finding a doubly balanced frame that is also equiangular will have implications for certain informationally complete quantum measurement operators. It is a major open problem in frame theory for which pairs (n, d) such frames exist [59]. The known examples are sporadic and based on group/number-theoretic constructions. We consider a related but more relaxed problem.

DEFINITION 4.20. *A doubly balanced frame is Grassmannian if its angle*

$$\theta(V) := \max_{i \neq j} \langle v_i, v_j \rangle^2$$

is minimized over all possible doubly balanced frames.

Doubly balanced frames with small angle are useful in constructing erasure codes [38, 58]. The original motivation of the Paulsen problem was to begin with some ϵ-nearly doubly balanced frame U that has small $\theta(U)$, and see if it could be "rounded" to a nearby doubly balanced frame V still having small $\theta(V)$. Bounding $\mathrm{dist}^2(U, V)$ is one way to achieve this goal.

In this section, we use the results in the spectral analysis to construct a doubly balanced frame with small angle. The idea is to start with a random frame U which is ϵ-nearly doubly balanced for small ϵ and has small $\theta(U)$ with high probability, and then use the results in spectral analysis to show that we can scale U to a doubly balanced frame V with $\theta(V) \approx \theta(U)$.

THEOREM 4.21. *For any* $n \geq \Omega(d^{4/3})$, *there exists a doubly balanced frame* $V = (v_1, \ldots, v_n)$ *where each* $v_i \in \mathbb{R}^d$ *with* $\|v_i\| = 1$ *and*

$$\theta(V) \leq O\left(\frac{\log n}{d} + \frac{d \log^3 d}{n}\right).$$

Remark 4.22. As a simple consequence of the improvement of Theorem 5.1 given in [20] (discussed in Remark 5.2), the conditions and conclusion of Theorem 4.21 can be improved as follows.

For any $n \geq \Omega(d \log^2 d)$, there exists a doubly balanced frame V where each $v_i \in S^{d-1}$ and

$$\theta(V) \leq O\left(\frac{\log n}{d} + \frac{d \log^2 d}{n}\right).$$

Proof. First, we generate a random frame $U = (u_1, \ldots, u_n)$ where each $u_i \in \mathbb{R}^d$ is an independent random unit vector with $\|u_i\| = 1$. By Lemma 5.4 and Theorem 5.1, U is ϵ-nearly doubly balanced for $\epsilon \leq O(\sqrt{d \log d / n})$ and satisfies the λ-spectral gap condition with $\lambda = \Omega(1)$ with probability at least 0.99. Next, we bound $\theta(U)$ using the following fact.

FACT 4.23 ([36]). *Let $x \in S^{d-1}$ be a fixed unit vector. For a random unit vector $u \sim S^{d-1}$,*

$$\mathbb{P}[\langle u, x \rangle^2 \geq t^2] \leq 2\sqrt{2} \exp(-t^2 d/4).$$

Choosing a large enough upper bound and applying union bound, it follows from the above fact and rotational invariance that

$$\mathbb{P}\left[\langle u_i, u_j \rangle^2 \geq \frac{12 \log n}{d}\right] \leq O\left(\exp\left(-\frac{12 d \log n}{4d}\right)\right) \leq O\left(n^{-3}\right)$$

$$\implies \quad \mathbb{P}\left[\theta(U) \geq \frac{12 \log n}{d}\right] \leq O(n^{-1}).$$

By Theorem 3.21 and the reduction in Lemma 4.12, there is a left scaling matrix $L \in \mathbb{R}^{d \times d}$ and a right diagonal scaling matrix $R \in \mathbb{R}^{n \times n}$ such that if we set $v_i = L u_i R_{ii}$, then the frame $V = (v_1, \ldots, v_n)$ is doubly balanced. By Theorem 3.22, the scaling solutions L, R satisfy

$$\|L - I\|_{\mathrm{op}} \leq \zeta \text{ and } \|R - I\|_{\mathrm{op}} \leq \zeta \quad \text{for } \zeta \leq O\left(\frac{\epsilon \log d}{\lambda}\right) \leq O\left(\sqrt{\frac{d \log^3 d}{n}}\right).$$

Using the arguments as in Lemma 3.20 (or Lemma B.17), we have

$$|\langle v_i, v_j \rangle - \langle u_i, u_j \rangle| = |\langle L u_i R_{ii}, L u_j R_{jj} \rangle - \langle u_i, u_j \rangle| \leq O(\zeta) \cdot \|u_i\|_2 \|u_j\|_2 = O(\zeta).$$

Therefore, we conclude that

$$\theta(V) \leq 2\theta(U) + O(\zeta^2) \leq O\left(\frac{\log n}{d} + \frac{d \log^3 d}{n}\right). \qquad \square$$

When $n = \Theta(d^2)$ the above theorem gives $\theta(V) \leq O(\log^3 d/d)$, and when $n = \Theta(d^2 \log^2 d)$ then the above theorem gives $\theta(V) \leq O(\log d/d)$.

4.3. Operator scaling. The operator scaling problem was used to compute the Brascamp–Lieb constant [23] and to compute the noncommutative rank of a symbolic matrix [22]. It is also used in [1] to solve the orbit intersection problem for the left-right group action.

4.3.1. Brascamp–Lieb constants. A Brascamp–Lieb datum is specified by an m-tuple $\mathbf{B} = \{B_j : \mathbb{R}^n \to \mathbb{R}^{n_j} \mid 1 \leq j \leq m\}$ of linear transformations and an m-tuple of exponents $\mathbf{p} = \{p_1, \ldots, p_m\}$. The Brascamp–Lieb constant $\mathrm{BL}(\mathbf{B}, \mathbf{p})$ of this datum

is defined as the smallest C such that for every m-tuple $\{f_j : \mathbb{R}^{n_j} \to \mathbb{R}_{\geq 0} \mid 1 \leq j \leq m\}$ of nonnegative integrable functions, we have

$$\int_{x \in \mathbb{R}^n} \prod_{j=1}^m \left(f_j(B_j x)\right)^{p_j} dx \leq C \prod_{j=1}^m \left(\int_{x_j \in \mathbb{R}^{n_j}} f_j(x_j) dx_j\right)^{p_j}.$$

For this inequality to be scale invariant in $\{f_1, \ldots, f_m\}$, we must have $\sum_j p_j n_j = n$. This is a common generalization of many useful inequalities; see [8, 23].

The important point we need is that the optimizers of any nondegenerate Brascamp–Lieb datum (i.e., the functions $f_1, \ldots, f_m$ for which the inequality is tight) is achieved by density functions of appropriately centered Gaussians [48], and this implies that the Brascamp–Lieb constant $\mathrm{BL}(\mathbf{B}, \mathbf{p})$ can be written as the following optimization problem:

$$\mathrm{BL}(\mathbf{B}, \mathbf{p}) = \left[\sup_{X_j \succ 0} \frac{\prod_{j=1}^m \left(\det(X_j)\right)^{p_j}}{\det\left(\sum_{j=1}^m p_j B_j^* X_j B_j\right)}\right]^{1/2},$$

which is closely related to the capacity of an operator.

An BL-datum is called geometric if we have

$$\sum_{j=1}^m p_j B_j^* B_j = I_n \quad \text{and} \quad B_j B_j^* = I_{n_j} \quad \text{for } 1 \leq j \leq m.$$

It is proved in [4, 5] that the BL-constant is one when the BL-datum is geometric. We will show that the BL-constant is small when the BL-datum is nearly geometric and satisfies a spectral condition, using the reduction in [23] from BL-constant to operator capacity and our capacity lower bound in Theorem 1.8.

Reduction. We describe the reduction in [23] from computing the BL-constant of a datum to computing the capacity of an operator. Let $p_j = c_j/d$ be rational numbers where c_j and d are integers. Given a BL-datum $(\mathbf{B}, \mathbf{p})$, a completely positive map $\Phi_{\mathcal{A}} : \mathbb{R}^{nd \times nd} \to \mathbb{R}^{n \times n}$ is constructed as follows. For intuition, think of the "intended" input matrix X to $\Phi_{\mathcal{A}}$ as a block diagonal matrix, with c_j blocks of $X_j \in \mathbb{R}^{n_j \times n_j}$ for $1 \leq j \leq m$, so that X is a square matrix with dimension $\sum_{j=1}^m c_j n_j = d \sum_{j=1}^m p_j n_j = dn$. For each $B_j \in \mathbb{R}^{n_j \times n}$ in $\mathbf{B}$, we create c_j matrices $\{A_{j1}, \ldots, A_{jc_j}\}$ in $\mathcal{A}$, where each $A_{ji} \in \mathbb{R}^{n \times dn}$ has a copy of $B_j/\sqrt{d}$ that acts only on the (j, i)th principle block of X (i.e., the ith copy of X_j in X) and all other entries of A_{ji} are zero. The operator $\mathcal{A}$ is defined by the Kraus operators $\cup_{j=1}^m \cup_{i=1}^{c_j} \{A_{ji}\}$, with the completely positive map

$$\Phi_{\mathcal{A}}(X) = \sum_{j=1}^m \sum_{i=1}^{c_j} A_{ji}^* X A_{ji} = \frac{1}{d} \sum_{j=1}^m \sum_{i=1}^{c_j} B_j^* X_{ji} B_j \quad \text{and} \quad \Phi_{\mathcal{A}}^*(Y) = \bigoplus_{j=1}^m \bigoplus_{i=1}^{c_j} \frac{1}{d} B_j Y B_j^*,$$

where X_{ji} is the (j, i)th principle block of X as described above, and the notation $\oplus$ denotes the direct sum of the matrices (i.e., putting each matrix in a diagonal block).

THEOREM 4.24 ([23]). *It follows from the reduction that*

$$\left(\frac{\mathrm{cap}(\mathcal{A})}{n}\right)^n = \left(\frac{1}{\mathrm{BL}(\mathbf{B}, \mathbf{p})}\right)^2.$$

Using this connection, it is shown in [23] that the Brascamp-Lieb constant $\mathrm{BL}(\mathbf{B}, \mathbf{p})$ can be computed by an operator scaling algorithm for $\mathcal{A}$.

Bounding BL-constants. Using Theorem 4.24, we would like to derive upper bounds on BL-constants using the capacity lower bound in Theorem 1.8 and show that for some random instances the BL-constant is small. To apply Theorem 1.8, we translate the definitions of ϵ-nearly doubly balanced operator and the λ-spectral gap conditions to the Brascamp–Lieb setting. Following the reduction from $\mathbf{B}, \mathbf{p}$ to $\mathcal{A}$, we have the following definitions from the corresponding definitions of the operator $\mathcal{A}$.

DEFINITION 4.25 (size of a datum). *The size of a BL-datum* $(\mathbf{B}, \mathbf{p})$ *is*

$$s(\mathbf{B}, \mathbf{p}) := p_j \sum_{j=1}^{m} \|B_j\|_F^2 .$$

The datum $(\mathbf{B}, \mathbf{p})$ is ϵ-nearly geometric if and only if the corresponding operator $\mathcal{A}$ is ϵ-nearly doubly balanced.

DEFINITION 4.26 (nearly geometric datum). *A datum* $\mathrm{BL}(\mathbf{B}, \mathbf{p})$ *is* ϵ-nearly geometric if

$$(1-\epsilon)\frac{s}{n}I_n \preceq \sum_{j=1}^{m} p_j B_j^* B_j \preceq (1+\epsilon)\frac{s}{n}I_n \quad \text{and} \quad (1-\epsilon)\frac{s}{n}I_{n_j} \preceq B_j B_j^* \preceq (1+\epsilon)\frac{s}{n}I_{n_j} \text{ for } 1 \le j \le m.$$

The datum $(\mathbf{B}, \mathbf{p})$ satisfies the λ-spectral gap condition if and only if the corresponding operator $\mathcal{A}$ satisfies the λ-spectral gap condition.

DEFINITION 4.27 (spectral gap of datum). *Let* $\bar{n} = \sum_{j=1}^{m} n_j$ *and* $\bar{B}^* \in \mathbb{R}^{n \times \bar{n}}$ *be the matrix*

$$\bar{B}^* := [B_1^*, B_2^*, \ldots, B_m^*].$$

Let $\bar{B}_j \in \mathbb{R}^{\bar{n} \times n}$ *be* $\bar{B}$ *with all but the jth block zeroed out, i.e.,* $\bar{B}_j^* := [0, \ldots, 0, B_j^*, 0, \ldots, 0]$. *The natural matrix representation* $M_{\mathbf{B}, \mathbf{p}} \in \mathbb{R}^{\bar{n}^2 \times n^2}$ *of the datum* $(\mathbf{B}, \mathbf{p})$ *is defined as*

$$M_{\mathbf{B}, \mathbf{p}} := \sum_{j=1}^{m} \sqrt{p_j} \cdot \bar{B}_j \otimes \bar{B}_j.$$

The datum $(\mathbf{B}, \mathbf{p})$ *is said to have a λ-spectral gap if*

$$\sigma_2(M_{\mathbf{B}, \mathbf{p}}) \le (1 - \lambda)\frac{s(\mathbf{B}, \mathbf{p})}{n}.$$

With these definitions, we can state the Brascamp–Lieb constant upper bound that follows from the capacity lower bound in Theorem 1.8.

COROLLARY 4.28. *Given a datum* $(\mathbf{B}, \mathbf{p})$ *with* $B_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$ *for* $1 \le j \le n$ *and* $\sum_{j=1}^{m} p_j n_j = n$, *if* $(\mathbf{B}, \mathbf{p})$ *is ϵ-nearly geometric and satisfies the λ-spectral gap condition with* $\lambda^2 \ge C\epsilon \log n$ *for some sufficiently large constant C, then*

$$\left(\frac{s}{n}\right)^{-n/2} \le \mathrm{BL}(\mathbf{B}, \mathbf{p}) \le \left(\left(\frac{s}{n}\right)\left(1 - \frac{4\epsilon^2}{\lambda}\right)\right)^{-n/2}.$$

Let's consider a concrete example to demonstrate the corollary.

Example 4.29. An interesting special case of the Brascamp–Lieb inequality is the rank one case $B_j = u_j^*$ where $u_j \in \mathbb{R}^d$ and $n_j = 1$ and $p_j = d/m$ for $1 \leq j \leq m$ which was studied in [5]. Consider a random rank one datum where each u_i is an independent random unit vector of $\|u_i\| = 1$. Following the reduction,

$$\text{cap}(\mathcal{A}) = \sup_{x \in \mathbb{R}^n : x > 0} \frac{d \left(\det \left(\sum_{j=1}^m x_j u_j u_j^* \right) \right)^{1/d}}{\left(\prod_{j=1}^m x_j \right)^{1/m}},$$

which is a form that is also studied in approximation algorithms [52]. Note that this is exactly the capacity of a frame $U = (u_1, \ldots, u_m)$ through the reduction in Lemma 4.12. By Theorem 5.1, if $m \geq \Omega(d^{4/3})$, then U is ϵ-nearly doubly balanced for $\epsilon \leq O(\sqrt{d \log d / m})$ and satisfies the λ-spectral gap condition with $\lambda = \Omega(1)$ with high probability. If we only used the capacity lower bound based on Δ that is given in [22] and [23], we get $\text{cap}(\mathcal{A}) \geq (1 - 4md\epsilon) s(U)$. Note this is nontrivial only for $\epsilon < \frac{1}{4md}$. On the other hand, applying Theorem 1.8 we are able to conclude that

$$\text{cap}(\mathcal{A}) \geq \left(1 - \frac{4\epsilon^2}{\lambda} \right) s(U) \geq \left(1 - \frac{4d \log d}{m} \right) m,$$

and from Corollary 4.28 the BL-constant for this datum is

$$1 \leq \text{BL}(\mathbf{B}, \mathbf{p}) \leq \left(1 - \frac{4d \log d}{m} \right)^{-m/2} = \exp(\Theta(d \log d)) = d^{\Theta(d)}.$$

This is independent on the number of vectors m and is much smaller than the worst case bound.

As another example, Hastings' result [37] implies for $\{U_1, \ldots, U_N\}$ random unitaries, the BL-datum $(\mathbf{U}, \frac{1}{N}\vec{1}_N)$ has small Brascamp–Lieb constant with high probability.

4.3.2. Rank nondecreasing operator. In [22, 21, 31], a polynomial time algorithm for computing the noncommutative rank of a symbolic matrix is designed using operator scaling. Given $\mathcal{A} = (A_1, \ldots, A_k)$ where each $A_i \in \mathbb{R}^{n \times n}$, let $Z_{\mathcal{A}} = \sum_{i=1}^k x_i A_i$ be the symbolic matrix defined by $\mathcal{A}$ over noncommutative variables $x_1, \ldots, x_k$, and the noncommutative rank nc-rank(Z) of Z is defined as the smallest r such that $Z = KM$ where K is of dimension $n \times r$ and M is of dimension $r \times n$ with entries in the "free skew field" of x (see [22, 21] for definitions). The algorithm in [22, 21, 31] is based on the following equivalent characterizations.

THEOREM 4.30 ([22, 21, 31]). *Given $\mathcal{A} = (A_1, \ldots, A_k)$ where each $A_i \in \mathbb{R}^{n \times n}$, the following conditions are equivalent:*
1. *The symbolic matrix $Z_{\mathcal{A}}$ is singular, i.e., nc-rank$(Z) < n$.*
2. *$\mathcal{A}$ has a shrunk subspace, i.e., there exists subspaces U, W with $\dim(W) < \dim(U)$ such that $A_i U \subseteq W$ for all $1 \leq i \leq k$.*
3. *The completely positive linear map $\Phi_{\mathcal{A}}$ is rank decreasing, i.e., there exists $P \succ 0$ and $\text{rank}(\Phi_{\mathcal{A}}(P)) < \text{rank}(P)$.*

The alternating scaling algorithm for operator scaling is used to check whether $\Phi_{\mathcal{A}}$ is rank nondecreasing. It is shown in [22, 21, 31] that $\Phi_{\mathcal{A}}$ is rank nondecreasing if and only if $\mathcal{A}$ can be scaled to ϵ-nearly balanced for $\epsilon \leq 1/\text{poly}(n)$, and so a polynomial

time algorithm for operator scaling can be used to compute the noncommutative rank of a symbolic matrix over the reals.

The shrunk subspace condition is closely related to the concept of Hall-blocker in matching theory. In the matrix case, it is shown in Lemma 4.5 that a matrix B satisfying the spectral condition is an almost regular bipartite expander graph, so there is no Hall-blocker and it always has a perfect matching as shown in Lemma 4.6. In the operator case, intuitively, the spectral condition is closely related to the notion of quantum expander (section 2.1), and so there should be no Hall-blocker as well. Theorem 1.5 implies that it is the case.

COROLLARY 4.31. *Given an operator $\mathcal{A}$ satisfying the conditions of Theorem* 1.5, $\Phi_{\mathcal{A}}$ *is rank nondecreasing and the corresponding symbolic matrix $Z_{\mathcal{A}}$ is nonsingular over reals.*

This is a new sufficient condition for an operator to be rank nondecreasing. We remark that the assumption can be weakened to $\lambda \geq 6\epsilon$ to get the same conclusion, but we omit the proof here.

4.3.3. The operator Paulsen problem. Given an ϵ-nearly doubly stochastic operator $\mathcal{A} = (A_1, \ldots, A_k)$ where each $A_i \in \mathbb{R}^{m \times n}$, the operator Paulsen problem asks us to find a doubly stochastic operator $\mathcal{B} = (B_1, \ldots, B_k)$ where each $B_j \in \mathbb{R}^{m \times n}$ with $\operatorname{dist}^2(\mathcal{A}, \mathcal{B}) := \sum_{i=1}^{k} \|A_i - B_i\|_F^2$. In [47], it was proved that $\operatorname{dist}^2(\mathcal{A}, \mathcal{B}) \leq O(mns\epsilon)$, and this result was used in [1] for the orbit intersection problem. For an operator $\mathcal{A}$ that satisfies the spectral gap condition with constant λ, Lemma 4.17 implies a much stronger bound that $\operatorname{dist}^2(\mathcal{A}, \mathcal{B}) \leq O(s\epsilon^2)$.

5. Spectral gap of random frames. In this section, we prove that a random frame is ϵ-nearly doubly stochastic for $\epsilon \ll 1/\ln d$ and satisfies the spectral gap condition for constant λ with high probability.

THEOREM 5.1. *If we generate n random unit vectors $v_1, \ldots, v_n$ in $\mathbb{R}^d$ with $n = \Omega(d^{4/3})$, then the resulting frame is ϵ-nearly doubly stochastic for $\epsilon \ll 1/\ln d$ and satisfies the spectral gap condition in Definition* 4.14 *with constant λ with probability at least* 0.99.

Remark 5.2. After the submission of this article, it was shown in [20] that the parameters in this theorem could be strictly improved. The authors of [20] used a version of Cheeger's inequality to show that generating $n \gtrsim d \log d$ random unit vectors produces a frame satisfying the spectral gap condition in Definition 4.14 with constant λ and failure probability $\leq \exp(-\Omega(n))$. They also used better matrix concentration results for random unit vectors to show that $n \gtrsim d/\epsilon^2$ random unit vectors produces an ϵ-nearly doubly balanced frame, again with failure probability $\leq \exp(-\Omega(n\epsilon^2))$. Therefore they were able to improve Theorem 5.1 to show $n = \Omega(d \log^2 d)$ random unit vectors is ϵ-nearly doubly stochastic for $\epsilon \ll 1/\ln d$ and satisfies the spectral gap condition in Definition 4.14 with constant λ with probability at least $1 - \exp(-\Omega(n/\log^2 d))$.

To generate a random unit vector $v \in \mathbb{R}^d$, we set each coordinate of v to be an independent random Gaussian variable $N(0, \frac{1}{d})$ for $1 \leq i \leq d$, and then we scale the vector to have norm one. The size of the frame is $s = \sum_{i=1}^{n} \|v_i\|_2^2 = n$. By construction, the frame $V := (v_1, \ldots, v_n)$ satisfies the equal norm condition.

In section 5.1, we will prove that V is ϵ-nearly doubly stochastic with high probability by using a standard matrix concentration bound. Then, in section 5.2, we will prove that the squared Gram matrix G in Definition 4.13 satisfies the spectral gap condition in Definition 4.14 with high probability by using the trace method.

5.1. Nearly doubly balanced condition by matrix concentration. By construction, each vector v_i has $\|v_i\|_2 = 1$ and $s = \sum_{i=1}^n \|v_i\|_2^2 = n$. So, for the nearly doubly stochastic condition, it remains to prove that $V = (v_1, \ldots, v_n)$ is ϵ-nearly Parseval for $\epsilon \ll 1/\log d$ with high probability when $n = \Omega(d^{4/3})$, i.e.,

$$(1-\epsilon)\frac{n}{d}I_d = (1-\epsilon)\frac{s}{d}I_d \preceq \sum_{i=1}^n v_i v_i^* \preceq (1+\epsilon)\frac{s}{d}I_d = (1+\epsilon)\frac{n}{d}I_d.$$

We establish this by using the following matrix Bernstein bound.

THEOREM 5.3 (matrix Bernstein [62]).   *Let $X_1, \ldots, X_n$ be independent random matrices in $\mathbb{R}^{d \times d}$. Assume that, for $1 \le i \le n$,*

$$\mathbb{E}X_i = 0 \quad \text{and} \quad \|X_i\|_{\mathrm{op}} \le L,$$

and

$$\nu := \max\left\{\left\|\sum_{i=1}^n \mathbb{E}(X_i X_i^*)\right\|_{\mathrm{op}}, \ \left\|\sum_{i=1}^n \mathbb{E}(X_i^* X_i)\right\|_{\mathrm{op}}\right\}.$$

Then, for all $\ell \ge 0$,

$$\mathbb{P}\left[\left\|\sum_{i=1}^n X_i\right\|_{\mathrm{op}} \ge \ell\right] \le 2d\exp\left(\frac{-\ell^2/2}{\nu + L\ell/3}\right).$$

LEMMA 5.4.   *If we generate n random unit vectors $v_1, \ldots, v_n$ in $\mathbb{R}^d$ with $n = O(d\log d/\epsilon^2)$, then*

$$(1-\epsilon)\frac{n}{d}I_d \preceq \sum_{i=1}^n v_i v_i^* \preceq (1+\epsilon)\frac{n}{d}I_d$$

with probability at least $1 - O(1/\operatorname{poly}(d))$.

Proof. To apply the matrix Bernstein bound, we consider the random matrix $X_i := v_i v_i^* - \frac{1}{d}I_d$ for $1 \le i \le n$. We check the assumptions in Theorem 5.3. First, as the covariance matrix of a Gaussian vector is a scaled identity matrix and we scale it so that $\operatorname{tr}(v_i v_i^*) = 1$, we have

$$\mathbb{E}[X_i] = \mathbb{E}\left[v_i v_i^* - \frac{1}{d}I_d\right] = \frac{1}{d}I_d - \frac{1}{d}I_d = 0.$$

Second, as each $v_i v_i^*$ is of rank one, the operator norm of X_i is achieved at v_i and

$$\|X_i\|_{\mathrm{op}} = \left\|v_i v_i^* - \frac{1}{d}I_d\right\|_{\mathrm{op}} = \left\|\left(v_i v_i^* - \frac{1}{d}I_d\right)v_i\right\|_2 = \left\|v_i - \frac{1}{d}v_i\right\|_2 = 1 - \frac{1}{d}.$$

Finally, as each X_i is Hermitian,

$$\mathbb{E}[X_i X_i^*] = \mathbb{E}[X_i^2] = \mathbb{E}\left[\left(v_i v_i^* - \frac{1}{d}I_d\right)^2\right] = \mathbb{E}[v_i v_i^*] - \frac{2}{d}\mathbb{E}[v_i v_i^*] + \frac{1}{d^2}I_d = \frac{1}{d}\left(1-\frac{1}{d}\right)I_d,$$

and thus

$$\nu = \left\|\sum_{i=1}^n \mathbb{E}[X_i X_i^*]\right\|_{\mathrm{op}} = \frac{n}{d}\left(1-\frac{1}{d}\right).$$

Therefore, we can bound the probability that the ϵ-Parseval condition is not satisfied by Theorem 5.3 with $\ell = \epsilon n/d$ and $L = 1 - 1/d$, which gives

$$
\mathbb{P}\left[\left\|\sum_{i=1}^{n} v_i v_i^* - \frac{n}{d} I_d\right\|_{\mathrm{op}} \geq \frac{\epsilon n}{d}\right] = \mathbb{P}\left[\left\|\sum_{i=1}^{n} X_i\right\|_{\mathrm{op}} \geq \frac{\epsilon n}{d}\right] \leq 2d \exp\left(\frac{-n\epsilon^2}{2(d-1)(1+\epsilon/3)}\right).
$$

Therefore, for $\epsilon \leq 1$, by setting $n \geq \Omega(d \log d/\epsilon^2)$, this failure probability is at most inverse polynomial in d. □

For our condition $\lambda^2 \gg \epsilon \log d$ to be satisfied, it is sufficient for $\lambda = \Omega(1)$ that we will show and $\epsilon \ll 1/\log d$, and Lemma 5.4 gives the following bound for the latter condition.

COROLLARY 5.5. *If we generate n random unit vectors $v_1, \ldots, v_n$ in $\mathbb{R}^d$ with $n = O(d \log^3 d)$, then*

$$
(1-\epsilon)\frac{s}{d}I_d \preceq \sum_{i=1}^{n} v_i v_i^* \preceq (1+\epsilon)\frac{s}{d}I_d
$$

for $\epsilon \ll 1/\log d$ with probability at least $1 - O(1/\mathrm{poly}(d))$.

5.2. Spectral gap condition by trace method. Our goal is to prove that

$$
\lambda_2(G) \leq (1-\lambda)^2 \cdot \frac{s^2}{dn} = (1-\lambda)^2 \cdot \frac{n}{d}
$$

when we generate $n = \Omega(d^{4/3})$ independent random unit vectors $v_1, \ldots, v_n$.

5.2.1. Trace method. As in most results from random matrix theory, we use the trace method to bound $\lambda_2(G)$.

LEMMA 5.6. *For any natural number k,*

$$
\mathbb{P}\left[\lambda_2(G) \geq (1-\lambda)^2 \cdot \frac{n}{d}\right] \leq \left(\mathbb{E}[\mathrm{tr}(G^k)] - \left(\frac{n}{d}\right)^k\left(1 + \frac{d-1}{n}\right)^k\right) \Big/ \left((1-\lambda)^{2k}\left(\frac{n}{d}\right)^k\right).
$$

Proof. Recall that G is positive semidefinite from section 4.2.2. Since all the eigenvalues of G are nonnegative, for any natural number k, $\lambda_2(G)^k \leq \mathrm{tr}(G^k) - \lambda_1(G)^k$ and thus $\mathbb{E}[\lambda_2(G)^k] \leq \mathbb{E}[\mathrm{tr}(G^k)] - \mathbb{E}[\lambda_1(G)^k]$. We bound the failure probability by applying Markov's inequality on the kth moment of λ_2 so that

$$
\mathbb{P}\left[\lambda_2(G) \geq (1-\lambda)^2 \cdot \frac{n}{d}\right] \leq \frac{\mathbb{E}[\lambda_2(G)^k]}{(1-\lambda)^{2k}\left(\frac{n}{d}\right)^k} \leq \frac{\mathbb{E}[\mathrm{tr}(G^k)] - \mathbb{E}[\lambda_1(G)^k]}{(1-\lambda)^{2k}\left(\frac{n}{d}\right)^k}.
$$

We lower bound the term $\mathbb{E}[\lambda_1(G)^k]$ by using the test vector $\vec{1}/\sqrt{n}$ so that

$$
\mathbb{E}[\lambda_1(G)^k] \geq \mathbb{E}\left[\left\langle \frac{\vec{1}}{\sqrt{n}}, G\frac{\vec{1}}{\sqrt{n}}\right\rangle^k\right] = \frac{1}{n^k} \cdot \mathbb{E}\left[\langle \vec{1}, G\vec{1}\rangle^k\right] \geq \frac{1}{n^k}\left(\mathbb{E}\left[\langle \vec{1}, G\vec{1}\rangle\right]\right)^k = \frac{1}{n^k}\langle \vec{1}, \mathbb{E}[G]\vec{1}\rangle^k,
$$

where the second inequality is by Jensen's inequality on the convex function $f(x) = x^k$ for integer $k \geq 1$. Note that

$$
\langle \vec{1}, \mathbb{E}[G]\vec{1}\rangle = \sum_{1 \leq i,j \leq n} \mathbb{E}\langle v_i, v_j\rangle^2 = \sum_{1 \leq i \leq n} \mathbb{E}\langle v_i, v_i\rangle^2 + \sum_{1 \leq i \neq j \leq n} \mathbb{E}\langle v_i, v_j\rangle^2 = n + n(n-1)\frac{1}{d},
$$

where the last equality follows from the independence of v_i and v_j for $i \neq j$ so that

$$\mathbb{E}\langle v_i, v_j \rangle^2 = \mathbb{E}\langle v_i v_i^*, v_j v_j^* \rangle = \left\langle \frac{1}{d}I_d, \frac{1}{d}I_d \right\rangle = \frac{1}{d}.$$

Putting the value of $\langle \vec{1}, \mathbb{E}[G]\vec{1} \rangle$ gives $\mathbb{E}[\lambda_1(G)^k] \geq (1 + (n-1)/d)^k$, and thus

$$\mathbb{P}\left[\lambda_2(G) \geq (1-\lambda)^2 \cdot \frac{n}{d} \right] \leq \left(\mathbb{E}[\mathrm{tr}(G^k)] - \left(1 + \frac{n-1}{d}\right)^k \right) \Big/ \left((1-\lambda)^{2k} \left(\frac{n}{d}\right)^k \right). \qquad \square$$

5.2.2. Expanding the trace. To use the bound in Lemma 5.6, we need to compute $\mathbb{E}(\mathrm{tr}(G^k))$. We expand the trace of G^k as

$$(5.1) \qquad \mathrm{tr}(G^k) = \sum_{1 \leq i_1,\ldots,i_k \leq n} \prod_{s=1}^{k} G_{i_s, i_{s+1}} = \sum_{1 \leq i_1,\ldots,i_k \leq n} \prod_{s=1}^{k} \langle v_{i_s}, v_{i_{s+1}} \rangle^2,$$

where the sum runs over all possible length k words with letters in $\{1,\ldots,n\}$ with $i_{k+1} := i_1$. We interpret each term in the summation as a length k closed walk in the complete graph of n vertices, where $(i_1,\ldots,i_k,i_1)$ are the vertices in the closed walk.

Let $\{e_1,\ldots,e_d\}$ be an arbitrary orthonormal basis of $\mathbb{R}^d$. To analyze the trace, we write $v_{i_s} = \sum_{a=1}^{d} \langle v_{i_s}, e_a \rangle e_a$ as a linear combination of the basis vectors, and

$$\langle v_{i_s}, v_{i_{s+1}} \rangle^2 = \left(\sum_{a=1}^{d} \langle v_{i_s}, e_a \rangle \langle v_{i_{s+1}}, e_a \rangle \right)^2 = \sum_{a=1}^{d} \sum_{b=1}^{d} \langle v_{i_s}, e_a \rangle \langle v_{i_{s+1}}, e_a \rangle \langle v_{i_s}, e_b \rangle \langle v_{i_{s+1}}, e_b \rangle.$$

Expanding each term in the product $\prod_{s=1}^{k} \langle v_{i_s}, v_{i_{s+1}} \rangle^2$ this way and further expand the product, we can write

$$(5.2)$$

$$\mathrm{tr}(G^k) = \sum_{1 \leq i_1,\ldots,i_k \leq n} \prod_{s=1}^{k} \left(\sum_{a=1}^{d} \sum_{b=1}^{d} \langle v_{i_s}, e_a \rangle \langle v_{i_{s+1}}, e_a \rangle \langle v_{i_s}, e_b \rangle \langle v_{i_{s+1}}, e_b \rangle \right)$$

$$= \sum_{1 \leq i_1,\ldots,i_k \leq n} \sum_{1 \leq a_1,\ldots,a_k \leq d} \sum_{1 \leq b_1,\ldots,b_k \leq d} \prod_{s=1}^{k} \langle v_{i_s}, e_{a_s} \rangle \langle v_{i_{s+1}}, e_{a_s} \rangle \langle v_{i_s}, e_{b_s} \rangle \langle v_{i_{s+1}}, e_{b_s} \rangle.$$

We interpret each a_s and b_s as a color on the edge (i_s, i_{s+1}) for $1 \leq s \leq k$. So, in this interpretation, the trace is summing over all possible closed k walks on the complete graph of n vertices, and all pairs of edge d-coloring $a, b : [k] \to [d]$ on the edges $(i_1, i_2),\ldots,(i_{k-1}, i_k),(i_k, i_1)$ in the closed k walk.

To calculate the expected value of the product terms in (5.2), we group the terms based on the vertices involved and use the following basic building block. The proof of the following lemma uses the normalization technique in the proof that $\mathrm{vol}(S^{d-1}) = 2\pi^{d/2}/\Gamma(d/2)$ in Ball's survey [4], where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$.

LEMMA 5.7. *Let* $\vec{q} = (q_1,\ldots,q_d) \in \mathbb{Z}_{\geq 0}^d$ *with* $q := \sum_{i=1}^{d} q_i$. *Then*

$$\xi(\vec{q}) := \mathbb{E}_{u \in S^{d-1}} \prod_{i=1}^{d} \langle u, e_i \rangle^{2q_i} = \frac{\prod_{i=1}^{d} (2q_i - 1)!!}{\prod_{j=0}^{q-1} (d + 2j)},$$

where $\ell!! = \ell(\ell-2)\cdots(3)(1)$ *for an odd number* ℓ.

Proof. Let $g \in \mathbb{R}^d$ be a random Gaussian vector where each coordinate is an independent Gaussian variable $g_i \sim N(0, 1)$. We will compute $\mathbb{E}_g \prod_{i=1}^d \langle g, e_i \rangle^{2q_i}$ in two ways to prove the lemma. On one hand,

$$\mathbb{E}_g \prod_{i=1}^d \langle g, e_i \rangle^{2q_i} = \prod_{i=1}^d \mathbb{E}_g g_i^{2q_i} = \prod_{i=1}^d (2q_i - 1)!!,$$

where the second equality follows from the formula for the even moments of a standard Gaussian variable (e.g., from Wikipedia). On the other hand, we can compute the same quantity by a change of variables to the polar coordinates. Using that the density function of g is $(2\pi)^{-d/2} \exp(-\|g\|_2^2 / 2)$,

$$
\begin{aligned}
\mathbb{E}_g \prod_{i=1}^d \langle g, e_i \rangle^{2q_i} &= (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \prod_{i=1}^d \langle g, e_i \rangle^{2q_i} \cdot \exp\left(-\frac{1}{2} \|g\|_2^2\right) dg \\
&= (2\pi)^{-\frac{d}{2}} \int_{r=0}^\infty \int_{v \in S^{d-1}} \prod_{i=1}^d \langle rv, e_i \rangle^{2q_i} \; e^{-\frac{1}{2}r^2} r^{d-1} dv dr \\
&= (2\pi)^{-\frac{d}{2}} \left(\int_0^\infty r^{2q+d-1} e^{-\frac{1}{2}r^2} dr\right) \left(\int_{v \in S^{d-1}} \prod_{i=1}^d \langle v, e_i \rangle^{2q_i} dv\right) \\
&= (2\pi)^{-\frac{d}{2}} \left(2^{q+\frac{d}{2}-1} \cdot \Gamma\left(\frac{d}{2} + q\right)\right) \left(\mathrm{vol}(S^{d-1}) \cdot \mathbb{E}_{v \in S^{d-1}} \prod_{i=1}^d \langle v, e_i \rangle^{2q_i}\right),
\end{aligned}
$$

where the factor r^{d-1} appears in the second equality because the sphere of radius r has area r^{d-1} times that of S^{d-1}, and the last equality follows by a change of variable $u = \frac{1}{2}r^2$ and $du = rdr$ so that

$$\int_0^\infty r^{2q+d-1} e^{-\frac{1}{2}r^2} dr = \int_0^\infty (2u)^{\frac{2q+d-2}{2}} e^{-u} du = 2^{q+\frac{d}{2}-1} \cdot \Gamma\left(\frac{d}{2} + q\right),$$

where the last equality follows from the definition of the Gamma function that $\Gamma(l) := \int_0^\infty u^{l-1} e^{-u} du$. By combining the two equalities for $\mathbb{E}_g \prod_{i=1}^d \langle g, e_i \rangle^{2q_i}$ and using the fact that $\mathrm{vol}(S^{d-1}) = 2\pi^{d/2}/\Gamma(d/2)$ (e.g., from Wikipedia), we have

$$
\begin{aligned}
\prod_{i=1}^d (2q_i - 1)!! &= (2\pi)^{-\frac{d}{2}} \left(2^{q+\frac{d}{2}-1} \cdot \Gamma\left(\frac{d}{2} + q\right)\right) \left(\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \cdot \mathbb{E}_{v \in S^{d-1}} \prod_{i=1}^d \langle v, e_i \rangle^{2q_i}\right) \\
&\implies \xi(\vec{q}) := \mathbb{E}_{v \in S^{d-1}} \prod_{i=1}^d \langle v, e_i \rangle^{2q_i} = \frac{1}{2^q} \cdot \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d}{2} + q\right)} \cdot \prod_{i=1}^d (2q_i - 1)!!.
\end{aligned}
$$

Using the fact that $\Gamma(l) = l \cdot \Gamma(l-1)$ and thus $\Gamma(\frac{d}{2} + q) = \Gamma(\frac{d}{2}) \cdot (\frac{d}{2} + q - 1) \cdot (\frac{d}{2} + q - 2) \cdots (\frac{d}{2})$, it implies that

$$\frac{2^q \cdot \Gamma(\frac{d}{2} + q)}{\Gamma(\frac{d}{2})} = (d + 2(q-1)) \cdot (d + 2(q-2)) \cdots (d) = \prod_{j=0}^{q-1} (d + 2j),$$

and the lemma follows. $\qquad\square$

By taking the expectation of (5.1),

$$(5.3) \qquad \mathbb{E}[\mathrm{tr}(G^k)] = \sum_{1 \le i_1, \dots, i_k \le n} \mathbb{E}_{v_{i_1}, \dots, v_{i_d} \in S^{d-1}} \prod_{s=1}^{k} \langle v_{i_s}, v_{i_{s+1}} \rangle^2.$$

For each closed k-walk $i_1, \dots, i_k, i_1$, we need to compute the expectation of the product term. For some specific closed k walks, it is easier to compute the expectation of the product term. In the next two subsubsections, we show how to compute the product terms when the closed k-walk forms a tree or a cycle.

5.2.3. Tree walk. The first simplification is that if there is any self-loop (i.e., $i_s = i_{s+1}$), then we can just remove the term $\langle v_{i_s}, v_{i_{s+1}} \rangle^2$ from the product as $\|v_{i_s}\|_2 = 1$ by our construction.

The next simplification is that if the closed k-walk looks like a tree, i.e., the edges $(i_1, i_2), \dots, (i_{k-1}, i_k), (i_k, i_1)$ formed a tree when self-loops are removed and parallel edges are identified to a single edge, then the terms corresponding to each edge in the tree can be computed independently using Lemma 5.7. This is because all nonneighbors in the tree are conditionally independent, and so we can iteratively fix all nonleaf vertices and compute the leaves independently.

LEMMA 5.8. *Let $H = (V, E)$ be the graph formed by the edges $(i_1, i_2), \dots, (i_{k-1}, i_k)$, (i_k, i_1) in a closed k-walk. Suppose H is a tree $T = (V, F)$ when self-loops are removed and parallel edges are identified to a single edge. For each edge $f = (i, j) \in F$, let q_f be the number of parallel edges of f in H. Then,*

$$\mathbb{E}_V \prod_{ij=f \in E} \langle v_i, v_j \rangle^2 = \mathbb{E}_V \prod_{ij=f \in F} \langle v_i, v_j \rangle^{2q_f} = \prod_{f \in F} \xi(q_f \chi_1),$$

where $\xi(q_f \chi_1)$ denotes $\xi((q_f, 0, \dots, 0))$ in Lemma 5.7.

Proof. We prove this by induction on $|V|$. When $|V| = 2$, the statement follows from the rotational invariance of the distribution, so that $\mathbb{E}_u \mathbb{E}_v \langle u, v \rangle^q = \mathbb{E}_u \langle u, e_1 \rangle^q = \xi(q)$ where e_1 is the first vector in the orthonormal basis $(e_1, \dots, e_d)$.

For the inductive step, let L be the set of the leaves of the tree T and $\delta(L)$ be the set of leaf edges in T. By conditional expectation and independence of v_i,

$$\mathbb{E}_V \prod_{ij=f \in E} \langle v_i, v_j \rangle^2 = \mathbb{E}_L \mathbb{E}_{V \setminus L} \prod_{ij=f \in E} \langle v_i, v_j \rangle^2$$

$$= \mathbb{E}_L \left[\prod_{ij=f \in \delta(L)} \langle v_i, v_j \rangle^2 \,\middle|\, \{v_i\}_{i \notin L} \right] \cdot \mathbb{E}_{V \setminus L} \left(\prod_{ij=f \notin \delta(L)} \langle v_i, v_j \rangle^2 \right).$$

Since $|V \setminus L| < |V|$, we can apply the induction hypothesis to obtain that the second term is equal to $\prod_{f \notin \delta(L)} \xi(2q_f)$. For the first term, note that each nonleaf vertex is fixed in the conditional expectation, and so by rotational invariance of the distribution and independence of v_i,

$$\mathbb{E}_L \left[\prod_{ij=f \in \delta(L)} \langle v_i, v_j \rangle^2 \,\middle|\, \{v_i\}_{i \notin L} \right] = \prod_{ij=f \in \delta(L)} \mathbb{E}_L \left[\langle v_i, v_j \rangle^2 \,\middle|\, \{v_i\}_{i \notin L} \right]$$

$$= \prod_{ij=f \in \delta(L)} \mathbb{E}_L \left[\langle e_1, v_j \rangle^2 \,\middle|\, \{v_i\}_{i \notin L} \right] = \prod_{f \in \delta(L)} \xi(q_f \chi_1),$$

where $\chi_1 \in \mathbb{R}^d$ is the vector with the first entry one and other entries zero. The lemma follows by combining the two terms. □

5.2.4. Cycle walk. We can also compute the expectation of a product term in (5.3) when the closed k-walk is a simple cycle, i.e., the edges $(i_1, i_2), \ldots, (i_{k-1}, i_k)$, (i_k, i_i) form a cycle and the vertices $i_1, \ldots, i_k$ are distinct.

LEMMA 5.9. *Suppose the edges* $(i_1, i_2), \ldots, (i_{k-1}, i_k), (i_k, i_i)$ *form a simple cycle. Then*

$$\mathbb{E}_{v_{i_1}, \ldots, v_{i_d} \in S^{d-1}} \prod_{s=1}^{k} \langle v_{i_s}, v_{i_{s+1}} \rangle^2 = \frac{1}{d^k} + \frac{d^2-1}{2} \left(\frac{2}{d(d+2)} \right)^k.$$

Proof. We use the expansion in (5.2) that

$$\mathbb{E}\left[\prod_{s=1}^{k} \langle v_{i_s}, v_{i_{s+1}} \rangle^2 \right]$$

$$= \sum_{1 \le a_1, \ldots, a_k \le d} \sum_{1 \le b_1, \ldots, b_k \le d} \mathbb{E}\left[\prod_{s=1}^{k} \langle v_{i_s}, e_{a_s} \rangle \langle v_{i_{s+1}}, e_{a_s} \rangle \langle v_{i_s}, e_{b_s} \rangle \langle v_{i_{s+1}}, e_{b_s} \rangle \right],$$

where $(e_1, \ldots, e_d)$ is an orthonormal basis in $\mathbb{R}^d$.

Since S^{d-1} is symmetric across any half-space, if any term $\langle v, e_j \rangle$ appears in a product term on the right-hand side with odd degree, then that product term is equal to zero. So, we only focus on those product terms where each $\langle v, e_j \rangle$ has even degree. Since the edges $(i_1, i_2), \ldots, (i_{k-1}, i_k), (i_k, i_i)$ form a simple cycle, each vertex i_s is involved in exactly four terms $\langle v_{i_s}, e_{a_s} \rangle, \langle v_{i_s}, e_{b_s} \rangle, \langle v_{i_s}, e_{a_{s-1}} \rangle, \langle v_{i_s}, e_{b_{s-1}} \rangle$. We consider two cases of the d-edge-colorings $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$.

The first case is when $a_1 \neq b_1$. Then, for $\langle v_{i_2}, e_{a_1} \rangle$ and $\langle v_{i_2}, e_{b_1} \rangle$ to have even degree, we must have $\{a_2, b_2\} = \{a_1, b_1\}$. The same argument applies to every vertex, and thus we must have $\{a_i, b_i\} = \{a_j, b_j\}$ for $i \neq j$, i.e., the same two colors appear in every edge in the simple cycle. There are two possibilities for each edge, either $a_i = a_j, b_i = b_j$ or $a_i = b_j, a_j = b_i$. So, for each two colors, there are exactly 2^k such product terms. For each such product term, there are two colors that appear twice on each vertex, and so each such product term is exactly $\xi(\chi_{1,2})^k$, where $\chi_{1,2} \in \mathbb{R}^d$ is the vector with the first two entries one and other entries zero. Therefore, the total contribution of these product terms is

$$\binom{d}{2} 2^k \xi(\chi_{1,2})^k = \binom{d}{2} 2^k \left(\frac{1}{d(d+2)} \right)^k.$$

The second case is when $a_1 = b_1$. Then, for the terms in i_2 to have even degree, we must have $a_2 = b_2$, which could be the same color as $a_1 = b_1$ or a different color. The same argument applies to every vertex, and thus we must have $a_i = b_i$ for $1 \le i \le k$, and so we can think of every edge in the cycle receiving one color from d colors. For each coloring, let l be the number of vertices with two different colors of degree two (and so $k - l$ is the number of vertices with one color of degree four); then its contribution to the sum is

$$\left(\xi(\chi_{1,2}) \right)^l \left(\xi(2\chi_1) \right)^{k-l} = \left(\frac{1}{d(d+2)} \right)^l \left(\frac{3}{d(d+2)} \right)^{k-l} = \frac{3^{k-l}}{d^k(d+2)^k}.$$

To count the number of such colorings, we use the following fact.

FACT 5.10. *The number of proper d-colorings of an l-cycle is $(d-1)^l+(-1)^l(d-1)$, where adjacent vertices receive different colors in a proper coloring. Since the line graph of an l-cycle is also an l-cycle, the number of proper d-edge-colorings of an l-cycle is also $(d-1)^l + (-1)^l(d-1)$.*

We would like to count the d-edge-colorings of a k-cycle with l vertices with different colors on its two edges and $k-l$ vertices with the same color on its two edges. Notice that once we fix the location of the l vertices with different colors, then the edges between any two such vertices must have the same color, and so we can think of the k-cycle as an l-cycle and each such coloring corresponds to a proper d-edge-coloring of an l-cycle. By enumerating the location of the l vertices and using Fact 5.10, the number of such d-edge-colorings is $\binom{k}{l} \cdot \left((d-1)^l + (-1)^l(d-1)\right)$. Therefore, the total contribution of the second case is equal to

$$\sum_{l=0}^{k} \binom{k}{l} \cdot \left((d-1)^l + (-1)^l(d-1)\right) \cdot \frac{3^{k-l}}{d^k(d+2)^k}$$

$$= \frac{1}{d^k(d+2)^k} \left(\sum_{l=0}^{k} \binom{k}{l}(d-1)^l 3^{k-l} + (d-1)\sum_{l=0}^{k} \binom{k}{l}(-1)^l 3^{k-l}\right)$$

$$= \frac{1}{d^k(d+2)^k} \left((d+2)^k + (d-1)2^k\right),$$

where the last equality is by the binomial theorem. Combining the two cases,

$$\mathbb{E}\left[\prod_{s=1}^{k}\langle v_{i_s}, v_{i_{s+1}}\rangle^2\right] = \frac{\binom{d}{2}2^k + (d+2)^k + (d-1)2^k}{d^k(d+2)^k} = \frac{1}{d^k} + \frac{d^2-1}{2}\left(\frac{2}{d(d+2)}\right)^k.$$

$\square$

5.2.5. Fourth moment analysis. We can use Lemmas 5.8 and 5.9 to compute $\mathrm{tr}(G^4)$.

LEMMA 5.11.

$$\mathbb{E}\,\mathrm{tr}(G^4) \le \frac{n^4}{d^4}\left(1 + \frac{d^4}{n^3} + \frac{18d^2}{n^2} + \frac{105}{n^2} + \frac{4d}{n} + \frac{34}{n} + \frac{8}{d^2}\right).$$

Proof. To compute $\mathbb{E}\,\mathrm{tr}(G^4)$, we only need to consider closed 4-walks $(i_1, i_2, i_3, i_4, i_1)$. We do a case analysis on the possible configurations of closed 4-walks.

1. There are four self-loops, i.e., $i_1 = i_2 = i_3 = i_4$, in which case the contribution is simply one as the vectors are of length one by construction. There are total n possibilities for the location of the self-loops, and so the total contribution in this case is $(L_4) := n$.

2. There are two self-loops and a single edge traversed two times. By Lemma 5.8, this graph contributes $\xi(2\chi_1) = 3/d(d+2)$. There are $\binom{4}{2}$ places to add two self-loops to a single edge and $n(n-1)$ possibilities for the two vertices of the edge, so the total contribution in this case is

$$(L_2E) := \frac{18n(n-1)}{d(d+2)} \le \frac{18n^2}{d^2}.$$

3. The only other case with two distinct vertices is that an edge is traversed four times, and its contribution is $\xi(4\chi_1)$ by Lemma 5.8. There are $n(n-1)$ for the location of the two vertices, and the total contribution in this case is

$$(E_2) := n(n-1) \cdot \xi(4\chi_1) = \frac{3 \cdot 5 \cdot 7 \cdot n(n-1)}{d(d+2)(d+4)(d+6)} \leq \frac{105n^2}{d^4}.$$

4. There is one self-loop and a 3-cycle. This graph contributes the same as a 3-cycle which is given by Lemma 5.9. There are four places to add the self-loop and $n(n-1)(n-2)$ possibilities for the three vertices of the triangle, so the total contribution in this case is

$$(LC_3) := 4n(n-1)(n-2) \left(\frac{1}{d^3} + \frac{d^2-1}{2} \left(\frac{2}{d(d+2)} \right)^3 \right) \leq \frac{4n^3}{d^3} + \frac{16n^3}{d^4}.$$

5. The only other case with three distinct vertices is two different edges sharing a single common vertex. By Lemma 5.8, this graph contributes $(\xi(2\chi_1))^2$. Note that there are two ways to combine, as the two edges could share the starting vertex or the middle vertex. There are $n(n-1)(n-2)$ for the locations of the three vertices, and so the total contribution is

$$(P_2) := 2n(n-1)(n-2)(\xi(2\chi_1))^2 = 2n(n-1)(n-2) \left(\frac{3}{d(d+2)} \right)^2 \leq \frac{18n^3}{d^4}.$$

6. Finally, the only case with four distinct vertices is a 4-cycle. There are $n(n-1)(n-2)(n-3)$ possibilities for the locations of the four vertices, and by Lemma 5.9 the total contribution is

$$(C_4) := n(n-1)(n-2)(n-3) \left(\frac{1}{d^4} + \frac{2^3(d^2-1)}{d^4(d+2)^4} \right) \leq \frac{n^4}{d^4} + \frac{8n^4}{d^6}.$$

Combining all the cases,

$$\text{tr}(G^4) = (L_4) + (L_2E) + (E_2) + (LC_3) + (P_2) + (C_4)$$
$$\leq n + \frac{18n^2}{d^2} + \frac{105n^2}{d^4} + \frac{4n^3}{d^3} + \frac{16n^3}{d^4} + \frac{18n^3}{d^4} + \frac{n^4}{d^4} + \frac{8n^4}{d^6}.$$

Taking the factor n^4/d^4 out proves the lemma. □

5.2.6. Proof of Theorem 5.1. We wrap up the fourth moment analysis to prove Theorem 5.1. Using Lemma 5.11 in Lemma 5.6, we have

$$\mathbb{P} \left[\lambda_2(G) \geq (1-\lambda)^2 \cdot \frac{n}{d} \right]$$
$$\leq \left(\mathbb{E}[\text{tr}(G^4)] - \left(\frac{n}{d} \right)^4 \left(1 + \frac{d-1}{n} \right)^4 \right) \Big/ \left((1-\lambda)^8 \left(\frac{n}{d} \right)^4 \right).$$
$$\leq \frac{1}{(1-\lambda)^8} \left(1 + \frac{d^4}{n^3} + \frac{18d^2}{n^2} + \frac{105}{n^2} + \frac{4d}{n} + \frac{34}{n} + \frac{8}{d^2} - \left(1 + \frac{d-1}{n} \right)^4 \right)$$
$$\leq \frac{1}{(1-\lambda)^8} \left(\frac{d^4}{n^3} + \frac{18d^2}{n^2} + \frac{105}{n^2} + \frac{4d}{n} + \frac{34}{n} + \frac{8}{d^2} - \frac{d-1}{n} \right),$$

where we used $(1 + (d-1)/n)^4 \geq 1 + (d-1)/n$.

For any constant $\lambda < 1$, by generating $n \gg d^{4/3}$ random unit vectors, the probability that $\lambda_2(G) > (1-\lambda)^2 n/d$ is at most $1/1000$ where the dominating term is d^4/n^3.

Also, by Corollary 5.5, by generating $n = d \log^3 d$ random unit vectors, the resulting frame is ϵ-nearly doubly stochastic with failure probability at most inverse polynomial in d.

Therefore, by generating $n \gg d^{4/3}$ random unit vectors, with probability at least 0.99, the resulting frame is ϵ-nearly doubly stochastic for $\epsilon \ll 1/\log d$ and $\lambda_2(G) \leq (1 - \lambda)^2 \cdot n/d$ for any constant $0 \leq \lambda < 1$. This proves Theorem 5.1.

Remark 5.12. We believe that the trace method can be improved to prove the same conclusion with only $O(d \operatorname{polylog} d)$ random unit vectors. Subsequent to our work, Franks and Moitra [20] show that $n = \Omega(d \operatorname{polylog} d)$ random unit vectors suffices to satisfy the conclusions of Theorem 1.11. A more detailed discussion of their results and techniques is given in Remark 5.2.

Appendix A. Operator scaling. The following is a proof that the continuous operator scaling algorithm is equivalent to the gradient flow that always moves in the direction of minimizing Δ.

LEMMA A.1. *Given an operator $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$ for $1 \leq i \leq k$, the direction defined by*

$$\frac{d}{dt} A_i := \left(s(\mathcal{A}) \cdot I_m - m \sum_{j=1}^{k} A_j A_j^* \right) A_i + A_i \left(s(\mathcal{A}) \cdot I_n - n \sum_{j=1}^{k} A_j^* A_j \right) \quad \text{for } 1 \leq i \leq k$$

minimizes the function

$$\Delta(\mathcal{A}) = \frac{1}{m} \left\| s(\mathcal{A}) \cdot I_m - m \sum_{i=1}^{k} A_i A_i^* \right\|_F^2 + \frac{1}{n} \left\| s(\mathcal{A}) \cdot I_n - n \sum_{i=1}^{k} A_i^* A_i \right\|_F^2.$$

Proof. As in Definition 2.14, we write

$$E(\mathcal{A}) = s(\mathcal{A}) \cdot I_m - m \sum_{i=1}^{k} A_i A_i^* \quad \text{and} \quad F(\mathcal{A}) = s(\mathcal{A}) \cdot I_n - n \sum_{i=1}^{k} A_i^* A_i.$$

Then

$$\Delta(\mathcal{A}) = \frac{1}{m} \operatorname{tr}(E(\mathcal{A})^2) + \frac{1}{n} \operatorname{tr}(F(\mathcal{A})^2) \quad \text{and} \quad \frac{d}{dt} A_i = E(\mathcal{A}) \cdot A_i + A_i \cdot F(\mathcal{A}).$$

Consider the directional derivative of $\Delta(\mathcal{A})$ at the direction of $\mathcal{H} = (H_1, \ldots, H_k)$ where each $H_i \in \mathbb{R}^{m \times n}$. For ease of notation, we write $E = E(\mathcal{A})$, $F = F(\mathcal{A})$ and $s = s(\mathcal{A})$ in the following, with the understanding that these are dependent on $\mathcal{A}$ and we are moving $\mathcal{A}$ in the direction $\mathcal{H}$.

$$\nabla_{\mathcal{H}} \Delta(\mathcal{A}) = \frac{1}{m} \operatorname{tr} \left(2E \cdot \nabla_{\mathcal{H}} E \right) + \frac{1}{n} \operatorname{tr} \left(2F \cdot \nabla_{\mathcal{H}} F \right)$$

$$= \frac{2}{m} \operatorname{tr} \left(E \cdot \nabla_{\mathcal{H}} s \cdot I_m - m \sum_{i=1}^{k} 2E \cdot \nabla_{\mathcal{H}} A_i \cdot A_i^* \right)$$

$$+ \frac{2}{n} \operatorname{tr} \left(F \cdot \nabla_{\mathcal{H}} s \cdot I_n - n \sum_{i=1}^{k} 2F A_i^* \cdot \nabla_{\mathcal{H}} A_i \right)$$

$$= \frac{2}{m} \operatorname{tr} \left(-m \sum_{i=1}^{k} 2EH_i A_i^* \right) + \frac{2}{n} \operatorname{tr} \left(-n \sum_{i=1}^{k} 2FA_i^* H_i \right)$$

$$= -4 \sum_{i=1}^{k} \operatorname{tr} \left(EH_i A_i^* + FA_i^* H_i \right)$$

$$= -4 \sum_{i=1}^{k} \langle EA_i + A_i F, H_i \rangle ,$$

where the third equality uses the fact that $\operatorname{tr}(E) = 0$ and $\operatorname{tr}(F) = 0$ as stated in Definition 2.14. It follows that the direction $H_i := EA_i + A_i F$ minimizes $\Delta(\mathcal{A})$. $\qquad \square$

The following is an alternative proof of Lemma 3.6 provided by John Watrous.

LEMMA A.2 (Watrous [64]). *If $\mathcal{A}$ is an ϵ-nearly doubly balanced operator, then the largest singular value of its matrix representation $M_\mathcal{A}$ in Definition* 1.4 *is*

$$\sigma_1(M_\mathcal{A}) \leq (1 + \epsilon) \frac{s(\mathcal{A})}{\sqrt{mn}}.$$

Proof. The proof is a generalization of the proof of Theorem 4.27 in [63]. As stated in Definition 2.7,

$$\sigma_1(M_\mathcal{A}) = \max_{Y \in \mathbb{R}^{n \times n}} \frac{\|\Phi(Y)\|_F}{\|Y\|_F} \leq \max_{Y \in \mathbb{C}^{n \times n}} \frac{\|\Phi(Y)\|_F}{\|Y\|_F},$$

where $\Phi(Y)$ is as defined in (2.1).

First, we bound the maximum for Hermitian matrix Y. Let $Y = \sum_{i=k}^{n} \lambda_k y_k y_k^*$ be an eigenvalue decomposition of H. Let

$$\rho_k := \Phi(y_k y_k^*) \quad \text{so that} \quad \Phi(Y) = \sum_{k=1}^{n} \Phi(\lambda_k y_k y_k^*) = \sum_{k=1}^{n} \lambda_k \rho_k \quad \text{and}$$

$$\Phi(I_n) = \sum_{k=1}^{n} \Phi(y_k y_k^*) = \sum_{k=1}^{n} \rho_k.$$

Then, by the Cauchy–Schwarz inequality and Hölder's inequality for Schatten norms for matrices,

$$\|\Phi(Y)\|_F^2 = \left\| \sum_{k=1}^{n} \lambda_k \rho_k \right\|_F^2 = \sum_{k=1}^{n} \sum_{j=1}^{n} \lambda_k \lambda_j \langle \rho_k, \rho_j \rangle$$

$$\leq \sqrt{\sum_{k=1}^{n} \sum_{j=1}^{n} \lambda_k^2 \langle \rho_k, \rho_j \rangle} \sqrt{\sum_{k=1}^{n} \sum_{j=1}^{n} \lambda_j^2 \langle \rho_k, \rho_j \rangle}$$

$$= \sum_{k=1}^{n} \sum_{j=1}^{n} \lambda_k^2 \langle \rho_k, \rho_j \rangle = \sum_{k=1}^{n} \lambda_k^2 \langle \rho_k, \Phi(I_n) \rangle \leq \sum_{k=1}^{n} \lambda_k^2 \|\rho_k\|_1 \|\Phi(I_n)\|_{\mathrm{op}} .$$

Since Φ is a positive map, $\rho_k = \Phi(y_k y_k^*) \succeq 0$ by Fact 2.10.2. It follows that the trace norm of ρ_k is simply the trace of ρ_k, and so

$$\|\rho_k\|_1 = \langle I_m, \rho_k \rangle = \langle I_m, \Phi(y_k y_k^*) \rangle = \langle \Phi^*(I_m), y_k y_k^* \rangle \leq \|\Phi^*(I_m)\|_{\mathrm{op}} \leq (1 + \epsilon) \frac{s}{n},$$

where the third equality is by Fact 2.10.3 and the last inequality follows from the assumption that $\mathcal{A}$ is ϵ-nearly doubly balanced. Therefore,

$$\|\Phi(Y)\|_F^2 \le \sum_{k=1}^n \lambda_i^2 \|\rho_k\|_1 \|\Phi(I_n)\|_{\mathrm{op}} \le (1+\epsilon)\frac{s}{n}\cdot(1+\epsilon)\frac{s}{m}\cdot\sum_{k=1}^n \lambda_k^2 = \frac{(1+\epsilon)^2 s^2}{mn}\|Y\|_F^2,$$

where the second inequality is from the assumption that $\mathcal{A}$ is ϵ-nearly doubly balanced.

For the non-Hermitian case, we use a standard reduction and write $Y = H + iK$ where $H = (Y + Y^*)/2$ and $K = (Y - Y^*)/2i$ are Hermitian matrices. Note that $\|Y\|_F^2 = \|H\|_F^2 + \|K\|_F^2$. As Φ is neccessarily Hermitian perserving, we also have $\|\Phi(Y)\|_F^2 = \|\Phi(H) + i\Phi(K)\|_F^2 = \|\Phi(H)\|_F^2 + \|\Phi(K)\|_F^2$. Therefore, as H and K are Hermitian,

$$\|\Phi(Y)\|_F^2 = \|\Phi(H)\|_F^2 + \|\Phi(K)\|_F^2 \le \frac{(1+\epsilon)^2 s^2}{mn}(\|H\|_F^2 + \|K\|_F^2) = \frac{(1+\epsilon)^2 s^2}{mn}\|Y\|_F^2.$$
$\square$

Appendix B. Matrix scaling. The aim of this section is to provide a self-contained proof of the linear convergence result in the simpler setting of matrix scaling. It can be read as an exposition of the main ideas in section 3.

In the matrix scaling problem, we are given a nonnegative matrix $B \in \mathbb{R}^{m\times n}$, and the goal is to find a left diagonal scaling matrix $L \in \mathbb{R}^{m\times m}$ and a right diagonal scaling matrix $R \in \mathbb{R}^{n\times n}$ such that LBR is doubly balanced, or report that such scaling matrices do not exist.

B.1. Definitions. In the following, we state the important definitions for the matrix scaling problem. Given a matrix $B \in \mathbb{R}^{m\times n}$, we define

$$(\text{B.1}) \qquad s(B) := \sum_{i=1}^m \sum_{j=1}^n B_{ij} \quad \text{and} \quad r_i(B) := \sum_{j=1}^n B_{ij} \quad \text{and} \quad c_j(B) := \sum_{i=1}^m B_{ij}$$

as the size, the ith row sum, and the jth column sum of the matrix B.

A matrix B is ϵ-nearly doubly balanced if

$$(\text{B.2})$$
$$(1-\epsilon)\frac{s(B)}{m} \le r_i(B) \le (1+\epsilon)\frac{s(B)}{m} \quad \text{and} \quad (1-\epsilon)\frac{s(B)}{n} \le c_j(B) \le (1+\epsilon)\frac{s(B)}{n}$$

for $1 \le i \le m$ and $1 \le j \le n$, and B is doubly balanced when $\epsilon = 0$.

The ℓ_2-error of B is defined as

$$(\text{B.3}) \qquad \Delta(B) := \Delta_r(B) + \Delta_c(B) \quad \text{where} \quad \Delta_r(B) := \frac{1}{m}\sum_{i=1}^m (s - mr_i)^2 \quad \text{and}$$

$$\Delta_c(B) := \frac{1}{n}\sum_{j=1}^n (s - nc_j)^2.$$

The spectral condition is the same as defined in Lemma 4.3.

DEFINITION B.1 (spectral gap condition for matrix). *A matrix $B \in \mathbb{R}^{m\times n}$ satisfies the λ-spectral gap condition if*

$$\sigma_2(B) \le (1-\lambda)\frac{s(B)}{\sqrt{mn}}.$$

B.2. Continuous matrix scaling. The matrix scaling problem is a special case of the operator scaling problem. Following the reduction in section 4.1, given a non-negative matrix $B \in R^{m \times n}$, we consider the matrix $A \in R^{m \times n}$ where the (i, j)th entry of A is

(B.4)
$$a_{ij} := \sqrt{B_{ij}}.$$

The continuous matrix scaling algorithm works on A and is defined by the following differential equation:

(B.5)
$$\frac{d}{dt} a_{ij} = (s(B) - mr_i(B) + s(B) - nc_j(B)) \cdot a_{ij}.$$

Many quantities change over time in the dynamical system. We use the superscript $^{(t)}$ to denote the quantity of interest at time t. Given a nonnegative matrix $B \in \mathbb{R}^{m \times n}$ as the input of the matrix scaling problem, the matrix A in (B.4) is the input of the continuous operator scaling algorithm at time $t = 0$, i.e., $A^{(0)} := A$ and $B^{(0)} := B$. Then $A^{(t)}$ changes over time following (B.5) and $B^{(t)}$ is defined as the matrix with $B_{ij}^{(t)} = (a_{ij}^{(t)})^2$. The dynamical system stops when $B^{(t)}$ is doubly balanced. It is proved in [47] that $\Delta^{(\infty)} = 0$.

We state some known results about the continuous matrix scaling algorithm for the analysis. First, the matrix A at any time is a scaling of the original matrix in the following form.

LEMMA B.2 (Lemma 4.2.10 in [47]).   *At time $T \geq 0$, define $L^{(T)} \in \mathbb{R}^{m \times m}$ and $R^{(T)} \in \mathbb{R}^{n \times n}$ as*

$$L^{(T)} := \mathrm{diag}\left(\exp\left(\int_0^T \left(s^{(t)} - mr_i^{(t)} \right) dt \right) \right) \quad \text{and}$$

$$R^{(T)} := \mathrm{diag}\left(\exp\left(\int_0^T \left(s^{(t)} - nc_j^{(t)} \right) dt \right) \right).$$

Then $A^{(T)} = L^{(T)} A^{(0)} R^{(T)}$.

In particular, if $\Delta^{(t)} = 0$, then $(L^{(t)})^2 \cdot B \cdot (R^{(t)})^2$ is doubly balanced, and $(L^{(t)})^2$ and $(R^{(t)})^2$ is a solution to the matrix scaling problem. This is how the continuous operator scaling algorithm finds a scaling solution.

From now on, the matrix of interest is $B^{(t)}$ and it evolves over time as $A^{(t)}$ changes in the dynamical system. For ease of notation, we will omit the matrix $B^{(t)}$ and sometimes also the superscript $^{(t)}$ on other quantities when they are clear from the context.

LEMMA B.3 (Lemma 3.6.1 in [47]).   *For an ϵ-nearly doubly balanced matrix B,*

$$\Delta \leq 2\epsilon^2 s^2.$$

LEMMA B.4 (Lemma 4.2.8 in [47]).   *For any time $t \geq 0$,*

$$\frac{d}{dt} s = -2\Delta.$$

LEMMA B.5 (Lemma 4.2.9 in [47]).   *For any time $t \geq 0$,*

$$\frac{d}{dt} \Delta = -4 \sum_{i=1}^m \sum_{j=1}^n (2s - mr_i - nc_j)^2 \cdot a_{ij}^2.$$

LEMMA B.6 (Proposition 4.3.1 in [47]). *Suppose there exists $\mu > 0$ such that for all $0 \leq t \leq T$,*

$$-\frac{d}{dt}\Delta^{(t)} \geq \mu\Delta^{(t)}.$$

Then

$$\Delta^{(T)} \leq \Delta^{(0)}e^{-\mu T} \quad \text{and} \quad s^{(0)} - s^{(T)} \leq \frac{2\Delta^{(0)}}{\mu}.$$

B.3. Overview. The proof overview is stated in section 1.5.2 in the matrix scaling setting, so we won't repeat here. It is easy to see from Lemma B.5 that

(B.6)
$$-\frac{1}{4}\frac{d}{dt}\Delta = \sum_{i=1}^{m}(s - mr_i)^2 r_i + \sum_{j=1}^{n}(s - nc_j)^2 c_j + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(s - mr_i)(s - nc_j)a_{ij}^2.$$

The structure is the same as in section 3 for the general operator setting. Our goal is to prove the following theorem.

THEOREM B.7 (linear convergence). *Given a nonnegative matrix $B \in \mathbb{R}^{m \times n}$ with $m \leq n$, if B is ϵ-nearly doubly balanced and B satisfies the λ-spectral gap condition in Definition B.1 with $\lambda^2 \geq C\epsilon \ln m$ for a sufficiently large constant C, then in the gradient flow,*

$$\Delta^{(t)} \leq \Delta^{(0)}e^{-\lambda s^{(0)}t} \quad \text{for any } t \geq 0.$$

In particular, the gradient flow converges to an η-nearly doubly balanced scaling in time $t = O(\frac{1}{\lambda}\log(\frac{m}{\eta}))$, and such a scaling always exists under our assumptions.

B.4. Lower bounding the quadratic terms. First, we prove a structural result bounding the maximum error of the rows and columns, which will also be useful in bounding the condition number of the scaling solution later. Then, we will use this structural result to lower bound the quadratic terms of $-\Delta'$.

PROPOSITION B.8. *If $B^{(0)}$ is ϵ-nearly doubly balanced, then for any $t \geq 0$,*

$$\left|s^{(t)} - mr_i^{(t)}\right| \leq (1 + \epsilon)s^{(0)} - s^{(t)} \quad \text{and} \quad \left|s^{(t)} - nc_j^{(t)}\right| \leq (1 + \epsilon)s^{(0)} - s^{(t)}$$

for $1 \leq i \leq m$ and $1 \leq j \leq n$.

Proof. We present a slightly informal proof, which can be made formal by using the envelope theorem stated in Theorem 3.3 as done in Proposition 3.2.

Let

$$g(t) = \max\left\{\max_{1 \leq i \leq m}\left\{\left|s^{(t)} - mr_i^{(t)}\right|\right\}, \max_{1 \leq j \leq n}\left\{\left|s^{(t)} - nc_j^{(t)}\right|\right\}\right\}$$

be the maximum violation of a row and a column at time t. Note that $g(0) \leq \epsilon s^{(0)}$ as $B^{(0)}$ is ϵ-nearly doubly balanced. We would like to show that for almost every time $\tau \geq 0$,

$$\frac{d}{d\tau}g(\tau) \leq 2\Delta^{(\tau)}.$$

This would imply the proposition as

$$g(t) = g(0) + \int_0^t \frac{d}{d\tau}g(\tau)d\tau \leq \epsilon s^{(0)} + \int_0^t 2\Delta^{(\tau)}d\tau = \epsilon s^{(0)} - \int_0^t \frac{d}{d\tau}s^{(\tau)}d\tau = (1 + \epsilon)s^{(0)} - s^{(t)},$$

where the second to last equality is by Lemma B.4.

To bound $\frac{d}{dt}g(t)$, we consider different cases of how the maximum of $g(t)$ is achieved. Suppose the maximum of $g(t)$ is achieved by column j and $s^{(t)} - nc_j^{(t)}$ is negative such that $g(t) = -s^{(t)} + nc_j^{(t)}$. The change of the jth column sum is

$$\frac{d}{dt}c_j^{(t)} = \frac{d}{dt}\sum_{i=1}^{m}\left(a_{ij}^{(t)}\right)^2 = 2\sum_{i=1}^{m}a_{ij}^{(t)}\cdot\frac{d}{dt}a_{ij}^{(t)} = 2\sum_{i=1}^{m}\left(a_{ij}^{(t)}\right)^2\left(s^{(t)} - mr_i^{(t)} + s^{(t)} - nc_j^{(t)}\right) \le 0,$$

where the last equality is by the definition of the dynamical system in (B.5), and the inequality is by our assumption that the maximum of $g(t)$ is achieved by column j so that $s^{(t)} - nc_j^{(t)} = -g(t)$ and $s^{(t)} - mr_i^{(t)} \le g(t)$ for all $1 \le i \le m$. It follows that

$$\frac{d}{dt}\left(-s^{(t)} + nc_j^{(t)}\right) = 2\Delta^{(t)} + \frac{d}{dt}nc_j^{(t)} \le 2\Delta^{(t)},$$

where the first equality is by Lemma B.4.

Similarly, suppose the maximum of $g(t)$ is achieved by column j and $s^{(t)} - nc_j^{(t)}$ is positive; we can show that

$$\frac{d}{dt}\left(s^{(t)} - nc_j^{(t)}\right) = -2\Delta^{(t)} - \frac{d}{dt}nc_j^{(t)} \le -2\Delta^{(t)}.$$

By symmetry of rows and columns, we can prove the same bounds for the change of the violation of the ith row sum. Therefore, in all four cases, the change of the maximum violation is at most $2\Delta^{(t)}$. Note that g can be written as the maximum of $m + n$ functions, one for each row and one for each column. We can then use the envelope theorem in Theorem 3.3 as done in Proposition 3.2 to prove formally that $g(t) = g(0) + \int_0^t \frac{d}{d\tau}g(\tau)d\tau$ to complete the proof.

It is possible to prove the proposition for the matrix case without using the envelope theorem as g is only the maximum of a finite number of functions, but in the operator case $g(t)$ is the maximum quadratic form of infinitely many unit vectors and we don't know of a proof without using the envelope theorem. $\square$

We have the following corollary about the row sums and the column sums by rewriting the conclusions of Proposition B.8.

PROPOSITION B.9. *If $B^{(0)}$ is ϵ-nearly doubly balanced, then for any $t \ge 0$, for $1 \le i \le m$ and $1 \le j \le n$,*

$$\frac{2s^{(t)} - (1+\epsilon)s^{(0)}}{m} \le r_i^{(t)} \le \frac{(1+\epsilon)s^{(0)}}{m} \quad \text{and} \quad \frac{2s^{(t)} - (1+\epsilon)s^{(0)}}{n} \le c_j^{(t)} \le \frac{(1+\epsilon)s^{(0)}}{n}.$$

We can use Proposition B.9 to lower bound the quadratic terms in (B.6).

LEMMA B.10. *If $B^{(0)}$ is ϵ-nearly doubly balanced, then for any $t \ge 0$,*

$$\sum_{i=1}^{m}\left(s^{(t)} - mr_i^{(t)}\right)^2\cdot r_i^{(t)} + \sum_{j=1}^{n}\left(s^{(t)} - nc_j^{(t)}\right)^2\cdot c_j^{(t)} \ge \left(2s^{(t)} - (1+\epsilon)s^{(0)}\right)\Delta^{(t)}.$$

Proof. Using Proposition B.9, the first term in (B.6) is

$$\sum_{i=1}^{m}\left(s^{(t)} - mr_i^{(t)}\right)^2\cdot r_i^{(t)} \ge \frac{2s^{(t)} - (1+\epsilon)s^{(0)}}{m}\sum_{i=1}^{m}\left(s^{(t)} - mr_i^{(t)}\right)^2 = \left(2s^{(t)} - (1+\epsilon)s^{(0)}\right)\Delta_r^{(t)}.$$

Similarly, the second term in (B.6) is

$$\sum_{j=1}^{n}\left(s^{(t)} - nc_j^{(t)}\right)^2\cdot c_j^{(t)} \ge \frac{2s^{(t)} - (1+\epsilon)s^{(0)}}{n}\sum_{j=1}^{n}\left(s^{(t)} - nc_j^{(t)}\right)^2 = \left(2s^{(t)} - (1+\epsilon)s^{(0)}\right)\Delta_c^{(t)}.$$

The lemma follows from $\Delta_r + \Delta_c = \Delta$ in (B.3). $\square$

B.5. Upper bounding the cross term. We will first bound the largest singular value of the matrix B for any ϵ-nearly doubly balanced matrix B. Then, we will use a spectral argument to upper bound the absolute value of the cross term in (B.6).

LEMMA B.11. *If $B \in \mathbb{R}^{m \times n}$ is ϵ-nearly doubly balanced, then*

$$\sigma_1(B) \leq (1+\epsilon)\frac{s(B)}{\sqrt{mn}}.$$

Proof. We use the fact that the square of the largest singular value of a non-negative matrix is at most the maximum column sum times the maximum row sum (see, e.g., page 223 of [40]). So,

$$\sigma_1^2(B) \leq \max_{1 \leq i \leq m} r_i(B) \cdot \max_{1 \leq j \leq n} c_j(B) \leq \frac{(1+\epsilon)s(B)}{m} \cdot \frac{(1+\epsilon)s(B)}{n} = \frac{(1+\epsilon)^2 s(B)^2}{mn},$$

where the second inequality follows from the assumption that B is ϵ-nearly doubly balanced. □

Lemma B.11 implies that $\vec{1}_n$ is an "approximate" first singular vector of B. By the spectral gap condition in Definition B.1, it will follow that any vector perpendicular to $\vec{1}_n$ has a "small" quadratic form, and this can be used to bound the cross term in Lemma B.6. The following lemma summarizes the spectral argument, which is the same as Lemma 3.7. Since Lemma 3.7 has no operators involved, we refer to the proof in section 3.2 and just restate the statement here for ease of reference.

LEMMA B.12. *Let $M \in \mathbb{R}^{m \times n}$. Let $p \in \mathbb{R}^m$ and $q \in \mathbb{R}^n$ be unit vectors. Suppose the following assumptions hold:*

$$\sigma_1(M)^2 \leq 1 + \delta_1 \quad \text{and} \quad \sigma_2(M)^2 \leq 1 - \delta_2 \quad \text{and} \quad p^* M q = 1.$$

Then, for any unit vectors $x \perp p$ and $y \perp q$, it holds that $|x^* M y| \leq 1 + \frac{2\delta_1 - \delta_2}{2}$.

We can use Lemma B.12 to bound the cross term in Lemma B.6.

LEMMA B.13. *If B satisfies the spectral condition in Definition B.1 with the additional assumption that $\sigma_1(B) \leq (1+\delta)s/\sqrt{mn}$ for $\delta \leq 1$, then*

$$2\left|\sum_{i=1}^{m}\sum_{j=1}^{n}(s - mr_i)(s - nc_j)a_{ij}^2\right| \leq \left(1 + \frac{6\delta - \lambda}{2}\right)s\Delta.$$

Proof. We apply Lemma B.12 with $M \in \mathbb{R}^{m \times n}$, $p, x \in \mathbb{R}^m$, and $q, y \in \mathbb{R}^n$ where

$$M = \frac{\sqrt{mn}}{s} \cdot B, \quad p = \frac{1}{\sqrt{m}} \cdot \vec{1}_m, \quad q = \frac{1}{\sqrt{n}} \cdot \vec{1}_n, \quad x_i = \frac{s - mr_i}{\sqrt{m\Delta_r}}, \quad y_j = \frac{s - nc_j}{\sqrt{n\Delta_c}}.$$

Clearly, p, q, x, y are unit vectors, and $x \perp p$ and $y \perp q$. We check the assumptions of Lemma B.12. By the additional assumption,

$$\sigma_1(M)^2 = \frac{mn}{s^2} \cdot \sigma_1(B)^2 \leq (1+\delta)^2 = 1 + 2\delta + \delta^2,$$

and so we can set $\delta_1 := 2\delta + \delta^2$. Similarly, by the spectral gap condition in Definition B.1,

$$\sigma_2(M)^2 = \frac{mn}{s^2} \cdot \sigma_2(B)^2 \leq (1-\lambda)^2 = 1 - 2\lambda + \lambda^2,$$

and so we can set $\delta_2 := 2\lambda - \lambda^2$. Also, we check that

$$p^* M q = \frac{\vec{1}_m^* B \vec{1}_n}{s} = \frac{1}{s} \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 = 1.$$

Therefore, we can conclude from Lemma B.1 that

$$1 + \frac{2\delta_1 - \delta_2}{2} \geq |x^* M y| = \left| \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{s - mr_i}{\sqrt{m\Delta_r}} \cdot \frac{\sqrt{mn}}{s} a_{ij}^2 \cdot \frac{s - nc_j}{\sqrt{n\Delta_c}} \right|$$

$$= \left| \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(s - mr_i)(s - nc_j)a_{ij}^2}{s\sqrt{\Delta_r \Delta_c}} \right|,$$

which implies that

$$\left| \sum_{i=1}^{m} \sum_{j=1}^{n} (s - mr_i)(s - nc_j)a_{ij}^2 \right| \leq \left(1 + \frac{4\delta + 2\delta^2 - 2\lambda + \lambda^2}{2} \right) s\sqrt{\Delta_r \Delta_c}$$

$$\leq \frac{1}{2} \left(1 + \frac{6\delta - \lambda}{2} \right) s\sqrt{\Delta},$$

where the last inequality follows from $\sqrt{\Delta_r \Delta_c} \leq (\Delta_r + \Delta_c)/2 = \Delta/2$ and $\delta \leq 1$ and $\lambda \leq 1$. □

B.6. Lower bounding the convergence rate. Putting the bounds in Lemmas B.10 and B.13 into (B.6), we obtain the following lower bound on the convergence rate of Δ at any time t.

PROPOSITION B.14. *If $B^{(0)}$ is ϵ-nearly doubly balanced and $B^{(t)}$ satisfies the spectral conditions that*

$$\sigma_1 \left(B^{(t)} \right) \leq (1 + \delta^{(t)}) \frac{s^{(t)}}{\sqrt{mn}} \quad \text{and} \quad \sigma_2 \left(B^{(t)} \right) \leq (1 - \lambda^{(t)}) \frac{s^{(t)}}{\sqrt{mn}}$$

for some continuous functions $\lambda^{(t)}, \delta^{(t)} \leq 1$, then

$$-\frac{1}{4} \frac{d}{dt} \Delta^{(t)} \geq \left(\left(1 + \frac{\lambda^{(t)} - 6\delta^{(t)}}{2} \right) s^{(t)} - (1 + \epsilon)s^{(0)} \right) \Delta^{(t)}.$$

Note that Proposition B.14 implies that the dynamical system has linear convergence at time $t = 0$. To see this, note that $\delta^{(0)} \leq \epsilon$ by Lemma B.11, and $\lambda^{(0)} = \lambda$ from Definition B.1, and therefore

$$-\frac{d}{dt} \Delta^{(0)} \geq 2(\lambda - 8\epsilon)s^{(0)} \Delta^{(0)}.$$

Under our assumption that $\lambda \gg \epsilon$, the dynamical system has linear convergence at time $t = 0$ with rate at least $\lambda s^{(0)}$.

To prove that the dynamical system has linear convergence with rate $\lambda s^{(0)}$ for all time $t \geq 0$, we will prove that the quantities in Proposition B.14 do not change much when we move from $A^{(0)}$ to $A^{(t)}$, i.e., $s^{(t)} \approx s^{(0)}$, $\delta^{(t)} \approx \delta^{(0)}$, and $\lambda^{(t)} \approx \lambda$.

To bound the change of the singular values of $B^{(t)}$, we will bound the condition number of the scaling solutions in the dynamical system in the next subsection, and then use these bounds to argue about the change of the singular values and establish Theorem B.7.

B.7. Condition number. Recall from Lemma B.2 that $A^{(T)} = L^{(T)} A^{(0)} R^{(T)}$ where

$$L^{(T)} = \text{diag}\left(\exp\left(\int_0^T \left(s^{(t)} - mr_i^{(t)}\right) dt\right)\right) \quad \text{and}$$

$$R^{(T)} = \text{diag}\left(\exp\left(\int_0^T \left(s^{(t)} - nc_j^{(t)}\right) dt\right)\right).$$

To bound the condition number of $L^{(T)}$ and $R^{(T)}$, we bound the integrals in the exponent. To bound the integral, we divide the time into two phases. In the first phase, we use Proposition B.8 to argue that $|s^{(t)} - mr_i^{(t)}| \approx |s^{(0)} - mr_i^{(0)}|$. In the second phase, we use that $\Delta^{(t)}$ is converging linearly to argue that $|s^{(t)} - mr_i^{(t)}| \leq \sqrt{m\Delta^{(t)}}$ is converging linearly. In the following lemma, we should think of g as the spectral gap parameter λ in Definition 1.4. The proof of the following lemma is almost identical to that of Lemma 3.16.

LEMMA B.15. *Suppose there exists $g > 0$ such that for all $0 \leq t \leq T$, it holds that*

$$-\frac{d}{dt}\Delta^{(t)} \geq gs^{(0)}\Delta^{(t)}.$$

If $B^{(0)}$ is ϵ-nearly doubly balanced for $\epsilon \leq g$, then

$$\max_i\{L_{ii}^{(T)}\} \leq \exp\left(O\left(\frac{\epsilon \ln m}{g}\right)\right) \quad \text{and} \quad \min_i\{L_{ii}^{(T)}\} \geq \exp\left(O\left(-\frac{\epsilon \ln m}{g}\right)\right).$$

Proof. To bound the condition number, we just need to bound $L_{ii}^{(T)}$ for each $1 \leq i \leq m$ as $L^{(T)}$ is a diagonal matrix. Using the form of $L^{(T)}$ described in Lemma B.2, we bound the absolute value of the integral

$$\left|\int_0^T (s^{(t)} - mr_i^{(t)}) dt\right| \leq \int_0^\tau \left|s^{(t)} - mr_i^{(t)}\right| dt + \int_\tau^T \left|s^{(t)} - mr_i^{(t)}\right| dt.$$

We split the integral into two terms. For the first term, we use Proposition B.8 to bound

$$\int_0^\tau \left|s^{(t)} - mr_i^{(t)}\right| dt \leq \int_0^\tau \left((1+\epsilon)s^{(0)} - s^{(t)}\right) dt \leq \tau\left(s^{(0)} - s^{(T)} + \epsilon s^{(0)}\right),$$

where the second inequality is by the fact that $s^{(t)}$ is nonincreasing from Lemma B.4. Applying Lemma B.6 with our assumption that $\mu = gs^{(0)}$, it follows that

$$\int_0^\tau \left|s^{(t)} - mr_i^{(t)}\right| dt \leq \tau\left(\frac{2\Delta^{(0)}}{gs^{(0)}} + \epsilon s^{(0)}\right) \leq \tau\left(\frac{4\epsilon^2 s^{(0)}}{g} + \epsilon s^{(0)}\right) \leq 5\tau\epsilon s^{(0)},$$

where the second inequality is by Lemma B.3, and the last inequality is by our assumption that $g \geq \epsilon$.

For the second term,

$$\int_\tau^T \left|s^{(t)} - mr_i^{(t)}\right| dt \leq \int_\tau^T \sqrt{m\Delta^{(t)}} dt \leq \sqrt{m\Delta^{(\tau)}} \int_\tau^T e^{-gs^{(0)}(t-\tau)/2} dt \leq \frac{2\sqrt{m\Delta^{(\tau)}}}{gs^{(0)}},$$

where the second inequality is from the inequality that $|s^{(t)} - mr_i^{(t)}| \leq \sqrt{m\Delta^{(t)}}$ from (B.3), and the third inequality follows from the assumption that Δ is converging linearly with $\mu = gs^{(0)}$; see Lemma B.6.

We choose

$$\tau = \frac{\ln m}{gs^{(0)}} \quad \Longrightarrow \quad e^{-gs^{(0)}\tau} \leq \frac{1}{m}.$$

This implies that

$$\Delta^{(\tau)} \leq \Delta^{(0)}e^{-gs^{(0)}\tau} \leq \frac{\Delta^{(0)}}{m} \leq \frac{2\epsilon^2(s^{(0)})^2}{m} \quad \Longrightarrow \quad \frac{2\sqrt{m\Delta^{(\tau)}}}{gs^{(0)}} \leq \frac{3\epsilon}{g},$$

and so the second term is at most $3\epsilon/g$. The first term is at most $5\tau\epsilon s^{(0)} \leq 5\epsilon \ln m/g$. Therefore, we conclude that

$$\exp\left(-\frac{8\epsilon \ln m}{g}\right) \leq \exp\left(-\int_0^T \left|s^{(t)} - mr_i^{(t)}\right| dt\right) \leq L_{i,i}^{(T)} \leq \exp\left(\int_0^T \left|s^{(t)} - mr_i^{(t)}\right| dt\right)$$

$$\leq \exp\left(\frac{8\epsilon \ln m}{g}\right). \qquad \square$$

We cannot use the same argument to bound $\kappa(R^{(T)})$, as it will only give us a bound with dependency on n (where we assumed $m \leq n$). Instead, we use the bound on $\kappa(L^{(T)})$ to derive a similar bound on $\kappa(L^{(T)})$. The proof of the following lemma is simpler than that of Lemma 3.18 in the operator case.

LEMMA B.16. *Suppose there exists $g > 0$ such that for all $0 \leq t \leq T$, it holds that*

$$-\frac{d}{dt}\Delta^{(t)} \geq gs^{(0)}\Delta^{(t)}.$$

If $B^{(0)}$ is ϵ-nearly doubly balanced for $\epsilon \leq g \leq 1$, then $\max_i\{L_{ii}^{(T)}\} \leq e^\ell$ and $\min_i\{L_{ii}^{(T)}\} \geq e^{-\ell}$ implies that

$$\max_j\{R_{jj}^{(T)}\} \leq e^\ell \cdot (1 + O(\epsilon)) \quad \text{and} \quad \min_j\{R_{jj}^{(T)}\} \geq e^{-\ell} \cdot (1 - O(\epsilon)).$$

Proof. By Lemma B.15,

$$\left(a_{ij}^{(T)}\right)^2 = \left(L_{i,i}^{(T)}\right)^2 \left(a_{ij}^{(0)}\right)^2 \left(R_{j,j}^{(T)}\right)^2 \geq e^{-2\ell} \left(a_{ij}^{(0)}\right)^2 \left(R_{j,j}^{(T)}\right)^2.$$

To upper bound $(R_{j,j}^{(T)})^2$, we consider the column sum by summing the above inequality over i to get

$$c_j^{(T)} \geq e^{-2\ell} c_j^{(0)} \left(R_{j,j}^{(T)}\right)^2.$$

This implies that

$$\left(R_{j,j}^{(T)}\right)^2 \leq e^{2\ell} c_j^{(T)}/c_j^{(0)} \leq e^{2\ell} \cdot \frac{(1+\epsilon)s^{(0)}}{n} \cdot \frac{n}{(1-\epsilon)s^{(0)}} \leq e^{2\ell}(1 + O(\epsilon)),$$

where the second inequality is by Proposition B.9 and that $B^{(0)}$ is ϵ-nearly doubly balanced.

Similarly, we can lower bound

$$\left(R_{j,j}^{(T)}\right)^2 \geq e^{-2\ell} c_j^{(T)} / c_j^{(0)} \geq e^{-2\ell} \cdot \frac{2s^{(T)} - (1+\epsilon)s^{(0)}}{n} \cdot \frac{n}{(1+\epsilon)s^{(0)}} \geq e^{-2\ell}(1 - O(\epsilon)),$$

where the last inequality uses the assumption that $\Delta^{(t)}$ is converging linearly to apply Lemma B.6 with $\mu = gs^{(0)}$ to obtain

$$s^{(0)} - s^{(T)} \leq \frac{2\Delta^{(0)}}{gs^{(0)}} \leq \frac{4\epsilon^2 s^{(0)}}{g} \leq 4\epsilon s^{(0)} \quad \implies \quad s^{(T)} \geq (1 - 4\epsilon) \cdot s^{(0)},$$

where we used Lemma B.3 and the assumption that $\epsilon \leq g$. $\quad\square$

B.8. Invariance of linear convergence. We will first use Lemma B.15 and Lemma B.16 to bound the change of the singular values of $B^{(t)}$. Then, we will combine the previous results to prove Theorem B.7 that $\Delta^{(t)}$ is converging linearly for all $t \geq 0$.

LEMMA B.17. *Given matrix* $A \in \mathbb{R}_+^{m \times n}$, *and left/right diagonal scalings* $\|L - I_m\|_{op} \leq \zeta$ *and* $\|R - I_n\|_{op} \leq \zeta$ *for some* $\zeta \leq 1$, *if we denote* $B := L^2 A R^2$ *then*

$$\forall k : |\sigma_k(B) - \sigma_k(A)| \leq O(\zeta) \cdot \|A\|_{op}.$$

Proof. We use Lemma 3.19 to bound the singular value change by the operator norm of the matrix change:

$$|\sigma_k(B) - \sigma_k(A)| = \left|\sigma_k((L)^2 A(R)^2) - \sigma_k(A)\right| \leq \left\|\left(L^{(t)}\right)^2 A \left(R^{(t)}\right)^2 - A\right\|_{op}.$$

We write $L = I + \widetilde{L}$ and $R = I + \widetilde{R}$, so that $\|\widetilde{R}\|_{op} \leq \zeta$ and $\|\widetilde{C}\|_{op} \leq \zeta$ by our assumptions. Then,

$$
\begin{aligned}
\left\|A - L^2 A R^2\right\|_{op} &= \left\|A - (I + \widetilde{L})^2 A (I + \widetilde{R})^2\right\|_{op} \\
&= \left\|2\widetilde{L}A + 2A\widetilde{R} + \widetilde{L}^2 A + A\widetilde{R}^2 + 2\widetilde{L}^2 A\widetilde{R} + 2\widetilde{L}A\widetilde{R}^2 + 4\widetilde{L}A\widetilde{R} + \widetilde{L}^2 A\widetilde{R}^2\right\|_{op} \\
&\leq O(\zeta) \|A\|_{op},
\end{aligned}
$$

where we used the triangle inequality and bound the sum of the eight operator norms, and used the fact that $\|XAY\|_{op} \leq \|X\|_{op} \|Y\|_{op} \|A\|_{op}$ for each term, and used the assumption that $\|\widetilde{L}\|_{op}, \|\widetilde{R}\|_{op} \leq \zeta \leq 1$ so that each term is at most $O(\zeta) \|A\|_{op}$. $\quad\square$

We are ready to put together the results to prove the following theorem, which implies Theorem B.7. The proof is almost the same as that of Theorem 3.21.

THEOREM B.18. *If* $B^{(0)}$ *is* ϵ-*nearly doubly balanced and* $B^{(0)}$ *satisfies the* λ-*spectral gap condition in Definition* B.1 *with* $\lambda^2 \geq C\epsilon \ln m$ *for a sufficiently large universal constant* C, *then for all* $t \geq 0$ *it holds that*

$$-\frac{d}{dt}\Delta^{(t)} \geq \lambda s^{(0)} \Delta^{(t)}.$$

Proof. For Proposition 3.9 let $\delta^{(t)}$ be defined according to the bounds achieved in Lemma B.11 and let $\lambda^{(t)}$ be the continuous function denoting the spectral gap of the scaling at time t. Note $\delta^{(0)} \leq \epsilon$ by Lemma B.11 and $\lambda^{(0)} = \lambda$ from Definition B.1. Let T be the supremum such that $s^{(t)} \geq (1-\epsilon)s^{(0)}$ and $\lambda^{(t)} - 6\delta^{(t)} \geq \frac{1}{2}(1-\epsilon)^{-1}(\lambda^{(0)} + 8\epsilon)$. Our goal is to prove that $\Delta^{(t)}$ is converging linearly for $0 \leq t \leq T$ and T is unbounded.

First, we show that $\Delta^{(t)}$ is converging linearly for $0 \leq t \leq T$. By Proposition B.14,

$$
\begin{aligned}
-\frac{d}{dt}\Delta^{(t)} &\geq 4\left(\left(1 + \frac{1}{2}\left(\lambda^{(t)} - 6\delta^{(t)}\right)\right)s^{(t)} - (1+\epsilon)s^{(0)}\right)\Delta^{(t)} \\
&\geq 4\left((1-\epsilon)\left(1 + \frac{1}{4}(1-\epsilon)^{-1}(\lambda^{(0)} + 8\epsilon)\right) - (1+\epsilon)\right)s^{(0)}\Delta^{(t)} \\
&= \lambda^{(0)}s^{(0)}\Delta^{(t)} = \lambda s^{(0)}\Delta^{(t)},
\end{aligned}
$$

where in the second inequality we used that $s^{(t)} \geq (1-\epsilon)s^{(0)}$ and $\lambda^{(t)} - 6\delta^{(t)} \geq \frac{1}{2}(1-\epsilon)^{-1}(\lambda^{(0)} + 8\epsilon)$ for $0 \leq t \leq T$.

Next, we argue that the size condition and the spectral gap condition will still be maintained beyond time T. For the size change, by Lemma B.6 with $\mu = \lambda s^{(0)}$,

$$
s^{(0)} - s^{(T)} \leq \frac{2\Delta^{(0)}}{\lambda s^{(0)}} \leq \frac{4\epsilon^2 s^{(0)}}{\lambda} \ll \epsilon s^{(0)},
$$

where the second inequality is by Lemma B.3 and the last inequality is by $\lambda \geq C\epsilon$ for a sufficiently large constant C.

For the change of the second largest singular value, by definition,

$$
\begin{aligned}
\sigma_2(B^{(T)}) - \sigma_2(B^{(0)}) &= \frac{(1-\lambda^{(T)})s^{(T)}}{\sqrt{mn}} - \frac{(1-\lambda^{(0)})s^{(0)}}{\sqrt{mn}} \\
&\geq \frac{(1-\lambda^{(T)})(1-\epsilon)s^{(0)}}{\sqrt{mn}} - \frac{(1-\lambda^{(0)})s^{(0)}}{\sqrt{mn}} \\
&= \frac{s^{(0)}}{\sqrt{mn}}\left(\lambda^{(0)} - (1-\epsilon)\lambda^{(T)} - \epsilon\right).
\end{aligned}
$$

On the other hand, we can upper bound $\sigma_2(B^{(T)}) - \sigma_2(B^{(0)})$ using condition numbers. Using Lemma B.15 with $g = \lambda$, $\max_i\{L_{ii}^{(T)}\} \leq \exp\left(O(\epsilon \ln m/\lambda)\right)$ and $\min_i\{L_{ii}^{(T)}\} \geq \exp\left(-O(\epsilon \ln m/\lambda)\right)$. Note that our assumption implies that

$$
O\left(\frac{\epsilon \ln m}{\lambda}\right) \leq O\left(\frac{\lambda}{C}\right) \ll 1 \quad \implies \quad \left\|L^{(T)} - I\right\|_{\mathrm{op}} \leq O\left(\frac{\lambda}{C}\right) \ll 1,
$$

where the implication is by the inequality $e^x - 1 \leq O(x)$ for x close to zero. Then, by Lemma B.16, we also have $\left\|R^{(T)} - I\right\|_{\mathrm{op}} \leq O\left(\lambda/C\right)$. Putting these bounds into ζ of Lemma B.17, we obtain

$$
\sigma_2(B^{(t)}) - \sigma_2(B^{(0)}) \leq O\left(\frac{\lambda}{C}\right) \cdot \left\|B^{(0)}\right\|_{\mathrm{op}} \leq O\left(\frac{\lambda}{C}\right)\frac{(1+\delta_1^{(0)})s^{(0)}}{\sqrt{mn}}.
$$

Combining the upper bound and lower bound and using $\delta_1^{(0)} \leq \epsilon$ from Lemma B.11, it follows that

$$
\lambda^{(T)} \geq \frac{\lambda - \epsilon - (1+\epsilon) \cdot O\left(\lambda/C\right)}{1-\epsilon} \geq \lambda - O\left(\frac{\lambda}{C}\right),
$$

where the last inequality is by the assumption that $\lambda \geq C\epsilon$.

For the change of the largest singular value, by Proposition B.9,

$$\frac{(1-3\epsilon)s^{(T)}}{m}I_m \preceq \frac{2s^{(T)} - (1+\epsilon)s^{(0)}}{m}I_m \preceq \mathrm{diag}\left(\left\{r_i^{(T)}\right\}_{i=1}^m\right) \preceq \frac{(1+\epsilon)s^{(0)}}{m}I_m$$
$$\preceq \frac{(1+3\epsilon)s^{(T)}}{m}I_m,$$

where the first and last inequalities use that $s^{(T)} \geq (1-\epsilon)s^{(0)}$. The same holds for $\mathrm{diag}(\{c_j^{(T)}\}_{j=1}^n)$ and these imply that $\mathcal{A}^{(T)}$ is 3ϵ-nearly doubly balanced. By Lemma B.11, this implies that $\delta^{(T)} \leq 3\epsilon$. Therefore,

$$\lambda^{(T)} - 6\delta^{(T)} \geq \lambda - O\left(\frac{\lambda}{C}\right) - 9\epsilon \geq \lambda - O\left(\frac{\lambda}{C}\right) \geq \frac{2}{3}\lambda > \frac{1}{2}(1-\epsilon)^{-1}(\lambda + 8\epsilon),$$

where the second to last inequality uses that C is a sufficiently large constant.

Since our dynamical system is continuous, we still have both conditions satisfied at time $T + \eta$ for some finite $\eta > 0$, which contradicts that T is the supremum that both conditions are satisifed. Therefore, T is unbounded and the linear convergence of Δ is maintained throughout the execution of the dynamical system. $\square$

REFERENCES

[1] Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, and A. Wigderson, *Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing*, in Proceedings of the 50th Annual ACM Symposium on Theory of Computing (STOC), 2018, pp. 172–181.

[2] Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson, *Much faster algorithms for matrix scaling*, in Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2017.

[3] M. F. Atiyah, *Convexity and commuting Hamiltonians*, Bull. Lond. Math. Soc., 14 (1982), pp. 1–5.

[4] K. Ball, *Volumes of sections of cubes and related problems*, in Geometric Aspects of Functional Analysis (1987-88), Lecture Notes in Math. 1376, Springer, Berlin, 1989, pp. 251–260.

[5] F. Barthe, *On a reverse form of the Brascamp-Lieb inequality*, Invent. Math., 134 (1998), pp. 335–361.

[6] A. Barvinok and A. Samorodnitsky, *Computing the partition function for perfect matchings in a hypergraph*, Combin. Probab. Comput., 20 (2011), pp. 815–835.

[7] A. Ben-Aroya, O. Schwartz, and A. Ta-Shma, *Quantum expanders: Motivation and construction*, Theory Comput., 6 (2010), pp. 47–79.

[8] J. Bennett, A. Carbery, M. Christ, and T. Tao, *The Brascamp-Lieb inequalities: Finiteness, structure, and extremals*, GAFA Geom. Funct. Anal., 17 (2008), pp. 1343–1415.

[9] R. Bhatia, *Positive Definite Matrices*, Princeton University Press, Princeton, NJ, 2007.

[10] H. Brascamp and E. Lieb, *Best constants in Young's inequality, its converse and its generalization to more than three functions*, Adv. Math., 20 (1976), pp. 151–172.

[11] P. Buergisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson, *Efficient algorithms for tensor scaling, quantum marginals and moment polytopes*, in Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2018.

[12] P. Buergisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson, *Towards a theory of non-commutative optimization: Geodesic 1st and 2nd order methods for moment maps and polytopes*, in Proceedings of the IEEE 60th Annual Symposium on Foundations of Computer Science, 2019, pp. 845–861.

[13] M. Choi, *Completely positive linear maps on complex matrices*, Linear Algebra Appl., 10 (1975), pp. 285–290.

[14] M. B. COHEN, A. MADRY, D. TSIPRAS, AND A. VLADU, *Matrix scaling and balancing via box constrained Newton's method and interior point methods*, in Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2017.

[15] M. CUTURI, *Sinkhorn distances: Lightspeed computation of optimal transport*, in Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS), 2013, pp. 2292–2300.

[16] G. P. EGORYCHEV, *The solution of van der Waerden's problem for permanents*, Adv. Math., 42 (1981), pp. 299–305.

[17] D. I. FALIKMAN, *Proof of the van der Waerden's conjecture on the permanent of a doubly stochastic matrix*, Mat. Zametki, 29 (1981), pp. 931–938.

[18] J. FORSTER, *A linear lower bound on the unbounded error probabilistic communication complexity*, J. Comput. System Sci., 65 (2002), pp. 612–625.

[19] J. FRANKLIN AND J. LORENZ, *On the scaling of multidimensional matrices*, Linear Algebra Appl., 114–115 (1989), pp. 717–735.

[20] W. C FRANKS AND A. MOITRA, *Rigorous guarantees for Tyler's M-estimator via quantum expansion*, in Proceedings of the 33rd Conference on Learning Theory, 2020, pp. 1601–1632.

[21] A. GARG, L. GURVITS, R. OLIVEIRA, AND A. WIGDERSON, *A deterministic polynomial time algorithm for non-commutative rational identity testing*, in Proceedings of the 57th Annual Symposium on Foundations of Computer Science (FOCS), 2016, pp. 109–117.

[22] A. GARG, L. GURVITS, R. OLIVEIRA, AND A. WIGDERSON, *Operator Scaling: Theory and Applications*, arXiv:1511.03730, 2015.

[23] A. GARG, L. GURVITS, R. OLIVEIRA, AND A. WIGDERSON, *Algorithmic and optimization aspects of Brascamp-Lieb inequalities, via operator scaling*, in Proceedings of the 49th Annual ACM Symposium on Theory of Computing (STOC), 2017, pp. 397–409.

[24] A. GARG AND R. OLIVEIRA, *Recent progress on scaling algorithms and applications*, Bull. Eur. Assoc. Theor. Comput. Sci. EATCS, 125 (2018).

[25] A. GOEL, M. KAPRALOV, AND S. KHANNA, *Perfect matchings in $O(n \log n)$ times in regular bipartite graphs*, in Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC), 2010, pp. 39–46.

[26] C. GONŹALEZ-GUILÉN, M. JUNGE, AND I. NECHITA, *Spectral gap of random quantum channels*, presented at Workshop on Probabilistic Techniques and Quantum Information Theory, 2017.

[27] V. GUILLEMIN AND S. STERNBERG, *Convexity properties of the moment mapping*, Invent. Math., 67 (1982), pp. 491–513.

[28] V. GUILLEMIN AND S. STERNBERG, *Convexity properties of the moment mapping,* II, Invent. Math., 77 (1984), pp. 533–546.

[29] V. GUILLEMIN AND R. SJAMAAR, *Convexity Properties of Hamiltonian Group Actions*, AMS, Providence, RI, 2005.

[30] L. GURVITS, *Hyperbolic polynomials approach to Van der Waerden/Schrijver-Valiant like conjectures: Sharper bounds, simpler proofs and algorithmic applications*, in Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC), 2006, pp. 417–426.

[31] L. GURVITS, *Classical complexity and quantum entanglement*, J. Comput. System Sci., 69 (2004), pp. 448–484.

[32] L. GURVITS AND A. SAMORODNITSKY, *A deterministic polynomial-time algorithm for approximating mixed discriminant and mixed volume*, in Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC), 2000, pp. 48–57.

[33] L. GURVITS AND P. N. YIANILOS, *The Deflation-Inflation Method for Certain Semidefinite Programming and Maximum Determinant Completion Problems*, Technical report, NECI, 1998.

[34] L. HAMILTON AND A. MOITRA, *The Paulsen Problem Made Simple*, arXiv:1809.04726, 2018.

[35] M. HARDT AND A. MOITRA, *Algorithms and hardness for robust subspace recovery*, in Proceedings of the 26th Annual Conference on Learning Theory (COLT), PMLR 30, 2013, pp. 354–375.

[36] N. HARMS, *Testing Halfspaces over Rotation-Invariant Distributions*, arXiv:1811.00139, 2018.

[37] M. B. HASTINGS, *Random unitaries give quantum expanders*, Phys. Rev. A, 76 (2007), 032315.

[38] R. B. HOLMES AND V. I. PAULSEN, *Optimal frames for erasures*, Linear Algebra Appl., 377 (2004), pp. 31–51.

[39] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, UK, 2012.

[40] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[41] M. Idel, *A Review of Matrix Scaling and Sinkhorn's Normal Form for Matrices and Positive Maps*, arXiv:1609.06349, 2016.

[42] B. Kalantari and L. Khachiyan, *On the complexity of nonnegative-matrix scaling*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 450–463.

[43] G. Kempf and L. Ness, *The length of vectors in representation spaces*, in Algebraic Geometry, Lecture Notes in Math. 233, Springer, New York, 1979, pp. 233–243.

[44] F. C. Kirwan, *Cohomology of Quotients in Symplectic and Algebraic Geometry*, Math. Notes 31, Princeton University Press, Princeton, NJ, 1984.

[45] F. Kirwan, *Convexity properties of the moment mapping,* III, Invent. Math., 77 (1984), pp. 547–552.

[46] M. Krivelevich and B. Sudakov, *Pseudo-random graphs*, in More Sets, Graphs and Numbers, Bolyai Soc. Math. Stud. 15, János Bolyai Mathematical Society, Budapest, 2006, pp. 199–262.

[47] T. C. Kwok, L. C. Lau, Y. T. Lee, and A. Ramachandran, *The Paulsen Problem, Continuous Operator Scaling, and Smoothed Analysis*, CoRR abs/1710.02587, 2017.

[48] E. Lieb, *Gaussian kernels have only Gaussian maximizers*, Invent. Math., 102 (1990), pp. 179–208.

[49] N. Linial, A. Samorodnitsky, and A. Wigderson, *A deterministic strongly polynomial algorithm for matrix scaling and approximate permanent*, in Proceedings of STOC, 1998, pp. 644–652.

[50] A. Madry, *Navigating central path with electrical flows: From flows to matchings, and back*, in Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2013, pp. 253–262.

[51] P. Milgrom and I. Segal, *Envelope theorems for arbitrary choice sets*, Econometrica, 70 (2010), pp. 583–601.

[52] A. Nikolov and M. Singh, *Maximizing determinants under partition constraints*, in Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC), 2016, pp. 192–201.

[53] A. Nemirovski and U. Rothblum, *On complexity of matrix scaling*, Linear Algebra Appl., 302 (1999), pp. 435–460.

[54] G. Peyré and M. Cuturi, *Computational Optimal Transport*, arXiv:1803.00567, 2018.

[55] M. Rudelson, A. Samorodnitsky, and O. Zeitouni, *Hafnians, perfect matchings and Gaussian matrices*, Ann. Probab., 44 (2016), pp. 2858–2888.

[56] R. Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, Ann. Math. Stat., 35 (1964), pp. 876–879.

[57] A. Slavík, *Product Integration, Its History and Applications*, Matfyzpress, Prague, 2007.

[58] M. Strohmer and R. W. Heath, Jr., *Grassmannian frames with applications to coding and communications*, Appl. Comput. Harmon. Anal., 14 (2003), pp. 257–275.

[59] M. A. Sustik, J. A. Tropp, I. S. Dhillon, R. W. Heath, Jr., *On the existence of equiangular tight frames*, Linear Algebra Appl., 426 (2007), pp. 619–635.

[60] T. Tao, *Topics in Random Matrix Theory*, Grad. Stud. Math. 132, AMS, Providence, RI, 2012.

[61] J. A. Tropp, I. S. Dhillon, R. W. Heath, Jr., and T. Strohmer, *Designing structured tight frames via an alternating projection method*, IEEE Trans. Inform. Theory, 51 (2005), pp. 188–209.

[62] J. A. Tropp, *An introduction to matrix concentration inequalities*, Found. Trends Mach. Learn., 8 (2015), pp. 1–230.

[63] J. Watrous, *The Theory of Quantum Information*, Cambridge University Press, Cambridge, UK, 2018.

[64] J. Watrous, *private communication*, 2018.