# Spectral Analysis of Matrix Scaling and Operator Scaling

Tsz Chiu Kwok
Institute for Theoretical Computer Science,
Shanghai University of Finance and Economics
Shanghai, China
Email: kwok@mail.sufe.edu.cn

Lap Chi Lau
School of Computer Science,
University of Waterloo
Waterloo, Canada
Email: lapchi@uwaterloo.ca

Akshay Ramachandran
School of Computer Science,
University of Waterloo
Waterloo, Canada
Email: a5ramach@uwaterloo.ca

*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.

*Index Terms*—Operator Scaling, Matrix scaling, Spectral analysis, Frame theory, Quantum expanders

## I. Introduction

In the matrix scaling problem, we are given a non-negative 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 [39] for a detailed survey.

The operator scaling problem is a significant generalization of the matrix scaling problem. Given a tuple of $m \times n$ real matrices $\mathcal{A} = (A_1, \ldots, A_k) \subset \mathbb{R}^{m \times n}$, 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 $A_i$, We will simply refer to $\mathcal{A}$ as an operator. The size of an operator $\mathcal{A}$ is $s(\mathcal{A}) := \sum_{i=1}^{k} \|A_i\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. An operator $\mathcal{A}$ is called $\epsilon$-nearly doubly balanced if

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

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

and is called doubly balanced when $\epsilon = 0$. The operator scaling problem is defined by Gurvits [29]. The objective is to scale the input operator so that it becomes doubly balanced with size one.

**Definition I.1** (Operator Scaling Problem).

**Input:** *An operator* $\mathcal{A} = (A_1, \ldots, A_k) \subset \mathbb{R}^{m \times n}$.

**Output:** *A left and right scaling matrix* $L \in \mathbb{R}^{m \times m}$, $R \in \mathbb{R}^{n \times n}$ *such that for* $\{B_i := LA_i R\}_{i=1}^{k}$

$$\sum_{i=1}^{k} B_i B_i^* = \frac{I_m}{m} \quad \text{and} \quad \sum_{i=1}^{k} B_i^* B_i = \frac{I_n}{n}$$

*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 IV-A for details.

The operator scaling problem generalizes matrix scaling and frame scaling and has many applications; see Section I-D and Section IV. Much work has been done in analyzing algorithms for these scaling problems and in understanding the scaling solutions and related quantities.

### A. Previous Algorithms

For matrix scaling, the most well-known algorithm is Sinkhorn's algorithm [54], which is a simple iterative algorithm that alternatively rescale the rows and rescale the columns. This algorithm is analyzed in [18] and it is shown that the alternating algorithm finds an $\eta$-nearly doubly stochastic scaling in time polynomial in $n$ and $1/\eta$.

The alternating scaling algorithm is generalized in [29] for the operator scaling problem. In this algorithm, we repeatedly 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 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. This alternating algorithm is partially analyzed in [29] and is fully analyzed in [19], [20].

**Theorem I.2** ( [18]–[20], [54]). *The alternating scaling algorithm returns an $\eta$-nearly doubly balanced scaling in $O(\mathrm{poly}(n, m, k, 1/\eta))$ iterations if such a scaling exists.*

This theorem is used in [19], [20] to give the first polynomial time algorithm for computing the non-commutative rank of a symbolic matrix, as it is sufficient to set $\eta$ to be inverse polynomial in $n$ to solve that problem exactly. For further applications, faster convergence of $\eta$ is required.

For matrix scaling, there are several algorithms with dependency on $\eta$ being $\log(1/\eta)$, including the ellipsoid method in [40], the interior point method in [51], and a strongly polynomial time combinatorial algorithm in [47]. The dependency on $n$ in these algorithms is at least $\Omega(n^{7/2})$ even for sparse matrices. Recently, two independent groups [2], [13] 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 I.3** ( [1], [2], [13]). *There is a second order method to return an $\eta$-nearly doubly balanced scaling in time $O(\mathrm{poly}(n, m, k, \log(1/\eta)))$ 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 $\kappa$ 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 [47]. For operator scaling, this theorem is used to obtain a polynomial time algorithm for an orbit intersection problem in invariant theory [1].

*B. Gradient Flow*

An important quantity in [1], [20], [29] to measure the progress of the algorithms is the $\ell_2$-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 $1 \leq i \leq k$, define

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

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 $i$-th row sum and the $j$-th 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 [45], where both operations are done simultaneously and continuously. The following differential equation describes how $\mathcal{A}$ changes over time:

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

In the matrix case, this continuous scaling 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 [45]. Its convergence rate is shown to be similar to that of the alternating scaling algorithm, with dependency on $\eta$ being $1/\eta$.

The continuous operator scaling algorithm can be understood as a natural first order method for operator scaling. It can be shown (see full version) that the dynamical system in continuous operator scaling is equivalent to the gradient flow of $\Delta(\mathcal{A})$ at each time. This shows a close connection between gradient descent and the alternating algorithm.

This gradient flow was studied in much greater generality in symplectic geometry and algebraic geometry (see [27], [41]). After a long line of work [3], [25], [26], [42], [43], Kirwan proved that the image of the moment map of a group action on a symplectic manifold is a convex polytope. To prove this, Kirwan uses the norm-square of the moment map (which in our setting is exactly $\Delta(\mathcal{A})$), and studies critical points of this function in order to understand the image of the moment map (where a point is critical for $\Delta(\mathcal{A})$ exactly when it is a fixed point of the gradient flow). The current result as well as the result in [45] can be seen as quantitative convergence analyses in the neighborhoods of fixed points of this natural gradient flow in the operator scaling setting.

*C. 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 $\eta$ 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 I.4** (Spectral Gap Condition)**.** *Given an operator* $\mathcal{A} = (A_1, \ldots, A_k)$ *where* $A_i \in \mathbb{R}^{m \times n}$ *for* $1 \leq i \leq k$*, 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* $\lambda$*-spectral gap if*

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

*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. [61]), 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

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

when $\mathcal{A}$ is $\epsilon$-nearly doubly balanced (Lemma III.5). The spectral gap condition is also studied under the name of "quantum expander" in [7], [35]. We will discuss more about this spectral gap condition in Section II-A 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

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

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 Section I-D1 and in Section IV-A.

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

**Theorem I.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* $\epsilon$*-nearly doubly balanced and* $\mathcal{A}$ *satisfies the* $\lambda$*-spectral gap condition in Definition I.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 a* $\eta$*-nearly doubly balanced scaling in time* $t = O\left(\frac{1}{\lambda} \log(\frac{m}{\eta})\right)$*, 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 $\eta$-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 I.6** (Gradient Descent)**.** *Under the assumptions in Theorem I.5, there is a gradient descent algorithm to return an* $\eta$*-nearly doubly balanced scaling in* $O\left(\frac{(n+m)^2}{\lambda} \log(\frac{m+n}{\eta})\right)$ *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. We believe that the answer is positive but we could not prove it at the moment.

*Condition Number:* The condition number of the scaling solutions $L, R$ are 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 I.3), but not much is known even in the simpler matrix scaling setting. Kalantari and Khachiyan [40] 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 I.7** (Condition Number)**.** *Under the assumptions in Theorem I.5, the condition number of the scaling solutions* $L \in \mathbb{R}^{m \times m}$ *and* $R \in \mathbb{R}^{n \times n}$ *satisfy*

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

The condition number of the scaling solutions is used in bounding the time complexity of the scaling algorithms using the second order method [1], [13], in analyzing an approximation algorithm for permanent [53], and in bounding the optimal transport cost [14], [52]. We will discuss the implications of Theorem I.7 to these applications in Section IV.

*Operator Capacity:* The capacity of an operator $\mathcal{A}$ is defined by Gurvits [29] as

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

The capacity of a matrix $B \in \mathbb{R}^{m \times n}$ has a simpler form (Section IV-A6) where

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

Optimization problems of this form are also studied in functional analysis [5] and in approximation algorithms [50].

In general, when $\mathcal{A}$ is $\epsilon$-nearly doubly balanced [20], [29], [45], it is proved that

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

Using a connection between the convergence rate of the gradient flow and the operator capacity developed in [45], we show a much stronger bound for operators that also satisfy the spectral gap condition.

**Theorem I.8** (Capacity). *Under the assumptions in Theorem I.5,*

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

The capacity of an operator is used in bounding the permanent of a matrix [47], the Brascamp-Lieb constant of an operator [21], and the total movement to a nearby doubly balanced operator [45]. We will discuss the implications of Theorem I.8 to these applications in Section I-D.

*D. Applications*

The matrix scaling and the operator scaling problem has many applications and we will discuss some implications of our results in this section.

*1) Matrix Scaling:* In the matrix scaling problem, we are given a non-negative 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 IV-A for definition), or report that such scaling matrices do not exist.

The matrix scaling problem is a special case of the operator scaling problem (Section IV-A1) and so the spectral analysis also applies. In the case of matrix scaling, the spectral condition in Definition I.4 is simply $\sigma_2(B) \leq (1 - \lambda)s(B)/\sqrt{mn}$ (Section IV-A2). 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 IV-A3). These imply that many random matrices will satisfy the condition in Theorem I.5 (Section IV-A4).

Our results has 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 IV-A5).

**Corollary I.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 $\epsilon$. 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 $\eta$-nearly perfect fractional matching in time $t = O\left(\frac{1}{\phi^2(G)} \log\left(\frac{|X|}{\eta}\right)\right)$.*

For permanent, the 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 [15], [16], [28]. The capacity

lower bound in Theorem I.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 I.10.** *If a non-negative matrix $B \in \mathbb{R}^{n \times n}$ is $\epsilon$-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. This implies a sub-exponential approximation of the permanent for this class of matrices [6]. See Section IV-A6 for details.

For optimal transportation distance, we can use the condition number result in Theorem IV-A7 to bound the Sinkhorn distance [14], [52], which is receiving increasing attention in computer vision and machine learning (Section IV-A7).

The condition number result in Theorem IV-A7 can also be used to show that the second-order method for matrix scaling [2], [13] as stated in Theorem I.3 is near linear time in the instances satisfying the spectral gap assumption.

*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 [17], machine learning [33], and in frame theory [32], [45].

The frame scaling problem is a special case of the operator scaling problem (Section IV-B1) and so the spectral analysis also applies. In the case of frame scaling, the spectral condition in Definition I.4 has a nice form (Section IV-B2): 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 $s := \sum_{i=1}^n \|u_i\|^2$. We show in the full version that this condition is satisfied for random frames with high probability.

**Theorem I.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 $\epsilon$-nearly doubly balanced for $\epsilon \ll 1/\log d$ and satisfies the spectral gap condition with constant $\lambda$ with high probability.*

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 $n$-by-$n$ all-one matrix. The matrix $J_n$ has the largest 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.

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

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

$$\|u_i\|_2^2 \in (1 \pm \epsilon)\frac{d}{n} \quad \text{for } 1 \leq i \leq n,$$

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 [45], 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 [32] 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 [36] 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 $\epsilon$, and almost all known constructions are random frames [36], [59]. 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 I.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 IV-B4).

**Theorem I.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 $\|v_i\| = 1$,*

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

*3) Operator Scaling:* The operator scaling problem was used to compute the Brascamp-Lieb constant [21]. 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 non-negative 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], [21]. It turns out that the functions $f_i$ for which the inequality is tight are density functions of Gaussians [46], and this implies the Brascamp-Lieb constant can be written in a form very similar to the capacity of an operator (see Section IV-C1). This is used in [21] 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 I.8.

**Corollary I.14.** *Given a datum $(\mathbf{B}, \mathbf{p})$ with $B_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$ for $1 \leq j \leq m$ and $\sum_{j=1}^{m} p_j n_j = n$, if $(\mathbf{B}, \mathbf{p})$ is $\epsilon$-nearly geometric and satisfies the $\lambda$-spectral gap condition with $\lambda^2 \geq C\epsilon \log n$ for some sufficiently large constant $C$ and $\sum_{j=1}^{m} p_j \|B_j\|_F^2 = n$, then*

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

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 $1 \leq j \leq m$ which was studied in [5]. In this case, the capacity of the operator $\mathcal{A}$ from the reduction (Section IV-C1) is

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

which is a form that is also studied in approximation algorithms [50]. Using the results in Section V and the above corollary, we can show that if each $u_i$ is an independent random unit vector and $m \geq \Omega(d^{4/3})$, then $m \geq \text{cap}(\mathcal{A}) \geq m\left(1 - 4d \log d/m\right)$ and $1 \leq \text{BL}(\mathbf{B}, \mathbf{p}) \leq d^{\Theta(d)}$; see Example IV.26. Note that this is independent of the number of vectors.

The operator scaling algorithm is used in [19], [20] to compute the non-commutative rank of a symbolic matrix. We show in Section IV-C2 that an operator satisfying the spectral gap condition has full non-commutative rank.

In solving the orbit intersection problem [1], the result of a generalization of the Paulsen problem to the operator setting in [45] was used. As in Theorem I.12, we prove a much stronger bound in Section IV-C3 on the squared distance when the operator satisfies the spectral gap condition.

*E. Techniques*

We are not aware of previous work on spectral analysis of matrix scaling and operator scaling. The closest work we are aware of is a recent work by Rudelson, Samorodnitsky and Zeitouni [53], 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) Comparisons with Previous Techniques:* The operator capacity defined by Gurvits [29] was used crucially as a potential function to analyze the discrete operator scaling algorithms in [20], [29] as well as the continuous operator scaling algorithm in [45].

A smoothed analysis of matrix scaling was presented in [45] for solving the Paulsen problem. It was shown that if most of the entries of an $m \times n$ matrix with $m \leq n$ is at least $\sigma^2$ for a large enough $\sigma$, 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 combinatorial assumption implies the spectral gap assumption in Definition I.4 with $\lambda \geq \Omega(\sigma^2)$ but not vice versa. Through a reduction from operator capacity to matrix capacity, the smoothed analysis can be extended to the frame setting but the proof was complicated, and it was not known whether the smoothed analysis can be extended to the general operator setting. The main difficulty is that there is no analogous combinatorial condition in the frame setting and in the operator setting to guarantee the linear convergence. This is an illustration of the difference between the matrix case and the noncommutative operator case, in which there is no natural basis to consider. In this paper, we have found a natural spectral condition to guarantee linear convergence directly in the general operator setting. 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 [45].

*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 $\Delta$. In the matrix scaling setting, it follow from Lemma 4.2.9 in [45] 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 non-negative matrix, and $s, r_i, c_j$ are the size, the $i$-th row sum and the $j$-th 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 $\epsilon$-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 $i$-th entry being $s - mr_i$ and $\vec{c} \in \mathbb{R}^n$ is the vector with the $j$-th 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.

### F. Organization

We first review some background about completely positive linear operators and the continuous operator scaling algorithm in Section II. We then present the main technical results in Section III and show various applications in Section IV. We provide a proof sketch in Section V that a random frame satisfies the spectral condition with high probability. Many proofs are omitted but all details are present in the full version.

## II. PRELIMINARIES

We first review in Section II-A some background in quantum information theory about completely positive maps and discuss the spectral gap condition stated in Definition I.4. Then, we review the known results about the continuous operator scaling algorithm in Section II-B

### A. Positive Linear Maps, Quantum Expanders

First, we define completely positive linear maps and their natural matrix representation in Section II-A1. Then, in Section II-A2, we present the spectral gap condition in Definition I.4 using this language, and compare to the notion of quantum expanders studied in the literature. Finally, we

introduce the Choi matrix in Section II-A3 and state some facts about tensors and completely positive maps that we will use in our proof.

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

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

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 II.1** (Completely Positive Map)**.** *A linear map $\Phi$ 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 $\Phi$ is completely positive if $\Phi \otimes I_l$ is positive for every natural number $l \ge 1$ (see [61] for more details).*

**Theorem II.2** (Choi [12])**.** *A linear map $\Phi$ is completely positive if and only if it can be written as the form described in (II.1).*

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

**Definition II.3** (Doubly Balanced Map)**.** *A linear map $\Phi$ is called unital if $\Phi(I_n) = I_m$. A linear map $\Phi$ is called trace preserving if $\Phi^*(I_m) = I_n$ (which implies that $\operatorname{tr}(\Phi(Y)) = \operatorname{tr}(Y)$ for any $Y \in \mathbb{R}^{n \times n}$). A linear map $\Phi$ 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 non-zero doubly balanced.

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

**Definition II.4** (Natural Matrix Representation)**.** *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*

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

*where $\operatorname{vec} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n^2}$ is the linear map satisfying $\operatorname{vec}(E_{i,j}) = e_i \otimes e_j$ for all $1 \le i, j \le n$, 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 $i$-th 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*

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

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

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

**Fact II.5** (Proposition 2.20 in [61])**.** *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 I.4 such that*

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

*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 I.4 can be presented as follows.

**Definition II.6** (Spectral Gap Condition of $\Phi$)**.** *Given an operator $\mathcal{A} = (A_1, \ldots, A_k) \subset \mathbb{R}^{m \times n}$, let*

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

*and $Y_1, y_1 := \operatorname{vec}(Y_1)$ as maximizers to the above. Let*

$$\sigma_2(\Phi_{\mathcal{A}}) := \max_{Y \perp Y_1} \frac{\|\Phi(Y)\|_F}{\|Y\|_F} = \max_{y \perp y_1} \frac{\|M_{\mathcal{A}} \cdot y\|_2}{\|y\|_2}$$

*The spectral gap condition in Definition I.4 is equivalent to $\sigma_2(\Phi_{\mathcal{A}}) \le (1 - \lambda)s(\mathcal{A})/\sqrt{mn}$.*

The concept of quantum expander was studied by Hastings [35] and Ben-Aroya, Schwartz, and Ta-Shma [7], which was stated using the above language with $m = n$.

**Definition II.7** (Quantum Expander [7], [35])**.** *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*

$$\max_{Y \perp I_n} \frac{\|\Phi(Y)\|_F}{\|Y\|_F} = \sigma_2(\Phi_{\mathcal{A}}) \le \frac{(1 - \lambda)s(\mathcal{A})}{n}.$$

In [7], [35], the map $\Phi$ 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 $\epsilon$-nearly doubly balanced, we will show in Lemma III.5 that $\sigma_1(\Phi_{\mathcal{A}}) \le (1 + \epsilon)s(\mathcal{A})/\sqrt{mn}$ and $I_n$ is an approximate optimizer. Therefore, in the case $m = n$, the spectral gap condition in Definition I.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 [58], almost all random non-negative matrices (from reasonable distributions) have a constant spectral gap, i.e. $\lambda$ is a constant. For random operators, Hastings [35] proved that the operator $\mathcal{A}$ has an almost Ramanujan spectral gap with $\lambda = 1 - 2\sqrt{k-1}/k$ 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 [24]. 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], [35] for some applications of quantum expanders.

*3) Choi Matrix and Useful Facts:* There is another matrix representation that is useful in studying completely positive linear maps.

**Definition II.8** (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 $\mathrm{tr}_n$ and $\mathrm{tr}_m$ 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. All but (4) are relatively straightforward.

**Fact II.9.** *In the following,* $\Phi_{\mathcal{A}}$ *is the completely positive map with Kraus operators* $\mathcal{A} = (A_1, \dots, A_k)$ *where each* $A_i \in \mathbb{R}^{m \times n}$.

1) *For any matrices* $A, X \in \mathbb{R}^{m \times m}$ *and* $B, Y \in \mathbb{R}^{n \times n}$,

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

$$\langle A \otimes B, X \otimes Y \rangle = \langle A, X \rangle \langle B, Y \rangle.$$

2) $\Phi_{\mathcal{A}}(Y) \succeq 0$ *for any* $Y \succeq 0$.
3) *For any* $X \in \mathbb{R}^{m \times m}$ *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* $L\mathcal{A}R := \{LA_1R, \dots, LA_kR\}$. *Then,*

$$\Phi_{L\mathcal{A}R}(I_n) = L \cdot \Phi_{\mathcal{A}}(RR^*) \cdot L^*$$

$$\Phi_{L\mathcal{A}R}^*(I_m) = R^* \cdot \Phi_{\mathcal{A}}^*(L^*L) \cdot R.$$

*B. Continuous Operator Scaling*

The continuous operator scaling algorithm was studied in [45]. 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.

*1) Operator Scaling:*

**Definition II.10** (Operator). *An operator* $\mathcal{A}$ *is defined by a tuple of* $m \times n$ *matrices* $\mathcal{A} = (A_1, \dots, A_k)$ *where* $A_i \in \mathbb{R}^{m \times n}$ *for* $1 \le i \le k$.

**Definition II.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 II.12** ($\epsilon$-nearly Doubly Balanced Operator). *An operator* $\mathcal{A}$ *is called* $\epsilon$-*nearly doubly balanced if*

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

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

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

**Definition II.13** ($\ell_2$-error). *Given an operator* $\mathcal{A}$, *define*

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

**Definition II.14** (Error Matrices). *We define the error matrices as*

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

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

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

*where the last equality is by Definition II.11. Also, we write*

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

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

**Lemma II.15** (Lemma 3.6.1 in [45]). *For an* $\epsilon$-*nearly doubly balanced operator* $\mathcal{A}$,

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

*2) Dynamical System:*

**Definition II.16** (Dynamical System). *The following dynamical system describes how $\mathcal{A}$ changes over time in the continuous operator scaling algorithm:*

$$\frac{d}{dt}A_i := EA_i + A_iF \quad \text{for } 1 \le i \le k.$$

It can be shown (full version) that the dynamical system is equivalent to the gradient flow with potential function $\Delta(\mathcal{A})$.

It is shown in [45] 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 II.17** (Lemma 3.4.2 in [45]). *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 [45]. It can also be seen as a consequence that the dynamical system is the gradient flow on $\Delta$.

**Lemma II.18** (Lemma 3.4.3 in [45]). *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 [45] for the smoothed analysis when the dynamical system has linear convergence.

**Lemma II.19** (Proposition 4.3.1 in [45]). *Suppose there exists $\mu > 0$ such that for all $0 \le t \le T$,*

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

*Then*

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

*3) Operator Capacity:*

**Definition II.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 [45] that the convergence rate of $\Delta$ can be used to derive a lower bound on operator capacity.

**Proposition II.21** (Proposition 4.3.1 in [45]). *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*

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

## III. Spectral Analysis of Operator Scaling

We present the main technical results in this section.

### A. Overview

The main goal is to show that the dynamical system in Definition II.16 has linear convergence. Let $\mathcal{A}$ be an $\epsilon$-nearly doubly balanced operator with $\lambda$-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 $\Delta$.

**Lemma III.1.** *The change of $\Delta$ is*

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

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 non-negative, and we call the term $2\langle Q, E \otimes F\rangle$ the cross term. The proof outline is the following:

1) In Section III-B, 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 III-C, 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 $-\Delta' \ge \lambda s\Delta$ when the spectral gap condition holds. To prove the linear convergence for all time $t \ge 0$, 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 III-E, and use it to conclude that the spectral gap condition and the linear convergence hold throughout in Section III-F.

In Section III-G and Section III-H, we use the results to prove Theorem I.7 and Theorem I.8 about condition number and operator capacity respectively.

Finally, in Section III-I, we explain how to discretize the gradient flow to obtain a discrete algorithm with linear convergence under the spectral assumption.

### B. 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 $t \ge 0$ in Proposition III.2, which will also be useful in bounding the condition number of the scaling solution in Section III-E. Then we will use this proposition to lower bound the quadratic terms.

**Proposition III.2.** *If $\mathcal{A}^{(0)}$ is $\epsilon$-nearly doubly balanced, then for any $t \geq 0$,*

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

*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 II.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 [49]. $\square$

We have the following corollary by rewriting the conclusions of Proposition III.2 using the definitions that $E^{(t)} = sI_m - m\Phi^{(t)}(I_n)$ and $F^{(t)} = sI_n - n\Phi^{(t)*}(I_m)$.

**Proposition III.3.** *If $\mathcal{A}^{(0)}$ is $\epsilon$-nearly doubly balanced, then for any $t \geq 0$,*

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

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

We can use Proposition III.3 to lower bound the quadratic terms in Lemma III.1.

**Lemma III.4.** *If $\mathcal{A}^{(0)}$ is $\epsilon$-nearly doubly balanced, then for any $t \geq 0$, the quadratic terms are at least*

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

*C. Upper Bounding the Cross Term*

We will first bound the largest singular value of the matrix $M_{\mathcal{A}}$ for any $\epsilon$-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 III.1.

Given a non-negative 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 [38]). The proof of this bound is generalized to prove the following.

John Watrous provided a different proof of Lemma III.5 by generalizing the proof of Theorem 4.27 in his book [61]. We include his proof in the Appendix of the full version.

**Lemma III.5.** *If $\mathcal{A}$ is an $\epsilon$-nearly doubly balanced operator, then the largest singular value of its matrix representation $M_{\mathcal{A}}$ in Definition I.4 is*

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

Lemma III.5 implies that $\mathrm{vec}(I_n)$ is an "approximate" first singular vector of $M_{\mathcal{A}}$. By the spectral gap condition in Definition I.4, it will follow that any vector perpendicular to $\mathrm{vec}(I_n)$ has a "small" quadratic form of $M_{\mathcal{A}}$, and this can be used to bound the cross term in Lemma III.1. The following

lemma summarizes the spectral argument, which will be used to bound the cross term in the next lemma.

**Lemma III.6.** *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*

$$|x^* A y| \leq 1 + \delta_1 - \delta_2.$$

We use Lemma III.6 to bound the cross term in Lemma III.1.

**Lemma III.7.** *If $\mathcal{A}$ satisfies the spectral gap condition in Definition I.4 with the additional assumption that $\sigma_1(M_{\mathcal{A}}) \leq (1+\delta)s(\mathcal{A})/\sqrt{mn}$ for $\delta \leq 1$, then*

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

*Proof Sketch.* Note that the cross term

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

where the first equality is by Fact II.9(3) and the second equality is by the definition of matrix representation in Definition II.4.

To prove the lemma, we apply Lemma III.6 with $\frac{\sqrt{mn}}{s}M_{\mathcal{A}}$,

$$p := \frac{1}{\sqrt{m}}\mathrm{vec}(I_m) \in \mathbb{R}^{m^2}, \quad q := \frac{1}{\sqrt{n}}\mathrm{vec}(I_n) \in \mathbb{R}^{n^2},$$

and

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

*D. Lower Bounding the Convergence Rate*

Putting the bounds in Lemma III.4 and Lemma III.7 into Lemma III.1, we obtain the following lower bound on the convergence rate of $\Delta$ at any time $t$.

**Proposition III.8.** *If $\mathcal{A}^{(0)}$ is $\epsilon$-nearly doubly balanced and the matrix representation $M_{\mathcal{A}^{(t)}}$ of $\mathcal{A}^{(t)}$ satisfies*

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

*then*

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

Note that Proposition III.8 implies that the dynamical system has linear convergence at time $t = 0$. To see this, note that $\delta^{(0)} \leq \epsilon$ by Lemma III.5, and $\lambda^{(0)} = \lambda$ from Definition I.4, and therefore

$$-\frac{d}{dt}\Delta^{(0)} \geq 4(\lambda - 4\epsilon)s^{(0)}\Delta^{(0)}.$$

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 prove that the quantities

in Proposition III.8 do not change much, 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 bound the condition number of the scaling solutions in the dynamical system in Section III-E, and then use these bounds to argue about the change of the singular values and establish Theorem I.5 in Section III-F.

### E. Scaling Solutions and Condition Numbers

Using the theory of product integration as presented in Slavik's book [55], we can find a closed form solution to the scaling given by the dynamical system and bound the condition number of the scaling solutions.

**Corollary III.9.** *For any* $T \geq 0$,

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

This corollary will be used to bound the condition number of $L^{(T)}$ in Lemma III.10, which will then be used to bound the condition number of $R^{(T)}$ in Lemma III.12.

To bound the condition number, we use Corollary III.9 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 III.2 to argue that $\| E^{(t)} \|_{\mathrm{op}} \approx \| E^{(0)} \|_{\mathrm{op}}$. In the second phase, we use that $\Delta^{(t)}$ is converging linearly to argue that $\| E^{(t)} \|_{\mathrm{op}} \leq \| E^{(t)} \|_F \leq \sqrt{m \Delta^{(t)}}$ is converging linearly. In the following lemma, we should think of $g$ as the spectral gap parameter in Definition I.4.

**Lemma III.10.** *Suppose there exists* $g > 0$ *such that for all* $0 \leq t \leq T$, *it holds that*

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

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

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

**Remark III.11.** *We have some examples indicating that the* $\log m$ *term in the condition number is necessary, but we do not have a formal proof for this lower bound at this time.*

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

**Lemma III.12.** *Suppose there exists* $g > 0$ *such that for all* $0 \leq t \leq T$, *it holds that*

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

*If* $\mathcal{A}^{(0)}$ *is* $\epsilon$*-nearly doubly balanced for* $\epsilon \leq g$, *and* $\epsilon, \ell \leq \frac{1}{2}$:

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

### F. Invariance of Linear Convergence

We will first use Lemma III.10 and Lemma III.12 to bound the change of the singular values of $M_{\mathcal{A}^{(t)}}$. Then, we will combine the previous results to prove Theorem I.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 III.13** (Theorem 3.3.16 in [37]). *Let* $A$ *and* $B$ *be two* $m \times n$ *matrices. For any* $1 \leq k \leq m$,

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

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

**Lemma III.14.** *For any* $t \geq 0$, *suppose* $\left\| L^{(t)} - I_m \right\|_{\mathrm{op}} \leq \zeta$ *and* $\left\| R^{(t)} - I_n \right\|_{\mathrm{op}} \leq \zeta$ *for some* $\zeta \leq 1$, *then*

$$|\sigma_k(M_{\mathcal{A}^{(t)}}) - \sigma_k(M_{\mathcal{A}^{(0)}})| \leq O(\zeta) \cdot \| M_{\mathcal{A}^{(0)}} \|_{\mathrm{op}}.$$

We are ready to put together the results to prove the following theorem which implies Theorem I.5.

**Theorem III.15.** *If* $\mathcal{A}^{(0)}$ *is* $\epsilon$*-nearly doubly balanced and* $\mathcal{A}^{(0)}$ *satisfies the* $\lambda$*-spectral gap condition in Definition I.4 with* $\lambda^2 \geq C \epsilon \ln m$ *for a sufficiently large constant* $C$, *then for all* $t \geq 0$ *it holds that*

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

*Proof.* Recall from Proposition III.8 the definitions of $\delta^{(t)}$ and $\lambda^{(t)}$, and $\delta^{(0)} \leq \epsilon$ by Lemma III.5 and $\lambda^{(0)} = \lambda$ from Definition I.4. Let $T$ be the supremum such that $s^{(t)} \geq (1 - \epsilon) s^{(0)}$ and $\lambda^{(t)} - 3\delta^{(t)} \geq \frac{1}{2} (\lambda^{(0)} - 3\delta^{(0)})$. 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 III.8,

$$\begin{aligned} -\frac{d}{dt} \Delta^{(t)} &\geq 4 \left( (1 + \lambda^{(t)} - 3\delta^{(t)}) s^{(t)} - (1 + \epsilon) s^{(0)} \right) \Delta^{(t)} \\ &= \left( 2(1 - \epsilon)(\lambda^{(0)} - 3\delta^{(0)}) - 8\epsilon \right) s^{(0)} \Delta^{(t)}, \end{aligned}$$

where we used that $s^{(t)} \geq (1 - \epsilon) s^{(0)}$ and $\lambda^{(t)} - 3\delta^{(t)} \geq \frac{1}{2} (\lambda^{(0)} - 3\delta^{(0)})$ for $0 \leq t \leq T$. Note that our assumption implies that $\lambda^{(0)} = \lambda \geq C \epsilon$ for a sufficiently large constant $C$ as $\lambda \leq 1$. Since $\delta^{(0)} \leq \epsilon$ from Lemma III.5, it follows that for any $0 \leq t \leq T$,

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

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 II.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 II.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,

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

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 III.10 with $g = \lambda$, our assumption implies that

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

where the implication is by the inequality $e^x - 1 \leq O(x)$ for $x$ close to zero. Then, by Lemma III.12, we also have $\left\|R^{(T)} - I\right\|_{\mathrm{op}} \leq O(\lambda/C)$. Putting these bounds into $\zeta$ of Lemma III.14, we obtain

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

Combining the upper bound and lower bound and using $\delta_1^{(0)} \leq \epsilon$ from Lemma III.5, 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 III.3,

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

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 $3\epsilon$-nearly doubly balanced. By Lemma III.5, this implies that $\delta^{(T)} \leq 3\epsilon$. Therefore,

$$\lambda^{(T)} - 3\delta^{(T)} \geq \lambda - O\left(\frac{\lambda}{C}\right) \gg \frac{1}{2}(\lambda - 3\delta^{(0)}),$$

where the second 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 $\eta > 0$, which contradicts that $T$ is the supremum that both conditions are satisfied. Therefore, $T$ is unbounded and the linear convergence of $\Delta$ is maintained throughout the execution of the dynamical system. $\qquad\square$

## G. Condition Number

With the invariance of the linear convergence, we can apply Lemma III.10 and Lemma III.12 to bound the condition number of the scaling solutions and prove Theorem I.7

**Theorem III.16.** *If $\mathcal{A}^{(0)}$ is $\epsilon$-nearly doubly balanced and $\mathcal{A}^{(0)}$ satisfies the $\lambda$-spectral gap condition in Definition I.4 with $\lambda^2 \geq C\epsilon \log m$ for a sufficiently large constant $C$, then for any $t \geq 0$,*

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

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

## H. Operator Capacity

Theorem I.8 follows easily from Theorem III.15.

**Theorem III.17.** *If $\mathcal{A}^{(0)}$ is $\epsilon$-nearly doubly balanced and $\mathcal{A}^{(0)}$ satisfies the $\lambda$-spectral gap condition in Definition I.4 with $\lambda^2 \geq C\epsilon \ln m$ for a sufficiently large constant $C$, then*

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

*Proof.* By Theorem III.15, $\Delta^{(t)}$ is linearly converging for all time $t$ with rate $\lambda s^{(0)}$. Apply Proposition II.21 with $\mu = \lambda s^{(0)}$,

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

where the second inequality is by Lemma II.15. $\qquad\square$

## I. 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 closely the continuous case, 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 II.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 $\alpha$, 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 $\alpha$ 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 $\alpha$ 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 $\Delta$ 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

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

that $\Delta$ is decreasing geometrically with rate $\lambda s$, 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 $\lambda s$ is maintained throughout the algorithm. Again, we go through the condition number of the error matrices, and use the arguments in Lemma III.14 to show that the change of the singular value is

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

and it follows that the $\lambda$-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 $\eta$-approximate solution in time $T$, the discrete algorithm will converge to an $\eta$-approximate solution in $T \cdot \Theta((m+n)^2)$ number of iterations, and the dependency on $\eta$ is $\log(1/\eta)$ by Theorem I.5.

**Remark III.18.** *The step size $\alpha = O((m+n)^{-2})$ is chosen for the same analysis as in the continuous to hold. It is an interesting open question whether the analysis can be extended to constant step size, in particular whether Sinkhorn's alternating algorithm has the same convergence rate as the gradient flow.*

## IV. APPLICATIONS

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

### A. Matrix Scaling

Given a non-negative 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 $i$-th row sum of $B$, and $c_j(B) := \sum_{i=1}^m B_{i,j}$ be the $j$-th column sum of $B$. A non-negative matrix is called $\epsilon$-nearly doubly balanced if for every $1 \leq i \leq m$ and for every $1 \leq j \leq n$,

$$(1-\epsilon)\frac{s(B)}{m} \leq r_i(B) \leq (1+\epsilon)\frac{s(B)}{m}$$

$$(1-\epsilon)\frac{s(B)}{n} \leq c_j(B) \leq (1+\epsilon)\frac{s(B)}{n},$$

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 IV.1** (Matrix Scaling Problem)**.** *We are given a non-negative 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:** It can be shown that the matrix scaling problem can be reduced to the operator scaling problem Then, we will see that the spectral condition has a simple form in Section IV-A2 , and there is a natural combinatorial condition that implies the spectral condition in Section IV-A3. We then argue that many random matrices will satisfy our condition in Section IV-A4. Finally, we see the implications of our results in several applications of matrix scaling, including bipartite matching in Section IV-A5, permanent lower bound in Section IV-A6, and optimal transportation in Section IV-A7.

*1) Reduction to Operator Scaling:* The matrix scaling problem is a special case of the operator scaling problem.

**Lemma IV.2.** *Given a non-negative 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 $1 \leq i \leq m$ and $1 \leq j \leq n$ 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 $\epsilon$-nearly doubly balanced if and only if $\mathcal{A}$ is $\epsilon$-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 $\mathcal{A}$.*

*2) Spectral Condition:* The spectral condition for operator scaling has a simple form for matrix scaling.

**Lemma IV.3.** *Using the reduction from Lemma IV.2, the spectral condition for operator scaling in Definition I.4 becomes*

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

*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 IV.4** (Weighted Bipartite Graph, Conductance)**.** *Given a non-negative 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*

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

*where $\text{vol}(S) := \sum_{i \in S} \sum_{j \in V} w_{ij}$.*

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 IV.5.** *If $B \in \mathbb{R}^{m \times n}$ is $\epsilon$-nearly doubly balanced for $\epsilon \leq 1/2$, then*

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

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

*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 standard probabilistic method. Also, $B$ is $\epsilon$-nearly doubly balanced for small $\epsilon$ by standard concentration inequality (e.g. $\epsilon = O(\sqrt{\log m/(pm)})$ in the above example). So, by Lemma IV.5, the $\lambda$ in Lemma IV.3 is $\Omega(1)$, which implies that the assumption $\lambda^2 \geq C\epsilon \ln m$ in Theorem I.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 $\eta$-nearly doubly balanced solution in time $t = O(\log(m/\eta))$.
2) The condition number of the scaling solution is $O(1)$ from Theorem I.7.
3) The capacity of the matrix is close to $s$ from Theorem I.8.

Indeed, the assumption $\lambda^2 \geq C\epsilon \ln m$ in Theorem I.5 should hold for a large class of random non-negative matrices where each entry is an independent random variable with reasonable distribution such as the chi-squared distribution [58], 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 IV.5, or to bound the second largest singular value directly using the trace method as in Section V.

*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 [47], 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 [23] (see also [48]).

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 IV.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 $\epsilon$. If $\phi(G)^4 \geq C\epsilon \ln |X|$ for some sufficiently large constant $C$, then the gradient flow converges to an $\eta$-nearly doubly balanced scaling (i.e. $\eta$-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 [2], [13] are near linear time algorithms for the instances in Corollary IV.6. This is because the condition number $\kappa$ of the scaling solution for those instances is a constant by Theorem I.7 and the algorithms in [2], [13] 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 IV.6.

*6) Permanent Lower Bound:* Given a matrix $A \in \mathbb{R}^{n \times n}$, the permanent is defined as

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

where $S_n$ is the set of all permutations of $n$ elements. Linial, Samorodnitsky, and Wigderson [47] used the matrix scaling algorithm to design a deterministic $e^n$-approximation algorithm for computing the permanent of a non-negative $n \times n$ 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 use 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 IV.3 (e.g. random matrices in Section IV-A4), we can use the capacity lower bound in Theorem I.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 IV.7** (Matrix Capacity). *Given a matrix $B \in \mathbb{R}^{m \times n}$, define*

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

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

**Lemma IV.8.** *Following the reduction in Lemma IV.2 from matrix scaling of $B$ to operator scaling of $\mathcal{A}$, we have that $\operatorname{cap}(B)$ in Definition IV.7 is equivalent to $\operatorname{cap}(\mathcal{A})$ in Definition II.20.*

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

**Corollary IV.9.** *If a non-negative matrix $B \in \mathbb{R}^{n \times n}$ is $\epsilon$-nearly doubly balanced with $s(B) = n$ and it satisfies the $\lambda$-spectral gap condition in Definition ?? with $\lambda^2 \geq C\epsilon \log n$ for some sufficiently large constant $C$, then*

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

*Proof.* Let $B \in R^{n \times n}$ be the input non-negative 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 [29], [31] defined the (unnormalized) capacity of $B \in \mathbb{R}^{n \times n}$ as

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

Note that $\overline{\operatorname{cap}}(LBR) = \det(L) \cdot \det(R) \cdot \overline{\operatorname{cap}}(B)$ and also $\operatorname{per}(LBR) = \det(L) \cdot \det(R) \cdot \operatorname{per}(B)$. Using the fact that $\overline{\operatorname{cap}}(A) = 1$ for a doubly stochastic matrix $A$ [20], [29],

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

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 $\epsilon$-nearly doubly balanced with $s(B) = n$ and $B$ satisfies the spectral gap condition in Definition **??**, then Theorem I.8 and Lemma IV.8 imply that

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

where $\mathcal{A}$ is the operator in the reduction from Lemma IV.2. Therefore, we conclude that

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

$\square$

**Example IV.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 IV.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.*

*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 [14] 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 $\beta$ is a parameter in the regularizer). Cuturi showed that the 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 [14]. 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 [52] and the references therein. Theorem I.5 provides a condition to establish the linear convergence observed, which is satisfied in many random matrices as discussed in Section IV-A4.

Also, it is of interest to bound the Sinkhorn distance, which is shown in [14], [52] 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 $\beta$ is the regularizer parameter. This result states that the distance is small if the condition number of the scaling solution is small. Theorem I.7 provides a condition to bound the condition number to bound the Sinkhorn distance.

*B. Frame Scaling*

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

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

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

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

**Definition IV.11** (Frame Scaling Problem)**.** *Given a frame $U = (u_1, \ldots, u_n)$ 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:** It can be shown that the frame scaling problem can be reduced to the operator scaling problem in Section IV-B1. Then, we will see that the spectral condition has a nice form in Section IV-B2, and explain that random frames will satisfy our condition in Section IV-B3. Finally, we show a significant implication of our results to the Paulsen problem in Section IV-B4 and a construction of doubly stochastic frame with small inner products in Section IV-B5.

*1) Reduction to Operator Scaling:* The frame scaling problem is a special case of the operator scaling problem.

**Lemma IV.12.** *Given a frame $U = (u_1, \ldots, u_n)$ where each $u_i \in \mathbb{R}^d$, let $\mathcal{A} = (A_1, \ldots, A_n)$ where each $A_i \in \mathbb{R}^{d \times n}$ for $1 \leq i \leq n$ is the matrix with the $i$-th column being $u_i$ and all other columns equal to zero. Then, $U$ is $\epsilon$-nearly doubly stochastic if and only if $\mathcal{A}$ is $\epsilon$-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 $\mathcal{A}$.*

*2) Spectral Condition:* The spectral condition for operator scaling is related to the following Hermitian matrix.

**Definition IV.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 $i$-th 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 I.4 translates to the following spectral condition for the squared Gram matrix in the frame scaling case.

**Lemma IV.14.** *Using the reduction from Lemma IV.12, the spectral condition becomes*

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

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

*3) Random Frames:* In Section V, we will prove that if we generate $\Omega(d^{4/3})$ random unit vectors, then the resulting frame is $\epsilon$-nearly doubly balanced for $\epsilon = O(1/\operatorname{poly}(d))$ and the $\lambda$ in Lemma IV.14 satisfies $\lambda = \Omega(1)$ with high probability. Hence, a random frame generated in this way will satisfy the condition $\lambda^2 \ge C\epsilon \ln d$ 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.

*4) The Paulsen Problem in Random Frames:* Given an $\epsilon$-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 $\epsilon$-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 $\epsilon$ but independent of $n$. Recently, this question was answered affirmatively in [45], showing that for any $\epsilon$-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 [32] 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 $\epsilon$-nearly doubly balanced frames for small $\epsilon$. 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 $\epsilon$-nearly doubly balanced for small $\epsilon$ by matrix concentration inequalities (see Section V-A). 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 $\epsilon$-nearly doubly balanced, there is a doubly balanced frame $V$ with $\operatorname{dist}^2(U, V) = O(d\epsilon^2)$ 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 $\epsilon$-nearly doubly balanced for $\epsilon \le O(1/\operatorname{poly}(d))$ and also satisfies the spectral gap condition in Lemma IV.14 with $\lambda = \Omega(1)$. Therefore, the assumption in Theorem I.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 V. We will prove the second step here. The following lemma states the result in [45] that we will use.

**Lemma IV.15** ( [45]). *The dynamical system in Definition II.16 will move the input operator $\mathcal{A}^{(0)}$ to a doubly balanced operator $\mathcal{A}^{(\infty)}$. For any time $T \ge 0$,*

$$\operatorname{dist}^2(\mathcal{A}^{(T)}, \mathcal{A}^{(0)}) \le \left( \int_0^T \sqrt{\sum_{i=1}^k \left\| \frac{d}{dt} A_i^{(t)} \right\|_F^2} \, dt \right)^2$$

$$= \frac{1}{4} \left( \int_0^T \sqrt{-\frac{d}{dt}\Delta^{(t)}} \, dt \right)^2$$

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

**Lemma IV.16.** *Given an operator $\mathcal{A} = (A_1, \ldots, A_k)$ where $A_i \in \mathbb{R}^{m \times n}$ with $m \le n$ for $1 \le i \le k$, if $\mathcal{A}$ is $\epsilon$-nearly doubly balanced and $\mathcal{A}$ satisfies the $\lambda$-spectral gap condition in Definition I.4 with $\lambda^2 \ge C\epsilon \ln m$ for a sufficiently large constant $C$, then*

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

Combining the two steps gives the following theorem.

**Theorem IV.17.** *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 $\operatorname{dist}^2(U, V) \le O(d\epsilon^2)$ if $U$ is $\epsilon$-nearly doubly balanced.*

*Proof.* By Theorem V.1, the random frame $U$ satisfies the spectral gap condition in Lemma IV.14 with constant $\lambda$ and $\epsilon \ll 1/\ln d$ with probability at least $0.99$. Note that Theorem V.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 $1 \le i \le n$. By the reduction in Lemma IV.12 and the spectral gap condition in Lemma IV.14, this implies that the condition $\lambda^2 \ge C\epsilon \ln d$ for operator scaling is satisfied and also $s(U) = d$. Therefore, by Lemma IV.16, the continuous operator scaling algorithm will move $U$ to a doubly balanced frame $V$ with $\operatorname{dist}^2(U, V) \le O(d\epsilon^2)$. $\qquad\square$

*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 IV.18.** *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 [57]. The known examples are sporadic and based on group/number-theoretic constructions. We consider a related but more relaxed problem.

**Definition IV.19.** *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 [36], [56]. The original motivation of the Paulsen problem was to begin with some $\epsilon$-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 $\epsilon$-nearly doubly balanced for small $\epsilon$ 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 IV.20.** *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*

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

*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 V.2 and Theorem V.1, $U$ is $\epsilon$-nearly doubly balanced for $\epsilon \leq O(\sqrt{d \log d / n})$ and satisfies the $\lambda$-spectral gap condition with $\lambda = \Omega(1)$ with probability at least 0.99. Next, we bound $\theta(U)$ using concentration:

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

By Theorem III.15 and the reduction in Lemma IV.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 $v_i = Lu_iR_{ii}$, then the frame $V = (v_1, \ldots, v_n)$ is doubly balanced. By Theorem III.16, the scaling solutions $L, R$ satisfy

$$\|L - I\|_{\mathrm{op}} \leq \zeta \text{ and } \|R - I\|_{\mathrm{op}} \leq \zeta$$

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 III.14, we have

$$|\langle v_i, v_j \rangle - \langle u_i, u_j \rangle| \leq O(\zeta) \cdot \|u_i\|_2 \|u_j\|_2 = O(\zeta).$$

Therefore, we conclude that

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

$\square$

For examples, 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 $\theta(V) \leq O(\log d / d)$.

### C. Operator Scaling

The operator scaling problem was used to the Brascamp-Lieb constant [21] and to compute the non-commutative rank of a symbolic matrix [20]. It is also used in [1] to solve the orbit intersection problem for the left-right group action.

*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 non-negative 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], [21].

The important point we need is that the optimizers of any non-degenerate 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 [46], 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 [21] from BL-constant to operator capacity and our capacity lower bound in Theorem I.8.

**Reduction:** We describe the reduction in [21] 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 $i$-th 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

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

$$\Phi_{\mathcal{A}}^*(Y) = \bigoplus_{j=1}^{m} \bigoplus_{i=1}^{c_j} \frac{1}{d} B_j Y B_j^*,$$

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 IV.21** ( [21]). *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 [21] 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 IV.21, we would like to derive upper bounds on BL-constants using the capacity lower bound in Theorem I.8, and show that for some random instances the BL-constant is small. To apply Theorem I.8, we translate the definitions of $\epsilon$-nearly doubly balanced operator and the $\lambda$-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 IV.22** (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 $\epsilon$-nearly geometric if and only if the corresponding operator $\mathcal{A}$ is $\epsilon$-nearly doubly balanced.

**Definition IV.23** (Nearly Geometric Datum). *A datum $\mathrm{BL}(\mathbf{B}, \mathbf{p})$ is $\epsilon$-nearly geometric if*

$$(1 - \epsilon) \frac{s}{n} I_n \preceq \sum_{j=1}^{m} p_j B_j^* B_j \preceq (1 + \epsilon) \frac{s}{n} I_n$$

$$(1 - \epsilon) \frac{s}{n} I_{n_j} \preceq B_j B_j^* \preceq (1 + \epsilon) \frac{s}{n} I_{n_j} \text{ for } 1 \leq j \leq m.$$

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

**Definition IV.24** (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 $j$-th 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*

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

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

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

With these definitions, we can state the Brascamp-Lieb constant upper bound that follows from the capacity lower bound in Theorem I.8.

**Corollary IV.25.** *Given a datum $(\mathbf{B}, \mathbf{p})$ with $B_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$ for $1 \leq j \leq n$ and $\sum_{j=1}^{m} p_j n_j = n$, if $(\mathbf{B}, \mathbf{p})$ is $\epsilon$-nearly geometric and satisfies the $\lambda$-spectral gap condition with $\lambda^2 \geq C \epsilon \log n$ for some sufficiently large constant $C$, then*

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

Let's consider a concrete example to demonstrate:

**Example IV.26.** *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,*

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

*which is a form that is also studied in approximation algorithms [50]. Note that this is exactly the capacity of a frame $U = (u_1, \ldots, u_m)$ through the reduction in IV.12. By Theorem V.1, if $m \geq \Omega(d^{4/3})$, then $U$ is $\epsilon$-nearly doubly balanced for $\epsilon \leq O(\sqrt{d \log d/m})$ and satisfies the $\lambda$-spectral gap condition with $\lambda = \Omega(1)$ with high probability. Therefore, we can apply Theorem I.8 to conclude that*

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

*and from Corollary IV.25 the BL-constant for this datum is*

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

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

As another example, Hastings' result [35] implies that a random operator where each $A_i$ is a random unitary has small Brascamp-Lieb constant with high probability.

*2) Rank Non-Decreasing Operator:* In [19], [20], [29], a polynomial time algorithm for computing the non-commutative 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 non-commutative variables $x_1, \ldots, x_k$, the non-commutative 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 [19], [20] for definitions). The algorithm in [19], [20], [29] is based on the following equivalent characterizations.

**Theorem IV.27** ( [19], [20], [29]). *Given $\mathcal{A} = (A_1, \ldots, A_k) \subset \mathbb{R}^{n \times n}$, the following conditions are equivalent.*

1) *Symbolic matrix $Z_\mathcal{A}$ is singular, i.e. ncrank$(Z) < n$.*
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 $1 \le i \le k$.*
3) *$\Phi_\mathcal{A}$ is rank decreasing, i.e. there exists $P \succ 0$ and rank$(\Phi_\mathcal{A}(P)) < $ rank$(P)$.*

The alternating scaling algorithm for operator scaling is used to check whether $\Phi_\mathcal{A}$ is rank non-decreasing. It is shown in [19], [20], [29] that $\Phi_\mathcal{A}$ is rank non-decreasing if and only if $\mathcal{A}$ can be scaled to $\epsilon$-nearly balanced for $\epsilon \le 1/\operatorname{poly}(n)$, and so a polynomial time algorithm for operator scaling can be used to compute the non-commutative 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 IV.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 IV.6. In the operator case, intuitively, the spectral condition is closely related to the notion of quantum expander (Section II-A), and so there should be no Hall-blocker as well. Theorem I.5 implies that it is the case.

**Corollary IV.28.** *For $\mathcal{A}$ satisfying the conditions of Theorem I.5, $\Phi_\mathcal{A}$ is rank-nondecreasing and the corresponding symbolic matrix $Z_\mathcal{A}$ is non-singular over reals.*

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

*3) The Operator Paulsen Problem:* Given an $\epsilon$-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 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 [45], it was proved that $\operatorname{dist}^2(\mathcal{A}, \mathcal{B}) \le O(mns\epsilon)$, 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 $\lambda$, Lemma IV.16 implies a much stronger bound that $\operatorname{dist}^2(\mathcal{A}, \mathcal{B}) \le O(s\epsilon^2)$.

## V. SPECTRAL GAP OF RANDOM FRAMES

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

**Theorem V.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 $\epsilon$-nearly doubly stochastic for $\epsilon \ll 1/\ln d$ and satisfies the spectral gap condition in Definition IV.14 with constant $\lambda$ with probability at least $0.99$.*

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 $1 \le i \le d$, 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 V-A, we will prove that $V$ is $\epsilon$-nearly doubly stochastic with high probability by using a standard matrix concentration bound. Then, in Section **??**, we will prove that the squared Gram matrix $G$ in Definition IV.13 satisfies the spectral gap condition in Definition IV.14 with high probability by using the trace method.

### A. Nearly Doubly Balanced Condition by 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 $\epsilon$-nearly Parseval for $\epsilon \ll 1/\log d$ with high probability when $n = \Omega(d^{4/3})$, i.e.

**Lemma V.2.** *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 $1 - O(1/\operatorname{poly}(d))$.*

We establish this by using the following matrix bound:

**Theorem V.3** (Matrix Bernstein [60]). *Let $X_1, \ldots, X_n$ be independent random matrices in $\mathbb{R}^{d \times d}$. Assume that*

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

*and*

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

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

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

*Proof Sketch.* We apply the concentration inequality on random matrix $X_i := v_i v_i^* - \frac{1}{d}I_d$ for $1 \le i \le n$. $\qquad\square$

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 V.2 gives the following bound for the latter condition.

**Corollary V.4.** *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 $1 - O(1/\operatorname{poly}(d))$.*

**Theorem V.5.** *Using the trace method, for any constant $\lambda$, by generating $n \gg d^{4/3}$ random unit vectors, the probability that $\lambda_2(G) > (1-\lambda)^2 n/d$ can be made arbitrarily small as the dominating term is $d^4/n^3$.*

**Remark V.6.** *We believe that the trace method can be improved to prove the same conclusion with only $O(d \operatorname{polylog} d)$ random unit vectors.*

### Acknowledgment

### References

[1] Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, 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), 172–181, 2018.

[2] Z. Allen-Zhu, Y. Li, R. Oliveira, 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*. Bulletin of the London Mathematical Society, Vol 14, Issue 1, Jan 1982.

[4] K. Ball. *Volumes of sections of cubes and related problems*. Geometric Aspects of Functional Analysis, 251–260, 1989.

[5] F. Barthe. *On a reverse form of the Brascamp-Lieb inequality*. Inventiones mathematicae 134(2), 335–361, 1998.

[6] A. Barvinok, A. Samorodnitsky. *Computing the partition function for perfect matchings in a hypergraph*. Combinatorics, Probability, and Computing, 20(6), 2011.

[7] A. Ben-Aroya, O. Schwartz, A. Ta-Shma. *Quantum expanders: motivation and construction*. Theory of Computing 6, 47–79, 2010.

[8] J. Bennett, A. Carbery, M. Christ, T. Tao. *The Brascamp-Lieb inequalities: finiteness, structure, and extremals*. GAFA Geom. funct. anal. (2008) 17: 1343.

[9] R. Bhatia. *Positive definite matrices*. Princeton University Press, 2007.

[10] H. Brascamp, E. Lieb. *Best constants in Young's inequality, its converse and its generalization to more than three functions*. Advances in Mathematics 20, 151–172, 1976.

[11] P. Buergisser, C. Franks, A. Garg, R. Oliveira, M. Walter, 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] M. Choi. *Completely positive linear maps on complex matrices*. Linear Algebra and its Applications, 285–290, 1975.

[13] M.B. Cohen, A. Madry, D. Tsipras, 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.

[14] M. Cuturi. *Sinkhorn distances: lightspeed computation of optimal transport*. In Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS), 2292–2300, 2013.

[15] G.P. Egorychev. *The solution of van der Waerdens problem for permanents*. Advances in Mathematics 42, 299–305, 1981.

[16] D.I. Falikman. *Proof of the van der Waerdens conjecture on the permanent of a doubly stochastic matrix*. Mat. Zametki 29(6), 931–938, 1981.

[17] J. Forster. *A linear lower bound on the unbounded error probabilistic communication complexity*. J Comput. Syst. Sci. (65), 612–625, 2002.

[18] J. Franklin, J. Lorenz. *On the scaling of multidimensional matrices*. Linear Algebra and Applications 114-115, 717–735, 1989.

[19] A. Garg, L. Gurvits, R. Oliveira, 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), 109–117, 2016.

[20] A. Garg, L. Gurvits, R. Oliveira, A. Wigderson. *Operator scaling: theory and applications*. In arXiv 1511.03730, 2017.

[21] A. Garg, L. Gurvits, R. Oliveira, 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), 397–409, 2017.

[22] A. Garg, R. Oliveira. *Recent progress on scaling algorithms and applications*. Bulletin of the EATCS 125, 2018.

[23] A. Goel, M. Kapralov, 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), 39-46, 2010.

[24] C. González-Guillén, M. Junge, I. Nechita. *Spectral gap of random quantum channels*. Workshop on "Probabilistic techniques and quantum information theory", 2017.

[25] V. Guillemin, S. Sternberg. *Convexity Properties of the Moment Mapping*. Inventiones Mathematicae 67(3), 1982.

[26] V. Guillemin, S. Sternberg. *Convexity properties of the moment mapping, II*. Invent. Math. 77, 533–546, 1984.

[27] V. Guillemin, R. Sjamaar. *Convexity Properties of Hamiltonian Group Actions*. AMS, 2005.

[28] 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), 417–426, 2006.

[29] L. Gurvits. *Classical complexity and quantum entanglement*. Journal of Computer and System Sciences 69(3), 448–484, 2004.

[30] L. Gurvits, 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), 48–57, 2000.

[31] L. Gurvits, P.N. Yianilos. *The deflation-inflation method for certain semidefinite programming and maximum determinant completion problems*. Technical Report, NECI, 1998.

[32] L. Hamilton, A. Moitra. *The Paulsen problem made simple*. In arXiv:1809.04726, 2018.

[33] M. Hardt, A. Moitra. *Algorithms and hardness for robust subspace recovery*. Proceedings of the 26th Annual Conference on Learning Theory (COLT) 30, 354–375, 2013.

[34] N. Harms. *Testing halfspaces over rotation-invariant distributions*. In arXiv:1811.00139, 2018.

[35] M.B. Hastings. *Random unitaries give quantum expanders*. Physical Review A 76, 032315, 2007.

[36] R.B. Holmes, V.I. Paulsen. *Optimal frames for erasures*. Linear Algebra and its Applications 377, 31–51, 2004.

[37] R.A. Horn, C.R. Johnson. *Matrix analysis*. Cambridge University Press, 2nd edition, 2012.

[38] R.A. Horn, C.R. Johnson. *Topics in matrix analysis*. Cambridge University Press, 1991.

[39] M. Idel. *A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps*. In arXiv:1609.06349, 2016.

[40] B. Kalantari, L. Khachiyan. *On the complexity of nonnegative-matrix scaling*. SIAM Journal on Matrix Analysis and Applications 18(2), 450–463, 1997.

[41] G. Kempf and L. Ness. *The length of vectors in representation spaces*. Algebraic geometry, page 233. Springer, 1979.

[42] F. C. Kirwan. *Cohomology of quotients in symplectic and algebraic geometry*. Mathematical Notes, vol. 31, Princeton Univ. Press, Princeton, 1984

[43] F. Kirwan. *Convexity properties of the moment mapping, III*. Invent. Math. 77, 547–552, 1984.

[44] M. Krivelevich, B. Sudakov. *Pseudo-random graphs*. (2006).

[45] T.C. Kwok, L.C. Lau, Y.T. Lee, A. Ramachandran. *The Paulsen problem, continuous operator scaling, and smoothed analysis*. CoRR abs/1710.02587, 2017. (Preliminary version in STOC 2018, 182–189.)

[46] E. Lieb. *Gaussian kernels have only Gaussian maximizers*. Inventiones Mathematicae 102, 179–208, 1990.

[47] N. Linial, A. Samorodnitsky, A. Wigderson. *A deterministic strongly polynomial algorithm for matrix scaling and approximate permanent*. STOC, 644–652, 1998.

[48] 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), 253–262, 2013.

[49] P. Milgrom, I. Segal. *Envelope theorems for arbitrary choice sets*. Econometrica, 70(2), 583–601, 2010.

[50] A. Nikolov, M. Singh. *Maximizing determinants under partition constraints*. In Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC), 192–201, 2016.

[51] A. Nemirovski, U. Rothblum. *On complexity of matrix scaling*. Linear Algebra and its Applications 302, 435–460, 1999.

[52] G. Peyré, M. Cuturi. *Computational Optimal Transport*. Draft book, ArXiv:1803.00567, 2018.

[53] M. Rudelson, A. Samorodnitsky, O. Zeitouni. *Hafnians, perfect matchings and Gaussian matrices*. Annals of Probability 44(4), 2858–2888, 2016.

[54] R. Sinkhorn. *A relationship between arbitrary positive matrices and doubly stochastic matrices*. The Annuals of Mathematical Statistics 35, 876–879, 1964.

[55] A. Slavík. *Product integration, its History and Applications*. Matfyzpress, 2007.

[56] T. Strohmer, R.W. Heath Jr.. *Grassmannian frames with applications to coding and communications*. Applied and Computational Harmonic Analysis 14(3), 257–275, 2003.

[57] M.A. Sustik, J.A. Tropp, I.S. Dhillon, R.W. Heath Jr. *On the existence of equiangular tight frames*. Linear Algebra and its Applications 426, 619–635, 2007.

[58] T. Tao. *Topics in random matrix theory*. Graduate Studies in Mathematics 132, 2012.

[59] J.A. Tropp, I.S. Dhillon, R.W. Heath Jr., T. Strohmer. *Designing structured tight frames via an alternating projection method*. IEEE Transactions on Information Theory 51(1), 188–209, 2005.

[60] J.A. Tropp. *An introduction to matrix concentration inequalities*. Foundations and Trends in Machine Learning 8(1-2), 1–230, 2015.

[61] J. Watrous. *The theory of quantum information*. Cambridge University Press, 2018.