The Paulsen Problem, Continuous Operator Scaling, and Smoothed Analysis[*]

Tsz Chiu Kwok
University of Waterloo
Canada

Lap Chi Lau
University of Waterloo
Canada

Yin Tat Lee[†]
University of Washington
USA

Akshay Ramachandran
University of Waterloo
Canada

ABSTRACT

The Paulsen problem is a basic open problem in operator theory: Given vectors $u_1, \ldots, u_n \in \mathbb{R}^d$ that are ϵ-nearly satisfying the Parseval's condition and the equal norm condition, is it close to a set of vectors $v_1, \ldots, v_n \in \mathbb{R}^d$ that exactly satisfy the Parseval's condition and the equal norm condition? Given $u_1, \ldots, u_n$, the squared distance (to the set of exact solutions) is defined as $\inf_v \sum_{i=1}^n \|u_i - v_i\|_2^2$ where the infimum is over the set of exact solutions. Previous results show that the squared distance of any ϵ-nearly solution is at most $O(\text{poly}(d, n, \epsilon))$ and there are ϵ-nearly solutions with squared distance at least $\Omega(d\epsilon)$. The fundamental open question is whether the squared distance can be independent of the number of vectors n.

We answer this question affirmatively by proving that the squared distance of any ϵ-nearly solution is $O(d^{13/2}\epsilon)$. Our approach is based on a continuous version of the operator scaling algorithm and consists of two parts. First, we define a dynamical system based on operator scaling and use it to prove that the squared distance of any ϵ-nearly solution is $O(d^2 n\epsilon)$. Then, we show that by randomly perturbing the input vectors, the dynamical system will converge faster and the squared distance of an ϵ-nearly solution is $O(d^{5/2}\epsilon)$ when n is large enough and ϵ is small enough. To analyze the convergence of the dynamical system, we develop some new techniques in lower bounding the operator capacity, a concept introduced by Gurvits to analyzing the operator scaling algorithm.

CCS CONCEPTS

• **Mathematics of computing** → **Mathematical analysis**; • **Theory of computation** → **Design and analysis of algorithms**;

---

[*]Full paper available at arXiv:1710.02587.

[†]Part of this work was done while visiting University of Waterloo

---

KEYWORDS

Paulsen problem, Operator scaling, Smoothed analysis, Frame theory, Dynamical system

1 INTRODUCTION

A set of n vectors $v_1, \ldots, v_n \in \mathbb{R}^d$ is called an equal norm Parseval frame if it satisfies the Parseval's condition and the equal norm condition:

$$\sum_{i=1}^n v_i v_i^T = I_d \quad \text{and} \quad \|v_i\|_2^2 = \frac{d}{n} \text{ for } 1 \leq i \leq n,$$

where I_d is the $d \times d$ identity matrix. A set of n vectors $u_1, \ldots, u_n \in \mathbb{R}^d$ is an ϵ-nearly equal norm Parseval frame if

$$(1 - \epsilon)I_d \leq \sum_{i=1}^n u_i u_i^T \leq (1 + \epsilon)I_d$$

and

$$(1 - \epsilon)\frac{d}{n} \leq \|u_i\|_2^2 \leq (1 + \epsilon)\frac{d}{n}.$$

Let $\mathcal{F}$ be the set of equal norm Parseval frames. Given a set of vectors $U = \{u_i\}_{i=1}^n$, the squared distance to the set of equal norm Parseval frame is defined as

$$\text{dist}^2(U, \mathcal{F}) = \inf_{V \in \mathcal{F}} \text{dist}^2(U, V) = \inf_{V \in \mathcal{F}} \sum_{i=1}^n \|u_i - v_i\|_2^2.$$

The Paulsen problem asks how close is an ϵ-nearly equal norm Parseval frame to an equal norm Parseval frame.

Definition 1.1 (the Paulsen problem). The Paulsen problem asks what is the best function $f(d, n, \epsilon)$ so that

$$\text{dist}^2(U, \mathcal{F}) \leq f(d, n, \epsilon)$$

for any set of d-dimensional vectors $U = \{u_1, \ldots, u_n\}$ that is ϵ-nearly equal norm Parseval.

The fundamental open question of the Paulsen problem is whether $f(d, n, \epsilon)$ can be independent of the number of vectors n and only dependent on the dimension d and the error ϵ [11, 13].

1.1 History and Motivations

The Paulsen problem has been open for over fifteen years despite receiving quite a bit of attention [7, 11–13]. It has been listed as a major open problem in frame theory in the literature (see e.g. [13, 16, 37]).

An equal norm Parseval frame (also known as an unit-norm tight frame) is a natural generalization of an orthonormal basis. It is used as an overcomplete basis (see the introductions in the books [15, 17, 39]) and has various applications in signal processing and communication theory, including noise and erasure reduction [8, 14, 30, 47], quantization robustness [5, 9, 25], and digital fingerprinting [38]. For some applications in signal processing [30, 44] and quantum information theory [40, 45], equal norm Parseval frames with additional properties such as Grassmannian frames (which minimize the maximal inner product) and equiangular frames (in which the inner products are the same) are needed to provide optimal performance.

These applications motivate the "frame design" questions of constructing equal norm Parseval frames. It is known that equal norm Parseval frames exist for any $d \leq n$. However, it is difficult to construct equal norm Parseval frames, with only a few algebraic constructions known (e.g. truncation of Discrete Fourier transform matrices, vertices of the Platonic solids, constructions from groups; see the survey [49]). On the other hand, it is known that the set of equal norm Parseval frames contains manifold of nontrivial dimensions [20], and so the algebraic methods only produce a few examples from the continuum of the set of all equal norm Parseval frames [12].

Besides algebraic constructions, researchers have also used numerical methods to construct equal norm Parseval frames. It is much easier to construct "nearly" equal norm Parseval frames as a set of random equal-norm vectors is nearly Parseval with high probability. Tropp et al. [46] proposed alternating projection algorithms to construct equal norm Parseval frames and equiangular frames from these nearly equal norm Parseval frames. They show positive experimental results and some partial convergence analysis. Holmes and Paulsen [30] studied the optimal parameters for Grassmannian frames which are even harder to construct. They construct some nearly equal norm Parseval frames with small maximal inner product, and ask the question whether these are good estimates of the optimal parameters for Grassmannian frames. This work led Paulsen to ask a number of people whether a nearly equal norm Parseval frame is always close to an equal norm Parseval frame (if so then their estimates are accurate), and eventually it is known as the Paulsen problem first formally stated in [7].

Proving a good upper bound for the Paulsen problem would give us a firm foundation to work with nearly equal norm Parseval frames, both in theory and in applications. Indeed, our method can be seen as a continuous version of the alternating projection algorithm of Tropp et al. [46], and our results can be viewed as a rigorous justification of the numerical approach for constructing equal norm Parseval frames. We hope that our techniques will be useful to the difficult open question of constructing equiangular equal norm Parseval frames, as Tropp et al. [46] also proposed an alternating projection algorithm for constructing these frames.

1.2 Previous Work on the Paulsen Problem

A compactness argument shows that the function f in the Paulsen problem must exist [13]. There are simple examples showing that $f(d, n, \epsilon) \geq d\epsilon$ [11].

Bodmann and Casazza [7] proved that $f(d, n, \epsilon) \leq O(d^4 n^{18} \epsilon^2)$ when d and n are relatively prime. Their approach is to analyze a dynamical system that improves the closeness to the equal norm condition while ensuring that the Parseval's condition is satisfied. Casazza, Fickus, and Mixon [12] proved that $f(d, n, \epsilon) \leq O(d^{42} n^{18} \epsilon^2)$ when d and n are relatively prime, and they extended this result to the general case and proved that $f(d, n, \epsilon) \leq O(d^{20/7} n^{2/7} \epsilon^{2/7})$. Their approach is to analyze a gradient descent algorithm that improves the closeness to the Parseval's condition while ensuring that the equal norm condition is satisfied.

Cahill and Casazza [11] showed that the Paulsen problem is equivalent to another fundamental problem in operator theory called the projection problem: Find the best function $g(d, n, \epsilon)$ such that the following holds. Given an n-dimensional orthonormal basis $e_1, \ldots, e_n \in \mathbb{R}^n$ and a projection P of rank d that satisfies

$$(1 - \epsilon)\frac{d}{n} \leq \|Pe_i\|_2^2 \leq (1 + \epsilon)\frac{d}{n} \text{ for } 1 \leq i \leq n,$$

there is a projection Q with $\|Qe_i\|_2^2 = \frac{d}{n}$ for $1 \leq i \leq n$ and

$$\sum_{i=1}^{n} \|Pe_i - Qe_i\|_2^2 \leq g(d, n, \epsilon).$$

Cahill and Casazza [11] proved that $f(d, n, \epsilon)$ and $g(d, n, \epsilon)$ are within a factor 2 of each other.

1.3 Results and Techniques

Our main result is a proof that the function in the Paulsen problem is independent of the number of vectors.

THEOREM 1.2. *For any set U of d-dimensional vectors that is an ϵ-nearly equal norm Parseval frame,*

$$\text{dist}^2(U, \mathcal{F}) \leq O(d^{13/2}\epsilon).$$

There is a very natural approach towards solving the Paulsen problem. Given an ϵ-nearly equal norm Parseval frame $u_1, \ldots, u_n \in \mathbb{R}^d$, we alternately fix the Parseval condition (by setting $u_i \leftarrow (\sum_{i=1}^{n} u_i u_i^T)^{-\frac{1}{2}} u_i$) and the equal norm condition (by scaling u_i so that $\|u_i\|_2^2 = \frac{d}{n}$), until both conditions are satisfied exactly and we keep track of the sum of the movement of these operations. We observe that this natural alternating algorithm is a special case of the operator scaling algorithm studied in [23, 26]. So, this alternating algorithm will converge under some mild condition [23, 26], and also there are closed-form formulas for the movement of each operation [13]. But the problem of this approach is that the sum of the movement could be very large, as the path to an exact solution could zig-zag between the alternating operations.

Our approach is based on a continuous version of the operator scaling algorithm [23, 26]. There are two main parts. To avoid the zig-zag movement, we define a dynamical system based on the (discrete) operator scaling algorithm, so that the two alternating operations are combined into one and the movement is continuous. The dynamical system satisfies some very nice identities. Using these identities and the concept of operator capacity defined by

Gurvits [26] in analyzing the convergence of the operator scaling algorithm, we can bound the total movement of our dynamical system given a nearly equal norm Parseval frame.

THEOREM 1.3 (INFORMAL). *Given a set of d-dimensional vectors $U = \{u_1, \ldots, u_n\}$ that forms an ϵ-nearly equal norm Parseval frame, there is a dynamical system that transforms U into $V = \{v_1, \ldots, v_n\}$ that forms an equal-norm Parseval frame and $\text{dist}^2(U, V) \leq O(d^2 n \epsilon)$.*

We prove Theorem 1.3 in the more general operator setting instead of just the frame setting as in the Paulsen problem. We believe that the operator setting is of independent interest, e.g. it is closely related to the Brascamp-Lieb constants that will be discussed in Subsection 1.4.

Using the dynamical system for an arbitrary ϵ-nearly equal norm Parseval frame, the analysis in Theorem 1.3 is tight and the dependency on n is unavoidable. Our intuition is that the set of ϵ-nearly equal norm Parseval frames with large total movement in our dynamical system is small. The second part in our approach is a smoothed analysis [43] of continuous operator scaling. We prove that if we randomly perturb an arbitrary ϵ-nearly equal norm Parseval frame appropriately (in a dependent manner), then the perturbed frame is simultaneously close to the original frame and an equal norm Parseval frame with high probability, by showing that its total movement in our dynamical system is independent of n.

THEOREM 1.4 (INFORMAL). *Given a set of d-dimensional vectors $U = \{u_1, \ldots, u_n\}$ that forms an ϵ-nearly equal norm Parseval frame with $n \gg d^4$ and $\epsilon \ll 1/d^{11/2}$, we can perturb U to obtain $\tilde{U}$ such that $\text{dist}^2(U, \tilde{U}) \leq O(d^{5/2}\epsilon)$ and furthermore $\text{dist}^2(\tilde{U}, \mathcal{F}) \leq O(\sqrt{d}\epsilon)$ by using the dynamical system in Theorem 1.3.*

This solves the problem when n is large enough. Together with Theorem 1.3, we obtain Threom 1.2 by using Theorem 1.3 when n is small. To prove Theorem 1.4, we develop some new techniques to analyze the convergence of the dynamical system in the perturbed frame. Gurvits [26] defined a notion called the operator capacity to analyze the operator scaling algorithm. Recently, the operator scaling algorithm is used to design a polynomial time algorithm to solve the non-commutative rank problem [22], while the key is a new lower bound on the operator capacity. We find an interesting connection between our dynamical system and the operator capacity. We use it to prove a better lower bound on the operator capacity in the perturbed instance, which implies a faster convergence rate in the perturbed instance. We discuss some implications of our results to related work on operator scaling in the following subsection.

1.4 Related Work on Frame Scaling, Operator Scaling, and Matrix Scaling

Scaling a frame into an equal norm Parseval frame, and more generally, scaling an operator into a doubly stochastic operator has various applications in theoretical computer science. Sometimes they go under different names such as radial isotropic positions in machine learning [29], and geometric conditions in Brascamp-Lieb inequalities [3, 4, 24].

An early application of frame scaling is discovered by Forster [21], who showed that a set of n vectors $v_1, \ldots, v_n \in \mathbb{R}^d$ can always be scaled to an equal norm Parseval frame if every subset of d vectors is linearly independent, and he used this result to derive a lower bound on the sign rank of the Hadamard matrix with applications in proving communication complexity lower bounds. We note that Forster's scaling result was proved earlier in a more general setting by Gurvits and Samorodnitsky [27] in their work of approximating mixed discriminants, and is also implicit in the work of Barthe [4] in proving Brascamp-Lieb inequalities. A recent application of frame scaling is found by Hardt and Moitra [29] in robust subspace discovery.

Operator scaling was introduced by Gurvits [26] in an attempt to design a deterministic polynomial time algorithm for polynomial identity testing, and he used it to solve the special case when the commutative rank of a symbolic matrix is equal to its non-commutative rank (e.g. this includes the linear matroid intersection problem over reals). Recently, Garg, Gurvits, Oliveira, and Wigderson [23] improved Gurvits' analysis to prove that the alternating algorithm for operator scaling can be used to compute the non-commutative rank of a symbolic matrix in polynomial time. Subsequently, the alternating algorithm for operator scaling is used by the same group [24] to obtain a polynomial time algorithm to compute the optimal constants in Brascamp-Lieb inequalities, which we will elaborate more below as it is related to our work.

The Brascamp-Lieb inequalities [10] and their reversed form established by Barthe [4] are general classes of inequalities with important applications in functional analysis and convex geometry (e.g. including Nelson's hypercontractivity inequality and the Brunn-Minkowski inequality as special cases). The optimal constants for thses inequalities are determined by Ball [3] assuming the geometric condition (which is a condition similar to that in John's ellipsoid theorem). Garg, Gurvits, Oliveira and Wigderson [24] show that the Brascamp-Lieb constants are equivalent to the capacity of an operator by a simple transformation, in which the geometric condition corresponds exactly to the doubly stochastic condition. Therefore, the algorithm in [23] can be employed to scale the input to satisfying the geometric condition so as to compute the optimal constant. For our smoothed analysis, we develop a new technique to proving a lower bound on the operator capacity and thus an upper bound on the Brascamp-Lieb constant. In particular, this implies improved bounds on the Brascamp-Lieb constants for perturbed instances in the rank-one case (which is the case that Brascamp and Lieb proved in [10]). See [24] and the references therein for applications of these bounds to non-linear Brascamp-Lieb inequalities.

Matrix scaling [41] is a well-studied special case of operator scaling. It has applications in numerical analysis, in approximating permanents [35] and in combinatorial geometry [19]. Very recently, much faster algorithms are developed for matrix scaling by two independent research groups [2, 18]. Cohen, Madry, Tsipras and Vladu [18] obtain an algorithm for matrix scaling with running time $\widetilde{O}(m \log \kappa \log^2(1/\epsilon))$, where m is the number of nonzeros in the input matrix, κ is the ratio between the largest and the smallest entries in the optimal scaling solution, and ϵ is the error parameter of the output. Note that the algorithm is near linear time when κ is bounded by a polynomial in m, but in general it could be exponentially large. Not much is known about upper bounding κ for specific instances, except when the input matrix is strictly

positive [33]. Our techniques for smoothed analysis provides a new way to bound κ; see Remark 4.3.6. In particular, this implies that the algorithm in [18] is near linear time in a pseudorandom instance as defined in Definition 4.3.2 (not necessarily strictly positive).

To summarize, our techniques developed in solving the Paulsen problem provides new tools in bounding the mathematical quantities involved in scaling problems such as the operator capacity and κ about optimal scaling solutions. See the second half of Subsection 4.1 for an overview of these techniques. These provide a new perspective to look at those quantities using the parameters in our dynamical system. Currently, our smoothed analysis is tailored for the Paulsen problem, but we believe that it can be extended to the more general operator setting and also to more natural conditions to prove useful results about other problems solved by scaling techniques.

Very recently, our result for the operator Paulsen problem was used in designing a deterministic polynomial time algorithm for the orbit intersection problem of the left-right group action in invariant theory [1].

1.5 Organization and Overview

In this overview, all references refer to the full version attached. We first review the background of operator scaling and see that the Paulsen problem is a special case in the operator framework in Section 2. We will also introduce the matrix scaling problem in Section 2 as this is a key intermediate problem in our proof of the second part. We omit Section 2 about the background in this ten-page version.

We divide the proof of Theorem 1.2 into two sections. In Section 3, we define our dynamical system based on operator scaling and prove Theorem 1.3. The results in this section works in the more general operator setting. We discover some nice formulas for the dynamical system to analyze its convergence. And we establish a close connection between the operator capacity lower bound and the squared distance bound for the Paulsen problem.

In Section 4, we analyze the perturbation step to prove Theorem 1.4. Using a known reduction that we will discuss in Section 2 and see the proof in Subsection 3.5, we reduce the operator capacity lower bound to a matrix capacity lower bound. Using some probabilistic arguments, we will show that by perturbing the vectors, the corresponding matrix will have some pseudorandom property. Then we use a combinatorial argument to show that the pseudorandom property will imply a fast convergence of our dynamical system for matrix scaling. Interestingly, we show that the fast convergence of our dynamical system will imply a stronger matrix capacity lower bound, and this leads to a bound on Paulsen problem without any dependency on the number of vectors. We remark that the perturbation step only applies in the frame setting.

The proof ideas described so far are of high level. We will give a more concrete technical overview in each section after the appropriate background is covered; see Subsection 3.3 and Subsection 4.1.

2 OPERATOR PAULSEN PROBLEM AND CONTINUOUS OPERATOR SCALING

We consider the following generalization of the Paulsen problem to the operator setting. We refer to a set of of matrices $\mathcal{U} = \{U_1, \ldots, U_k\}$ as an operator.

Definition 2.1 (doubly balanced and doubly stochastic operator). An operator $\mathcal{V} = \{V_1, \ldots, V_k\}$ where $V_i \in \mathbb{R}^{m \times n}$ for $1 \leq i \leq k$ is called doubly balanced if

$$\sum_{i=1}^{k} V_i V_i^T = cn I_m$$

and

$$\sum_{i=1}^{k} V_i^T V_i = cm I_n$$

for some scalar $c \geq 0$, and it is called doubly stochastic in the case when $c = 1/n$.

Definition 2.2 (ϵ-nearly doubly stochastic operator). An operator $\mathcal{U} = \{U_1, \ldots, U_k\}$ where $U_i \in \mathbb{R}^{m \times n}$ for $1 \leq i \leq k$ is called ϵ-nearly doubly stochastic if

$$(1 - \epsilon) I_m \leq \sum_{i=1}^{k} U_i U_i^T \leq (1 + \epsilon) I_m$$

and

$$(1 - \epsilon) \frac{m}{n} I_n \leq \sum_{i=1}^{k} U_i^T U_i \leq (1 + \epsilon) \frac{m}{n} I_n.$$

Definition 2.3 (distance and squared distance). Given $\mathcal{U} = \{U_1, \ldots, U_k\}$ and $\mathcal{V} = \{V_1, \ldots, V_k\}$ where $U_i, V_i \in \mathbb{R}^{m \times n}$ for $1 \leq i \leq k$, the squared distance and the distance between $\mathcal{U}$ and $\mathcal{V}$ are defined as

$$\text{dist}^2(\mathcal{U}, \mathcal{V}) := \sum_{i=1}^{k} \|U_i - V_i\|_F^2$$

and

$$\text{dist}(\mathcal{U}, \mathcal{V}) := \sqrt{\text{dist}^2(\mathcal{U}, \mathcal{V})},$$

where $\|.\|_F$ is the Frobenius norm of the matrix.

Definition 2.4 (the operator Paulsen problem). Given $\mathcal{U} = \{U_1, \ldots, U_k\}$ where $U_i \in \mathbb{R}^{m \times n}$ for $1 \leq i \leq k$ that is ϵ-nearly doubly stochastic, the operator Paulsen problem asks what is the best function $h(k, n, m, \epsilon)$ so that

$$\inf_{\mathcal{V}} \text{dist}^2(\mathcal{U}, \mathcal{V}) \leq h(k, n, m, \epsilon),$$

where the infimum $\mathcal{V}$ is over the sets of matrices which are doubly stochastic.

The main theorem in this section is that $h(k, n, m, \epsilon) \leq O(m^2 n \epsilon)$. This will imply Theorem 1.3, using a reduction from frame scaling to operator scaling.

Organization and Overview: We first define a dynamical system based on the operator scaling algorithm. Then, we present our main technical result that given $\mathcal{U}^{(0)}$ that is ϵ-nearly doubly stochastic, the dynamical system will produce $\mathcal{U}^{(\infty)}$ that is doubly balanced with $\text{dist}^2(\mathcal{U}^{(0)}, \mathcal{U}^{(\infty)}) \leq O(m^2 n \epsilon)$, assuming some

formulas for the dynamical system and a lower bound on the operator capacity. The formulas and the capacity are in Subsection 3.4 and 3.5 in the full version attached. The details about the operator Paulsen problem and the Paulsen problem are in Subsection 3.6 and 3.7 in the full verison.

2.1 Continuous Operator Scaling

Our idea is to define a continuous version of the operator scaling algorithm so that the two alternating steps are combined into one step and the movement is continuous, and we will bound the distance $\text{dist}(U^{(0)}, U^{(T)})$ by the total movement $\int_0^T \sqrt{\sum_{i=1}^n \left\| \frac{d}{dt} U_i^{(t)} \right\|_2^2} dt$. We define our dynamical system in the more general operator setting. There is a time t in the evolution of the matrices, but we will drop the superscript to ease our notation whenever it is clear from the context.

Definition 2.5 (size of an operator). Given $\mathcal{U} = (U_1, \ldots, U_n)$, let

$$s(\mathcal{U}) := \sum_{i=1}^k \|U_i\|_F^2 = \text{tr}(\sum_{i=1}^k U_i U_i^T) = \text{tr}(\sum_{i=1}^k U_i^T U_i)$$

be the size of the operator. We use the shorthand s when the system $\mathcal{U}$ is clear from the context.

Our dynamical system is defined by the following differential equation.

Definition 2.6 (dynamical system from operator scaling). The following differential equation describes how $\mathcal{U}^{(t)} = \{U_1^{(t)}, \ldots, U_k^{(t)}\}$ changes over time:

$$\frac{d}{dt} U_i := (sI_m - m \sum_{j=1}^k U_j U_j^T) U_i + U_i(sI_n - n \sum_{j=1}^k U_j^T U_j)$$

for $1 \le i \le k$.

The following definition is the key parameter in our analysis. We can think of ϵ as an ℓ_∞-error bound of the input, and the following Δ as an ℓ_2-error bound. Indeed, we will work with Δ as the error measure in all our proofs, and only use the relation that $\Delta \le 2m^2\epsilon^2$ to draw the conclusion.

Definition 2.7 (Δ of an operator). We measure the progress of our dynamical system by:

$$\Delta(\mathcal{U}) = \frac{1}{m} \text{tr}[(sI_m - m \sum_{i=1}^k U_i U_i^T)^2] + \frac{1}{n} \text{tr}[(sI_n - n \sum_{i=1}^k U_i^T U_i)^2],$$

which is zero if and only if $\mathcal{U}$ is doubly balanced. The two conditions are scaled appropriately so that m and n are symmetric. We use the shorthand $\Delta^{(t)}$ for $\Delta(\mathcal{U}^{(t)})$ when it is clear.

Another motivation for our dynamical system is that it moves in the direction that minimizes $\Delta(\mathcal{U})$.

2.2 Total Movement of Dynamical System

We present our main technical result of this section in this subsection. Assuming some formulas for the dynamical system and a lower bound on the operator capacity, we will show that given $\mathcal{U}^{(0)}$ that is ϵ-nearly stochastic, the dynamical system will produce

$\mathcal{U}^{(\infty)}$ that is doubly balanced with $\text{dist}^2(\mathcal{U}^{(0)}, \mathcal{U}^{(\infty)}) \le O(m^2 n\epsilon)$. First, using triangle inequality, we will bound the squared distance of $\mathcal{U}$ to the set of doubly balanced operators by the total movement in our dynamical system.

LEMMA 2.8. *Let $\mathcal{U}^{(0)}$ be the input operator to the dynamical system and $\mathcal{U}^{(T)}$ be the operator in the dynamical system at time T, we have*

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

To analyze the convergence of the dynamical system, the operator capacity is important.

Definition 2.9 (operator capacity [23, 26]). Given an operator $\mathcal{U} = \{U_1, \ldots, U_k\}$ where each $U_i \in \mathbb{R}^{m \times n}$, we define the capacity of $\mathcal{U}$ as

$$\text{cap}(\mathcal{U}) = \inf_{X \ge 0} \frac{m \det(\sum_{i=1}^k U_i X U_i^T)^{1/m}}{\det(X)^{1/n}}.$$

In the following, we state the facts that we need for our proof.

Proof Steps: It turns out that there are very nice formulas of our dynamical system which can be used to bound the total movement. We will prove

(i) in Lemma 3.4.2 of the full version that

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

which in particular implies that the size of the operator is decreasing over time;

(ii) in Lemma 3.4.3 of the full version that

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

which in particular implies that $\Delta^{(t)}$ is decreasing over time;

(iii) in Lemma 3.4.5 of the full version that $\text{cap}(\mathcal{U}^{(t)})$ is unchanged over time;

(iv) and in Lemma 3.3.3 and Theorem 3.5.16 of the full version that $s(\mathcal{U}) \ge \text{cap}(\mathcal{U}) \ge s(\mathcal{U}) - mn\sqrt{\Delta(\mathcal{U})}$, which implies that $s^{(T)} \ge \text{cap}(\mathcal{U}^{(T)}) = \text{cap}(\mathcal{U}^{(t)}) \ge s^{(t)} - mn\sqrt{\Delta^{(t)}}$ for $T \ge t \ge 0$.

With these, we can analyze the total movement of the dynamical system before $\Delta^{(T)} \le \Delta^{(0)}/2$.

PROPOSITION 2.10. *For $t \ge 0$ with $\Delta^{(t)} > 0$, let T be the first time that $\Delta^{(T)} = \Delta^{(t)}/2$. Then*

$$T \le t + \frac{mn}{\sqrt{\Delta^{(t)}}} \quad \text{and} \quad \text{dist}^2(\mathcal{U}^{(T)}, \mathcal{U}^{(t)}) \le 2mn\sqrt{\Delta^{(t)}}.$$

PROOF. The assumption implies that $\Delta^{(\tau)} > \Delta^{(t)}/2$ for $t \le \tau < T$, and thus it follows from point (i) that

$$\frac{d}{d\tau} s^{(\tau)} = -2\Delta^{(\tau)} < -\Delta^{(t)} \quad \text{for } t \le \tau < T.$$

The capacity lower bound thus allows us to conclude that $T \leq t + \frac{mn}{\sqrt{\Delta^{(t)}}}$, as otherwise

$$
\begin{aligned}
s^{(T)} &= s^{(t)} + \int_t^T \frac{d}{d\tau} s^{(\tau)} d\tau \\
&< s^{(t)} - (T-t)\Delta^{(t)} \\
&< s^{(t)} - mn\sqrt{\Delta^{(t)}},
\end{aligned}
$$

contradicting point (iv). Therefore,

$$
\begin{aligned}
\text{dist}(\mathcal{U}^{(T)}, \mathcal{U}^{(t)}) &\leq \int_t^T \sqrt{\sum_{i=1}^k \left\| \frac{d}{d\tau} U_i^{(\tau)} \right\|_F^2} \, d\tau \\
&= 2 \int_t^T \sqrt{-\frac{d}{d\tau} \Delta^{(\tau)}} \, d\tau \\
&\leq 2\sqrt{\int_t^T (-\frac{d}{d\tau} \Delta^{(\tau)}) d\tau \cdot \int_t^T 1 d\tau} \\
&= 2\sqrt{-(\Delta^{(T)} - \Delta^{(t)})(T-t)} \\
&\leq \sqrt{2mn\sqrt{\Delta^{(t)}}},
\end{aligned}
$$

where the first inequality is by Lemma 2.8, the first equality is by point (ii), the second inequality is by Cauchy-Schwarz, and the last inequality is by the bound on T above and the assumption that $\Delta^{(T)} = \Delta^{(t)}/2$. Squaring both sides proves the lemma. □

Using this argument repeatedly will give us a decreasing geometric sequence and we obtain the main technical result in this section.

THEOREM 2.11. *Given any operator* $\mathcal{U}^{(0)} = \{U_1^{(0)}, \ldots, U_k^{(0)}\}$ *where* $U_i \in \mathbb{R}^{m \times n}$ *for* $1 \leq i \leq k$, *the dynamical system in Definition 2.6 will move* $\mathcal{U}^{(0)}$ *to* $\mathcal{U}^{(\infty)}$ *such that*

$$
\Delta(\mathcal{U}^{(\infty)}) = 0 \quad \text{and} \quad \text{dist}^2(\mathcal{U}^{(\infty)}, \mathcal{U}^{(0)}) \leq O(mn\sqrt{\Delta^{(0)}}).
$$

3 IMPROVED BOUND THROUGH SMOOTHED ANALYSIS

We can understand what was done in the previous section as a reduction from the Paulsen problem to proving capacity lower bound: If $\text{cap}(U^{(t)}) \geq s(U^{(t)}) - p(d, n)\sqrt{\Delta(U^{(t)})}$ for all t, then $\text{dist}^2(U, V) \leq O(p(d,n)\sqrt{\Delta(U^{(0)})}) \leq O(p(d,n) \cdot d\epsilon)$. We discuss in the full version that the worst case analysis of $p(d,n) = O(dn)$ in the previous section cannot be improved.

Smoothed Analysis: Our intuition is that the instances with $\text{cap}(U) \approx s(U) - dn\sqrt{\Delta(U)}$ are rare. So our idea is to perturb the input instance and to prove a stronger capacity lower bound on the perturbed instance. Using some probabilistic arguments, we will prove that with high probability a perturbation W of U satisfies $\Delta(W) \approx \Delta(U)$ and $\text{cap}(W) \gg \text{cap}(U)$. To prove the lower bound of the capacity in the perturbed instance W, we have developed an interesting method to prove matrix capacity lower bound using the results in our dynamical system.

3.1 Overview of the Smoothed Analysis

Perturbation: The perturbation is informally described as follows. For each $1 \leq i \leq n$, let g_i be a d-dimensional vector where each entry is an independent Gaussian random variable $N(0, \sigma^2)$ with mean zero and variance σ^2. We let $y_i = P_L(g_i)$ and $w_i = u_i + y_i$ for $1 \leq i \leq n$, where L is a subspace of co-dimension $d^2 + n$ and P_L is the orthogonal projection acting on $\mathbb{R}^{d \times n}$ to the subspace L. For technical reasons, we will normalize the vectors so that they have equal norm. We will choose $\sigma^2 \approx \frac{\sqrt{d\Delta(U)}}{n}$. We will explain the choice of σ^2 later in this subsection.

Analysis of Perturbed Instances: We will prove that the perturbed instance $W = \{w_1, \ldots, w_n\}$ has the following properties with high probability.

(i) The squared distance between U and W is small in Proposition 4.8.6 of the full version: $\text{dist}^2(U, W) \leq O(dn\sigma^2) \leq O(d^{3/2}\sqrt{\Delta(U)})$, where the second inequality is by our choice of σ^2.

(ii) Assuming n is large enough and Δ is small enough, we bound the increase of Δ in Proposition 4.8.7 of the full version that $\Delta(W) \leq O(\Delta(U))$. This is the place where we need to use the subspace L in the perturbation process to ensure that $\Delta(W)$ is bounded, and is also the bottleneck of the current proof that requires the assumptions on n and Δ.

(iii) Assuming n is large enough, we establish an improved capacity lower bound in Theorem 4.8.8 of the full version that $\text{cap}(W) \geq s(W) - O(\sqrt{\frac{\Delta(W)}{d}})$. This is the heart of the smoothed analysis, where we have removed the dependency on n in the capacity lower bound.

Paulsen Problem: From point (iii) and the reduction of the Paulsen problem to capacity lower bound discussed earlier, we expect that we can set $p(d, n) = 1/\sqrt{d}$ and $\Delta(W) = O(\Delta(U))$ to bound the squared distance after the perturbation to be $O(\sqrt{\Delta(U)/d}) = O(\sqrt{d}\epsilon)$, independent of n when n is large enough. This is eventually what we will prove. We discuss in the full version that there is a subtlety that we need to do the perturbation step many times. The precise step-by-step procedure to move from the initial frame to an equal norm Parseval frame is described in Procedure 4.8.1 of Subsection 4.8 of the full version.

Projection: A natural attempt for the perturbation is to add independent noise to each coordinate for each vector. Unfortunately, it does not work as $\Delta(W)$ would become much bigger than $\Delta(U)$ with high probability. The linear subspace L consists of $d^2 + n$ linear constraints which are added to enforce that the "cross terms/first order terms" become zero to ensure that point (ii) holds. This comes with the price of the additional assumption that $n \gg d^2$ for point (iii) to hold as the linear subspace L has co-dimension $d^2 + n$.

Matrix Capacity Lower Bound from Dynamical System. Most of the work in Theorem 1.4 is to prove point (iii). There are two main ingredients.

Pseudorandom Property: The first ingredient is to identify a pseudorandom property for a frame to have a stronger capacity lower bound. Instead of doing it directly, we follow a known reduction as described in Proposition 3.5.8 in the full version to consider the corresponding matrix A. The pseudorandom property

that we will use of a $d \times n$ matrix A is that every column has at least one entry with value at least $\Omega(\sigma^2)$ and every row has almost all entries with value at least $\Omega(\sigma^2)$. We will prove that after we do the perturbation on the vectors as described, the corresponding matrix A has the pseudorandom property with high probability. The proof of this lemma is quite technical, and this is the step that we could not prove in the operator setting, and also we need the assumption that $n \gg d^2$ for the proof to go through.

Bounding Matrix Capacity using Dynamical System: The second ingredient is a new method to prove matrix capacity lower bound. In Subsection 2.2, we have seen that the capacity lower bound provides an indirect way to argue that $\Delta^{(t)}$ will converge to zero. We prove the reverse direction to establish matrix capacity lower bound, that a fast convergence of $\Delta^{(t)}$ to zero implies a good capacity lower bound.

To do this, we define a matrix version of the Paulsen problem, and also a dynamical system from matrix scaling to solve this problem. We will show that the dynamical system will satisfy the same formulas as outlined in Subsection 2.2. Assuming the pesudorandom property of a matrix holds in the beginning, we will show in Proposition 4.3.5 in the full version that it will hold during the execution of the dynamical system. Proposition 4.3.5 requires a lower bound on σ^2 for the proof to go through, and this is the reason for our choice of σ^2, which is the bottleneck of the current proof as it requires a large movement in the perturbation process.

A key step is a combinatorial argument in Subsection 4.4 of the full version that proves that there exists an absolute constant κ such that

$$-\frac{d}{dt}\Delta^{(t)} \gtrsim \kappa \sigma^2 n \Delta^{(t)} \text{ for all } t \geq 0$$
$$\implies \Delta^{(t)} \lesssim \exp(\kappa \sigma^2 n t) \cdot \Delta^{(0)} \text{ for all } t \geq 0,$$

assuming the pesudorandom property of the matrix holds throughout the execution of the dynamical system. This can be used to lower bound the matrix capacity using the following relations:

$$s^{(0)} - \text{cap}(A) = s^{(0)} - s^{(\infty)}$$
$$= -\int_0^\infty \frac{d}{dt} s \, dt$$
$$= \int_0^\infty 2\Delta^{(t)} dt$$
$$\lesssim \Delta^{(0)} \int_0^\infty 2\exp(-\kappa\sigma^2 nt)dt$$
$$= \frac{\Delta^{(0)}}{\kappa\sigma^2 n},$$

where the first equality uses the capacity bound in the previous section and the third equality is by an identity analogous to that in previous section. This implies that

$$\text{cap}(A) \gtrsim s(A) - \frac{\Delta(A)}{\kappa\sigma^2 n}$$
$$\implies \text{cap}(W) \gtrsim s(W) - \frac{\Delta(W)}{\kappa\sigma^2 n},$$

where the implication follows from the reduction. This completes the outline of the new method in proving capacity lower bound and the proof of point (iii).

REFERENCES

[1] Zeyuan Allen-Zhu, Ankit Garg, Yuanzhi Li, Rafael Oliveira, Avi Wigderson. *Operator Scaling via Geodesically Convex Optimization, Invariant Theory and Polynomial Identity Testing.* In Proceedings of the 50th Annual ACM Symposium on Theory of Computing (STOC), 2018.
[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] K. Ball. *Volumes of sections of cubes and related problems.* Geometric Aspects of Functional Analysis, 251–260, 1989.
[4] F. Barthe. *On a reverse form of the Brascamp-Lieb inequality.* Inventiones mathematicae 134(2), 335–361, 1998.
[5] J.J. Benedetto, A.M. Powell, Ö, Yilmaz. *Second order sigma-delta quantization of finite frame expansions.* Applied and Computational Harmonic Analysis 20, 126–148, 2006.
[6] B. Birnir. *Dynamical Systems Theory.* Course notes, UCSB, 2006.
[7] B. Bodmann, P.G. Casazza. *The road to equal-norm Parseval frames.* Journal of Functional Analysis 258(2), 397–420, 2010.
[8] B.G. Bodmann, V.I. Paulsen. *Frames, graphs and erasures.* Linear Algebra and its Applications 404, 118–146, 2005.
[9] B.G. Bodmann, V.I. Paulsen. *Frame paths and error bounds for sigma-delta quantization.* Applied and Computational Harmonic Analysis 22, 176–197, 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] J. Cahill, P.G. Casazza. *The Paulsen Problem in Operator Theory.* Operators and Matrices 7(1), 117–130, 2013.
[12] P.G. Casazza, M. Fickus, D. Mixon. *Auto-tuning unit norm frames.* Applied and Computational Harmonic Analysis 32(1), 1–15, 2012.
[13] P.G. Casazza. *The Kadison-Singer and Paulsen problems in finite frame theory.* In "Finite frames: theory and applications", 2013.
[14] P.G. Casazza, J. Kovacević. *Equal-Norm Tight Frames with Erasures.* Advances in Computational Mathematics 18, 387–430, 2003.
[15] P.G. Casazza, G. Kutyniok. *Finite frames: theory and applications.* Birkhäuser Basel, 2013.
[16] P.G. Casazza, R.G. Lynch. *A brief introduction to Hilbert space frame theory and its applications.* AMS Short Course, 2015, San Antonio (arxiv 1509.07347).
[17] O. Christensen. *An introduction to frames and Riesz bases.* Birkhäuser Boston, 2003.
[18] 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.
[19] Z. Dvir, A. Garg, R. Oliveira, J. Solymosi. *Rank bounds for design matrices with block entries and geometric applications.* In arXiv 1610.08923, 2016.
[20] K. Dykema, N. Strawn. *Manifold structure of spaces of spherical tight frames.* Int. J. Pure Appl. Math. 28, 217–256, 2006.
[21] J. Forster. *A linear lower bound on the unbounded error probabilistic communication complexity.* J Comput. Syst. Sci. (65), 612–625, 2002.
[22] 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.
[23] A. Garg, L. Gurvits, R. Oliveira, A. Wigderson. *Operator scaling: theory and applications.* In arXiv 1511.03730, 2017.
[24] 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.
[25] V.K. Goyal, J. Kovacević, J.A. Kelner. *Quantized frame expansions with erasures.* Applied and Computational Harmonic Analysis 10, 203–233, 2001.
[26] L. Gurvits. *Classical complexity and quantum entanglement.* Journal of Computer and System Sciences 69(3), 448–484, 2004.

[27] 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.

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

[29] 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.

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

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

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

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

[34] B. Laurent, P. Massart. *Adaptive estimation of a quadratic functional by model selection.* Annals of Statistics 28(5), 1302–1338, 2000.

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

[36] L. Lovász, S. Vempala. *The geometry of logconcave functions and sampling algorithms.* Random Structures and Algorithms 30(3), 307–358, 2007.

[37] D.G. Mixon. *Four open problems in frame theory.* Frames and Algebraic and Combinatorial Geometry, 2015.

[38] D.G. Mixon, C.J. Quinn, N. Kiyavash, M. Fickus. *Fingerprinting with equiangular tight frames.* IEEE Transactions on Information Theory 59, 1855–1865, 2013.

[39] K.A. Okoudjou. *Finite frame theory: a complete introduction to overcompleteness.* Proceedings of Symposia in Applied Mathematics 73, 2016.

[40] J.M. Renes, R. Blume-Kohout, A.J. Scott, C.M. Caves. *Symmetric informationally complete quantum measurements.* Journal of Mathematical Physics 45(6), 2171, 2004.

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

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

[43] D.A. Spielman, S.H. Teng. *Smoothed analysis: an attempt to explain the behavior of algorithms in practice.* Communications of the ACM 52(10), 76–84, 2009.

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

[45] 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.

[46] 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.

[47] R. Vershynin. *Frame expansions with erasures: an approach through the noncommutative operator theory.* Applied and Computational Harmonic Analysis 18, 167–176, 2005.

[48] R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices.* Compressed sensing, Cambridge University Press, 210–268, 2012.

[49] S. Waldron. *Group Frames.* In "Finite frames: theory and applications", 2013.