The Paulsen problem, continuous operator scaling, and smoothed analysis

Lap Chi Lau, University of Waterloo

Joint work with Tsz Chiu Kwok (Waterloo),
Yin Tat Lee (Washington),
Akshay Ramachandran (Waterloo)

Outline

Part I: Paulsen problem

- Motivation from frame theory

Part II: Continuous operator scaling

- Operator scaling, alternating algorithm, reduction
- Analysis of dynamical system

Part III: Smoothed analysis

- Proof outline, capacity lower bound

Part IV: Discussions

Frames

Frame: a collection of vectors $u_1, u_2, \dots, u_n \in \mathbb{R}^d$ that spans $\mathbb{R}^d$

Equal norm: if $\|u_i\|_2 = \|u_j\|_2$ for all i, j.

Parseval: if $\sum_{i=1}^{n} u_i u_i^T = I_d$.

An equal norm Parseval frame is an *overcomplete* basis:

$$\sum_{i=1}^{n} \langle x, u_i \rangle \; u_i = x \quad \forall\, x \in \mathbb{R}^d$$

It has applications in signal processing, communication theory, and quantum information theory.

Motivation

Equal norm Parseval frames are difficult to construct with only a few known algebraic constructions.

[Holmes-Paulsen 04] were interested in constructing Grassmaniann frames, equal norm Parseval frames with minimal $\max_{i,j} \langle u_i, u_j \rangle^2$, which are even more difficult to construct.

It is easier to construct "approximate" equal norm Parseval frames (e.g. random unit vectors, optimal packing of lines).

Question: Can we turn an "approximate" frame into an equal norm Parseval frame by just moving the vectors "slightly"?

The Paulsen Problem

What is the best function $f(n, d, \epsilon)$ such that for any $u_1, \ldots, u_n \in \mathbb{R}^d$ with

$$(1-\epsilon)\frac{d}{n} \le \|u_i\|_2^2 \le (1+\epsilon)\frac{d}{n} \quad \forall\, 1 \le i \le n \qquad (\epsilon - \textbf{nearly equal norm})$$

$$(1-\epsilon)I_d \preccurlyeq \sum_{i=1}^{n} u_i u_i^T \preccurlyeq (1+\epsilon)I_d \qquad (\epsilon - \textbf{nearly Parseval}),$$

there exist $v_1, \ldots, v_n \in \mathbb{R}^d$ with

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

such that

$$\sum_{i=1}^{n} \|u_i - v_i\|_2^2 \le f(n, d, \epsilon)?$$

Previous work

[Bodmann-Casazza, 10] $f(d,n,\epsilon) \leq O(d^{42}\, n^{18}\, \epsilon^{2})$ when $\mathbf{gcd}(d,n) = 1.$

- dynamical system improves on equal norm while keeping Parseval.

[Casazza-Fickus-Mixon, 12] $f(d,n,\epsilon) \leq O\left(d^{20/7}\, n^{2/7}\, \epsilon^{2/7}\right)$

- gradient descent improves on Parseval while keeping equal norm.

There are examples showing that $f(d,n,\epsilon) \geq d\epsilon$.

Question: Can the bound be independent of n?

Main Result

$$\boxed{\textbf{Theorem.}\quad f(d,n,\epsilon) \le O\left(d^{13/2}\,\epsilon\right)}$$

The proof has two parts.

First, we define a dynamical system based on operator scaling,

and show that $f(d,n,\epsilon) \le O(d^2\, n\, \epsilon)$.

Then, we do a smoothed analysis to remove the dependency on n.

***[Hamilton, Moitra 18]** $f(d,n,\epsilon) \le O(d^2\epsilon)$

Outline

Part I: Paulsen problem

- Motivation from frame theory

Part II: Continuous operator scaling

- Operator scaling, alternating algorithm, reduction
- Analysis of dynamical system

Part III: Smoothed analysis

- Proof outline, capacity lower bound

Part IV: Discussions

Alternating Algorithm

How to move an approximate frame to satisfy the two conditions exactly?

The problems is difficult with two conditions. It is easy with one condition.

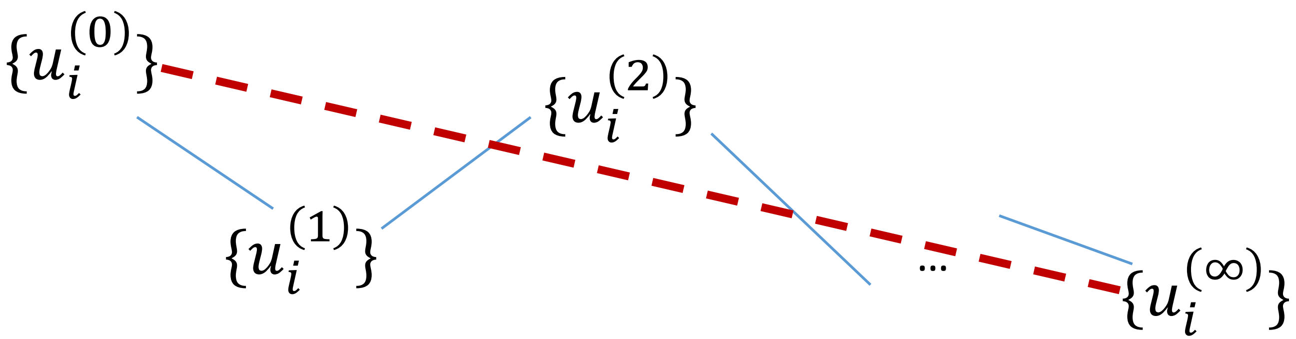
- To satisfy the equal norm condition, we just rescale the vectors.
- To satisfy the Parseval condition, we can set

$$u_i \leftarrow \left(\sum_{i=1}^{n} u_i u_i^T\right)^{-\frac{1}{2}} u_i \qquad \text{so that} \qquad \sum_{i=1}^{n} u_i u_i^T = I_d.$$

A natural algorithm is to alternate between these two steps and hope that it will converge to a solution satisfying both conditions.

First Idea

Our starting point is to bound the distance by the total movement in the alternating algorithm (assuming it converges):



This is a special case of the alternating algorithm for operator scaling, which was analyzed in [Gurvits 04, Garg-Gurvits-Oliveira-Wigderson 16].

Operator Scaling

An operator is a collection of matrices $U_1, \dots, U_k \in \mathbb{R}^{m \times n}$.

[Gurvits 04]

Given $U_1, \dots, U_k \in \mathbb{R}^{m \times n}$, we would like to find $L \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{n \times n}$ such that if we define $V_i = LU_iR$ for $1 \le k \le n$ then

$$\sum_{i=1}^{k} V_i V_i^T = cnI_m \qquad \text{and} \qquad \sum_{i=1}^{k} V_i^T V_i = cmI_n$$

for some constant c.

We say an operator satisfying the two conditions <u>doubly balanced</u>.

Alternating Algorithm

Repeat the following two steps [Gurvits 04]:

- To satisfy the condition $\sum_{i=1}^{k} U_i U_i^T = I_m$, we set

$$U_i \leftarrow \left(\sum_{j=1}^{k} U_j U_j^T\right)^{-\frac{1}{2}} U_i$$

- To satisfy the condition $\sum_{i=1}^{k} U_i^T U_i = I_n$, we set

$$U_i \leftarrow U_i \left(\sum_{j=1}^{k} U_j^T U_j\right)^{-\frac{1}{2}}$$

A natural algorithm is to alternate between these two steps and hope that it will converge to a solution satisfying both conditions.

A simple reduction from frame scaling to operator scaling:

$$u_i \in \mathbb{R}^d \quad \rightarrow \quad U_i \equiv \begin{pmatrix} | & | & | \\ 0 & u_i & 0 \\ | & | & | \end{pmatrix} \in \mathbb{R}^{d\times n}$$

- The condition $\sum_{i=1}^{n} U_i U_i^T = I_d$ is the Parseval condition $\sum_{i=1}^{n} u_i u_i^T = I_d$.
- The condition $\sum_{i=1}^{n} U_i^T U_i = \frac{d}{n} I_n$ is the equal norm condition

$$\begin{pmatrix} \|u_1\|_2^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \|u_n\|_2^2 \end{pmatrix} = \frac{d}{n} I_n.$$

So we focus on this more general setting in this part of the talk.

The Operator Paulsen Problem

What is the best function $g(m, n, k, \epsilon)$ s.t. for any $U_1, \dots, U_k \in \mathbb{R}^{m \times n}$ with

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

there exist $V_1, \dots, V_k \in \mathbb{R}^{m \times n}$ with

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

such that

$$\sum_{i=1}^{k} \|U_i - V_i\|_F^2 \le g(m, n, k, \epsilon) \boxed{\le m^2 n \epsilon}$$

Applications

Matrix Scaling:

- Preconditioning for linear solvers [Osborne 60]
- Optimal transportation [Wilson 69]
- Bipartite matching
- Deterministic approximation of permanents [Linial-Samorodnitsky-Wigderson 00]

Frame Scaling:

- Sign rank lower bound [Forster 02]
- Robust subspace recovery [Hardt-Moitra 13]
- Paulsen problem

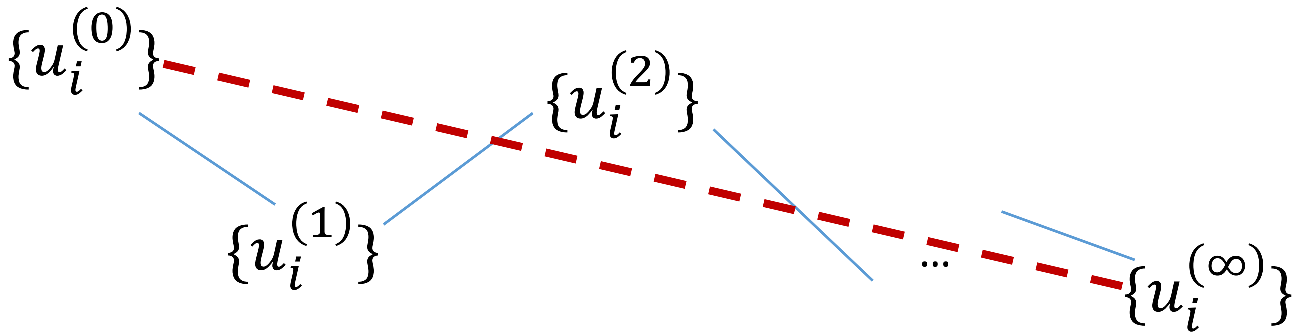
PSD scaling:

- Approximation of mixed discriminants [Gurvits-Samorodnitsky 02]

Operator Scaling:

- Computing non-commutative rank [Garg-Gurvits-Oliveira-Wigderson 16]
- Computing Brascamp-Lieb constants [Garg-Gurvits-Oliveira-Wigderson 17]
- Orbit intersection problem [AllenZhu-Garg-Li-Oliveira-Wigderson 18]

Issues in First Idea



There are examples which do not converge:

$$\binom{1}{0}, \binom{1}{0}, \binom{0}{1} \Leftrightarrow \binom{\sqrt{2}/2}{0}, \binom{\sqrt{2}/2}{0}, \binom{0}{1}$$

Even if it converges, the path could zig-zag a lot and

the total movement is much larger than the distance.

Error Measure

[Gurvits 04]

$$\boxed{\Delta = \frac{1}{m}\left\| sI_m - m\sum_{j=1}^{k} U_j U_j^T \right\|_F^2 + \frac{1}{n}\left\| sI_n - n\sum_{j=1}^{k} U_j^T U_j \right\|_F^2}$$

where $s = \sum_{i=1}^{k} ||U_i||_F^2$ is the **size** of the operator.

- Δ is zero if and only if the operator is doubly balanced.
- Can show that $\Delta \leq m^2\epsilon^2$.
- Focus on proving the total movement is $\leq mn\sqrt{\Delta} \leq m^2 n\epsilon$.

The dynamical system is moving in the direction that minimizes Δ.

Continuous Operator Scaling

Dynamical System: Do both steps simultaneously and continuously.

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

where $s = \sum_{i=1}^{k} \|U_i\|_F^2$ is the <u>**size**</u> of the operator.

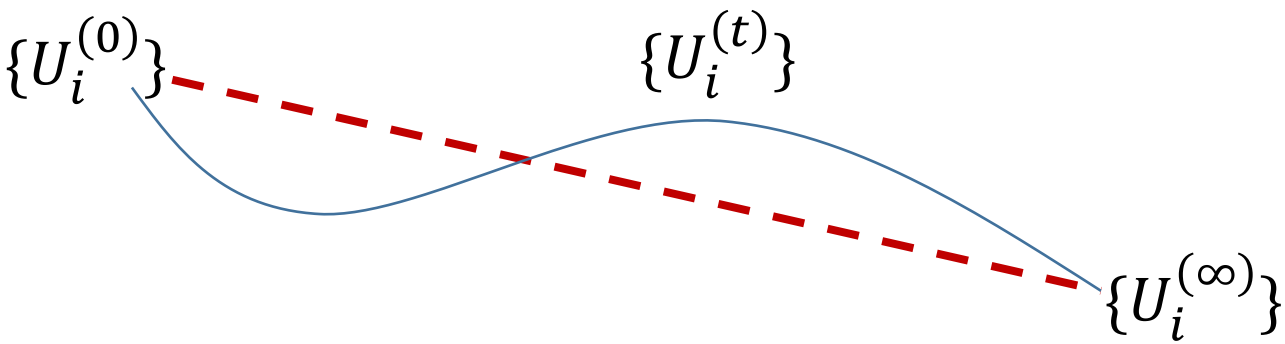
We find some nice identities to analyze the convergence.

Lemma 1. $\frac{d}{dt} s^{(t)} = -\Delta^{(t)}$.

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

Claim. The dynamical system converges to a doubly balanced operator.

Total Movement



We again bound the final distance by the path length.

$$\left(\sum_{i=1}^{k} \left\| U_i^{(\infty)} - U_i^{(0)} \right\|_F^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{k} \left\| \int_0^\infty \frac{d}{dt} U_i^{(t)}\, dt \right\|_F^2\right)^{\frac{1}{2}}$$

distance

$$\leq \int_0^\infty \left(\sum_{i=1}^{k} \left\| \frac{d}{dt} U_i^{(t)} \right\|_F^2\right)^{\frac{1}{2}} dt \quad = \int_0^\infty \sqrt{-\frac{d}{dt}\Delta^{(t)}}\, dt$$

local movement

(triangle inequality) (Lemma 2)

Let T be the first time that $\Delta^{(T)} = \Delta^{(0)}/2$.

$$\left(\int_0^T \sqrt{-\frac{d}{dt}\Delta^{(t)}}\, dt\right)^2 \le \left(\int_0^T 1\, dt\right)\left(\int_0^T -\frac{d}{dt}\Delta^{(t)}\, dt\right) \le T\,\Delta^{(0)}.$$

We can complete the movement bound by a geometric sum argument.
So it remains to bound the **half time.**

Note Lemma 1 implies for all time up to T:

$$\boxed{\frac{d}{dt}\, s^{(t)} = -\Delta^{(t)} \le -\Delta^{(0)}/2}$$

Capacity

[Gurvits 04] Potential function to analyze operator scaling

$$\mathrm{cap}(\{U_i\}) = \inf_{X \in \mathbb{R}^{n\times n}, X \succ 0} \frac{m \det\left(\sum_{i=1}^{k} U_i X U_i^T\right)^{\frac{1}{m}}}{\det(X)^{\frac{1}{n}}}$$

Lemma 3. Capacity is unchanged over time.

Lemma 4. $s^{(t)} \geq \mathrm{cap}^{(t)} \geq s^{(t)} - mn\sqrt{\Delta^{(t)}}$.

We adapt the proof of Lemma 4 from [GGOW 16].

One implication is that $s^{(\infty)} = \mathrm{cap}^{(\infty)} = \mathrm{cap}^{(0)}$.

Bounding Half Time

<u>Half time</u>. Want to upper bound the first time T so that $\Delta^{(T)} = \Delta^{(0)}/2$.

Lemma 3. Capacity is unchanged over time.

+

Lemma 4. $s^{(t)} \geq \mathrm{cap}^{(t)} \geq s^{(t)} - mn\sqrt{\Delta^{(t)}}$.

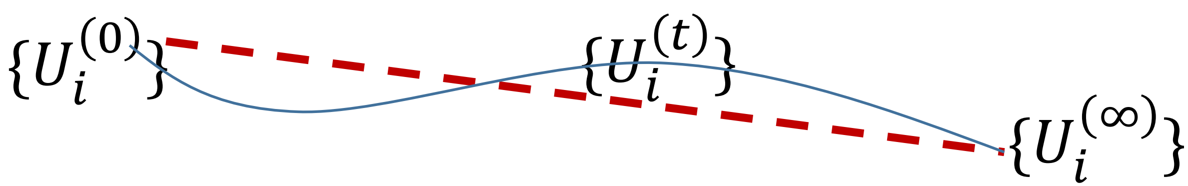
⇒ $s^{(T)} \geq \mathrm{cap}^{(T)} = \mathrm{cap}^{(0)} \geq s^{(0)} - mn\sqrt{\Delta^{(0)}}$

⇒ size of the operator decreases by at most $mn\sqrt{\Delta^{(0)}}$

Lemma 1. $\frac{d}{dt}\, s^{(t)} = -\Delta^{(t)}$. ⇒ size decreases by at least $\frac{1}{2}\,\Delta^{(0)} T$

⇒ $\boxed{T \leq \frac{2mn}{\sqrt{\Delta}}}$ ⇒ $\boxed{\text{total movement} \leq T\Delta \leq mn\sqrt{\Delta}.}$

Summary of Analysis



local movement

squared distance

$$\sum_{i=1}^{k} \left\| U_i^{(\infty)} - U_i^{(0)} \right\|_F^2 \le \left(\int_0^\infty \left(\sum_{i=1}^{k} \left\| \frac{d}{dt} U_i^{(t)} \right\|_F^2 \right)^{\frac{1}{2}} dt \right)^2$$

$$\le \left(\int_0^\infty \sqrt{-\frac{d}{dt} \Delta^{(t)}}\, dt \right)^2$$

Lemma 2

half time, geometric sum, Cauchy-Schwarz

$$\le T \Delta^{(0)}$$

$$\le mn\sqrt{\Delta^{(0)}}$$

Capacity argument, Lemma 1

L_2 vs L_∞

$$\le m^2 n \epsilon.$$

Outline

Part I: Paulsen problem

- Motivation from frame theory

Part II: Continuous operator scaling

- Operator scaling, alternating algorithm, reduction
- Analysis of dynamical system

Part III: Smoothed analysis

- Proof outline, capacity lower bound

Part IV: Discussions

Capacity and Total Movement

Part II can be understood as a reduction from total movement to capacity lower bound:

$$\mathrm{cap} \geq s - f(d,n,\Delta) \quad \underset{\text{part II}}{\Longrightarrow} \quad \mathrm{dist}^2 \leq f(d,n,\Delta).$$

In Part II, we proved $f(d,n,\Delta) \leq dn\sqrt{\Delta}$.

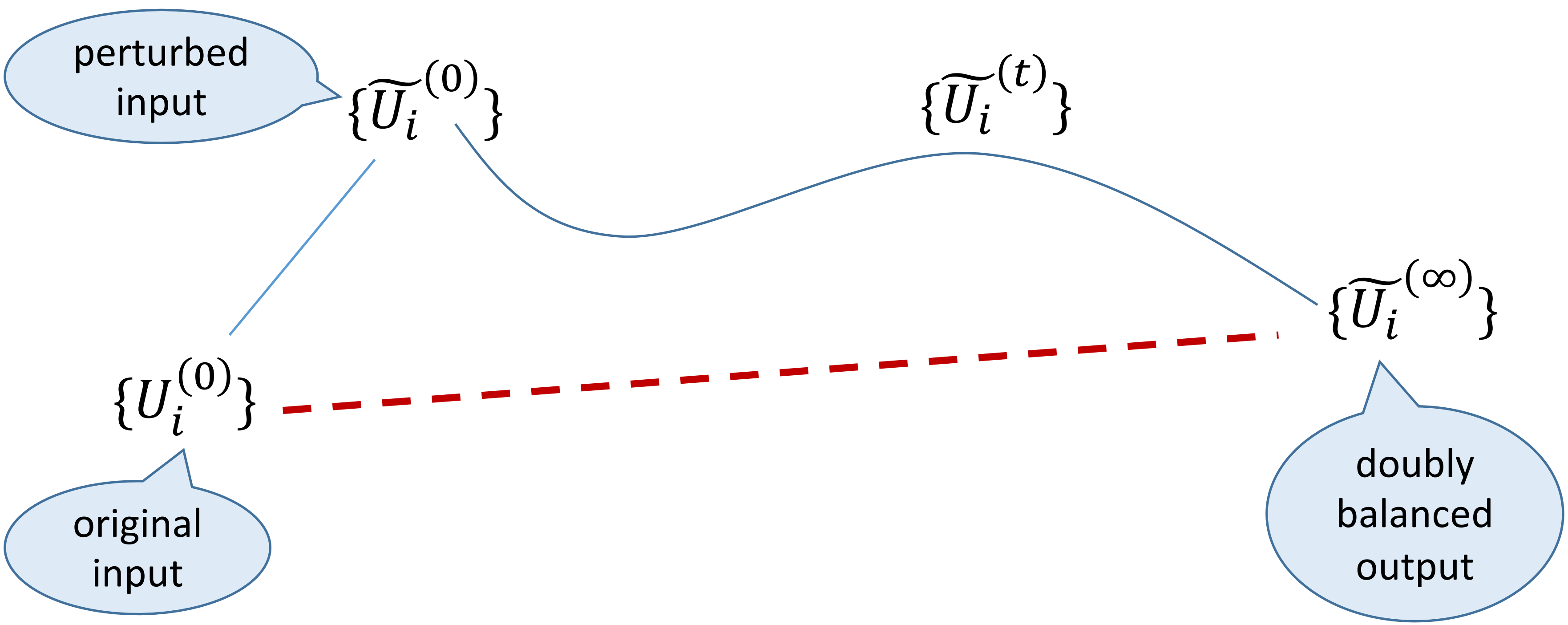
In Part III, we prove that $f(d,n,\Delta) \leq d^c\sqrt{\Delta}$ in “perturbed” instances.

Remark: Smoothed analysis only works in the frame setting, not (yet) in the operator setting.

Smoothed Analysis

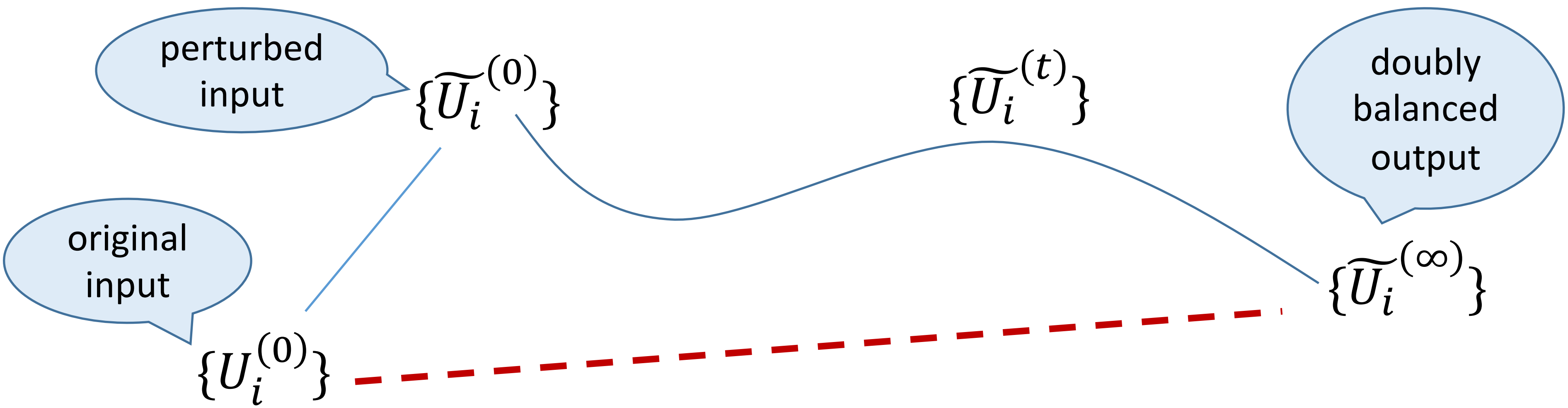
Intuition: operators with small capacity are rare.

Idea: perturb an operator, and apply the dynamical system.



Plan

movement in dynamical system $\leq f\left(d, n, \Delta(\widetilde{U})\right)$



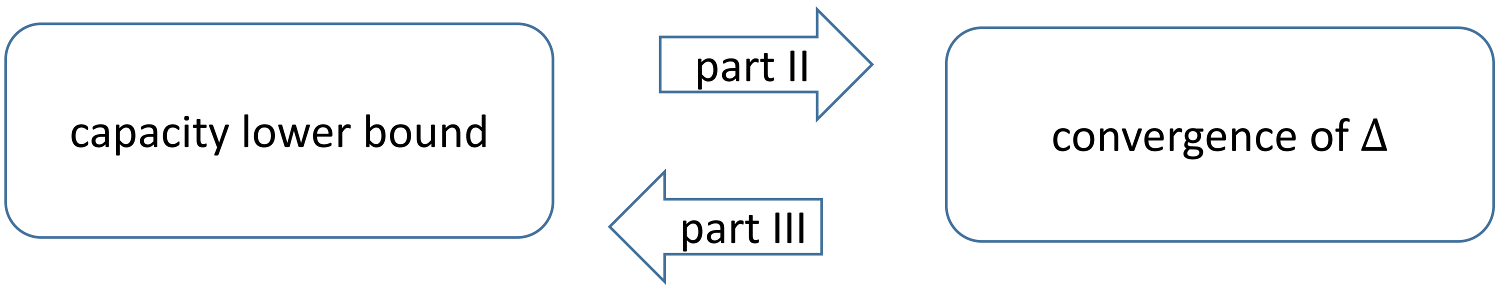
1. Upper bound the perturbation movement, i.e. $dist^2\left(\left\{U_i^{(0)}\right\}, \left\{\widetilde{U_i}^{(0)}\right\}\right)$.

2. Error won't increase too much, i.e. $\Delta(\widetilde{U}) \approx \Delta(U)$.

3. Improved capacity in perturbed instances, i.e. $f(d, n, \Delta) \leq d^c\sqrt{\Delta}$.

New Method in Capacity Lower Bound

New method: We use our dynamical system to bound matrix capacity.



$$\boxed{-\frac{d}{dt}\Delta \geq \mu\Delta \quad \Rightarrow \quad \mathrm{cap} \geq s - \frac{\Delta}{\mu}.}$$

So we need to show the fast convergence for the perturbed instances.

Outline

Part I: Paulsen problem

- Motivation from frame theory

Part II: Continuous operator scaling

- Operator scaling, alternating algorithm, reduction
- Analysis of dynamical system

Part III: Smoothed analysis

- Proof outline, capacity lower bound

Part IV: Discussions

Open Problems

New tools in bounding the mathematical quantities in scaling problems.

1. Bounding the condition number of scaling solutions. *
   - Used in fast algorithms for scaling problems.

2. Bounding (non-uniform) operator capacity
   - Equivalent in bounding Brascamp-Lieb constants.

3. Smoothed analysis of operator scaling

4. Generalization to Tensor scaling etc.