Rigidity of near-optimal superdense coding protocols

by

Xingyu Zhou

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Combinatorics and Optimization (Quantum Information)

Waterloo, Ontario, Canada, 2023

© Xingyu Zhou 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Rigidity in quantum information theory refers to the stringent constraints underlying optimal or near-optimal performance in certain quantum tasks. This property plays a crucial role in verifying untrusted quantum devices and holds significance for secure quantum protocols. Previous work by Nayak and Yuen [18] demonstrated that all optimal superdense coding protocols are locally equivalent to the canonical Bennett-Wiesner protocol. For higher-dimensional superdense coding protocols, [18] showed they may exist only in a relaxed form, and Farkas, Kaniewski and Nayak [6] showed there are infinitely many dimensions $d \geq 4$ such that the rigidity does not exist even in the relaxed form.

Our work is dedicated to establishing the rigidity properties of near-optimal superdense coding protocols. Specifically, we explore scenarios where Alice can employ finite but arbitrary ancilla qubits for encoding, Bob can perform positive operator-valued measure (POVM) for decoding and can answer with error. In such contexts, we prove that any near-optimal superdense coding must be locally equivalent to a superdense coding protocol close to the canonical Bennett-Wiesner protocol.

In the search for extending the result to higher dimensional superdense coding protocols, we find a method to orthogonalize any two unitary matrices in the same space. However, the question of whether it is feasible to orthogonalize more than two $d \times d$ unitary matrices when $d > 2$ remains an intriguing yet unresolved matter.

Acknowledgements

I am profoundly grateful to my dedicated supervisor, Ashwin Nayak, for his unwavering guidance, mentorship, and commitment to my academic growth. His expertise and support have been instrumental in shaping the direction and quality of my research.

I also want to express my deepest gratitude to my mother, Ying Jiang, whose boundless belief in my abilities and unwavering encouragement provided the emotional strength I needed to persevere through the challenges of this demanding academic pursuit.

I am genuinely grateful for Shalev Ben-David and William Slofstra's critical feedback and constructive insights on this thesis.

I would also like to acknowledge my feline companion Yuanbao for his soothing presence.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

List of Figures vii

1 Introduction and preliminaries 1

1.1 Introduction 1

1.2 Preliminary 2

1.3 Our results 3

2 Distinguishing n quantum states in $\mathbb{C}^n$ 7

3 Rigidity of superdense coding on the initial state 15

4 Discussion on Alice's ancilla qubits 20

4.1 Structure of the unitary matrices applied by Alice 21

4.2 Eliminating Alice's ancilla 23

5 Rigidity of near-optimal superdense coding protocols 27

5.1 Orthogonalizing 2×2 unitary matrices 27

5.2 Rigidity of near-optimal superdense coding protocols 36

6 Orthogonalizing two unitary matrices in general **39**

6.1 Orthogonalizing two unitary operators by "rotating" eigenvalues 39

6.2 Rotating vectors in $\mathbb{R}^2$ to sum up to $\vec{0}$ 41

6.3 Orthogonalizing two unitary matrices 61

References **64**

List of Figures

1.1 An illustration of Definition 1.1. The circuit is from [18]. 4

6.1 $\arg(\rho' \exp(\mathrm{i}\theta) - s) > \arg(\rho \exp(\mathrm{i}\theta) - s)$ for $\rho > \rho' > 0$. 47

6.2 $\arg(\rho \exp(\mathrm{i}\theta') - s) > \arg(\rho \exp(\mathrm{i}\theta) - s)$ for $\pi \geq \theta' > \theta \geq 0$. 48

6.3 $\omega = \arccos\left(\dfrac{2\cos^2(\sqrt{s}) - s}{r}\right) \leq \arccos\left(\dfrac{2\cos^2(\sqrt{s}) - s}{2}\right)$. 49

6.4 $|\rho \exp(\mathrm{i}|\theta|)| = |\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|$. 50

6.5 Illustration of each while-loop iteration in Algorithm 2. 59

Chapter 1

Introduction and preliminaries

1.1 Introduction

Rigidity in quantum information theory refers to a fascinating property observed in certain quantum systems and tasks. When we say a protocol is "rigid," it means that achieving optimal or even near-optimal performance by that protocol demands highly specific and strict constraints. One example is the rigidity property of the CHSH game [3]. The canonical optimal quantum protocol that wins the game with probability $\cos\left(\frac{\pi}{8}\right)^2 \approx 0.85$ requires the two parties (Alice and Bob) playing the game to share an Einstein-Podolsky-Rosen (EPR) state $\left(\text{i.e., } \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)\right)$, and they each measure in a basis depending on the received bit. Miller and Shi [16] showed that any protocol that achieves a winning probability close to the optimal probability $\cos\left(\frac{\pi}{8}\right)^2$ is, up to local isometries, close to the canonical optimal strategy. Several other works demonstrate the optimal strategy is often uniquely characterized, or in the case of near-optimal strategies, there is little room for variation.

The concept of rigidity is essential because it provides a clear understanding of the limitations and possibilities in quantum information processing tasks. Moreover, it becomes particularly valuable when dealing with untrusted quantum devices since it allows us to verify the correctness of a quantum system based solely on its observable behaviour, without having to trust the intricate mechanisms of the device. This property plays a vital role in building secure and reliable quantum protocols and applications.

The investigation into rigidity in quantum information processing traces its origins to the works of Mayers and Yao [14, 15], who laid the foundation for the concept of device-independent quantum cryptography. The central notion behind their study is that classical users can ascertain the correctness of untrusted quantum hardware by checking merely the classical input-output statistics of the protocol versus that of an optimal non-local game. Subsequently, non-local game rigidity has been used in diverse domains such as quantum cryptography [24, 4], complexity theory [9], and quantum information [22]. However, very few works consider the rigidity properties other than non-local games. Notable ones include the rigidity of quantum random access codes [23, 5], and rigidity of superdense coding [18, 6].

In this work, we focus on proving the rigidity properties of near-optimal superdense coding protocols. In 1992, Bennett and Wiesner [1] proposed the superdense coding protocol: Alice and Bob each initially own one qubit of an EPR state. Then, Alice is given a classical message $i \in [4]$. Depending on i, Alice applies a Pauli operator to her part of the shared initial state and sends that qubit to Bob. Bob then performs a projective measurement to perfectly distinguish the state and recover the classical message i. Nayak and Yuen [18] showed that any optimal superdense coding protocol is locally equivalent to the superdense coding protocol by Bennett and Wiesner. For higher-dimensional superdense coding protocols, for $d > 2$, Nayak and Yuen [18] showed there are multiple non-equivalent superdense coding schemes, even if maximally entangled states of local dimension d are used, each given by an orthogonal unitary basis (OUB). So rigidity may only hold up to the choice of an OUB. For $d > 3$, Farkas, Kaniewski and Nayak [6] showed if entangled states of local dimension larger than d are allowed, there are schemes which are provably not locally equivalent to those that use the additional entanglement as shared randomness and an OUB of $d \times d$ matrices depending on the randomness. A natural question to ask is whether there are rigidity properties when the superdense coding is performing non-optimally.

In section 1.2, we introduce notations for the rest of this work. In section 1.3, we formally explain the settings of the problem, summarize the rigidity properties of near-optimal superdense coding protocols, and show the key techniques used in the proofs.

1.2 Preliminary

In this work, for standard quantum computing notations, we refer readers to [19] for reference. Most other notations are defined at their first use, and in their successive uses, we believe the readers can find the definition within a page or two. We list the remaining notations that are used throughout this work:

For a complex number $c \coloneqq \rho \exp(\mathrm{i}\theta)$ where $\rho \geq 0$ and $\theta \in (-\pi, \pi]$, denote its real part as $\mathrm{Re}\{c\}$ and its imaginary part as $\mathrm{Im}\{c\}$. Define $\arg(c) \coloneqq \theta$ when $\rho > 0$, and define $\arg(0) \coloneqq 0$. For non-zero $x, y \in \mathbb{C}$, define $\angle(x, y) \coloneqq \arccos\left(\frac{\mathrm{Re}\{xy^*\}}{|x||y|}\right) \in [0, \pi]$.

Given a complex Euclidean space $\mathcal{X}$, let $\dim(\mathcal{X})$ denote its dimension. Let $\mathcal{H}$ represent a Hilbert space. If we put a superscript over $\mathcal{H}$, that superscript represents the dimension of the Hilbert space, and if we put a subscript below $\mathcal{H}$, that subscript represents a specific sub-space's label.

Define the maximally entangled state with local dimension d as $|\Phi_d\rangle \coloneqq \frac{1}{\sqrt{d}} \sum_{i \in [d]} |i\rangle \otimes |i\rangle \in \mathcal{H}^d \otimes \mathcal{H}^d$ where the $\otimes$ represents the tensor product.

Let $\mathbb{I}$ be the identity matrix. A subscript of a lower-case letter or a number represents the dimension of the identity matrix (i.e. $\mathbb{I}_d$ is a $d \times d$ identity matrix), and a subscript of an upper-case letter represents the label of the space the identity matrix is acting on. Let $\mathrm{diag}(x_1, \cdots, x_n)$ represent an $n \times n$ diagonal matrix with $x_1, \cdots, x_n$ along the main diagonal, and 0 everywhere else. Let $\mathcal{L}(\mathcal{X})$ denote the set of all $\dim(\mathcal{X}) \times \dim(\mathcal{X})$ complex matrices. Let $\mathcal{U}(d)$ denote the set of all $d \times d$ unitary matrices, and let $\mathcal{SU}(d)$ denote the set of all $d \times d$ unitary matrices with determinant 1. For any $\mathcal{S} \subset \mathbb{R}_{\geq 0}$, define $\mathcal{U}_{\mathcal{S}}(d) \coloneqq \{U : U \in \mathbb{C}^{d \times d}, \exists s \in \mathcal{S}, UU^\dagger = U^\dagger U = s^2 \mathbb{I}_d\}$.

For two matrices $M, N \in \mathbb{C}^{n \times n}$, define $\langle M, N \rangle \coloneqq \frac{1}{n} \mathrm{Tr}\big(M^\dagger N\big)$. Define the Frobenius norm $\|M\|_{\mathrm{F}} \coloneqq \sqrt{\mathrm{Tr}(M^\dagger M)}$. If M and N are positive semi-definite operators on the same space, define the fidelity as $\mathrm{F}(M, N) \coloneqq \mathrm{Tr}\left(\sqrt{\sqrt{M} N \sqrt{M}}\right)$.

1.3 Our results

We use Definition 2.1 from [18] to define a superdense protocol. We restate it as follows:

Definition 1.1. *(Superdense coding protocol). We say $(\tau, (U_i))$ is a (d, ϵ)-superdense coding protocol if the following conditions are met: Let $\mathcal{H}_A \coloneqq \mathcal{H}_{A'} \otimes \mathcal{H}_{A''}$ and $\mathcal{H}_B$ be a finite-dimensional Hilbert space with $\dim(\mathcal{H}_{A''}) = \dim(\mathcal{H}_B) = d$, and $\dim(\mathcal{H}_{A'})$ is finite but arbitrary. Alice and Bob initially share a density matrix $\tau \in \mathcal{L}(\mathcal{H}_A \otimes \mathcal{H}_B)$. Alice receives* *an input $i \in [d^2]$ which is given uniformly at random, and then performs unitary operator $U_i \in \mathcal{L}(\mathcal{H}_A)$ on τ. Then, Alice sends her qubit(s) A'' to Bob. At this point, Bob has the state $\rho_i \coloneqq \mathrm{Tr}_{A'}((U_i \otimes \mathbb{I}_B)\tau(U_i^\dagger \otimes \mathbb{I}_B))$. Bob uses a POVM $(M_i)_{i\in[d^2]}$ such that*

$$\frac{1}{d^2}\sum_{i=1}^{d^2} \mathrm{Tr}(\rho_i M_i) \geq 1 - \epsilon.$$

In addition, we say $(\tau, (U_i))$ is a (d, ϵ)-worst case superdense coding protocol if

$$\min_{i\in[d^2]}\{\mathrm{Tr}(\rho_i M_i)\} \geq 1 - \epsilon.$$

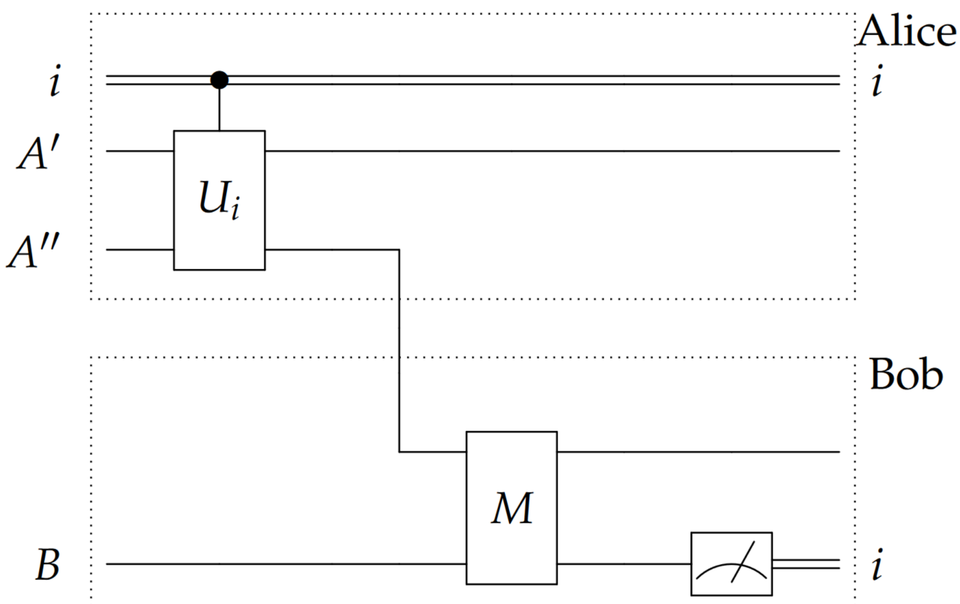

Figure 1.1: An illustration of Definition 1.1. The circuit is from [18].

See Figure 1.1 for illustration. Notice that the setting for Bennett and Wiesner protocol is a special case when $d = 2$, $\epsilon = 0$, $\dim(\mathcal{H}_{A'}) = 0$, and Bob uses a projective measurement. Their setting is extended in Definition 1.1 so that

- We allow a higher dimensional shared entanglement.

- We allow Bob to answer with an error ϵ.
- We allow Alice to use ancilla qubits.
- We allow Bob to perform positive operator-valued measure (POVM).

In Definition 1.1, we allow Alice to have ancilla qubits with an arbitrary finite dimension. However, on Bob's side, the dimension ($\dim(\mathcal{H}_B)$) is restricted to d. This is because we would like to restrict the amount of the shared entanglement between Alice and Bob while not limiting the power of Bob's measurement. Fundamentally, using a POVM can be seen as performing a projective measurement in a larger space. Thus, Bob effectively has ancilla qubits that are used exclusively for implementing the POVM but not for shared entanglement.

By adapting Definition 2.4 in [18], we define the local equivalence as follows:

Definition 1.2. *(Local equivalence). Suppose $(\tau, (U_i))$ and $(\tau', (V_i))$ are both (d, ϵ)-superdense coding protocols. They are locally equivalent if and only if there exists a unitary matrix $W \in \mathcal{L}(\mathcal{H}_A)$ such that*

$$\tau' = (W \otimes \mathbb{I}_B)\tau(W^\dagger \otimes \mathbb{I}_B),$$

and

$$V_i = U_i W^\dagger, \qquad \forall i \in [d^2].$$

Also, define the closeness between unitary matrices U and V when acting on density matrix τ and then tracing out the A part as

$$\mathrm{S}_{\tau,A}(U, V) := \mathrm{F}(\mathrm{Tr}_A(U\tau U^\dagger), \mathrm{Tr}_A(V\tau V^\dagger)).$$

Intuitively, if $\mathrm{S}_{\tau,A}(U, V)$ is high, then U and V are interchangeable with little effect to a superdense coding protocol with initial state τ.

In this work, we prove the following main theorem:

Theorem 1.3. *Any (d, ϵ)-superdense coding protocol $(\tau', (V_i))$ is locally equivalent to (d, ϵ)-superdense coding protocol $(\tau, (U_i))$, such that there exists a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ and unitary matrices $W_i \in \mathcal{L}(\mathcal{H}_{A''})$ with*

$$\mathrm{F}\left(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|\right) \geq 1 - (21 + 6\sqrt{6})\epsilon = 1 - O(\epsilon),$$

and

$$\frac{1}{d^2}\sum_{i=1}^{d^2} \mathrm{S}_{\tau,A'}(U_i \otimes \mathbb{I}_B, \mathbb{I}_{A'} \otimes W_i \otimes \mathbb{I}_B) \geq 1 - (106 + 28\sqrt{6})\epsilon = 1 - O(\epsilon).$$

Notice the above theorem works for any dimension $d \geq 2$. In the case when $d = 2$, we further get:

Theorem 1.4. *There exists $c > 0$ such that any $(2, \epsilon)$-superdense coding protocol $(\tau', (V_i))$ with $\epsilon < c$ is locally equivalent to $(2, \epsilon)$-superdense coding protocol $(\tau, (U_i))$ which satisfies the following properties: there exists a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ and pair-wise orthogonal $(\tilde{W}_i)_{i \in [4]} \subset \mathcal{U}(2)$ (i.e., $\langle \tilde{W}_i, \tilde{W}_j \rangle = \delta_{ij}$, and δ_{ij} is the Kronecker delta), such that*

$$\mathrm{F}\left(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|\right) \geq 1 - (21 + 6\sqrt{6})\epsilon = 1 - O(\epsilon),$$

and

$$\frac{1}{4} \sum_{i=1}^{4} \mathrm{S}_{\tau, A'}(U_i \otimes \mathbb{I}_B, \mathbb{I}_{A'} \otimes \tilde{W}_i \otimes \mathbb{I}_B) \geq 1 - (394 + 108\sqrt{6})\epsilon = 1 - O(\epsilon).$$

The proof of the stricter rigidity results when $d = 2$ relies heavily on being able to nicely orthogonalize 2×2 unitary matrices. This is done by finding a Hilbert space isomorphism between real and non-negative scalings of $\mathcal{SU}(2)$ and $\mathbb{R}^4$. We can then make use of vector orthogonalization algorithms to orthogonalize 2×2 unitary matrices. When $d > 2$, such a nice property does not hold. In the search for extending the previous result, we found a way to orthogonalize any two $d \times d$ unitary matrices while perturbing one only slightly:

Theorem 1.5. *Suppose we have $U_1, U_2 \in \mathcal{U}(d)$ for any $d \geq 2$ such that*

$$|\langle U_1, U_2 \rangle| = \left| \frac{1}{d} \operatorname{Tr}\left(U_1^\dagger U_2\right) \right| \leq \epsilon,$$

then, there exists $U \in \mathcal{U}(d)$ such that $\langle U_1, UU_2 \rangle = 0$ and

$$\|UU_2 - U_2\|_{nhs}^2 = \|U - \mathbb{I}_d\|_{nhs}^2 \leq 196\epsilon = O(\epsilon),$$

where $\|M\|_{nhs} := \sqrt{\frac{1}{d} \operatorname{Tr}(M^\dagger M)}$, for any $M \in \mathbb{C}^{d \times d}$.

The key idea is reducing the problem into "rotating" the eigenvalues of $U_1^\dagger U_2$, or into rotating unit vectors in $\mathbb{R}^2$, so that the vectors sum up to 0. We additionally ensure that the total rotation angle is small. An upper bound on the total rotation angle is shown by analyzing an algorithm. The question of whether it is feasible to similarly orthogonalize more than two unitary matrices when $d > 2$ remains an intriguing yet unresolved matter.

Chapter 2

Distinguishing n quantum states in $\mathbb{C}^n$

The goal of this chapter is to prove a useful result in distinguishing n quantum states in $\mathbb{C}^n$, and this will help to prove Theorem 3.2 in chapter 3.

The setting for the main result of this chapter (Lemma 2.6) is as follows: suppose we sample a density matrix τ from $(\tau_i)_{i\in[n]} \subset \mathcal{L}(\mathbb{C}^n)$ uniformly at random. Suppose there exists a POVM $(M_i)_{i\in[n]} \subset \mathcal{L}(\mathbb{C}^n)$ such that we can distinguish τ with an average success probability at least $1-\epsilon$ $\left(\text{i.e. } \frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}(M_i\tau_i) \geq 1-\epsilon\right)$. Informally, Lemma 2.6 proves that the τ_i's on average are close to pure states, and the POVM is close to a projective measurement.

Before proving Lemma 2.6, we first show a few other results about the closeness of quantum states, that is: suppose we have three density matrices ρ_1, ρ_2 and ρ_3. If the fidelity between ρ_1 and ρ_2 is high, and the fidelity between ρ_2 and ρ_3 is high, then we show the fidelity between ρ_1 and ρ_3 is high. The results here can be seen as an adaption of Lemma 3.3 in [17].

Lemma 2.1. *Suppose there are density matrices ρ_1, ρ_2, ρ_3 such that the fidelity $\mathrm{F}(\rho_1,\rho_2)^2 \geq 1-\epsilon_1$ and $\mathrm{F}(\rho_2,\rho_3)^2 \geq 1-\epsilon_2$, then $\mathrm{F}(\rho_1,\rho_3)^2 \geq 1-\epsilon_1-\epsilon_2-2\sqrt{\epsilon_1\epsilon_2}$.*

Proof. Suppose $\mathrm{F}(\rho_1,\rho_2)^2 = 1-\delta_1 \geq 1-\epsilon_1$ and $\mathrm{F}(\rho_2,\rho_3)^2 = 1-\delta_2 \geq 1-\epsilon_2$. By [20, 8, 21], the function $\mathrm{C}(\rho,\delta) \coloneqq \sqrt{1-\mathrm{F}(\rho,\delta)^2}$ is a metric.

By the triangle inequality of a metric,

$$\mathrm{C}(\rho_1,\rho_3) \leq \mathrm{C}(\rho_1,\rho_2) + \mathrm{C}(\rho_2,\rho_3)$$

$$
\begin{aligned}
&= \sqrt{1-(1-\delta_1)} + \sqrt{1-(1-\delta_2)} \\
&= \sqrt{\delta_1} + \sqrt{\delta_2}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{F}(\rho_1, \rho_3)^2 =& 1 - \mathrm{C}(\rho_1, \rho_3)^2 \\
\geq& 1 - (\sqrt{\delta_1} + \sqrt{\delta_2})^2 \\
=& 1 - \delta_1 - \delta_2 - 2\sqrt{\delta_1 \delta_2} \\
\geq& 1 - \epsilon_1 - \epsilon_2 - 2\sqrt{\epsilon_1 \epsilon_2}.
\end{aligned}
$$

□

Corollary 2.2. *Suppose there are density matrices* ρ_1, ρ_2, ρ_3 *such that* $\mathrm{F}(\rho_1, \rho_2)^2 \geq 1 - \epsilon_1$ *and* $\mathrm{F}(\rho_2, \rho_3)^2 \geq 1 - \epsilon_2$*, then* $\mathrm{F}(\rho_1, \rho_3)^2 \geq 2\,\mathrm{F}(\rho_1, \rho_2)^2 + 2\,\mathrm{F}(\rho_2, \rho_3)^2 - 3$.

Proof. Suppose $\mathrm{F}(\rho_1, \rho_2)^2 = 1 - \delta_1$ and $\mathrm{F}(\rho_2, \rho_3)^2 = 1 - \delta_2$. As proved earlier,

$$
\begin{aligned}
\mathrm{F}(\rho_1, \rho_3)^2 = 1 - \mathrm{C}(\rho_1, \rho_3)^2 \geq& 1 - \delta_1 - \delta_2 - 2\sqrt{\delta_1 \delta_2} \\
\geq& 1 - 2\delta_1 - 2\delta_2 \\
=& 2\,\mathrm{F}(\rho_1, \rho_2)^2 + 2\,\mathrm{F}(\rho_2, \rho_3)^2 - 3,
\end{aligned}
$$

where the second inequality is by Cauchy-Schwartz inequality. □

Corollary 2.3. *Suppose there are arbitrary pure states* $|a_1\rangle, |a_2\rangle, |b\rangle$*. Then,* $|\langle a_1|a_2\rangle|^2 \geq 2|\langle a_1|b\rangle|^2 + 2|\langle a_2|b\rangle|^2 - 3$.

Proof. The square of the inner product between two pure states equals the square of fidelity of corresponding density matrices of the two pure states. Thus, this is a direct consequence of Corollary 2.2. □

Lemma 2.4. *Suppose there are density matrices* ρ_1, ρ_2, ρ_3 *such that* $\mathrm{F}(\rho_1, \rho_2) \geq 1 - \epsilon_1$ *and* $\mathrm{F}(\rho_2, \rho_3) \geq 1 - \epsilon_2$*, then* $\mathrm{F}(\rho_1, \rho_3) \geq 1 - \epsilon_1 - \epsilon_2 - 2\sqrt{\epsilon_1 \epsilon_2}$.

Proof. The function $\mathrm{B'}(\rho, \delta) \coloneqq \sqrt{1 - \mathrm{F}(\rho, \delta)}$ is also a metric since it is the standard Bures metric multiplied by a $\dfrac{1}{\sqrt{2}}$ factor. The remaining of this proof is almost identical to the proof of Lemma 2.1. By triangle inequality,

$$
\mathrm{B'}(\rho_1, \rho_3) \leq \mathrm{B'}(\rho_1, \rho_2) + \mathrm{B'}(\rho_2, \rho_3) \leq \sqrt{1-(1-\epsilon_1)} + \sqrt{1-(1-\epsilon_2)} = \sqrt{\epsilon_1} + \sqrt{\epsilon_2}.
$$

Therefore,

$$\mathrm{F}(\rho_1, \rho_3) = 1 - \mathrm{B'}(\rho_1, \rho_3)^2 \geq 1 - \epsilon_1 - \epsilon_2 - 2\sqrt{\epsilon_1 \epsilon_2}.$$

□

Corollary 2.5. *Suppose there are density matrices ρ_1, ρ_2, ρ_3 such that $\mathrm{F}(\rho_1, \rho_2) \geq 1 - \epsilon_1$ and $\mathrm{F}(\rho_2, \rho_3) \geq 1 - \epsilon_2$, then $\mathrm{F}(\rho_1, \rho_3) \geq 2\,\mathrm{F}(\rho_1, \rho_2) + 2\,\mathrm{F}(\rho_2, \rho_3) - 3$.*

Proof. Suppose $\mathrm{F}(\rho_1, \rho_2) = 1 - \delta_1$ and $\mathrm{F}(\rho_2, \rho_3) = 1 - \delta_2$. As proved earlier,

$$\begin{aligned}\mathrm{F}(\rho_1, \rho_3) = 1 - \mathrm{B'}(\rho_1, \rho_3)^2 &\geq 1 - \delta_1 - \delta_2 - 2\sqrt{\delta_1 \delta_2} \\ &\geq 1 - 2\delta_1 - 2\delta_2 \\ &= 2\,\mathrm{F}(\rho_1, \rho_2) + 2\,\mathrm{F}(\rho_2, \rho_3) - 3.\end{aligned}$$

□

Now, we prove the main result of this chapter.

Lemma 2.6. *Suppose we sample a density matrix τ from $(\tau_i)_{i \in [n]} \subset \mathcal{L}(\mathbb{C}^n)$ uniformly at random. Suppose there exists a POVM $(M_i)_{i \in [n]} \subset \mathcal{L}(\mathbb{C}^n)$ such that we can distinguish τ with an average success probability at least $1 - \epsilon$ $\left(\text{i.e. } \frac{1}{n} \sum_{i=1}^{n} \mathrm{Tr}(M_i \tau_i) \geq 1 - \epsilon\right)$. Then, we can construct an orthonormal basis $(|\zeta_i\rangle)_{i \in [n]}$ of $\mathbb{C}^n$ and pure states $(|\psi_i\rangle)_{i \in [n]} \subset \mathbb{C}^n$ such that $\frac{1}{n} \sum_{i=1}^{n} \langle \zeta_i | M_i | \zeta_i \rangle \geq 1 - 2\epsilon$, $\frac{1}{n} \sum_{i=1}^{n} \langle \psi_i | \tau_i | \psi_i \rangle \geq 1 - 2\epsilon$, and $\frac{1}{n} \sum_{i=1}^{n} |\langle \zeta_i | \psi_i \rangle|^2 \geq 1 - 12\epsilon$.*

Proof. Suppose the spectral decomposition of M_i is given by $M_i = \sum_{j=1}^{n} \lambda_{i,j} |\phi_{i,j}\rangle\langle\phi_{i,j}|$. Without loss of generality, assume $\lambda_{i,j} \geq \lambda_{i,k}$ for all $i, j, k \in [n]$ and $j < k$.

Define $p_{i,j} := \mathrm{Tr}(\tau_i |\phi_{i,j}\rangle\langle\phi_{i,j}|)$. Since τ_i is positive semi-definite, $p_{i,j} \geq 0$ for all $i, j \in [n]$. Also, $\sum_{j=1}^{n} |\phi_{i,j}\rangle\langle\phi_{i,j}| = \mathbb{I}_n$, so $\sum_{j=1}^{n} p_{i,j} = \mathrm{Tr}\left(\tau_i \sum_{j=1}^{n} |\phi_{i,j}\rangle\langle\phi_{i,j}|\right) = \mathrm{Tr}(\tau_i \mathbb{I}_n) = \mathrm{Tr}(\tau_i) = 1$ for all $i \in [n]$, and $(p_{i,j})_{j \in [n]}$ forms a probability distribution over $[n]$.

Since we can distinguish τ with average success probability at least $1 - \epsilon$,

$$n(1 - \epsilon) \leq \sum_{i=1}^{n} \mathrm{Tr}(\tau_i M_i) = \sum_{i=1}^{n} \mathbb{E}_{p_i} \lambda_{i,\cdot} \leq \sum_{i=1}^{n} \lambda_{i,1}. \tag{2.1}$$

Also, as $\sum_{i=1}^{n} M_i = \mathbb{I}_n$,

$$\begin{aligned} \sum_{i=1}^{n} \lambda_{i,1} + \sum_{i=1}^{n}\sum_{j=2}^{n} \lambda_{i,j} &= \operatorname{Tr}\left(\sum_{i=1}^{n} M_i\right) = \operatorname{Tr}(\mathbb{I}_n) = n \\ \implies \qquad \sum_{i=1}^{n}\sum_{j=2}^{n} \lambda_{i,j} &\le n - n(1-\epsilon) = n\epsilon. \end{aligned}$$

At this point, we have proved that the largest eigenvalues of the M_i sum to at least $n(1-\epsilon)$, and the sum of the remaining eigenvalues is at most $n\epsilon$. This shows that M_i is close to $|\phi_{i,1}\rangle\langle\phi_{i,1}|$ on average. However, $(|\phi_{i,1}\rangle)_{i\in[n]}$ may not be pairwise orthogonal. To prove that (M_i) is close to a projective measurement, we orthogonalize the states $(|\phi_{i,1}\rangle)_{i\in[n]}$. Define

$$A \coloneqq \sum_{i=1}^{n} \sqrt{\lambda_{i,1}}\, |i\rangle\langle\phi_{i,1}|\,,$$

and

$$N \coloneqq AA^\dagger = \left(\sum_{i=1}^{n} \sqrt{\lambda_{i,1}}\, |i\rangle\langle\phi_{i,1}|\right)\left(\sum_{j=1}^{n} \sqrt{\lambda_{j,1}}\, |\phi_{j,1}\rangle\langle j|\right).$$

Suppose A has singular value decomposition $U\Sigma V^\dagger$ where U, V are unitary and Σ is diagonal and positive semi-definite. We hope to show rows of $UV^\dagger$ are a "good" orthogonalization of $(|\phi_{i,1}\rangle)_{i\in[n]}$. This in the literature is called Löwdin's symmetric orthogonalization [13].

By the singular value decomposition, $N = U\Sigma V^\dagger V\Sigma U^\dagger = U\Sigma^2U^\dagger$. Since

$$A^\dagger A = \sum_{i=1}^{n} \lambda_{i,1}\, |\phi_{i,1}\rangle\langle\phi_{i,1}| \preceq \sum_{i=1}^{n} M_i = \mathbb{I}_n,$$

all eigenvalues of $A^\dagger A$ are less than 1. Since $A^\dagger A$ and $AA^\dagger = N$ have the same non-zero eigenvalues, and they are both positive semi-definite, $0 \preceq N \preceq \mathbb{I}_n$, so $0 \preceq \Sigma^2 \preceq \mathbb{I}_n$ which further implies $0 \preceq \Sigma^2 \preceq \Sigma \preceq \mathbb{I}_n$. Thus,

$$\begin{aligned} \left\|A - UV^\dagger\right\|_{\mathrm{F}}^2 &= \|\Sigma - \mathbb{I}_n\|_{\mathrm{F}}^2 \\ &= \operatorname{Tr}((\Sigma - \mathbb{I}_n)(\Sigma - \mathbb{I}_n)) \\ &= \operatorname{Tr}\big(\Sigma^2\big) - 2\operatorname{Tr}(\Sigma) + \operatorname{Tr}(\mathbb{I}_n) \end{aligned}$$

$$
\begin{aligned}
&\leq \mathrm{Tr}\left(\Sigma^2\right) - 2\,\mathrm{Tr}\left(\Sigma^2\right) + \mathrm{Tr}(\mathbb{I}_n) \\
&= \mathrm{Tr}(\mathbb{I}_n) - \mathrm{Tr}\left(\Sigma^2\right) \\
&= n - \mathrm{Tr}(N) \\
&= n - \sum_{i=1}^{n} \langle i|N|i\rangle \\
&= n - \sum_{i=1}^{n} \lambda_{i,1} \\
&\leq n\epsilon,
\end{aligned}
$$

and the last inequality is simply by Inequality 2.1.

Define $\langle\zeta_i| \coloneqq \langle i|\, UV^\dagger$ which is the i-th row of $UV^\dagger$. Since $UV^\dagger$ is unitary, $(\langle\zeta_i|)_{i\in[n]}$ is an orthonormal basis. Now, we want to bound $\frac{1}{n}\left(\sum_{i=1}^{n} \langle\zeta_i|M_i|\zeta_i\rangle\right)$.

$$
\begin{aligned}
n\epsilon \geq \left\|A - UV^\dagger\right\|_{\mathrm{F}}^2 &= \sum_{i=1}^{n} \left\|\sqrt{\lambda_{i,1}}\,\langle\phi_{i,1}| - \langle\zeta_i|\right\|^2 \\
&= \sum_{i=1}^{n} \left(\lambda_{i,1}\,\|\langle\phi_{i,1}|\|^2 + \|\langle\zeta_i|\|^2 - 2\sqrt{\lambda_{i,1}}\,\mathrm{Re}\{\langle\phi_{i,1}|\zeta_i\rangle\}\right) \\
&= \sum_{i=1}^{n} \left(\lambda_{i,1} + 1 - 2\sqrt{\lambda_{i,1}}\,\mathrm{Re}\{\langle\phi_{i,1}|\zeta_i\rangle\}\right) \\
&\geq \sum_{i=1}^{n} \left(\lambda_{i,1} + 1 - 2\sqrt{\lambda_{i,1}}|\langle\phi_{i,1}|\zeta_i\rangle|\right).
\end{aligned}
$$

The above inequality implies

$$
\begin{aligned}
& \sum_{i=1}^{n} 2\sqrt{\lambda_{i,1}}|\langle\phi_{i,1}|\zeta_i\rangle| \geq n - n\epsilon + \sum_{i=1}^{n} \lambda_{i,1} \geq n(1-\epsilon) + n(1-\epsilon) = 2n(1-\epsilon) \\
\implies \quad & \sum_{i=1}^{n} \sqrt{\lambda_{i,1}}|\langle\phi_{i,1}|\zeta_i\rangle| \geq n(1-\epsilon) \\
\implies \quad & \left(\frac{1}{n}\sum_{i=1}^{n} \sqrt{\lambda_{i,1}}|\langle\phi_{i,1}|\zeta_i\rangle|\right)^2 \geq (1-\epsilon)^2 \geq 1 - 2\epsilon
\end{aligned}
$$

$$\implies \quad \frac{1}{n}\left(\sum_{i=1}^{n}\lambda_{i,1}|\langle\phi_{i,1}|\zeta_i\rangle|^2\right) \geq \left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\lambda_{i,1}}|\langle\phi_{i,1}|\zeta_i\rangle|\right)^2 \geq 1-2\epsilon \tag{2.2}$$

$$\implies \quad \frac{1}{n}\left(\sum_{i=1}^{n}\langle\zeta_i|M_i|\zeta_i\rangle\right) \geq \frac{1}{n}\left(\sum_{i=1}^{n}\lambda_{i,1}|\langle\phi_{i,1}|\zeta_i\rangle|^2\right) \geq 1-2\epsilon, \tag{2.3}$$

where the second last implication (Equation 2.2) is due to convexity of the function $f(x) = x^2$, and the last implication is by the spectral decomposition of M_i. This proves the POVM is close to a projective measurement with projectors $P_i \coloneqq |\zeta_i\rangle\langle\zeta_i|$.

The next step is to show the τ_i's are close to pure states. We first show $\sum_{i=1}^{n} p_{i,1}$ is large:

$$n(1-\epsilon) \leq \sum_{i=1}^{n}\mathrm{Tr}(\tau_i M_i) = \sum_{i=1}^{n}\sum_{j=1}^{n}p_{i,j}\lambda_{i,j} \leq \sum_{i=1}^{n}p_{i,1} + \sum_{i=1}^{n}\sum_{j=2}^{n}\lambda_{i,j}$$

$$\implies \quad \sum_{i=1}^{n}\mathrm{Tr}(\tau_i\,|\phi_{i,1}\rangle\langle\phi_{i,1}|) = \sum_{i=1}^{n}p_{i,1} \geq n(1-\epsilon) - n\epsilon \geq n(1-2\epsilon). \tag{2.4}$$

Then, we use an approach similar to the beginning of this entire proof. Suppose the spectral decomposition of τ_i is $\sum_{j=1}^{n}\eta_{i,j}\,|\psi_{i,j}\rangle\langle\psi_{i,j}|$. Define $q_{i,j} \coloneqq |\langle\psi_{i,j}|\phi_{i,1}\rangle|^2$. Since $(|\psi_{i,j}\rangle)_{j\in[n]}$ is orthonormal for every $i \in [n]$, $q_{i,j}$ is a probability distribution over $j \in [n]$. Denote this probability distribution as q_i. Without loss of generality, assume $\eta_{i,j} \geq \eta_{i,k}$ for all $i, j, k \in [n]$ and $j < k$. Then,

$$n(1-2\epsilon) \leq \sum_{i=1}^{n}p_{i,1} = \sum_{i=1}^{n}\mathbb{E}_{q_i}(\eta_{i,\cdot}) \leq \sum_{i=1}^{n}\eta_{i,1} = \sum_{i=1}^{n}\langle\psi_{i,1}|\tau_i|\psi_{i,1}\rangle\,.$$

We can define $|\psi_i\rangle \coloneqq |\psi_{i,1}\rangle$ for all $i \in [n]$ so that

$$\frac{1}{n}\sum_{i=1}^{n}\langle\psi_i|\tau_i|\psi_i\rangle \geq 1-2\epsilon,$$

which proves that the τ_i's are close to pure states $|\psi_i\rangle$ on average.

The final step is to bound $\frac{1}{n}\sum_{i=1}^{n}|\langle\zeta_i|\psi_i\rangle|^2$. From Equation 2.4,

$$n(1-2\epsilon) \leq \sum_{i=1}^{n}\mathrm{Tr}(\tau_i\,|\phi_{i,1}\rangle\langle\phi_{i,1}|) = \sum_{i=1}^{n}\sum_{j=1}^{n}\eta_{i,j}|\langle\phi_{i,1}|\psi_{i,j}\rangle|^2$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \left|\langle \phi_{i,1}|\psi_{i,1}\rangle\right|^2 + \sum_{i=1}^{n}\sum_{j=2}^{n} \eta_{i,j} \\
&= \sum_{i=1}^{n} \left|\langle \phi_{i,1}|\psi_{i}\rangle\right|^2 + \left(n - \sum_{i=1}^{n} \eta_{i,1}\right) \\
&\leq \sum_{i=1}^{n} \left|\langle \phi_{i,1}|\psi_{i}\rangle\right|^2 + 2n\epsilon,
\end{aligned}
$$

so $\frac{1}{n}\sum_{i=1}^{n} \left|\langle \phi_{i,1}|\psi_i\rangle\right|^2 \geq 1 - 4\epsilon$. Combine this result, Corollary 2.3 and Equation 2.3, we get

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \left|\langle \zeta_i|\psi_i\rangle\right|^2 \geq & \frac{1}{n}\sum_{i=1}^{n} \left(2|\langle \zeta_i|\phi_{i,1}\rangle|^2 + 2|\langle \psi_i|\phi_{i,1}\rangle|^2 - 3\right) \\
\geq & 2(1-2\epsilon) + 2(1-4\epsilon) - 3 \\
= & 1 - 12\epsilon.
\end{aligned}
$$

This finishes the proof of the lemma. □

Another result that will be used later is as follows:

Corollary 2.7. *Suppose we sample a density matrix τ from $(\tau_i)_{i\in[n]} \subset \mathcal{L}(\mathbb{C}^n)$ uniformly at random. Suppose there exists a POVM $(M_i)_{i\in[n]} \subset \mathcal{L}(\mathbb{C}^n)$ such that we can distinguish τ with an average success probability at least $1-\epsilon$ $\left(\text{i.e. } \frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}(M_i\tau_i) \geq 1-\epsilon\right)$. Then, there are pure states $(|\psi_i\rangle)_{i\in[n]}$ such that $\frac{1}{n}\sum_{i=1}^{n} \langle \psi_i|\tau_i|\psi_i\rangle \geq 1-2\epsilon$, and $\frac{1}{n}\sum_{i=1}^{n} \langle \psi_i|M_i|\psi_i\rangle \geq 1-3\epsilon$.*

Proof. Following the notation in Lemma 2.6, $\frac{1}{n}\sum_{i=1}^{n} \langle \psi_i|\tau_i|\psi_i\rangle = \sum_{i=1}^{n} \eta_{i,1} \geq 1 - 2\epsilon$ was proved. Then, by the intermediate results in Lemma 2.6,

$$
\begin{aligned}
n(1-\epsilon) \leq \sum_{i=1}^{n} \mathrm{Tr}(\tau_i M_i) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \eta_{i,j} \langle \psi_{i,j}|M_i|\psi_{i,j}\rangle \\
&\leq \sum_{i=1}^{n} \langle \psi_{i,1}|M_i|\psi_{i,1}\rangle + \sum_{i=1}^{n}\sum_{j=2}^{n} \eta_{i,j}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \langle \psi_i | M_i | \psi_i \rangle + \left(n - \sum_{i=1}^{n} \eta_{i,1} \right) \\
&\leq \sum_{i=1}^{n} \langle \psi_i | M_i | \psi_i \rangle + 2n\epsilon,
\end{aligned}
$$

so $\displaystyle \frac{1}{n} \sum_{i=1}^{n} \langle \psi_i | M_i | \psi_i \rangle \geq 1 - 3\epsilon.$ $\square$

Chapter 3

Rigidity of superdense coding on the initial state

Define the maximally entangled state $|\Phi_d\rangle$ with local dimension d as

$$|\Phi_d\rangle \coloneqq \frac{1}{\sqrt{d}} \sum_{i\in[d]} |i\rangle \otimes |i\rangle \in \mathcal{H}^d \otimes \mathcal{H}^d.$$

For any bipartite pure state $|\Psi\rangle \in \mathcal{H}_X^d \otimes \mathcal{H}_Y^d$, it is maximally entangled if and only if $\mathrm{Tr}_X(|\Psi\rangle\langle\Psi|) = \mathrm{Tr}_Y(|\Psi\rangle\langle\Psi|) = \dfrac{\mathbb{I}}{d}$.

In this chapter, we exploit the structure of the initial state τ of any (d,ϵ)-superdense coding protocol $(\tau, (U_i))$. In Theorem 3.2 at the end of this chapter, we prove there exists a unitary matrix $W \in \mathcal{L}(\mathcal{H}_A)$ and a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ such that the fidelity

$$\mathrm{F}\left(\tau, (W \otimes \mathbb{I}_B)(\sigma \otimes |\Phi_d\rangle\langle\Phi_d|)(W^\dagger \otimes \mathbb{I}_B)\right) \geq 1 - O(\epsilon).$$

We first prove a lemma that will be used in Theorem 3.2:

Lemma 3.1. *For any maximally entangled state $|\Psi\rangle \in \mathcal{H}_X^d \otimes \mathcal{H}_Y^d$, there exists $U \in \mathcal{U}(d)$ such that $|\Psi\rangle = (U \otimes \mathbb{I})\,|\Phi_d\rangle$.*

Proof. Since $|\Psi\rangle$ is maximally-entangled, $\mathrm{Tr}_Y(|\Psi\rangle\langle\Psi|) = \dfrac{\mathbb{I}}{d} = \dfrac{1}{d}\displaystyle\sum_{i=1}^{d} |i\rangle\langle i|$ is maximally-mixed. Any purification of $\mathrm{Tr}_Y(|\Psi\rangle\langle\Psi|)$ in $\mathcal{H}_X^d \otimes \mathcal{H}_Y^d$ is in the form of $\dfrac{1}{\sqrt{d}}\displaystyle\sum_{i=1}^{d} |u_i\rangle \otimes |i\rangle$

for some orthonormal basis $\{|u_i\rangle\}_{i\in[d]}$ of $\mathcal{H}_X^d$. Let $U \coloneqq \sum_{i=1}^{d} |u_i\rangle\langle i|$. U is unitary and $(U \otimes \mathbb{I})\,|\Phi_d\rangle = \frac{1}{\sqrt{d}} \sum_{i\in[d]} |u_i\rangle\langle i|i\rangle \otimes |i\rangle = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} |u_i\rangle \otimes |i\rangle$. □

Then, we prove the main result of this chapter.

Theorem 3.2. *For any $(\tau, (U_i))$ that is a (d,ϵ)-superdense coding protocol, there exists a unitary matrix $W \in \mathcal{L}(\mathcal{H}_A)$ and a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ such that*

$$\mathrm{F}\left(\tau, (W \otimes \mathbb{I}_B)(\sigma \otimes |\Phi_d\rangle\langle\Phi_d|)(W^\dagger \otimes \mathbb{I}_B)\right) \geq 1 - (21 + 6\sqrt{6})\epsilon.$$

Proof. Suppose after Alice applies U_i on $\mathcal{H}_A$, the state on $\mathcal{H}_A \otimes \mathcal{H}_B$ becomes $\tau_i \coloneqq (U_i \otimes \mathbb{I}_B)\tau(U_i^\dagger \otimes \mathbb{I}_B)$, and denote the state on $\mathcal{H}_{A''} \otimes \mathcal{H}_B$ as ρ_i which is $\mathrm{Tr}_{A'}(\tau_i)$. Denote the POVM Bob uses as $(M_i)_{i\in[d^2]}$.

We first prove $\mathrm{Tr}_A(\tau)$ is close to maximally mixed. By Corollary 2.7, there exist pure states $(|\psi_i\rangle)_{i\in[d^2]}$ such that $\frac{1}{n}\sum_{i=1}^{n} \langle\psi_i|\rho_i|\psi_i\rangle \geq 1-2\epsilon$, and $\frac{1}{n}\sum_{i=1}^{n} \langle\psi_i|M_i|\psi_i\rangle \geq 1-3\epsilon$. Then,

$$\begin{aligned}
\mathrm{F}(|\psi_i\rangle\langle\psi_i|\,, M_i) &= \mathrm{Tr}\Big(\sqrt{\sqrt{|\psi_i\rangle\langle\psi_i|}M_i\sqrt{|\psi_i\rangle\langle\psi_i|}}\Big)\\
&= \mathrm{Tr}\Big(\sqrt{|\psi_i\rangle\langle\psi_i|\,M_i\,|\psi_i\rangle\langle\psi_i|}\Big)\\
&= \sqrt{\langle\psi_i|M_i|\psi_i\rangle}\,\mathrm{Tr}\Big(\sqrt{|\psi_i\rangle\langle\psi_i|}\Big)\\
&= \sqrt{\langle\psi_i|M_i|\psi_i\rangle} \in [0,1].
\end{aligned}$$

Let p be the uniform distribution over $[d^2]$. By the monotonicity of fidelity under partial trace, and joint-concavity of fidelity [25],

$$\begin{aligned}
\mathrm{F}\left(\mathbb{E}_p \mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \frac{\mathbb{I}}{d}\right) &= \mathrm{F}(\mathbb{E}_p \mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \mathbb{E}_p \mathrm{Tr}_{A''}(M_i))\\
&\geq \mathbb{E}_p\,\mathrm{F}(\mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \mathrm{Tr}_{A''}(M_i))\\
&\geq \mathbb{E}_p\,\mathrm{F}(|\psi_i\rangle\langle\psi_i|\,, M_i)\\
&\geq \mathbb{E}_p(\mathrm{F}(|\psi_i\rangle\langle\psi_i|\,, M_i)^2)\\
&= \mathbb{E}_p\,\langle\psi_i|M_i|\psi_i\rangle
\end{aligned}$$

$$\geq 1 - 3\epsilon. \tag{3.1}$$

Similarly,

$$\begin{aligned}
\mathrm{F}\left(\mathbb{E}_p \operatorname{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \operatorname{Tr}_A(\tau)\right) =& \mathrm{F}(\mathbb{E}_p \operatorname{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \mathbb{E}_p \operatorname{Tr}_A(\tau_i)) \\
\geq& \mathbb{E}_p \mathrm{F}(\operatorname{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \operatorname{Tr}_{A''}(\rho_i)) \\
\geq& \mathbb{E}_p \mathrm{F}(|\psi_i\rangle\langle\psi_i|, \rho_i) \\
\geq& \mathbb{E}_p (\mathrm{F}(|\psi_i\rangle\langle\psi_i|, \rho_i)^2) \\
=& \mathbb{E}_p \langle\psi_i|\rho_i|\psi_i\rangle \\
\geq& 1 - 2\epsilon. \qquad (3.2)
\end{aligned}$$

Apply Lemma 2.4 on Inequalities 3.1 and 3.2,

$$\mathrm{F}\left(\operatorname{Tr}_A(\tau), \frac{\mathbb{I}}{d}\right) \geq 1 - 3\epsilon - 2\epsilon - 2\sqrt{3\epsilon 2\epsilon} = 1 - (5 + 2\sqrt{6})\epsilon, \tag{3.3}$$

and this finishes the proof that $\operatorname{Tr}_A(\tau)$ is close to maximally mixed.

Since

$$\mathbb{E}_p \langle\psi_i|\rho_i|\psi_i\rangle \geq 1 - 2\epsilon,$$

there exists $k \in [d^2]$ such that

$$\mathrm{F}(\rho_k, |\psi_k\rangle\langle\psi_k|) = \sqrt{\langle\psi_k|\rho_k|\psi_k\rangle} \geq 1 - 2\epsilon.$$

Now, we prove $|\psi_k\rangle\langle\psi_k|$ is close to maximally-entangled. As $\operatorname{Tr}_A(\tau) = \operatorname{Tr}_A(\tau_k) = \operatorname{Tr}_{A''}(\rho_k)$,

$$\begin{aligned}
\mathrm{F}(\operatorname{Tr}_A(\tau), \operatorname{Tr}_{A''}(|\psi_k\rangle\langle\psi_k|)) =& \mathrm{F}(\operatorname{Tr}_{A''}(\rho_k), \operatorname{Tr}_{A''}(|\psi_k\rangle\langle\psi_k|)) \\
\geq& \mathrm{F}(\rho_k, |\psi_k\rangle\langle\psi_k|) \\
\geq& 1 - 2\epsilon. \qquad (3.4)
\end{aligned}$$

Apply Lemma 2.4 on Inequalities 3.3 and 3.4,

$$\begin{aligned}
\mathrm{F}\left(\operatorname{Tr}_{A''}(|\psi_k\rangle\langle\psi_k|), \frac{\mathbb{I}}{d}\right) \geq& 1 - 2\epsilon - (5 + 2\sqrt{6})\epsilon - 2\sqrt{2\epsilon(5 + 2\sqrt{6})\epsilon} \\
=& 1 - (7 + 2\sqrt{6})\epsilon - 2\sqrt{(2 + \sqrt{6})^2}\epsilon
\end{aligned}$$

$$=1-(11+4\sqrt{6})\epsilon.$$

Since $|\psi_k\rangle$ is a purification of $\mathrm{Tr}_{A''}(|\psi_k\rangle\langle\psi_k|)$ on $\mathcal{H}_{A''}\otimes\mathcal{H}_{B''}$, and any purification of $\frac{\mathbb{I}}{d}\in\mathcal{L}(\mathcal{H}_{B''})$ on $\mathcal{H}_{A''}\otimes\mathcal{H}_{B''}$ is maximally-entangled (proved in Lemma 3.1), by Uhlmann's theorem, there exists a unitary matrix $V\in\mathcal{L}(\mathcal{H}_{A''})$ such that

$$\left|\langle\Phi_d|\,(V^\dagger\otimes\mathbb{I}_B)\,|\psi_k\rangle\right| = \mathrm{F}\left(\mathrm{Tr}_{A''}(|\psi_k\rangle\langle\psi_k|),\frac{\mathbb{I}}{d}\right)\geq 1-(11+4\sqrt{6})\epsilon, \tag{3.5}$$

and this finishes the proof that $|\psi_k\rangle\langle\psi_k|$ is close to maximally-entangled.

Then, we prove the main part of the theorem. Let $|\chi_k\rangle$ be a purification of τ_k on $\mathcal{H}_R\otimes\mathcal{H}_A\otimes\mathcal{H}_{B''}$ where $\mathcal{H}_R$ is some extra space for purification. In addition, $|\chi_k\rangle$ is also a purification of $\mathrm{Tr}_{A'}(\tau_k)=\rho_k$. Any purification of $|\psi_k\rangle$ on $\mathcal{H}_R\otimes\mathcal{H}_A\otimes\mathcal{H}_{B''}$ can be expressed as $|\xi\rangle\otimes|\psi_k\rangle$ for some pure state $|\xi\rangle\in\mathcal{H}_R\otimes\mathcal{H}_{A'}$. Again, by Uhlmann's theorem, there exists $\left|\tilde{\xi}\right\rangle\in\mathcal{H}_R\otimes\mathcal{H}_{A'}$ such that

$$\begin{aligned}
\mathrm{F}\left(|\chi_k\rangle\langle\chi_k|\,,\left|\tilde{\xi}\right\rangle\left\langle\tilde{\xi}\right|\otimes|\psi_k\rangle\langle\psi_k|\right) =&\mathrm{F}(\rho_k,|\psi_k\rangle\langle\psi_k|)\\
=&\sqrt{\langle\psi_k|\rho_k|\psi_k\rangle}\\
\geq&\sqrt{1-2\epsilon}\\
\geq&1-2\epsilon.
\end{aligned}$$

If we define $\sigma:=\mathrm{Tr}_R\left(\left|\tilde{\xi}\right\rangle\left\langle\tilde{\xi}\right|\right)\in\mathcal{L}(\mathcal{H}_{A'})$, then

$$\begin{aligned}
\mathrm{F}(\tau_k,\sigma\otimes|\psi_k\rangle\langle\psi_k|) =&\mathrm{F}\left(\mathrm{Tr}_R\left(|\chi_k\rangle\langle\chi_k|\right),\mathrm{Tr}_R\left(\left|\tilde{\xi}\right\rangle\left\langle\tilde{\xi}\right|\otimes|\psi_k\rangle\langle\psi_k|\right)\right)\\
\geq&\mathrm{F}\left(|\chi_k\rangle\langle\chi_k|\,,\left|\tilde{\xi}\right\rangle\left\langle\tilde{\xi}\right|\otimes|\psi_k\rangle\langle\psi_k|\right)\\
\geq&1-2\epsilon.
\end{aligned}\tag{3.6}$$

Then, with V as in Equation 3.5 above,

$$\begin{aligned}
&\mathrm{F}(\sigma\otimes((V\otimes\mathbb{I}_B)\,|\Phi_d\rangle\langle\Phi_d|\,(V^\dagger\otimes\mathbb{I}_B)),\sigma\otimes|\psi_k\rangle\langle\psi_k|)\\
=&\mathrm{F}\left(\mathrm{Tr}_R\left(\left|\tilde{\xi}\right\rangle\left\langle\tilde{\xi}\right|\otimes((V\otimes\mathbb{I}_B)\,|\Phi_d\rangle\langle\Phi_d|\,(V^\dagger\otimes\mathbb{I}_B))\right),\mathrm{Tr}_R\left(\left|\tilde{\xi}\right\rangle\left\langle\tilde{\xi}\right|\otimes|\psi_k\rangle\langle\psi_k|\right)\right)\\
\geq&\mathrm{F}\left(\left|\tilde{\xi}\right\rangle\left\langle\tilde{\xi}\right|\otimes((V\otimes\mathbb{I}_B)\,|\Phi_d\rangle\langle\Phi_d|\,(V^\dagger\otimes\mathbb{I}_B)),\left|\tilde{\xi}\right\rangle\left\langle\tilde{\xi}\right|\otimes|\psi_k\rangle\langle\psi_k|\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{F}((V \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (V^\dagger \otimes \mathbb{I}_B), |\psi_k\rangle\langle\psi_k|) \\
&= \left| \langle\Phi_d| (V^\dagger \otimes \mathbb{I}_B) |\psi_k\rangle \right| \\
&\geq 1 - (11 + 4\sqrt{6})\epsilon. \qquad (3.7)
\end{aligned}
$$

Apply Lemma 2.4 on Inequalities 3.6 and 3.7,

$$
\begin{aligned}
&\mathrm{F}\left(\tau_k, \sigma \otimes ((V \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (V^\dagger \otimes \mathbb{I}_B))\right) \\
\geq& 1 - 2\epsilon - (11 + 4\sqrt{6})\epsilon - 2\sqrt{2\epsilon(11 + 4\sqrt{6})\epsilon} \\
=& 1 - (13 + 4\sqrt{6})\epsilon - 2\sqrt{(4 + \sqrt{6})^2\epsilon} \\
=& 1 - (21 + 6\sqrt{6})\epsilon.
\end{aligned}
$$

If we define the unitary matrix $W \coloneqq U_k^\dagger(\mathbb{I}_{A'} \otimes V) \in \mathcal{L}(\mathcal{H}_A)$, then

$$
\begin{aligned}
&\mathrm{F}\left(\tau, (W \otimes \mathbb{I}_B)(\sigma \otimes |\Phi_d\rangle\langle\Phi_d|)(W^\dagger \otimes \mathbb{I}_B)\right) \\
=& \mathrm{F}\left(\tau, (U_k^\dagger \otimes \mathbb{I}_B)(\sigma \otimes ((V \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (V^\dagger \otimes \mathbb{I}_B)))(U_k \otimes \mathbb{I}_B)\right) \\
=& \mathrm{F}\left((U_k \otimes \mathbb{I}_B)\tau(U_k^\dagger \otimes \mathbb{I}_B), \sigma \otimes ((V \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (V^\dagger \otimes \mathbb{I}_B))\right) \\
=& \mathrm{F}\left(\tau_k, \sigma \otimes ((V \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (V^\dagger \otimes \mathbb{I}_B))\right) \\
\geq& 1 - (21 + 6\sqrt{6})\epsilon.
\end{aligned}
$$

□

Chapter 4

Discussion on Alice's ancilla qubits

For any (d, ϵ)-superdense coding protocol $(\tau, (U_i))$, we learned the structure of the shared initial state τ in chapter 3. Specifically, Theorem 3.2 shows there exists a unitary matrix $W \in \mathcal{L}(\mathcal{H}_A)$ and a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ such that

$$\mathrm{F}\left(\tau, (W \otimes \mathbb{I}_B)(\sigma \otimes |\Phi_d\rangle\langle\Phi_d|)(W^\dagger \otimes \mathbb{I}_B)\right) \geq 1 - (21 + 6\sqrt{6})\epsilon.$$

Without repeatedly referring to the unitary matrix W later, we use Definition 1.2 for local equivalency. Intuitively, up to Alice's local freedom, we can think of the initial state τ as being conjugated by W. When Alice needs to apply U_i to her part of τ, she first applies $W^\dagger$ to cancel the conjugation and then applies U_i. $(\tau, (U_i))$ and $(\tau', (V_i))$ are equivalent because $(U_i W^\dagger)(W \tau W^\dagger)(U_i W^\dagger)^\dagger = U_i \tau U_i^\dagger$ for all $i \in [d^2]$. Thus, the two protocols achieve identical performance when Bob uses the same measurement.

Using this definition, Theorem 3.2 can also be stated as: Any (d, ϵ)-superdense coding protocol $(\tau', (V_i))$ is locally equivalent to (d, ϵ)-superdense coding protocol $(\tau, (U_i))$, such that there exists a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ with

$$\mathrm{F}\left(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|\right) \geq 1 - (21 + 6\sqrt{6})\epsilon.$$

In section 4.1, we get some information about the structure of the unitary matrices $(U_i)_{i \in [d^2]}$ in $(\tau, (U_i))$ mentioned in the above paragraph. This is a natural extension of the previously proved Theorem 3.2. Then, in section 4.2, we discuss the implications of the results from section 4.1. Specifically, we try to answer why we can further restrict Alice to have no ancilla qubits and assume the shared initial state is exactly the maximally entangled state $|\Phi_d\rangle$.

4.1 Structure of the unitary matrices applied by Alice

For simplicity, define the closeness between unitary matrices U and V when acting on density matrix τ and then tracing out the A part as

$$\mathrm{S}_{\tau,A}(U,V) \coloneqq \mathrm{F}(\mathrm{Tr}_A(U\tau U^\dagger), \mathrm{Tr}_A(V\tau V^\dagger)).$$

We get the following result:

Theorem 1.3. *Any (d,ϵ)-superdense coding protocol $(\tau', (V_i))$ is locally equivalent to (d,ϵ)-superdense coding protocol $(\tau, (U_i))$, such that there exists a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ and unitary matrices $W_i \in \mathcal{L}(\mathcal{H}_{A''})$ with*

$$\mathrm{F}\left(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|\right) \geq 1 - (21 + 6\sqrt{6})\epsilon = 1 - O(\epsilon),$$

and

$$\frac{1}{d^2}\sum_{i=1}^{d^2} \mathrm{S}_{\tau,A'}(U_i \otimes \mathbb{I}_B, \mathbb{I}_{A'} \otimes W_i \otimes \mathbb{I}_B) \geq 1 - (106 + 28\sqrt{6})\epsilon = 1 - O(\epsilon).$$

Proof. Let $(\tau, (U_i))$ be a superdense coding protocol that is locally equivalent to $(\tau', (V_i))$ and $\mathrm{F}\left(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|\right)$ is maximized.

We continue to use the notation from Theorem 3.2. Let $\tau_i \coloneqq (U_i \otimes \mathbb{I}_B)\tau(U_i^\dagger \otimes \mathbb{I}_B)$ and $\rho_i \coloneqq \mathrm{Tr}_{A'}(\tau_i)$. Denote the POVM Bob uses as $(M_i)_{i\in[d^2]}$. Let p be the uniform distribution over support $[d^2]$. The following results are proved in Theorem 3.2:

- $\mathrm{F}\left(\mathrm{Tr}_A(\tau), \dfrac{\mathbb{I}}{d}\right) \geq 1 - (5 + 2\sqrt{6})\epsilon.$
- There exists $(|\psi_i\rangle)_{i\in[d^2]}$ such that $\mathbb{E}_p \mathrm{F}(\rho_i, |\psi_i\rangle\langle\psi_i|) \geq 1 - 2\epsilon$.
- $\mathrm{F}\left(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|\right) \geq 1 - (21 + 6\sqrt{6})\epsilon$. This finishes the proof of the first part of this theorem.

For the second part of the proof, since $\mathrm{Tr}_A(\tau) = \mathrm{Tr}_A(\tau_i) = \mathrm{Tr}_{A''}(\rho_i), \forall i \in [d^2]$,

$$\begin{aligned}\mathbb{E}_p \mathrm{F}(\mathrm{Tr}_A(\tau), \mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|)) =& \mathbb{E}_p \mathrm{F}(\mathrm{Tr}_{A''}(\rho_i), \mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|)) \\ \geq& \mathbb{E}_p \mathrm{F}(\rho_i, |\psi_i\rangle\langle\psi_i|)\end{aligned}$$

$$\geq 1 - 2\epsilon.$$

By Corollary 2.5 and linearity of expectation,

$$\begin{aligned}\mathbb{E}_p \,\mathrm{F}\left(\mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \frac{\mathbb{I}}{d}\right) \geq & \mathbb{E}_p \left(2\,\mathrm{F}\left(\mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \mathrm{Tr}_A(\tau)\right) + 2\,\mathrm{F}\left(\mathrm{Tr}_A(\tau), \frac{\mathbb{I}}{d}\right) - 3\right) \\ \geq & 2(1-2\epsilon) + 2(1-(5+2\sqrt{6})\epsilon) - 3 \\ = & 1 - (14+4\sqrt{6})\epsilon.\end{aligned}$$

$|\psi_i\rangle$ is a purification of $\mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|)$ on $\mathcal{H}_{A''} \otimes \mathcal{H}_{B''}$, and any purification of $\frac{\mathbb{I}}{d} \in \mathcal{L}(\mathcal{H}_{B''})$ on $\mathcal{H}_{A''} \otimes \mathcal{H}_{B''}$ is maximally-entangled. By Uhlmann's theorem, there exist unitary matrices $W_i \in \mathcal{L}(\mathcal{H}_{A''})$ such that

$$\left| \langle \Phi_d | \left(W_i^\dagger \otimes \mathbb{I}_B\right) |\psi_i\rangle \right| = \mathrm{F}\left(\mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \frac{\mathbb{I}}{d}\right), \forall i \in [d^2].$$

Therefore, by Corollary 2.5

$$\begin{aligned}& \mathbb{E}_p \,\mathrm{F}\left(\rho_i, (W_i \otimes \mathbb{I}_B)\, |\Phi_d\rangle\langle\Phi_d|\, (W_i^\dagger \otimes \mathbb{I}_B)\right) \\ \geq & \mathbb{E}_p \left(2\,\mathrm{F}\left(\rho_i, |\psi_i\rangle\langle\psi_i|\right) + 2\,\mathrm{F}\left(|\psi_i\rangle\langle\psi_i|, (W_i \otimes \mathbb{I}_B)\, |\Phi_d\rangle\langle\Phi_d|\, (W_i^\dagger \otimes \mathbb{I}_B)\right) - 3\right) \\ \geq & 2(1-2\epsilon) + 2(1-(14+4\sqrt{6})\epsilon) - 3 \\ = & 1 - (32+8\sqrt{6})\epsilon.\end{aligned}$$

As $\mathrm{F}(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|) \geq 1 - (21+6\sqrt{6})\epsilon$,

$$\begin{aligned}& \mathrm{F}(\mathrm{Tr}_{A'}(\tau), |\Phi_d\rangle\langle\Phi_d|) \\ = & \mathrm{F}(\mathrm{Tr}_{A'}(\tau), \mathrm{Tr}_{A'}(\sigma \otimes |\Phi_d\rangle\langle\Phi_d|)) \\ \geq & \mathrm{F}(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|) \\ \geq & 1 - (21+6\sqrt{6})\epsilon,\end{aligned}$$

and we get

$$\frac{1}{d^2} \sum_{i=1}^{d^2} \mathrm{S}_{\tau, A'}(U_i \otimes \mathbb{I}_B, \mathbb{I}_{A'} \otimes W_i \otimes \mathbb{I}_B)$$

$$
\begin{aligned}
&=\mathbb{E}_p \operatorname{F}\left(\operatorname{Tr}_{A'}((U_i \otimes \mathbb{I}_B)\tau(U_i^\dagger \otimes \mathbb{I}_B)), \operatorname{Tr}_{A'}((\mathbb{I}_{A'} \otimes W_i \otimes \mathbb{I}_B)\tau(\mathbb{I}_{A'} \otimes W_i^\dagger \otimes \mathbb{I}_B))\right) \\
&=\mathbb{E}_p \operatorname{F}\left(\rho_i, (W_i \otimes \mathbb{I}_B) \operatorname{Tr}_{A'}(\tau)(W_i^\dagger \otimes \mathbb{I}_B)\right) \\
&\geq \mathbb{E}_p \left(\begin{array}{l} 2\operatorname{F}\left(\rho_i, (W_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (W_i^\dagger \otimes \mathbb{I}_B)\right) + \\ 2\operatorname{F}\left((W_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (W_i^\dagger \otimes \mathbb{I}_B), (W_i \otimes \mathbb{I}_B) \operatorname{Tr}_{A'}(\tau)(W_i^\dagger \otimes \mathbb{I}_B)\right) - 3 \end{array} \right) \\
&=\mathbb{E}_p \left(2\operatorname{F}\left(\rho_i, (W_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (W_i^\dagger \otimes \mathbb{I}_B)\right) + 2\operatorname{F}\left(|\Phi_d\rangle\langle\Phi_d|, \operatorname{Tr}_{A'}(\tau)\right) - 3\right) \\
&\geq 2(1-(32+8\sqrt{6})\epsilon) + 2(1-(21+6\sqrt{6})\epsilon) - 3 \\
&=1-(106+28\sqrt{6})\epsilon,
\end{aligned}
$$

where the second line is by definition of $\operatorname{S}_{\tau,A'}$, and the fourth line is by Corollary 2.5. □

4.2 Eliminating Alice's ancilla

Theorem 1.3 implies Alice's ancilla qubits do not help the protocol. Up to local equivalence, the shared initial state τ is close to a bipartite state $\sigma \otimes |\Phi_d\rangle\langle\Phi_d|$ where the σ is on Alice's ancilla qubits, and the U_i's are essentially only acting on $|\Phi_d\rangle$, which is similar to apply W_i without using the ancilla qubits. However, we have yet to exploit the full structure of the unitary operators (U_i). In the following chapters, we show more structures about the unitary operators assuming Alice has no ancilla qubits, and the shared initial state is exactly $|\Phi_d\rangle\langle\Phi_d|$. The last step of this section is to show why such simplification is valid.

The following version of the union bound will be used in Lemma 4.2 later:

Lemma 4.1. *Let p be any probability distribution with support $[n]$. Suppose $\sum_{i=1}^{n} p_i x_i \geq 1-\epsilon$ and $\sum_{i=1}^{n} p_i y_i \geq 1-\delta$ with $\epsilon, \delta, x_i, y_i \in [0,1], \forall i \in [n]$, then $\sum_{i=1}^{n} p_i x_i y_i \geq 1-\epsilon-\delta$.*

Proof. Since $x_i, y_i \in [0,1]$ for all $i \in [n]$,

$$
\begin{aligned}
&\sum_{i=1}^{n} p_i(1-x_i)(1-y_i) \geq 0 \\
\iff \quad &\sum_{i=1}^{n} p_i - p_i x_i - p_i y_i + p_i x_i y_i \geq 0
\end{aligned}
$$

$$\begin{aligned}
\iff \qquad & \sum_{i=1}^{n} p_i x_i y_i \geq \sum_{i=1}^{n} p_i x_i + \sum_{i=1}^{n} p_i y_i - \sum_{i=1}^{n} p_i \\
\iff \qquad & \sum_{i=1}^{n} p_i x_i y_i \geq (1-\epsilon) + (1-\delta) - 1 = 1 - \epsilon - \delta.
\end{aligned}$$

□

Lemma 4.2. *In addition to all the properties stated in Theorem 1.3, $(\sigma \otimes |\Phi_d\rangle\langle\Phi_d|, (\mathbb{I}_{A'} \otimes W_i))$ is a $(d, (56+16\sqrt{6})\epsilon)$-superdense coding protocol.*

Proof. We continue to use the notation from Theorem 3.2. Let

$$\rho_i \coloneqq \mathrm{Tr}_{A'}((U_i \otimes \mathbb{I}_B)\tau(U_i^\dagger \otimes \mathbb{I}_B)).$$

Denote the POVM Bob uses as $(M_i)_{i\in[d^2]}$. Let the spectral decomposition of M_i be $\sum_{j=1}^{n} \lambda_{i,j} |\phi_{i,j}\rangle\langle\phi_{i,j}|$. Without loss of generality, assume $\lambda_{i,j} \geq \lambda_{i,k}$ for all $i,j,k \in [n]$ and $j < k$. Let p be the uniform distribution over $[d^2]$. The following properties were previously proved:

- $$\mathbb{E}_p \lambda_{i,1} \geq 1 - \epsilon, \tag{4.1}$$

  proved in Lemma 2.6.

- There exists $(|\psi_i\rangle)_{i\in[d^2]}$ such that $\mathbb{E}_p \, \mathrm{F}(|\psi_i\rangle\langle\psi_i|, \rho_i)^2 \geq 1 - 2\epsilon$, proved in Theorem 3.2, and $\mathbb{E}_p \langle\psi_i|M_i|\psi_i\rangle \geq 1 - 3\epsilon$, proved in Corollary 2.7.

- $\mathrm{F}\left(\mathrm{Tr}_{A''}(\rho_i), \frac{\mathbb{I}}{d}\right) \geq 1 - (5+2\sqrt{6})\epsilon, \forall i \in [d^2]$, proved in Theorem 3.2.

- Define $\langle\xi_i| \coloneqq \langle\Phi_d| (W_i^\dagger \otimes \mathbb{I}_B)$, then $|\langle\xi_i|\psi_i\rangle| = \mathrm{F}\left(\mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \frac{\mathbb{I}}{d}\right), \forall i \in [d^2]$, proved in Theorem 1.3.

We first prove $\mathbb{E}_p \left(|\langle\xi_i|\psi_i\rangle|^2\right)$ is large. By Corollary 2.2 and linearity of expectation,

$$\mathbb{E}_p \left(|\langle\xi_i|\psi_i\rangle|^2\right)$$

$$
\begin{aligned}
&= \mathbb{E}_p \left(\mathrm{F}\left(\mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \frac{\mathbb{I}}{d} \right)^2 \right) \\
&\geq \mathbb{E}_p \left(2\,\mathrm{F}\left(\mathrm{Tr}_{A''}(|\psi_i\rangle\langle\psi_i|), \mathrm{Tr}_{A''}(\rho_i) \right)^2 + 2\,\mathrm{F}\left(\mathrm{Tr}_{A''}(\rho_i), \frac{\mathbb{I}}{d} \right)^2 - 3 \right) \\
&\geq \mathbb{E}_p \left(2\,\mathrm{F}\left(|\psi_i\rangle\langle\psi_i| , \rho_i \right)^2 + 2\,\mathrm{F}\left(\mathrm{Tr}_{A''}(\rho_i), \frac{\mathbb{I}}{d} \right)^2 - 3 \right) \\
&\geq 2(1-2\epsilon) + 2(1-(5+2\sqrt{6})\epsilon)^2 - 3 \\
&\geq 2(1-2\epsilon) + 2(1-(10+4\sqrt{6})\epsilon) - 3 \\
&= 1-(24+8\sqrt{6})\epsilon.
\end{aligned} \tag{4.2}
$$

Then, we prove our final result. Notice M_i may not have trace 1, so we cannot apply Corollary 2.2 directly. Expanding M_i's spectral decomposition gives us the following:

$$
\begin{aligned}
&\mathbb{E}_p\left(\langle \xi_i | M_i | \xi_i \rangle \right) \\
=& \frac{1}{d^2} \sum_{i=1}^{d^2} \sum_{j=1}^{d^2} \lambda_{i,j} |\langle \phi_{i,j} | \xi_i \rangle|^2 \\
\geq& \frac{1}{d^2} \sum_{i=1}^{d^2} \sum_{j=1}^{d^2} \lambda_{i,j} \left(2|\langle \phi_{i,j} | \psi_i \rangle|^2 + 2|\langle \psi_i | \xi \rangle|^2 - 3 \right) \\
\geq& 2\left(\frac{1}{d^2} \sum_{i=1}^{d^2} \langle \psi_i | \left(\sum_{j=1}^{d^2} \lambda_{i,j} \, |\phi_{i,j}\rangle\langle\phi_{i,j}| \right) |\psi_i\rangle \right) + 2\left(\sum_{i=1}^{d^2} \frac{1}{d^2} \lambda_{i,1} |\langle \psi_i | \xi_i \rangle|^2 \right) \\
&- 3\left(\frac{1}{d^2} \sum_{i=1}^{d^2} \sum_{j=1}^{d^2} \lambda_{i,j} \right) \\
\geq& 2\left(\frac{1}{d^2} \sum_{i=1}^{d^2} \langle \psi_i | M_i | \psi_i \rangle \right) + 2(1-\epsilon-(24+8\sqrt{6})\epsilon) - 3\left(\frac{1}{d^2} \sum_{i=1}^{d^2} \mathrm{Tr}(M_i) \right) \\
\geq& 2(1-3\epsilon) + 2(1-(25+8\sqrt{6})\epsilon) - \frac{3\,\mathrm{Tr}(\mathbb{I}_{d^2})}{d^2} \\
=& 1-(56+16\sqrt{6})\epsilon,
\end{aligned}
$$

where the third line is due to Corollary 2.3, and the third last line is by applying Lemma 4.1 to Inequality 4.1 and Inequality 4.2. □

Corollary 4.3. *In addition to all the properties stated in Theorem 1.3,* $(|\Phi_d\rangle\langle\Phi_d|, (W_i))$ *is a* $(d, (56+16\sqrt{6})\epsilon)$*-superdense coding protocol.*

Proof. Since

$$\mathrm{Tr}_{A'}((\mathbb{I}_{A'} \otimes W_i \otimes \mathbb{I}_B)(\sigma \otimes |\Phi_d\rangle\langle\Phi_d|)(\mathbb{I}_{A'} \otimes W_i^\dagger \otimes \mathbb{I}_B)) = (W_i \otimes \mathbb{I}_B)\,|\Phi_d\rangle\langle\Phi_d|\,(W_i^\dagger \otimes \mathbb{I}_B),$$

Alice's ancilla qubits have no effect in $(\sigma \otimes |\Phi_d\rangle\langle\Phi_d|, (\mathbb{I}_{A'} \otimes W_i))$. Therefore, Lemma 4.2 directly implies $(|\Phi_d\rangle\langle\Phi_d|, (W_i))$ is a $(d, (56+16\sqrt{6})\epsilon)$-superdense coding protocol. □

Therefore, any (d,ϵ)-superdense coding protocol $(\tau', (V_i))$ is locally equivalent to (d,ϵ)-superdense coding protocol $(\tau, (U_i))$, such that $\mathrm{Tr}_{A'}(\tau)$ is close to $|\Phi_d\rangle\langle\Phi_d|$, there exists unitary matrices $W_i \in \mathcal{L}(\mathcal{H}_{A''})$, and U_i is close to $\mathbb{I}_{A'} \otimes W_i$ when acting on τ and then tracing out the A' part for each $i \in [d^2]$. Further, $(|\Phi_d\rangle\langle\Phi_d|, (W_i))$ is a $(d, O(\epsilon))$-superdense coding protocol. If we can then show some structure of the (W_i) even only considering $(|\Phi_d\rangle\langle\Phi_d|, (W_i))$, we automatically get results about the structure of (U_i) by Theorem 1.3. We will show how the last reduction is formally done in section 5.2.

Chapter 5

Rigidity of near-optimal superdense coding protocols

In this chapter, we try to obtain some further results on the unitary matrices (W_i) in Theorem 1.3 in the case when the dimension $d = 2$. We show any near-optimal superdense coding protocol, up to local equivalence, is close to the standard Bennett-Wiesner superdense coding protocol in Theorem 1.4.

5.1 Orthogonalizing 2×2 unitary matrices

By the discussions in section 4.2, we consider $(|\Phi_d\rangle\langle\Phi_d|, (W_i))$ first, that is, the shared initial state is exactly $|\Phi_d\rangle$, and Alice has no ancilla qubits, so the unitary matrices $W_i \in \mathcal{L}(\mathcal{H}_{A''})$. Under this simplification, at the end of this section, we will show when $d = 2$, the unitary matrices (W_i) have a "good" orthogonalization.

We start with proving some general results that may also be useful for dimensions $d > 2$.

Lemma 5.1. *For any $M \in \mathbb{C}^{d\times d}$, we have* $\mathrm{Tr}(M) = d \langle\Phi_d| (M \otimes \mathbb{I}_d) |\Phi_d\rangle$.

Proof.

$$\langle\Phi_d| (M \otimes \mathbb{I}_d) |\Phi_d\rangle = \frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{d} (\langle i| M |j\rangle \otimes \langle i| \mathbb{I}_d |j\rangle) = \frac{1}{d}\sum_{i=1}^{d} \langle i| M |i\rangle = \frac{1}{d}\mathrm{Tr}(M).$$

□

Notice that the matrix M might not be unitary in the above proof.

Lemma 5.2. *For any pure state* $|\Psi\rangle \in \mathbb{C}^d \otimes \mathbb{C}^d$*, there exists a matrix* $M \in \mathbb{C}^{d\times d}$ *such that* $|\Psi\rangle = (M \otimes \mathbb{I}_d)\,|\Phi_d\rangle$ *and* $\mathrm{Tr}\big(M^\dagger M\big) = d$.

Although M may not be unitary and cannot act on states in the usual quantum computing setting, the above lemma will be useful for later proofs.

Proof. Suppose $|\Psi\rangle = \sum_{i,j=1}^{d} m_{i,j}\,|i\rangle \otimes |j\rangle$ for complex coefficients $m_{i,j}$. $\langle\Psi|\Psi\rangle = 1$ implies $\sum_{i,j=1}^{d} |m_{i,j}|^2 = 1$. Define $M \coloneqq \sqrt{d}\sum_{i,j=1}^{d} m_{i,j}\,|i\rangle\!\langle j|$. We show $|\Psi\rangle$ is essentially a vectorization of M multiplied by a $\sqrt{d}$ factor.

$$
\begin{aligned}
(M \otimes \mathbb{I}_d)\,|\Phi_d\rangle &= \frac{1}{\sqrt{d}}\sum_{j=1}^{d}(M\,|j\rangle) \otimes |j\rangle \\
&= \sum_{j=1}^{d}\left(\sum_{i=1}^{d} m_{i,j}\,|i\rangle\!\langle j|j\rangle\right) \otimes |j\rangle \\
&= \sum_{i,j=1}^{d} m_{i,j}\,|i\rangle \otimes |j\rangle \\
&= |\Psi\rangle\,.
\end{aligned}
$$

For the last part of the proof, by Lemma 5.1,

$$
\mathrm{Tr}\big(M^\dagger M\big) = d\,\langle\Phi_d|\,(M^\dagger M \otimes \mathbb{I})\,|\Phi_d\rangle = d\,\langle\Psi|\Psi\rangle = d.
$$

□

Define

$$
\vec{\sigma} \coloneqq [\mathrm{i}\mathbb{I}_2, \mathrm{X}, \mathrm{Y}, \mathrm{Z}]^\intercal\,,
$$

and notice that $\mathbb{I}$ has coefficient i, the imaginary square-root of -1. The X, Y, and Z are Pauli matrices.

Lemma 5.3. *For any* $U \in \mathcal{U}(2)$*, there exists unique* $\alpha \in [0, \pi)$*, and a unique unit vector* $\vec{v} = [v_I, v_X, v_Y, v_Z]^\intercal \in \mathbb{R}^4$ *such that* $U = \mathrm{e}^{\mathrm{i}\alpha}\,(\vec{v}\cdot\vec{\sigma}) = \mathrm{e}^{\mathrm{i}\alpha}\,(v_I\mathrm{i}\mathbb{I} + v_X\mathrm{X} + v_Y\mathrm{Y} + v_Z\mathrm{Z})$.

Proof. By Equation 4.8 in [19] on page 175,

$$U = \mathrm{e}^{\mathrm{i}\beta} R_{\vec{n}}(\theta) = \mathrm{e}^{\mathrm{i}\beta}\left(\cos\left(\frac{\theta}{2}\right)\mathbb{I}_2 - \mathrm{i}\sin\left(\frac{\theta}{2}\right)(n_X \mathrm{X} + n_Y \mathrm{Y} + n_Z \mathrm{Z})\right),$$

for some $\beta, \theta \in \mathbb{R}$, and a unit vector $\vec{n} = [n_X, n_Y, n_Z]^\intercal \in \mathbb{R}^3$. This is the well-known Bloch sphere representation of 2×2 unitary operators.

Define $\vec{u} \coloneqq [u_I, u_X, u_Y, u_Z]^\intercal$ where $u_I \coloneqq \cos\left(\frac{\theta}{2}\right)$, $u_X \coloneqq \sin\left(\frac{\theta}{2}\right)n_X$,$u_Y \coloneqq \sin\left(\frac{\theta}{2}\right)n_Y$, and $u_Z \coloneqq \sin\left(\frac{\theta}{2}\right)n_Z$. Notice $\vec{u} \in \mathbb{R}^4$, $|\vec{u}|^2 = \cos^2\left(\frac{\theta}{2}\right) + \sin^2\left(\frac{\theta}{2}\right)|\vec{n}|^2 = 1$, and

$$\begin{aligned}
&\mathrm{e}^{\mathrm{i}\beta}\left(\cos\left(\frac{\theta}{2}\right)\mathbb{I}_2 - \mathrm{i}\sin\left(\frac{\theta}{2}\right)(n_X \mathrm{X} + n_Y \mathrm{Y} + n_Z \mathrm{Z})\right)\\
=&\mathrm{e}^{\mathrm{i}\left(\beta - \frac{\pi}{2}\right)}\left(\cos\left(\frac{\theta}{2}\right)\mathrm{i}\mathbb{I}_2 + \sin\left(\frac{\theta}{2}\right)n_X\mathrm{X} + \sin\left(\frac{\theta}{2}\right)n_Y\mathrm{Y} + \sin\left(\frac{\theta}{2}\right)n_Z\mathrm{Z}\right)\\
=&\mathrm{e}^{\mathrm{i}\left(\beta - \frac{\pi}{2}\right)}\left(\vec{u} \cdot \vec{\sigma}\right).
\end{aligned}$$

Then, there exists a unique $k \in \mathbb{Z}$ and $\alpha \in [0, \pi)$ such that $k\pi + \alpha = \beta - \frac{\pi}{2}$. If the k is even, define $\vec{v} \coloneqq \vec{u}$, and if the k is odd, define $\vec{v} \coloneqq -\vec{u}$. It is straightforward to check $\mathrm{e}^{\mathrm{i}\left(\beta - \frac{\pi}{2}\right)}\left(\vec{u} \cdot \vec{\sigma}\right) = \mathrm{e}^{\mathrm{i}\alpha}\left(\vec{v} \cdot \vec{\sigma}\right)$.

If there exists $\alpha' \in [0, \pi)$ and $\vec{v'} \in \mathbb{R}^4$ such that $\mathrm{e}^{\mathrm{i}\alpha'}\left(\vec{v'} \cdot \vec{\sigma}\right) = \mathrm{e}^{\mathrm{i}\alpha}\left(\vec{v} \cdot \vec{\sigma}\right)$, as $\{\mathrm{i}\mathbb{I}_2, \mathrm{X}, \mathrm{Y}, \mathrm{Z}\}$ forms a basis for $\mathbb{C}^{2\times 2}$ and $\vec{v} \neq \vec{0}$, we must have $\mathrm{e}^{\mathrm{i}\alpha'}\vec{v'} = \mathrm{e}^{\mathrm{i}\alpha}\vec{v}$, or equivalently $\vec{v'} = \mathrm{e}^{\mathrm{i}(\alpha - \alpha')}\vec{v}$. Since $\vec{v}, \vec{v'} \in \mathbb{R}^4$, $\mathrm{e}^{\mathrm{i}(\alpha - \alpha')} \in \mathbb{R}$, so $\alpha - \alpha' = n\pi$ for some $n \in \mathbb{Z}$. As $\alpha, \alpha' \in [0, \pi)$, $\alpha - \alpha' \in (-\pi, \pi)$, and n must be equal to 0. Therefore, $\alpha = \alpha'$ and $\vec{v} = \vec{v'}$, and the uniqueness is proved. $\square$

For any $\mathcal{S} \subset \mathbb{R}_{\geq 0}$, define

$$\mathcal{U}_{\mathcal{S}}(d) \coloneqq \{U : U \in \mathbb{C}^{d\times d}, \exists s \in \mathcal{S}, UU^\dagger = U^\dagger U = s^2\mathbb{I}_d\},$$

and we can think of it as a set containing constant scalings of all $d \times d$ unitary matrices.

Lemma 5.4. *For any $U \in \mathcal{U}_{\{k\}}(2)$ where $k \geq 0$, there exists a unique $\alpha \in [0, \pi)$ except when $k = 0$, and a unique vector $\vec{v} = [v_I, v_X, v_Y, v_Z]^\intercal \in \mathbb{R}^4$ such that $|\vec{v}| = k$ and $U = \mathrm{e}^{\mathrm{i}\alpha}\left(\vec{v} \cdot \vec{\sigma}\right)$.*

Proof. When $k = 0$, U has to be the zero matrix. For any $\alpha \in [0, \pi)$, $\vec{v} = \vec{0}$.

When $k > 0$, $U' := \dfrac{U}{k}$ is unitary. By Lemma 5.3, there exists a unique $\alpha' \in [0, \pi)$ and $\vec{v'} \in \mathbb{R}^4$ such that $U' = \mathrm{e}^{\mathrm{i}\alpha'}\left(\vec{v'} \cdot \vec{\sigma}\right)$. Then, if we let $\alpha := \alpha'$ and $\vec{v} := k\vec{v'}$, $U = \mathrm{e}^{\mathrm{i}\alpha}\left(\vec{v} \cdot \vec{\sigma}\right)$. The uniqueness of α and $\vec{v}$ follows from the proof of Lemma 5.3. $\square$

For any $U, V \in \mathcal{U}_{\mathbb{R}_{\geq 0}}(d)$, define the equivalence relationship $U \sim V$ if and only if there exists $\alpha \in \mathbb{R}$ such that $\mathrm{e}^{\mathrm{i}\alpha}U = V$. It is straightforward to verify this equivalence relationship $\sim$ is well-defined. Notice that $\mathcal{U}_{\{1\}}(d)/\sim$ is isomorphic to $\mathcal{SU}(d)$

Let $f : \mathbb{C}^4 \to \mathbb{C}^{2\times 2}$ be defined as $f(\vec{v}) := \vec{v} \cdot \vec{\sigma}$.

For any matrices $U, V \in \mathbb{C}^{2\times 2}$, define

$$\langle U, V\rangle := \frac{1}{2}\operatorname{Tr}\bigl(U^\dagger V\bigr).$$

Lemma 5.5. *The restriction of f to $\mathbb{R}^4$ is a Hilbert space isomorphism between $\mathbb{R}^4$ and $\mathcal{U}_{\mathbb{R}_{\geq 0}}(2)/\sim$.*

Proof. f is clearly a linear map by its definition.

f is surjective by Lemma 5.4 and the definition of f.

For any $\vec{u}, \vec{v} \in \mathbb{R}^4$, call their entries as $\vec{u} = [u_I, u_X, u_Y, u_Z]^\intercal$, and $\vec{v} = [v_I, v_X, v_Y, v_Z]^\intercal$, then

$$\begin{aligned}
&\langle f(\vec{u}), f(\vec{v})\rangle \\
=&\langle \vec{u} \cdot \vec{\sigma}, \vec{v} \cdot \vec{\sigma}\rangle \\
=&\frac{1}{2}\operatorname{Tr}\Bigl(\Bigl(-u_I \mathrm{i}\mathbb{I}_2^\dagger + u_X \mathrm{X}^\dagger + u_Y \mathrm{Y}^\dagger + u_Z \mathrm{Z}^\dagger\Bigr)\left(v_I \mathrm{i}\mathbb{I}_2 + v_X \mathrm{X} + v_Y \mathrm{Y} + v_Z \mathrm{Z}\right)\Bigr) \\
=&\frac{1}{2}(u_I v_I + u_X v_X + u_Y v_Y + u_Z v_Z)\operatorname{Tr}(\mathbb{I}) \\
=&u_I v_I + u_X v_X + u_Y v_Y + u_Z v_Z \\
=&\langle \vec{u}, \vec{v}\rangle,
\end{aligned}$$

where the third equality above is because $\mathbb{I}, \mathrm{X}, \mathrm{Y}, \mathrm{Z}$ are Hermitian, unitary, and mutually orthogonal with respect to the trace inner product. $\square$

One may ask why proving surjection is sufficient. It can be checked that the three above conditions imply f is a bijection.

Corollary 5.6. *f is a Hilbert space isomorphism between $\mathbb{C}^4$ and $\mathbb{C}^{2\times 2}$.*

Proof. The proof is almost identical to the proof of Lemma 5.5. The surjection part can be verified as $\{\mathrm{i}\mathbb{I}_2, \mathrm{X}, \mathrm{Y}, \mathrm{Z}\}$ forms a basis for $\mathbb{C}^{2\times 2}$. $\square$

By Lemma 5.5, to orthogonalize unitary matrices in $\mathcal{U}(2)$ or to orthogonalize maximally-entangled states in $\mathcal{H}^2 \otimes \mathcal{H}^2$, we can equivalently orthogonalize unit vectors in $\mathbb{R}^4$, and we can make use of known (linear) vector orthogonalization algorithms. Consider the following example:

- Suppose we start with two maximally-entangled states $(\mathrm{T}\otimes\mathbb{I}_2)\left|\Phi_2\right\rangle$ and $(\mathrm{H}\otimes\mathbb{I}_2)\left|\Phi_2\right\rangle$, where $\mathrm{T} \coloneqq \mathrm{e}^{-\mathrm{i}\frac{3\pi}{8}}\left(\cos\left(\frac{\pi}{8}\right)\mathrm{i}\mathbb{I}_2 + \sin\left(\frac{\pi}{8}\right)\mathrm{Z}\right)$. Define $\vec{v}_\mathrm{T} \coloneqq \left[\cos\left(\frac{\pi}{8}\right), 0, 0, \sin\left(\frac{\pi}{8}\right)\right]^\intercal$ and $\theta_\mathrm{T} \coloneqq -\frac{3\pi}{8}$ so that $f(\vec{v}_\mathrm{T}) \sim \mathrm{T} = \mathrm{e}^{\mathrm{i}\theta_\mathrm{T}} f(\vec{v}_\mathrm{T})$. Similarly, $\mathrm{H} = \mathrm{e}^{\mathrm{i}0}\left(\frac{1}{\sqrt{2}}\mathrm{X} + \frac{1}{\sqrt{2}}\mathrm{Z}\right)$. Define $\vec{v}_\mathrm{H} \coloneqq \left[0, \frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right]^\intercal$ and $\theta_\mathrm{H} \coloneqq 0$ so that $f(\vec{v}_\mathrm{H}) = \mathrm{H} = \mathrm{e}^{\mathrm{i}\theta_\mathrm{H}} f(\vec{v}_\mathrm{H})$.

  We use Löwdin's symmetric orthogonalization [13] to orthogonalize $\vec{v}_\mathrm{T}$ and $\vec{v}_\mathrm{H}$. If the singular value decomposition of $[\vec{v}_\mathrm{T}\vec{v}_\mathrm{H}]$ is $U\Sigma V^\dagger$, then,

$$\left[\vec{v'}_\mathrm{T}\ \vec{v'}_\mathrm{H}\right] \coloneqq U\begin{bmatrix}\mathbb{I}_2\\0\end{bmatrix}V^\dagger \approx \begin{bmatrix}0.950690 & -0.131072\\-0.100318 & 0.727627\\0 & 0\\0.293470 & 0.673335\end{bmatrix},$$

  and we get two orthonormal unit vectors $\vec{v'}_\mathrm{T}$ and $\vec{v'}_\mathrm{H}$. If we define $U_1 \coloneqq \mathrm{e}^{\mathrm{i}\theta_\mathrm{T}}\vec{v'}_\mathrm{T}\cdot\vec{\sigma}$, $U_2 \coloneqq \mathrm{e}^{\mathrm{i}\theta_\mathrm{H}}\vec{v'}_\mathrm{H}\cdot\vec{\sigma}$,

$$\frac{1}{2}\operatorname{Tr}\left(U_1^\dagger U_2\right) = \langle U_1, U_2\rangle = \mathrm{e}^{\mathrm{i}(\theta_\mathrm{H}-\theta_\mathrm{T})}\langle\vec{v'}_\mathrm{T}, \vec{v'}_\mathrm{H}\rangle = 0.$$

  So by Lemma 5.1,

$$\left\langle\Phi_2\right|(U_1^\dagger\otimes\mathbb{I}_2)(U_2\otimes\mathbb{I}_2)\left|\Phi_2\right\rangle = \left\langle\Phi_2\right|(U_1^\dagger U_2\otimes\mathbb{I}_2)\left|\Phi_2\right\rangle = \frac{1}{2}\operatorname{Tr}\left(U_1^\dagger U_2\right) = 0,$$

  and $(U_1\otimes\mathbb{I})\left|\Phi_2\right\rangle$ and $(U_2\otimes\mathbb{I})\left|\Phi_2\right\rangle$ are orthonormal maximally-entangled states. We also have

$$\begin{cases}\left\langle\Phi_2\right|(U_1^\dagger\otimes\mathbb{I})(\mathrm{T}\otimes\mathbb{I})\left|\Phi_2\right\rangle = \left\langle\Phi_2\right|(U_1^\dagger\mathrm{T}\otimes\mathbb{I})\left|\Phi_2\right\rangle = \langle U_1, \mathrm{T}\rangle = \langle\vec{v'}_\mathrm{T}, \vec{v}_\mathrm{T}\rangle,\\ \left\langle\Phi_2\right|(U_2^\dagger\otimes\mathbb{I})(\mathrm{H}\otimes\mathbb{I})\left|\Phi_2\right\rangle = \left\langle\Phi_2\right|(U_2^\dagger\mathrm{H}\otimes\mathbb{I})\left|\Phi_2\right\rangle = \langle U_2, \mathrm{H}\rangle = \langle\vec{v'}_\mathrm{H}, \vec{v}_\mathrm{H}\rangle.\end{cases}$$

Therefore, minimizing $\left|\langle \vec{v'}_{\mathrm{T}}, \vec{v}_{\mathrm{T}} \rangle\right|$ is equivalent to minimizing $|\langle U_1, \mathrm{T} \rangle|$ or

$$\left|\langle \Phi_2 | \, (U_1^\dagger \otimes \mathbb{I})(\mathrm{T} \otimes \mathbb{I}) \, |\Phi_2\rangle\right|,$$

and minimizing $\left|\langle \vec{v'}_{\mathrm{H}}, \vec{v}_{\mathrm{H}} \rangle\right|$ is equivalent to minimizing $|\langle U_2, \mathrm{H} \rangle|$ or

$$\left|\langle \Phi_2 | \, (U_2^\dagger \otimes \mathbb{I})(\mathrm{H} \otimes \mathbb{I}) \, |\Phi_2\rangle\right|.$$

Thus, such a Hilbert space isomorphism allows us to build algorithms for unitary matrices in $\mathcal{U}(2)$ and maximally-entangled states in $\mathcal{H}^4$ from known algorithms for vectors.

Now we try to find a bound on the absolute value of the inner product of vectors that guarantees linear independence. We need this result because later results depend on linear independence.

Lemma 5.7. *Suppose there are n unit vectors $(\vec{v}_i)_{i\in[n]} \subset \mathbb{C}^n$. If for all $i \neq j$, $|\langle \vec{v}_i, \vec{v}_j \rangle| < \dfrac{1}{n-1}$, then $(\vec{v}_i)_{i\in[n]}$ are linearly independent.*

Proof. $(\vec{v}_i)_{i\in[n]}$ are linearly independent if and only if the Gram matrix $G \coloneqq (\langle \vec{v}_i, \vec{v}_j \rangle)_{i,j}$ has full rank. The notation here means the entry of G on the i-th row and the j-th column is $\langle \vec{v}_i, \vec{v}_j \rangle$.

Let $R_i \coloneqq \sum_{j\in[n]\setminus\{i\}} |G_{i,j}|$. By our assumption, $R_i < (n-1)\dfrac{1}{n-1} = 1$ for all $i \in [n]$. By Gershgorin circle theorem [7], each eigenvalue of G is at least $G_{i,i} - R_i > 1 - 1 = 0$. Therefore, G has full rank and $(\vec{v}_i)_{i\in[n]}$ are linearly independent. $\square$

Lemma 5.8. *The condition $|\langle \vec{v}_i, \vec{v}_j \rangle| < \dfrac{1}{n-1}$ for all $i \neq j$ in Lemma 5.7 is optimal.*

Proof. We show if we relax the condition to $|\langle \vec{v}_i, \vec{v}_j \rangle| \leq \dfrac{1}{n-1}$ for all $i \neq j$, then there are unit vectors $(\vec{v}_i)_{i\in[n]} \subset \mathbb{R}^n$ that are linearly dependent.

Consider the n elementary basis vectors $\vec{e_i} \in \mathbb{R}^n$. The convex hull of the endpoints is a regular polygon and lies inside a hyperplane H of $\mathbb{R}^n$ given by $x_1 + x_2 + \cdots + x_n = 1$. The center of this polygon is

$$\vec{c} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \vec{e_i} = \begin{bmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \end{bmatrix}.$$

Let $\vec{v_i} = \dfrac{\vec{e_i} - \vec{c}}{\|\vec{e_i} - \vec{c}\|}$, then

$$\vec{v_i} = \frac{\vec{e_i} - \vec{c}}{\|\vec{e_i} - \vec{c}\|} = \frac{\vec{e_i} - \sum_{j \in [n]} \vec{e_j}/n}{\sqrt{\frac{(n-1)^2}{n^2} + (n-1)\frac{(-1)^2}{n^2}}} = \sqrt{\frac{n-1}{n}}\vec{e_i} - \sum_{j \in [n] \setminus \{i\}} \frac{1}{\sqrt{n(n-1)}}\vec{e_j}.$$

Clearly, $\mathrm{span}(\{\vec{v_i}\}_{i \in [n]}) < n$ because all vectors are parallel to the hyperplane H of $\mathbb{R}^n$. However, for $i, j \in [n]$ and $i \neq j$,

$$\begin{aligned} \langle \vec{v_i}, \vec{v_j} \rangle =& \sqrt{\frac{n-1}{n}}\left(-\frac{1}{\sqrt{n(n-1)}}\right)(\langle \vec{e_i}, \vec{e_i} \rangle + \langle \vec{e_j}, \vec{e_j} \rangle) + \sum_{k \in [n] \setminus \{i,j\}} \left(-\frac{1}{\sqrt{n(n-1)}}\right)^2 \\ =& -\frac{1}{n-1}. \end{aligned}$$

□

Lemma 5.9. *Let $(|\Phi_d\rangle\langle\Phi_d|, (W_i))$ be a (d, ϵ)-superdense coding protocol. If*

$$\epsilon < \frac{1}{2d^2(d^2-1)^2},$$

then the states $((W_i \otimes \mathbb{I}_d)|\Phi_d\rangle)_{i \in [d^2]}$ are linearly independent.

Proof. Suppose Bob uses POVM $(M_i)_{i \in [d^2]}$. Let

$$s_i := \mathrm{Tr}\Big(M_i(W_i \otimes \mathbb{I}_d)|\Phi_d\rangle\langle\Phi_d|(W_i^\dagger \otimes \mathbb{I}_d)\Big) = 1 - \epsilon_i,$$

be the success probability for some $\epsilon_i \in [0, 1]$ when Alice receives i.

Since $\dfrac{1}{d^2}\displaystyle\sum_{i=1}^{d^2} s_i = \dfrac{1}{d^2}\sum_{i=1}^{d^2}(1 - \epsilon_i) \geq 1 - \epsilon$, for any $i, j \in [d^2]$ with $i \neq j$, we have $\epsilon_i + \epsilon_j \leq d^2\epsilon$. By Appendix A.9 of [10], for any $i, j \in [d^2]$ and $i \neq j$,

$$\begin{aligned} \Big|\langle\Phi_d|(W_i^\dagger \otimes \mathbb{I}_d)(W_j \otimes \mathbb{I}_d)|\Phi_d\rangle\Big| \leq& \sqrt{\epsilon_i(1-\epsilon_j)} + \sqrt{\epsilon_j(1-\epsilon_i)} \\ \leq& \sqrt{2(\epsilon_i(1-\epsilon_j) + \epsilon_j(1-\epsilon_i))} \\ =& \sqrt{2(\epsilon_i + \epsilon_j - 2\epsilon_i\epsilon_i)} \end{aligned}$$

$$
\begin{aligned}
&\leq \sqrt{2(\epsilon_i + \epsilon_j)} \\
&\leq \sqrt{2d^2\epsilon}.
\end{aligned}
$$

If $\sqrt{2d^2\epsilon} < \dfrac{1}{d^2-1}$, or equivalently $\epsilon < \dfrac{1}{2d^2(d^2-1)^2}$, then by Lemma 5.7, $((W_i \otimes \mathbb{I}_d)\,|\Phi_d\rangle)_{i\in[d^2]}$ are linearly independent. $\square$

Corollary 5.10. *Let $(|\Phi_d\rangle\langle\Phi_d|\,, (W_i))$ be a (d,ϵ)-worst case superdense coding protocol. If*

$$\epsilon < \frac{1}{2(d^2-1)^2},$$

then the states $((W_i \otimes \mathbb{I}_d)\,|\Phi_d\rangle)_{i\in[d^2]}$ are linearly independent.

Proof. Similar to the proof of Lemma 5.9, this time, we have

$$s_i \geq 1 - \epsilon, \forall i \in [d^2].$$

Then, for any $i, j \in [d^2]$ and $i \neq j$,

$$\left| \langle\Phi_d|\,(W_i^\dagger \otimes \mathbb{I}_d)(W_j \otimes \mathbb{I}_d)\,|\Phi_d\rangle \right| \leq \sqrt{2(\epsilon+\epsilon)} \leq \sqrt{2\epsilon}.$$

If $\sqrt{2\epsilon} < \dfrac{1}{d^2-1}$, or equivalently $\epsilon < \dfrac{1}{2(d^2-1)^2}$, then by Lemma 5.7, $((W_i \otimes \mathbb{I}_d)\,|\Phi_d\rangle)_{i\in[d^2]}$ are linearly independent. $\square$

Corollary 5.11. *Let $(|\Phi_2\rangle\langle\Phi_2|\,, (W_i))$ be a $(2,\epsilon)$-superdense coding protocol. If $\epsilon < \dfrac{1}{72}$, then the states $((W_i \otimes \mathbb{I}_d)\,|\Phi_d\rangle)_{i\in[d^2]}$ are linearly independent.*

Proof. This is obtained by applying Lemma 5.9 when $d = 2$. $\square$

Lemma 5.12. *Suppose we have n pure states $(|\phi_i\rangle)_{i\in[n]} \subset \mathbb{R}^n$ (not $\mathbb{C}^n$). Consider any POVM $(M_i)_{i\in[n]} \subset \mathbb{C}^{n\times n}$ that maximizes $\sum_{i=1}^{d} \langle\phi_i|M_i|\phi_i\rangle$. If $(|\phi_i\rangle)_{i\in[n]}$ are linearly independent, then there exists orthonormal $(|\psi_i\rangle)_{i\in[n]} \subset \mathbb{R}^n$ such that*

$$\sum_{i=1}^{d} |\langle\psi_i|\phi_i\rangle|^2 = \sum_{i=1}^{d} \langle\phi_i|M_i|\phi_i\rangle\,.$$

Proof. Solving the semi-definite program from [26] to find the optimal POVM to distinguish the real states $(|\phi_i\rangle)_{i\in[n]}$ yields real and symmetric solutions $(M'_i)_{i\in[n]}$. Since $(|\phi_i\rangle)_{i\in[n]}$ are pure and linearly independent, by the main result of [11], $(M'_i)_{i\in[n]}$ are pairwise orthogonal rank 1 projectors. Therefore, there exists orthonormal $(|\psi_i\rangle)_{i\in[n]} \subset \mathbb{R}^n$ such that $M'_i = |\psi_i\rangle\langle\psi_i|$. □

Theorem 5.13. *Let $(|\Phi_2\rangle\langle\Phi_2|, (W_i))$ be a $(2,\epsilon)$-superdense coding protocol. If $\epsilon < \frac{1}{72}$, then there exists pair-wise orthogonal $(\tilde{W}_i)_{i\in[4]} \subset \mathcal{U}(2)$ (i.e. $\langle \tilde{W}_i, \tilde{W}_j\rangle = \delta_{ij}$), such that*

$$\frac{1}{4}\sum_{i=1}^{4}\left|\langle \tilde{W}_i, W_i\rangle\right|^2 \geq 1-\epsilon.$$

Proof. Denote the optimal POVM performed by Bob as $(M_i)_{i\in[4]}$. By Corollary 5.11, $((W_i\otimes\mathbb{I}_2)|\Phi_2\rangle)_{i\in[4]}$ are linearly independent. Thus, by [11], $(M_i)_{i\in[4]}$ are rank 1 projectors that are pairwise orthogonal (i.e., $M_iM_j = \delta_{i,j}M_i, \forall i,j \in [4]$), so there exists orthonormal $(|\psi_i\rangle)_{i\in[4]}$ such that $M_i = |\psi_i\rangle\langle\psi_i|, \forall i \in [4]$. By Lemma 5.2, there exists matrices $(N_i)_{i\in[4]} \subset \mathbb{C}^{2\times 2}$ such that $(N_i\otimes\mathbb{I}_2)|\Phi_2\rangle = |\psi_i\rangle$ and $\langle N_i, N_i\rangle = 1$ for all $i \in [4]$.

Consider the unique vectors $(\vec{w}_i)_{i\in[4]} \subset \mathbb{R}^4$ and $(\theta_i)_{i\in[4]} \subset [0,\pi)$ such that $\mathrm{e}^{\mathrm{i}\theta_i}\vec{w}_i\cdot\vec{\sigma} = W_i$ as given by Lemma 5.4. Since $\langle N_i, N_i\rangle = 1$ and $\{\mathrm{i}\mathbb{I}_2, \mathrm{X}, \mathrm{Y}, \mathrm{Z}\}$ forms a basis for $\mathbb{C}^{2\times 2}$, there exists $(\vec{n}_i)_{i\in[4]} \subset \mathbb{C}^4$ such that $\vec{n}_i\cdot\vec{\sigma} = N_i$. By Corollary 5.6, we must have $\langle \vec{n}_i, \vec{n}_i\rangle = \langle N_i, N_i\rangle = 1$. Then, by the condition of success probability,

$$\begin{aligned}
4(1-\epsilon) &\leq \sum_{i=1}^{4} \langle\Phi_2|(W_i^\dagger\otimes\mathbb{I}_2)M_i(W_i\otimes\mathbb{I}_2)|\Phi_2\rangle \\
&= \sum_{i=1}^{4} \langle\Phi_2|(W_i^\dagger\otimes\mathbb{I}_2)|\psi_i\rangle\langle\psi_i|(W_i\otimes\mathbb{I}_2)|\Phi_2\rangle \\
&= \sum_{i=1}^{4} \left|\langle\Phi_2|(W_i^\dagger\otimes\mathbb{I}_2)(N_i\otimes\mathbb{I}_2)|\Phi_2\rangle\right|^2 \\
&= \sum_{i=1}^{4} |\langle W_i, N_i\rangle|^2 \\
&= \sum_{i=1}^{4} |\langle \vec{w}_i, \vec{n}_i\rangle|^2,
\end{aligned}$$

where the second last equality is due to Lemma 5.1, and the last equality is due to Corollary 5.6. By Lemma 5.12, there exists real orthonormal vectors $(\vec{n_i'})_{i\in[4]} \subset \mathbb{R}^4$ such that

$$\sum_{i=1}^{4} \left|\langle \vec{w}_i, \vec{n_i'}\rangle\right|^2 = \sum_{i=1}^{4} |\langle \vec{w}_i, \vec{n}_i\rangle|^2 \geq 4(1-\epsilon).$$

Therefore, if we define $\tilde{W}_i = \mathrm{e}^{\mathrm{i}\theta_i}\vec{n_i'} \cdot \vec{\sigma}$, by Lemma 5.3, $\tilde{W}_i$ is unitary. By definition of f and Lemma 5.5,

$$\langle \tilde{W}_i, \tilde{W}_j\rangle = \mathrm{e}^{\mathrm{i}(\theta_j-\theta_i)}\langle \vec{n_i'}, \vec{n_j'}\rangle = \delta_{i,j}, \forall i,j \in [4],$$

and

$$\frac{1}{4}\sum_{i=1}^{4} \left|\langle \tilde{W}_i, W_i\rangle\right|^2 = \frac{1}{4}\sum_{i=1}^{4} \left|\mathrm{e}^{\mathrm{i}(\theta_i-\theta_i)}\langle \vec{w}_i, \vec{n_j'}\rangle\right|^2 = \frac{1}{4}\sum_{i=1}^{4} \left|\langle \vec{w}_i, \vec{n_j'}\rangle\right|^2 \geq 1-\epsilon.$$

□

5.2 Rigidity of near-optimal superdense coding protocols

Now, we have all the tools to prove the rigidity of any near-optimal superdense coding protocol when $d = 2$. This shows the discussions in section 4.2 formally:

Theorem 1.4. *There exists $c > 0$ such that any $(2,\epsilon)$-superdense coding protocol $(\tau', (V_i))$ with $\epsilon < c$ is locally equivalent to $(2,\epsilon)$-superdense coding protocol $(\tau, (U_i))$ which satisfies the following properties: there exists a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ and pair-wise orthogonal $(\tilde{W}_i)_{i\in[4]} \subset \mathcal{U}(2)$ (i.e., $\langle \tilde{W}_i, \tilde{W}_j\rangle = \delta_{ij}$, and δ_{ij} is the Kronecker delta), such that*

$$\mathrm{F}\left(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|\right) \geq 1 - (21 + 6\sqrt{6})\epsilon = 1 - O(\epsilon),$$

and

$$\frac{1}{4}\sum_{i=1}^{4} \mathrm{S}_{\tau,A'}(U_i \otimes \mathbb{I}_B, \mathbb{I}_{A'} \otimes \tilde{W}_i \otimes \mathbb{I}_B) \geq 1 - (394 + 108\sqrt{6})\epsilon = 1 - O(\epsilon).$$

Proof. The construction of the protocol $(\tau, (U_i))$ is directly from Theorem 1.3, and there exists a density matrix $\sigma \in \mathcal{L}(\mathcal{H}_{A'})$ such that

$$\mathrm{F}\left(\tau, \sigma \otimes |\Phi_d\rangle\langle\Phi_d|\right) \geq 1 - (21 + 6\sqrt{6})\epsilon,$$

and the first part of the proof is done.

Theorem 1.3 also implies that there exist unitary matrices $W_i \in \mathcal{L}(\mathcal{H}_{A''})$ such that if we define $\rho_i := \mathrm{Tr}_{A'}((U_i \otimes \mathbb{I}_B)\tau(U_i^\dagger \otimes \mathbb{I}_B))$ and let p be the uniform distribution over $[4]$, then

$$\mathbb{E}_p \mathrm{F}\left(\rho_i, (W_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (W_i^\dagger \otimes \mathbb{I}_B)\right) \geq 1 - (32 + 8\sqrt{6})\epsilon,$$

$$\mathrm{F}(\mathrm{Tr}_{A'}(\tau), |\Phi_d\rangle\langle\Phi_d|) \geq 1 - (21 + 6\sqrt{6})\epsilon,$$

and by Corollary 4.3, $(|\Phi_d\rangle\langle\Phi_d|, (W_i))$ is a $(d, (56 + 16\sqrt{6})\epsilon)$-superdense coding protocol.

Let $c := \dfrac{1}{72(56 + 16\sqrt{6})}$. When $(56+16\sqrt{6})\epsilon < \dfrac{1}{72}$ (i.e., $\epsilon < c$), by Theorem 5.13, there exists orthogonal $(\tilde{W}_i)_{i\in[d^2]} \subset \mathcal{U}(2)$ (i.e. $\langle \tilde{W}_i, \tilde{W}_j\rangle = \delta_{ij}$), such that

$$\frac{1}{4}\sum_{i=1}^{4} \left|\langle\Phi_2| (\tilde{W}_i{}^\dagger \otimes \mathbb{I}_2)(W_i \otimes \mathbb{I}_2) |\Phi_2\rangle\right|^2 = \frac{1}{4}\sum_{i=1}^{4}\left|\langle \tilde{W}_i, W_i\rangle\right|^2 \geq 1 - (56 + 16\sqrt{6})\epsilon,$$

where the equality comes from Lemma 5.1. Combining all above results and using Corollary 2.5 twice, we have

$$\begin{aligned}
&\frac{1}{d^2}\sum_{i=1}^{d^2} \mathrm{S}_{\tau,A'}(U_i \otimes \mathbb{I}_B, \mathbb{I}_{A'} \otimes \tilde{W}_i \otimes \mathbb{I}_B) \\
=&\mathbb{E}_p \mathrm{F}\left(\mathrm{Tr}_{A'}((U_i \otimes \mathbb{I}_B)\tau(U_i^\dagger \otimes \mathbb{I}_B)), \mathrm{Tr}_{A'}((\mathbb{I}_{A'} \otimes \tilde{W}_i \otimes \mathbb{I}_B)\tau(\mathbb{I}_{A'} \otimes \tilde{W}_i{}^\dagger \otimes \mathbb{I}_B))\right) \\
=&\mathbb{E}_p \mathrm{F}\left(\rho_i, (\tilde{W}_i \otimes \mathbb{I}_B)\,\mathrm{Tr}_{A'}(\tau)(\tilde{W}_i{}^\dagger \otimes \mathbb{I}_B)\right) \\
\geq&\mathbb{E}_p \Big(2\,\mathrm{F}\left(\rho_i, (\tilde{W}_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (\tilde{W}_i{}^\dagger \otimes \mathbb{I}_B)\right) \\
&\qquad + 2\,\mathrm{F}\left((\tilde{W}_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (\tilde{W}_i{}^\dagger \otimes \mathbb{I}_B), (\tilde{W}_i \otimes \mathbb{I}_B)\,\mathrm{Tr}_{A'}(\tau)(\tilde{W}_i{}^\dagger \otimes \mathbb{I}_B)\right) - 3\Big) \\
\geq&\mathbb{E}_p \Big(4\,\mathrm{F}\left(\rho_i, (W_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (W_i^\dagger \otimes \mathbb{I}_B)\right) \\
&\qquad + 4\,\mathrm{F}\left((W_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (W_i^\dagger \otimes \mathbb{I}_B), (\tilde{W}_i \otimes \mathbb{I}_B) |\Phi_d\rangle\langle\Phi_d| (\tilde{W}_i{}^\dagger \otimes \mathbb{I}_B)\right) - 6 \\
&\qquad + 2\,\mathrm{F}\left(|\Phi_d\rangle\langle\Phi_d|, \mathrm{Tr}_{A'}(\tau)\right) - 3\Big) \\
\geq&4(1 - (32 + 8\sqrt{6})\epsilon) + 4(1 - (56 + 16\sqrt{6})\epsilon) + 2(1 - (21 + 6\sqrt{6})\epsilon) - 9 \\
=&1 - (394 + 108\sqrt{6})\epsilon.
\end{aligned}$$

□

Therefore, by Theorem 1.1 in [18], any near-optimal superdense coding protocol, up to local equivalence, is close to the standard Bennett-Wiesner superdense coding protocol.

Chapter 6

Orthogonalizing two unitary matrices in general

Notice that chapter 5 only works for dimension $d = 2$. In the higher dimension case when $d > 2$, the same method does not work because $\mathcal{SU}(d)$ or even scalings of it (defined as $\mathcal{U}_{\mathbb{R}_{\geq 0}}(d)/\sim$) when $d > 2$ is not isomorphic to a vector space. In the attempt to solve the problem when $d > 2$, we find a way to orthogonalize 2 arbitrary $d \times d$ unitary matrices with any $d > 2$.

In this chapter, we first explain how this orthogonalization of 2 unitary matrices is reduced to another simpler problem of rotating 2D vectors such that the sum of vectors is $\vec{0}$ while the total angle of rotation is small. The reduction is shown in section 6.1, then we solve that simpler problem in section 6.2, and derive the final result in section 6.3.

6.1 Orthogonalizing two unitary operators by “rotating" eigenvalues

Suppose we have $U_1, U_2 \in \mathcal{U}(d)$ for any $d \geq 2$ such that

$$|\langle U_1, U_2\rangle|^2 = \left|\frac{1}{d}\operatorname{Tr}\left(U_1^\dagger U_2\right)\right|^2 \leq \epsilon.$$

This implies $\left|\operatorname{Tr}\left(U_1^\dagger U_2\right)\right| \leq d\sqrt{\epsilon}$.

Suppose we modify U_2 by multiplying a unitary matrix U such that $\langle U_1, UU_2\rangle = 0$ and U is close to the identity matrix $\mathbb{I}_d$. This orthogonalizes U_1 and U_2 with a small change to U_2 because

$$\|UU_2 - U_2\|_{\mathrm{F}} = \|U - \mathbb{I}_d\|_{\mathrm{F}} .$$

Define $U' \coloneqq U_1^\dagger U U_1$. Notice

$$\|U' - \mathbb{I}_d\|_{\mathrm{F}} = \left\|U_1^\dagger U U_1 - U_1^\dagger U_1\right\|_{\mathrm{F}} = \left\|U_1^\dagger (U - \mathbb{I}_d) U_1\right\|_{\mathrm{F}} = \|U - \mathbb{I}_d\|_{\mathrm{F}} ,$$

and

$$\langle U_1, UU_2\rangle = \frac{1}{d}\operatorname{Tr}\left(U_1^\dagger U U_2\right) = \frac{1}{d}\operatorname{Tr}\left(U_1^\dagger U U_1 U_1^\dagger U_2\right) = \frac{1}{d}\operatorname{Tr}\left(U' U_1^\dagger U_2\right).$$

Equivalently, if we can find such unitary U' that is close to $\mathbb{I}_d$ and $\operatorname{Tr}\left(U' U_1^\dagger U_2\right) = 0$, then it gives us the U which is also close to $\mathbb{I}_d$ and orthogonalizes U_1 and U_2.

Let the spectral decomposition of $U_1^\dagger U_2$ be $\tilde{U} D \tilde{U}^\dagger$ where $\tilde{U}$ is unitary and D is diagonal. Values on the diagonal are eigenvalues of $U_1^\dagger U_2$, and by properties of unitary matrices, each eigenvalue has modulus 1. So we can rewrite $D \coloneqq \operatorname{diag}(\mathrm{e}^{\mathrm{i}\theta_1}, \mathrm{e}^{\mathrm{i}\theta_2}, \cdots, \mathrm{e}^{\mathrm{i}\theta_d})$. Since

$$|\operatorname{Tr}(D)| = \left|\operatorname{Tr}\left(\tilde{U} D \tilde{U}^\dagger\right)\right| = \left|\operatorname{Tr}\left(U_1^\dagger U_2\right)\right| \le d\sqrt{\epsilon},$$

we have $\left|\sum_{j=1}^{d} \mathrm{e}^{\mathrm{i}\theta_j}\right| \le d\sqrt{\epsilon}$.

We can construct U' in the form of $\tilde{U} D' \tilde{U}^\dagger$ where $D' \coloneqq \operatorname{diag}(\mathrm{e}^{\mathrm{i}\omega_1}, \mathrm{e}^{\mathrm{i}\omega_2}, \cdots, \mathrm{e}^{\mathrm{i}\omega_d})$. Then,

$$\operatorname{Tr}\left(U' U_1^\dagger U_2\right) = \operatorname{Tr}\left(\tilde{U} D' \tilde{U}^\dagger \tilde{U} D \tilde{U}^\dagger\right) = \operatorname{Tr}(D'D) = \sum_{j=1}^{d} \mathrm{e}^{\mathrm{i}(\theta_j + \omega_j)},$$

and

$$\|U' - \mathbb{I}_d\|_{\mathrm{F}} = \left\|\tilde{U} D' \tilde{U}^\dagger - \tilde{U}\tilde{U}^\dagger\right\|_{\mathrm{F}} = \|D' - \mathbb{I}_d\|_{\mathrm{F}} = \sqrt{\sum_{j=1}^{d} |\mathrm{e}^{\mathrm{i}\omega_j} - 1|^2}.$$

Informally, the original problem is reduced to the following problem: Suppose we have unit vectors $\vec{v_1}, \cdots, \vec{v_d} \in \mathbb{R}^2$ (representing $\mathrm{e}^{\mathrm{i}\theta_1}, \cdots, \mathrm{e}^{\mathrm{i}\theta_d}$). We want to make small rotations to them $\left(\text{each } \mathrm{e}^{\mathrm{i}\theta_j} \text{ is rotated to } \mathrm{e}^{\mathrm{i}(\theta_j+\omega_j)}\right)$ such that the vectors sum up to the zero vector

$\left(\sum_{j=1}^{d} \mathrm{e}^{\mathrm{i}(\theta_j+\omega_j)} = 0\right)$ while the sum of the angles of rotation $\left(\sum_{j=1}^{d} |\omega_j|\right)$ is small. We will show later that it is sufficient to bound $\sum_{j=1}^{d} |\omega_j|$ from above to get an upper bound for $\left|\sqrt{\sum_{j=1}^{d} |\mathrm{e}^{\mathrm{i}\omega_j} - 1|^2}\right|$. We will first solve this reduced problem in section 6.2, and then use the result to solve the original problem in section 6.3.

6.2 Rotating vectors in $\mathbb{R}^2$ to sum up to $\vec{0}$

In this section, we solve the problem proposed at the end of section 6.1. The idea is as follows: suppose the unit vectors $\vec{v_1}, \cdots, \vec{v_d} \in \mathbb{R}^2$ do not sum up to $\vec{0}$. Let $\vec{s} \coloneqq \sum_{i=1}^{d} \vec{v_i}$. There are two cases:

1. If there exists $i \in [d]$ such that $\vec{v_i}$'s component orthogonal to $\vec{s}$ is not "too small," then we can rotate $\vec{v_i}$ by a tiny angle to reduce $|\vec{s}|$. The direction of rotation depends on the cross product between $\vec{v_i}$ and $\vec{s}$. We will show this in Lemma 6.2.

2. If $\vec{v_i}$'s component orthogonal to $\vec{s}$ is "small" for all $i \in [d]$, then we can find two vectors $\vec{v_{j_1}}$ and $\vec{v_{j_2}}$, rotate them by a small amount, and make the sum of all vectors equal to $\vec{0}$. We will show this in Lemmas 6.3, 6.5 and 6.6.

We can keep reducing $|\vec{s}|$ as described in case 1 and update the sum $\vec{s}$ until $\vec{s} = \vec{0}$ or we reach case 2, and in the latter case, we perform the described fix to make $\vec{s} = \vec{0}$. In the analysis, we propose an algorithm to do case 1 discretely in Theorem 6.9. The algorithm either halts and gives us the correct solution with a small total rotation, or it runs indefinitely, and we prove the intermediate vectors produced by the algorithm converge to a correct solution with a small total rotation.

In the remainder of this section, we still use complex numbers to represent vectors in $\mathbb{R}^2$. Lemma 6.1 shows how to compute the inner product and cross product of the vectors in terms of the complex numbers.

Lemma 6.1. *For $x, y \in \mathbb{C}$, suppose $x = a + b\mathrm{i}$ and $y = c + d\mathrm{i}$, then* $\mathrm{Re}\{xy^*\}$ *equals the inner product between vectors* $\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix} \in \mathbb{R}^2$*, and* $\mathrm{Im}\{x^*y\}$ *equals the last component of the cross product between vectors* $\begin{bmatrix} a \\ b \\ 0 \end{bmatrix}, \begin{bmatrix} c \\ d \\ 0 \end{bmatrix} \in \mathbb{R}^3$*.*

Proof. $\mathrm{Re}\{xy^*\} = \mathrm{Re}\{(a + b\mathrm{i})(c - d\mathrm{i})\} = ac + bd = \begin{bmatrix} a \\ b \end{bmatrix} \cdot \begin{bmatrix} c \\ d \end{bmatrix}$.

$\mathrm{Im}\{x^*y\} = \mathrm{Im}\{(a - b\mathrm{i})(c + d\mathrm{i})\} = ad - bc$, and $\begin{bmatrix} a \\ b \\ 0 \end{bmatrix} \times \begin{bmatrix} c \\ d \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ ad - bc \end{bmatrix}$. $\square$

For non-zero $x, y \in \mathbb{C}$, define

$$\angle(x, y) \coloneqq \arccos\left(\frac{\mathrm{Re}\{xy^*\}}{|x||y|}\right) \in [0, \pi].$$

By the property of vector inner product and Lemma 6.1, $\angle(x, y)$ equals the angle between x and y viewed as non-zero vectors on the complex plane.

Lemma 6.2 formally explains case 1 at the beginning of the section 6.2.

Lemma 6.2. *Suppose $s \in (0, 1]$, $\theta \in [\sqrt{s}, \pi - \sqrt{s}]$, and $0 < \Delta < 0.28s\sqrt{s}$, then*

$$|s + \exp(\mathrm{i}(\theta + \Delta)) - \exp(\mathrm{i}\theta)| < s - \frac{\sqrt{s}}{2}\Delta.$$

Proof.

$$\begin{aligned}
&|s + \exp(\mathrm{i}(\theta + \Delta)) - \exp(\mathrm{i}\theta)|^2 \\
=&(s + \exp(\mathrm{i}(\theta + \Delta)) - \exp(\mathrm{i}\theta))(s + \exp(-\mathrm{i}(\theta + \Delta)) - \exp(-\mathrm{i}\theta)) \\
=&s^2 + s(\exp(\mathrm{i}(\theta + \Delta)) + \exp(-\mathrm{i}(\theta + \Delta))) - s(\exp(\mathrm{i}\theta) + \exp(-\mathrm{i}\theta)) \\
&+ \exp(\mathrm{i}(\theta + \Delta)) \exp(-\mathrm{i}(\theta + \Delta)) + \exp(\mathrm{i}\theta) \exp(-\mathrm{i}\theta) \\
&- (\exp(\mathrm{i}\Delta) + \exp(-\mathrm{i}\Delta)) \\
=&s^2 + s(2\cos(\theta + \Delta) - 2\cos(\theta)) + 2 - 2\cos(\Delta) \\
=&s^2 - 4s\sin\left(\theta + \frac{\Delta}{2}\right)\sin\left(\frac{\Delta}{2}\right) + 2 - 2\cos(\Delta), && (6.1)
\end{aligned}$$

where the last equality is due to the trigonometric identity

$$\cos(a-b) - \cos(a+b) = 2\sin(a)\sin(b). \tag{6.2}$$

Since $s \in (0,1]$, we have $0 < \Delta < 0.28s\sqrt{s} < \sqrt{s}$. Therefore, $\theta + \frac{\Delta}{2} \in \left(\sqrt{s}, \pi - \sqrt{s} + \frac{\Delta}{2}\right] \subset \left(\sqrt{s}, \pi - \frac{\sqrt{s}}{2}\right)$, and $\sin\left(\theta + \frac{\Delta}{2}\right) \geq \sin\left(\pi - \sqrt{s} + \frac{\Delta}{2}\right)$. Continuing from Equation 6.1, we have

$$\begin{aligned}
&|s + \exp(\mathrm{i}(\theta+\Delta)) - \exp(\mathrm{i}\theta)|^2 \\
\leq & s^2 - 4s\sin\left(\pi - \sqrt{s} + \frac{\Delta}{2}\right)\sin\left(\frac{\Delta}{2}\right) + \Delta^2 && \text{By } \cos(\Delta) \geq 1 - \frac{\Delta^2}{2}, \forall \Delta \in \mathbb{R} \\
= & s^2 - 2s\left(\cos\left(\sqrt{s} - \Delta\right) - \cos\left(\sqrt{s}\right)\right) + \Delta^2 && \text{By the trigonometric identity 6.2} \\
= & s^2 - 2s\left(\int_{\sqrt{s}-\Delta}^{\sqrt{s}} \sin(t)\,\mathrm{d}t\right) + \Delta^2 \\
\leq & s^2 - 2s(\sqrt{s} - (\sqrt{s} - \Delta))\sin\left(\sqrt{s} - \Delta\right) + \Delta^2 \\
\leq & s^2 - 2s\Delta(\sqrt{s} - \Delta)\sin(1) + \Delta^2 \\
= & s^2 - 2s\Delta\sin(1)\sqrt{s} + 2s\Delta^2\sin(1) + \Delta^2,
\end{aligned} \tag{6.3}$$

where the second last inequality follows because $\sin(t) \geq \sin\left(\sqrt{s} - \Delta\right)$ when $t \in [\sqrt{s} - \Delta, \sqrt{s}] \subset [0,1]$, and the last inequality is because $\sin(t) \geq t\sin(1)$ when $t \in [0,1]$.

Note that if the right hand side of Equation 6.3 is less than $\left(s - \frac{\sqrt{s}}{2}\Delta\right)^2$, the proof is complete. We have

$$\begin{aligned}
& s^2 - 2s\Delta\sin(1)\sqrt{s} + 2s\Delta^2\sin(1) + \Delta^2 < s^2 - s\sqrt{s}\Delta + \frac{s}{4}\Delta^2 \\
\iff \quad & \Delta^2((2\sin(1) - 1/4)s + 1) < \Delta(2\sin(1) - 1)s\sqrt{s} \\
\iff \quad & \Delta < \frac{(2\sin(1) - 1)s\sqrt{s}}{(2\sin(1) - 1/4)s + 1}.
\end{aligned}$$

Since $s \leq 1$, $(2\sin(1) - 1/4)s + 1 \leq 2\sin(1) - 1/4 + 1$,

$$\frac{(2\sin(1) - 1)s\sqrt{s}}{(2\sin(1) - 1/4)s + 1} \geq \frac{(2\sin(1) - 1)}{2\sin(1) - 1/4 + 1}s\sqrt{s} > 0.28s\sqrt{s}.$$

Therefore, when $\Delta < 0.28s\sqrt{s}$, $|s + \exp(\mathrm{i}(\theta+\Delta)) - \exp(\mathrm{i}\theta)| < s - \frac{\sqrt{s}}{2}\Delta$. $\square$

Lemmas 6.3, 6.5 and 6.6 combined formally explains case 2 at the beginning of the section 6.2. Specifically, denote $s := \sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)$. If $1 \geq |s| > 0$ and $\angle(s, \exp(\mathrm{i}\theta_i)) \leq \sqrt{|s|}$ or $\geq \pi - \sqrt{|S|}$ for all $i \in [d]$, by Lemma 6.3, there are $j, k \in [d]$ such that

$$\angle(s, \exp(\mathrm{i}\theta_j)), \angle(s, \exp(\mathrm{i}\theta_k)) \leq \sqrt{|s|}.$$

Then, by Lemma 6.6, there exists ω_j and ω_k such that $|\omega_j| + |\omega_k| \leq 10\sqrt{|s|}$ and

$$\exp(\mathrm{i}(\theta_j + \omega_j)) + \exp(\mathrm{i}(\theta_k + \omega_k)) + \sum_{i \in [d] \setminus \{j,k\}} \exp(\mathrm{i}\theta_i) = 0.$$

Lemma 6.3. *Suppose the angles* $\theta_1, \cdots, \theta_d \in \left[-\frac{\pi}{3}, \frac{\pi}{3}\right] \cup \left[\frac{2\pi}{3}, \frac{4\pi}{3}\right]$ *are sorted in non-decreasing order. Suppose further that* $d \geq 2$ *and* $\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)$ *is real and positive. Then,* $\theta_1, \theta_2 \in \left[-\frac{\pi}{3}, \frac{\pi}{3}\right]$.

Proof. By design, $\mathrm{Re}\{\exp(\mathrm{i}\theta_i)\}$ is either in $\left[\frac{1}{2}, 1\right]$ (when $\theta_i \in \left[-\frac{\pi}{3}, \frac{\pi}{3}\right]$) or in $\left[-1, -\frac{1}{2}\right]$ (when $\theta_i \in \left[\frac{2\pi}{3}, \frac{4\pi}{3}\right]$). Suppose all θ_i are in $\left[\frac{2\pi}{3}, \frac{4\pi}{3}\right]$, then $\mathrm{Re}\left\{\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right\} \leq -\frac{d}{2} < 0$.

Suppose only θ_1 is in $\left[-\frac{\pi}{3}, \frac{\pi}{3}\right]$, then $\mathrm{Re}\left\{\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right\} \leq 1 - \frac{d-1}{2}$, and

- When $d = 2$, to make $\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i) = \exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2)$ real, the imaginary parts of $\exp(\mathrm{i}\theta_1)$ and $\exp(\mathrm{i}\theta_2)$ must cancel. Either $\exp(\mathrm{i}\theta_2) = \exp(\mathrm{i}\theta_1)^*$ or $\exp(\mathrm{i}\theta_1) = -\exp(\mathrm{i}\theta_2)$. When $\exp(\mathrm{i}\theta_2) = \exp(\mathrm{i}\theta_1)^*$, $\theta_2 = -\theta_1 \in \left[-\frac{\pi}{3}, \frac{\pi}{3}\right]$, and it violates our assumption that only θ_1 is in $\left[-\frac{\pi}{3}, \frac{\pi}{3}\right]$. When $\exp(\mathrm{i}\theta_1) = -\exp(\mathrm{i}\theta_2)$, $\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i) = 0$ which is not positive, so this is a contradiction.

- When $d > 2$, $\operatorname{Re}\left\{\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right\} \leq 1 - \frac{d-1}{2} \leq 0$ and is not positive, so this is a contradiction.

Therefore, we must have both θ_1 and θ_2 in $\left[-\frac{\pi}{3}, \frac{\pi}{3}\right]$. □

The following inequality is used multiple times later, so we prove it as a lemma:

Lemma 6.4. $\arccos\left(\frac{2\cos^2(\sqrt{s}) - s}{2}\right) \leq 2\sqrt{s}$ *for any* $s \in [0,1]$.

Proof. When $s \in [0,1]$,

$$\begin{aligned}
& 2\sqrt{s} \geq \arccos\left(\frac{2\cos^2(\sqrt{s}) - s}{2}\right) \\
\iff \quad & 2\cos\big(2\sqrt{s}\big) - (2\cos^2(\sqrt{s}) - s) \leq 0 \\
\iff \quad & 2\cos\big(2\sqrt{s}\big) - (\cos\big(2\sqrt{s}\big) + 1 - s) \leq 0 \\
\iff \quad & \cos\big(2\sqrt{s}\big) - 1 + s \leq 0. \qquad (6.4)
\end{aligned}$$

Equation 6.4 holds for all $s \in [0,1]$ if and only if $\cos(2s) - 1 + s^2 \leq 0$ holds for all $s \in [0,1]$ because $f(x) = x^2$ is a bijection from $[0,1]$ to itself.

$\frac{\mathrm{d}}{\mathrm{d}s}\left(\cos(2s) - 1 + s^2\right) = 2s - 2\sin(2s)$ and $\frac{\mathrm{d}^2}{\mathrm{d}s^2}\left(\cos(2s) - 1 + s^2\right) = 2 - 4\cos(2s)$. The second derivative is negative when $s \in \left[0, \frac{\pi}{6}\right)$, and is positive in $s \in \left(\frac{\pi}{6}, 1\right]$, so the first derivative is 0 when $s = 0$, decreases between $s = 0$ and $s = \frac{\pi}{6}$, and increases between $s = \frac{\pi}{6}$ and $s = 1$. Thus, $\cos(2s) - 1 + s^2 \leq \max\{\cos(0) - 1 + 0^2, \cos(2) - 1 + 1^2\} = 0$ when $s \in [0,1]$. Therefore, $\arccos\left(\frac{2\cos^2(\sqrt{s}) - s}{2}\right) \leq 2\sqrt{s}$ for any $s \in [0,1]$. □

For any complex number $\rho\exp(\mathrm{i}\theta)$ with $\theta \in (-\pi, \pi]$, define

$$\arg(\rho\exp(\mathrm{i}\theta)) \coloneqq \theta,$$

when $\rho > 0$, and define

$$\arg(0) \coloneqq 0.$$

Lemma 6.5. *For an arbitrary* $s \in [0,1]$ *and* $\theta_1, \theta_2 \in [-\sqrt{s}, \sqrt{s}]$*, we have* $\arg(\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s) \in [-2\sqrt{s}, 2\sqrt{s}]$.

Proof. Denote $\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2)$ as $\rho \exp(\mathrm{i}\theta)$ with $\rho \geq 0$ and $\theta \in (-\pi, \pi]$, so

$$\arg(\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s) = \arg(\rho \exp(\mathrm{i}\theta) - s).$$

By the parallelogram rule for vector addition, $\theta \in [\min\{\theta_1, \theta_2\}, \max\{\theta_1, \theta_2\}] \subset [-\sqrt{s}, \sqrt{s}]$, and

$$\rho^2 = (\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2))(\exp(-\mathrm{i}\theta_1) + \exp(-\mathrm{i}\theta_2)) = 2 + 2\cos(\theta_1 - \theta_2).$$

Since $\theta_1, \theta_2 \in [-\sqrt{s}, \sqrt{s}]$, $\theta_1 - \theta_2 \in [-2\sqrt{s}, 2\sqrt{s}] \subset [-2, 2]$, so

$$\cos(\theta_1 - \theta_2) \in [\cos\big(2\sqrt{s}\big), 1],$$

and

$$\rho^2 \in [2 + 2\cos\big(2\sqrt{s}\big), 4].$$

Since $2 + 2\cos\big(2\sqrt{s}\big) = 2 + 2\cos^2(\sqrt{s}) - 2\sin^2(\sqrt{s}) = 4\cos^2(\sqrt{s})$, $\rho \in [2\cos\big(\sqrt{s}\big), 2]$. One crucial property that will be used later is $\rho \geq 2\cos\big(\sqrt{s}\big) \geq 2\cos(1) > 1 \geq s$.

With the bound on θ and ρ, we prove the following inequality geometrically:

$$|\arg(\rho \exp(\mathrm{i}\theta) - s)| \leq \left|\arg\left(2\cos\big(\sqrt{s}\big)\exp\big(\mathrm{i}\sqrt{s}\big) - s\right)\right|.$$

Suppose we fix θ and vary ρ, as in Figure 6.1, the shorter the height of the parallelogram, the larger the angle its diagonal incident with the origin makes to the real axis. That is, if $\rho > \rho' > 0$, then $\theta_2 > \theta_1$, where $\theta_1 \coloneqq \arg(\rho \exp(\mathrm{i}\theta) - s)$ and $\theta_2 \coloneqq \arg(\rho' \exp(\mathrm{i}\theta) - s)$.

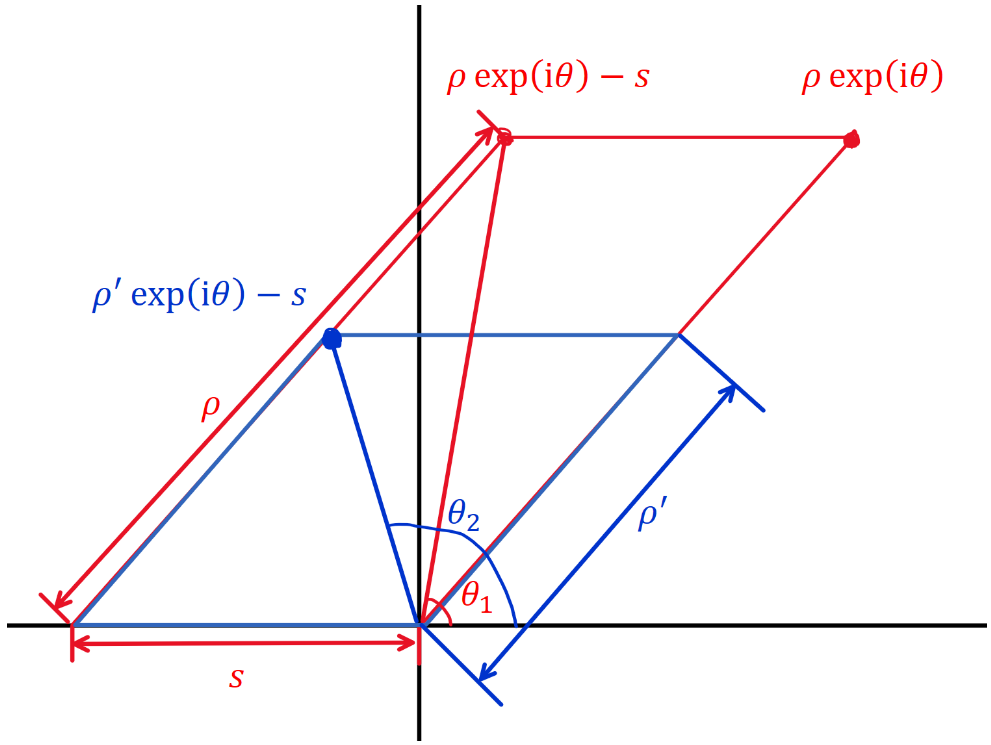

Figure 6.1: $\arg(\rho'\exp(\mathrm{i}\theta)-s) > \arg(\rho\exp(\mathrm{i}\theta)-s)$ for $\rho > \rho' > 0$.

For the case when θ is negative, by symmetry, we have

$$|\arg(\rho\exp(\mathrm{i}\theta)-s)| = |\arg(\rho\exp(-\mathrm{i}\theta)-s)|.$$

So we obtain the inequality

$$|\arg(\rho'\exp(\mathrm{i}\theta)-s)| > |\arg(\rho\exp(\mathrm{i}\theta)-s)|,$$

for any $\rho' \in (0,\rho)$.

If we fix ρ and vary θ, as in Figure 6.2, $\rho\exp(\mathrm{i}\theta)-s$ represents a point on a circle of radius ρ centered at $-s$. If $\pi \geq \theta' > \theta \geq 0$, then $\arg(\rho\exp(\mathrm{i}\theta')-s) > \arg(\rho\exp(\mathrm{i}\theta)-s)$.

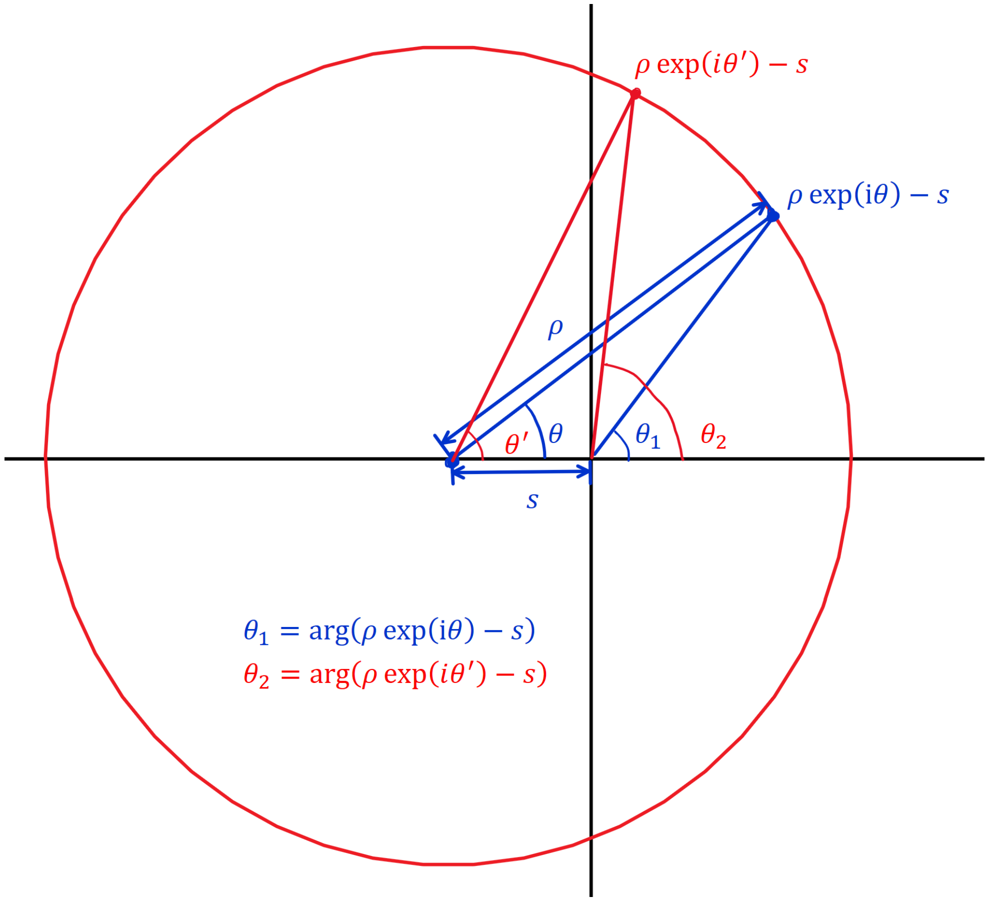

Figure 6.2: $\arg(\rho\exp(\mathrm{i}\theta') - s) > \arg(\rho\exp(\mathrm{i}\theta) - s)$ for $\pi \geq \theta' > \theta \geq 0$.

Similarly, in the case when θ or θ' is negative, if $|\theta'| > |\theta|$, then $|\arg(\rho\exp(\mathrm{i}\theta') - s)| > |\arg(\rho\exp(\mathrm{i}\theta) - s)|$. Hence, as we explain below,

$$\begin{aligned}
|\arg(\rho\exp(\mathrm{i}\theta) - s)| &\leq \left|\arg(2\cos\left(\sqrt{s}\right)\exp(\mathrm{i}\theta) - s)\right| \\
&\leq \left|\arg\left(2\cos\left(\sqrt{s}\right)\exp\left(\mathrm{i}\sqrt{s}\right) - s\right)\right| \\
&\leq \arccos\left(\frac{2\cos^2(\sqrt{s}) - s}{2}\right) \\
&\leq 2\sqrt{s}.
\end{aligned}$$

Here, the last inequality is by Lemma 6.4, and the third inequality is by the definition of arg function and the monotonicity of arccos function. The latter inequality can be seen more clearly with the help of Figure 6.3 below. Let $\omega := \arg\left(2\cos(\sqrt{s})\exp(\mathrm{i}\sqrt{s}) - s\right)$. From Figure 6.3, $\cos(\omega) = \dfrac{2\cos^2(\sqrt{s}) - s}{r} \geq \dfrac{2\cos^2(\sqrt{s}) - s}{2}$, so $\omega \leq \arccos\left(\dfrac{2\cos^2(\sqrt{s}) - s}{2}\right)$.

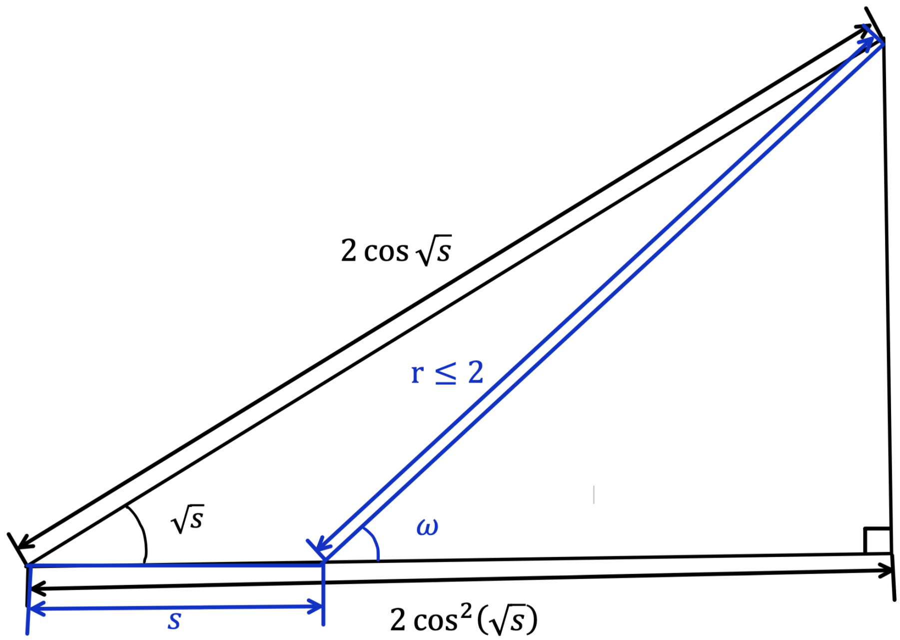

Figure 6.3: $\omega = \arccos\left(\dfrac{2\cos^2(\sqrt{s}) - s}{r}\right) \leq \arccos\left(\dfrac{2\cos^2(\sqrt{s}) - s}{2}\right)$.

$\square$

Lemma 6.6. *For an arbitrary $s \in [0,1]$, suppose $\theta_1, \theta_2 \in [-\sqrt{s}, \sqrt{s}]$. Then, there exists ω_1, ω_2, such that $|\omega_1| + |\omega_2| \leq 10\sqrt{s}$, and* $\exp(\mathrm{i}(\theta_1 + \omega_1)) + \exp(\mathrm{i}(\theta_2 + \omega_2)) = \exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s$.

Proof. For any $a, b \in \mathbb{C}$, we say a is aligned with b if $ab = 0$ or $\angle(a,b) = 0$. We first find ω_1' and ω_2' such that both $\exp(\mathrm{i}(\theta_1 + \omega_1'))$ and $\exp(\mathrm{i}(\theta_2 + \omega_2'))$ are aligned with $\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s$. Using Lemma 6.5,

$$|\arg(\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s)| \leq 2\sqrt{s},$$

and as $|\theta_1|, |\theta_2| \leq \sqrt{s}$, the angles

$$|\omega_1'|, |\omega_2'| \leq (2+1)\sqrt{s} = 3\sqrt{s}. \tag{6.5}$$

Since $\exp(\mathrm{i}(\theta_1 + \omega_1'))$ and $\exp(\mathrm{i}(\theta_2 + \omega_2'))$ are aligned,

$$|\exp(\mathrm{i}(\theta_1 + \omega_1')) + \exp(\mathrm{i}(\theta_2 + \omega_2'))| = 2.$$

Next, we derive bounds on $|\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|$. Define $\rho := |\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2)|$ and $\theta := \arg(\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2))$. By the proof of Lemma 6.3, $\rho \in [2\cos\big(\sqrt{s}\big), 2]$ and $\theta \in [-\sqrt{s}, \sqrt{s}]$. Applying the law of cosine to the triangle in Figure 6.4,

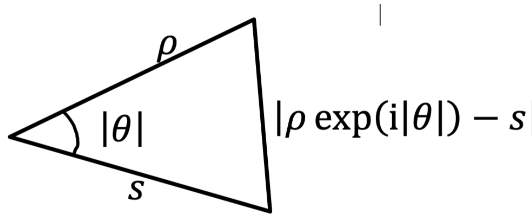

Figure 6.4: $|\rho \exp(\mathrm{i}|\theta|)| = |\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|$.

we get

$$|\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|^2 = \rho^2 + s^2 - 2\rho s \cos\theta.$$

Since

$$\frac{\partial}{\partial \rho}\left(\rho^2 + s^2 - 2\rho s \cos\theta\right) = 2(\rho - s\cos\theta)$$

and $s\cos\theta \leq 1$, but $\rho \geq 2\cos\sqrt{s} \geq 2\cos(1) > 1$, if we fix s and θ, $|\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|^2$ increases as ρ increases. Then,

$$\rho^2 + s^2 - 2\rho s\cos\theta \geq (2\cos\sqrt{s})^2 + s^2 - 4s\cos\sqrt{s} = (2\cos\sqrt{s} - s)^2,$$

$$\rho^2 + s^2 - 2\rho s\cos\theta \leq 2^2 + s^2 - 4s\cos\sqrt{s} \leq 4,$$

and the last inequality uses $4\cos\big(\sqrt{s}\big) \geq 2\cos\big(\sqrt{s}\big) \geq 2\cos(1) > 1 \geq s$ when $s \in [0,1]$. Therefore,

$$2 \geq |\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s| \geq 2\cos\big(\sqrt{s}\big) - s > 0,$$

and the last inequality guarantees $2\cos\big(\sqrt{s}\big) - s$ is always positive when s increases from 0 to 1. This condition is necessary as the sign change may potentially ruin the alignment because we consider two vectors pointing in opposite directions not aligned. Instead of writing $\max\{2\cos(\sqrt{s}) - s, 0\}$, we can thus write the quantity $2\cos(\sqrt{s}) - s$.

Since $\exp(\mathrm{i}(\theta_1 + \omega_1'))$ is aligned with $\exp(\mathrm{i}(\theta_2 + \omega_2'))$, and both of them are aligned with $\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s$ for any $x \in [0, \pi/2]$,

$$\exp(\mathrm{i}(\theta_1 + \omega_1' + x)) + \exp(\mathrm{i}(\theta_2 + \omega_2' - x))$$

is also aligned with $\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s$. In addition, if we make

$$|\exp(\mathrm{i}(\theta_1 + \omega_1' + x)) + \exp(\mathrm{i}(\theta_2 + \omega_2' - x))| = |\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|,$$

then $\exp(\mathrm{i}(\theta_1 + \omega_1' + x)) + \exp(\mathrm{i}(\theta_2 + \omega_2' - x)) = \exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s$ because if two complex numbers viewed as vectors are aligned, and they have the same length, then they must be equal.

Since $\exp(\mathrm{i}(\theta_1 + \omega_1'))$ and $\exp(\mathrm{i}(\theta_2 + \omega_2'))$ are aligned,

$$\begin{aligned}
&|\exp(\mathrm{i}(\theta_1 + \omega_1' + x)) + \exp(\mathrm{i}(\theta_2 + \omega_2' - x))| \\
=&|\exp(\mathrm{i}(\theta_1 + \omega_1' + x)) + \exp(\mathrm{i}(\theta_1 + \omega_1' - x))| \\
=&|\exp(\mathrm{i}x) + \exp(-\mathrm{i}x)| \\
=&2\cos(x).
\end{aligned}$$

To make $|\exp(\mathrm{i}(\theta_1 + \omega_1' + x)) + \exp(\mathrm{i}(\theta_2 + \omega_2' - x))| = |\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|$, we choose x so that

$$2\cos(x) = |\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|.$$

A solution for x exists because when $x \in [0, \pi/2]$, $2\cos(x)$ takes all values in $[0, 2]$ which contains $|\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s|$. Furthermore, by Lemma 6.4,

$$\begin{aligned}
& 2\cos(x) = |\exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s| \geq 2\cos(\sqrt{s}) - s \\
\implies \quad & x \leq \arccos\left(\frac{2\cos(\sqrt{s}) - s}{2}\right) \leq \arccos\left(\frac{2\cos^2(\sqrt{s}) - s}{2}\right) \leq 2\sqrt{s}.
\end{aligned}$$

If we define $\omega_1 := \omega_1' + x$ and $\omega_2 := \omega_2' - x$, $\exp(\mathrm{i}(\theta_1 + \omega_1)) + \exp(\mathrm{i}(\theta_2 + \omega_2)) = \exp(\mathrm{i}\theta_1) + \exp(\mathrm{i}\theta_2) - s$, and by Inequality 6.5, $|\omega_1| + |\omega_2| \leq 2(3 + 2)\sqrt{s} = 10\sqrt{s}$. $\square$

The following lemma helps prove the convergence of the algorithm when it runs indefinitely. The idea comes from the differential inequality

$$\begin{cases} \frac{\mathrm{d}f}{\mathrm{d}x} \le -c\sqrt{f(x)}, \\ c > 0, \\ f(x) \ge 0, \forall x \in \mathbb{R}, \end{cases} \tag{6.6}$$

that the solution to 6.6 decreases to 0 quickly. One can verify it by solving the differential equation with the inequality in 6.6 replaced by an equality, and applying Bihari–LaSalle inequality [12, 2]. Furthermore, using the definitions from below, B_N's can be considered as the total cost of the algorithm until the N-th iteration and the a_N's can be considered as the objective value at the N-th iteration. Lemma 6.7 shows that if the objective values a_N decreases quickly in each iteration of the algorithm, then the total cost B_N will not be too large when a_N becomes small or 0.

Lemma 6.7. *Let $((a_i, b_i) : i \in \mathbb{N}_{\ge 1})$ be a sequence in $\mathbb{R}_{\ge 0} \times \mathbb{R}_{>0}$ such that $0 \le a_{i+1} \le a_i - \frac{b_i\sqrt{a_i}}{2}$ for all $i \in \mathbb{N}_{\ge 1}$. For an arbitrary $N \in \mathbb{N}_{\ge 1}$, define $B_N := \sum_{i=1}^{N-1} b_i$ when $N > 1$ and $B_1 = 0$. For any $N \ge 1$, if $B_N \le 4\sqrt{a_1}$, then $a_N \le \frac{(4\sqrt{a_1} - B_N)^2}{16}$, or if $B_N > 4\sqrt{a_1}$, then $a_N = 0$.*

Proof. We prove this lemma by induction. When $N = 1$, $B_1 = 0 \le 4\sqrt{a_1}$, and $a_1 \le \frac{(4\sqrt{a_1} - 0)^2}{16} = a_1$.

For an arbitrary $N \in \mathbb{N}_{\ge 1}$, suppose if $B_N \le 4\sqrt{a_1}$, then $a_N \le \frac{(4\sqrt{a_1} - B_N)^2}{16}$, or if $B_N \ge 4\sqrt{a_1}$, then $a_N = 0$.

- If $B_N \le 4\sqrt{a_1}$, then
  - If $a_N > 0$, $a_N \le \frac{(4\sqrt{a_1} - B_N)^2}{16}$ implies $B_N \le 4(\sqrt{a_1} - \sqrt{a_N})$, and $0 \le a_{N+1} \le a_N - \frac{b_N\sqrt{a_N}}{2}$ implies $b_N \le 2\sqrt{a_N}$. Therefore, $B_{N+1} = B_N + b_N \le 4\sqrt{a_1} - 2\sqrt{a_N} \le 4\sqrt{a_1}$, and

$$a_{N+1} \le a_N - \frac{b_N\sqrt{a_N}}{2}$$

$$
\begin{aligned}
&\leq \frac{(4\sqrt{a_1}-B_N)^2}{16} - \frac{b_n}{2}\sqrt{\frac{(4\sqrt{a_1}-B_N)^2}{16}} \\
&= \frac{(4\sqrt{a_1}-B_N)^2}{16} - \frac{b_n(4\sqrt{a_1}-B_N)}{8} \\
&= \frac{(4\sqrt{a_1}-B_N)(4\sqrt{a_1}-B_N-2b_n)}{16} \\
&= \frac{(4\sqrt{a_1}-B_N-b_n+b_n)(4\sqrt{a_1}-B_N-b_n-b_n)}{16} \\
&< \frac{(4\sqrt{a_1}-B_N-b_n)^2}{16} \\
&= \frac{(4\sqrt{a_1}-B_{N+1})^2}{16}.
\end{aligned}
$$

- If $a_N = 0$, then $0 \leq a_{N+1} \leq a_N - \dfrac{b_N\sqrt{a_N}}{2}$, so $a_{N+1} = 0$. If $B_{N+1} \leq 4\sqrt{a_1}$, then $a_{N+1} = 0 \leq \dfrac{(4\sqrt{a_1}-B_{N+1})^2}{16}$. If $B_{N+1} > 4\sqrt{a_1}$, we have $a_{N+1} = 0$.

• If $B_N > 4\sqrt{a_1}$, then $a_N = 0$, $a_{N+1} = 0$.

□

Corollary 6.8. *Let $((a_i, b_i) : i \in \mathbb{N}_{\geq 1})$ be a sequence in $\mathbb{R}_{\geq 0} \times \mathbb{R}_{>0}$ such that $0 \leq a_{i+1} \leq a_i - \dfrac{b_i\sqrt{a_i}}{2}$ for all $i \in \mathbb{N}_{\geq 1}$. If $N := \inf\{i : a_i = 0, i \in \mathbb{N}_{\geq 1}\}$ exists, then $B_N \leq 4\sqrt{a_1}$.*

Proof. Suppose such N exists. If $N = 1$, then $B_1 = 0 \leq 4\sqrt{a_1}$. If $N > 1$, we have $a_{N-1} > 0$, so $B_{N-1} \leq 4\sqrt{a_1}$ by Lemma 6.7. By the proof of Lemma 6.7 for the case when $B_{N-1} \leq 4\sqrt{a_1}$ and $a_{N-1} > 0$, $B_N = B_{N-1} + b_{N-1} \leq 4\sqrt{a_1} - 2\sqrt{a_{N-1}} \leq 4\sqrt{a_1}$. □

Theorem 6.9. *Suppose we have angles $\theta_1, \cdots, \theta_d \in \mathbb{R}$ such that $\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| \leq 1$, then there exists $\omega_1, \cdots, \omega_d \in \mathbb{R}$ such that $\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i)) = 0$ and*

$$\sum_{i=1}^{d} |\omega_i| < 14\sqrt{\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right|}.$$

Proof. All variables in Algorithm 1 are global variables and are initialized to 0. We run Algorithm 1 below by calling MAIN on line 24 with arguments $\theta_1, \cdots, \theta_d$. The high-level idea of each function is as follows:

- MAIN finds a small Δ and then call ADJUST.
- In ADJUST, if there exists $j \in [d]$ such that $\exp(\mathrm{i}\theta_j)$'s component orthogonal to $\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)$ is not "too small," then we increase/decrease θ_i by Δ so that $\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right|$ becomes smaller. This process is repeated until such j does not exist or Δ is too large. If such j does not exist, we call FIX. If Δ is too large, we return to MAIN and update Δ to a smaller value.
- FIX is called when $\exp(\mathrm{i}\theta_j)$'s component orthogonal to $\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)$ is "small" for all $j \in [d]$. By Lemmas 6.3 and 6.6, we can find $j_1, j_2 \in [d]$ and modify θ_{j_1} and θ_{j_2} by a small amount, so that $\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right|$ becomes 0 immediately.

Note that we do not actually change θ_i for a clearer proof later. We use ω_i to represent the total change to θ_i.

Algorithm 1

1: **procedure** FIX
2: $\quad \exists(\omega'_i)_{i\in[d]}$ `such that` $\sum_{i=1}^{n} |\omega'_i| \le 10\sqrt{S_k}$ `and` $\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i + \omega'_i)) = 0$
3: $\quad$ ▷ See explanation later on how to construct ω'_i
4: $\quad$ **for all** $i \in [d]$ **do**
5: $\quad\quad \omega_i \leftarrow \omega_i + \omega'_i$
6: $\quad$ **end for** $\quad$ ▷ After the for-loop, $\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i)) = 0$
7: **end procedure**
8: **procedure** ADJUST
9: $\quad$ **while** $\Delta < 0.28|S_k|\sqrt{|S_k|}$ **do** $\quad$ ▷ Check if Δ is still small enough
10: $\quad\quad$ **if** $\exists j \in [d], \angle(\exp(\mathrm{i}(\theta_j + \omega_j)), S_k) \in (\sqrt{|S_k|}, \pi - \sqrt{|S_k|})$ **then**
11: $\quad\quad\quad$ **if** $\mathrm{Im}\{S_k^* \exp(\mathrm{i}(\theta_j + \omega_j))\} > 0$ **then**

12: ▷ Check cross product to determine the rotation direction
13: $\omega_j \leftarrow \omega_j + \Delta$
14: **else**
15: $\omega_j \leftarrow \omega_j - \Delta$
16: **end if**
17: $\Omega_k \leftarrow \Delta$ ▷ Record the amount of change at iteration k
18: **else**
19: Fix
20: **end if**
21: $k \leftarrow k+1,\ S_k \leftarrow \sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i))$
22: **end while**
23: **end procedure**
24: **procedure** Main$(\theta_1, \cdots, \theta_n)$
25: $k \leftarrow 1, |S_1| \leftarrow \sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)$ ▷ $S_1 \leq 1$ by assumption
26: **while** $|S_k| > 0$ **do**
27: $\Delta \leftarrow \min\left\{\frac{1}{2^k}, \frac{|S_k|^{\frac{3}{2}}}{10}\right\}$ ▷ A small (positive) value less than $0.28|S_k|^{\frac{3}{2}}$
28: Adjust ▷ After the procedure call, $|S_k| < \left(\frac{\Delta}{0.28}\right)^{\frac{2}{3}}$
29: **end while**
30: **end procedure**

We verify some properties of Algorithm 1:

- If the if-condition on line 10 evaluates to true, and if we further assume $S_k \leq 1$, by the while-loop condition on line 9 and Lemma 6.2, $|S_{k+1}| \leq |S_k| - \frac{\sqrt{|S_k|}}{2}\Omega_k$. Since $|S_1| \leq 1$, a short induction can show until FIX on line 19 is called for the first time, $(|S_k|)$ is a decreasing sequence in $[0, 1]$.

- When FIX on line 19 is called for the first time, the previous point shows $|S_k| \leq 1$. The if-statement on line 10 and Lemma 6.3 further show we can find $j_1, j_2 \in [d]$ such that $\angle(\exp(\mathrm{i}(\theta_{j_1} + \omega_{j_1})), S_k), \angle(\exp(\mathrm{i}(\theta_{j_2} + \omega_{j_2})), S_k) \in [0, \sqrt{|S_k|}]$. Then, by Lemma 6.6, we can change ω_{j_1} and ω_{j_2} by at most $10\sqrt{|S_k|}$ to make $\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i)) =$

0. This explains the assertion on line 2.

- When we return from FIX back to line 19, $k \leftarrow k+1$ and $S_k \leftarrow 0$, the algorithm will terminate after jumping out of the two while-loops on line 9 and on line 26 as $\Delta > 0$ and $S_k = 0$.

- If FIX on line 19 is never called and assume the while-condition on line 9 evaluates to true, by Lemmas 6.2 6.7 and Corollary 6.8, $\sum_{i=1}^{k-1} \Omega_i$ has to be smaller than $4(\sqrt{|S_1|} - \sqrt{|S_k|})$ for all k. After each iteration of the while-loop on line 9, k increases by 1, and $\sum_{i=1}^{k-1} \Omega_i$ increases by Δ, so eventually, $\sum_{i=1}^{k-1} \Omega_i$ will exceed $4(\sqrt{|S_1|} - \sqrt{|S_k|}) \le 4\sqrt{|S_1|}$, and the while-condition on line 9 will evaluate to false.

- The previous two points show the while-loop on line 9 terminates regardless of whether FIX on line 19 is called.

If Algorithm 1 eventually terminates with $k = K$, define $\omega_i^{(j)}$ as the ω_i at the snapshot when S_j is computed, then at the end, as we explain below,

$$\sum_{i=1}^{d} |\omega_i| = \sum_{i=1}^{d} \left|\omega_i^{(K)}\right| \le \sum_{k'=1}^{K-1} \sum_{i=1}^{d} \left|\omega_i^{(k'+1)} - \omega_i^{(k')}\right| \le 4\sqrt{|S_1|} + 10\sqrt{|S_1|} = 14\sqrt{|S_1|}.$$

The second inequality holds because

$$\sum_{i=1}^{d} \left|\omega_i^{(k+1)} - \omega_i^{(k)}\right| \le \Omega_k, \qquad \forall 1 \le k < K-1,$$

and

- If FIX on line 19 is not called,

$$\sum_{i=1}^{d} \left|\omega_i^{(K)} - \omega_i^{(K-1)}\right| \le \Omega_{K-1},$$

  and by Lemma 6.7 and Corollary 6.8 on sequence $(|S_i|, \Omega_i)$, $\sum_{i=1}^{K-1} \Omega_i \le 4\sqrt{|S_1|}$.

- If FIX on line 19 is called in the end,

$$\sum_{i=1}^{d} \left|\omega_i^{(K)} - \omega_i^{(K-1)}\right| \le 10\sqrt{|S_1|},$$

and by Lemma 6.7 and Corollary 6.8 on sequence $(|S_i|, \Omega_i)$, $\sum_{i=1}^{K-2} \Omega_i \le 4\sqrt{|S_1|}$.

If the algorithm does not terminate, then FIX on line 19 is never called. When Δ is updated on line 27, since $\frac{|S_k|^{\frac{3}{2}}}{10} < 0.28|S_k|^{\frac{3}{2}}$, whenever ADJUST on line 28 is called, the while-condition on line 9 is satisfied initially, so k increases after calling ADJUST on line 28 and goes to infinity as the algorithm runs indefinitely.

$\Delta \in \left(0, \frac{1}{2^k}\right]$, and the while-condition on line 9 ensures that after calling ADJUST on line 28, $|S_k| \in \left(0, \left(\frac{\Delta}{0.28}\right)^{\frac{2}{3}}\right] \subset \left(0, \left(\frac{1}{0.28 \cdot 2^k}\right)^{\frac{2}{3}}\right]$. When $k = k'$, for any $k_2 > k_1 \ge k'$, by Lemma 6.7,

$$\begin{aligned}
\sum_{i=1}^{d} \left|\omega_i^{(k_2)} - \omega_i^{(k_1)}\right| &\le \sum_{j=k_1}^{k_2-1} \sum_{i=1}^{d} \left|\omega_i^{(j+1)} - \omega_i^{(j)}\right| \\
&\le \left(\sum_{j=k_1}^{k_2-1} \Omega_j\right) \\
&\le 4\sqrt{|S_{k_1}|} \\
&\le 4\sqrt{|S_{k'}|} \\
&\le 4\sqrt{\left(\frac{1}{0.28 \cdot 2^{k'}}\right)^{\frac{2}{3}}}.
\end{aligned}$$

So the sequence $\left(\left(\omega_i^{(j)}\right)_{i\in[d]}\right)_{j\in\mathbb{N}^+}$ is a Cauchy-sequence in compact set $\left[-20\sqrt{|S_1|}, 20\sqrt{|S_1|}\right]^d$, and it has a limit. Since the function $f(\omega_1, \cdots, \omega_d) = \left|\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i))\right|$ is continuous,

$$f\left(\lim_{k\to\infty} \omega_1^{(k)}, \cdots, \lim_{k\to\infty} \omega_d^{(k)}\right) = \lim_{k\to\infty} f\left(\omega_1^{(k)}, \cdots, \omega_d^{(k)}\right) = \lim_{k\to\infty} |S_k| = 0.$$

Therefore, if we define $\omega_i := \lim_{k\to\infty} \omega_i^{(k)}$ for all $i \in [d]$, then by Lemma 6.7,

$$\sum_{i=1}^{d} |\omega_i| = \sum_{i=1}^{d} \left| \lim_{k\to\infty} \omega_i^{(k)} \right| \le \lim_{k\to\infty} \sum_{j=1}^{k-1} \sum_{i=1}^{d} \left| \omega_i^{(j+1)} - \omega_i^{(j)} \right| \le \left(\lim_{k\to\infty} \sum_{j=1}^{k-1} \Omega_k \right) \le 4\sqrt{|S_1|}.$$

□

Lemma 6.10. *Suppose we have angles $\theta_1, \cdots, \theta_d \in \mathbb{R}$ such that $\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| > 1$. Then, there exist $\omega_1, \cdots, \omega_d \in \mathbb{R}$ such that $\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i)) = 0$ and $\sum_{i=1}^{d} |\omega_i| < \frac{\pi}{2}\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| + \frac{\pi}{2} + 14$.*

Proof. All variables in Algorithm 2 are initialized to 0. We run Algorithm 2 below by calling MAIN with arguments $\theta_1, \cdots, \theta_d$.

Algorithm 2

---

procedure MAIN$(\theta_1, \cdots, \theta_d)$
    $S \leftarrow \sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)$
    while $|S| > 1$ **do** ▷ This while-loop's functionality is explained later
        $j \leftarrow \arg\min_{k\in[d]} \{\angle(\exp(\mathrm{i}(\theta_k + \omega_k)), S)\}$
        $\alpha \leftarrow \angle(\exp(\mathrm{i}(\theta_j + \omega_j)), S)$
        if $\mathrm{Im}\{S^* \exp(\mathrm{i}(\theta_j + \omega_j))\} > 0$ **then**
            $\omega_j \leftarrow \omega_j + \pi - 2\alpha$
        else
            $\omega_j \leftarrow \omega_j - \pi + 2\alpha$
        end if
        $S \leftarrow \sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i))$
    end while
end procedure

In each iteration of the while-loop, we find $j \in [d]$ such that $\exp(\mathrm{i}(\theta_j + \omega_j))$ has the largest signed projection length onto S. Since $S = \sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i))$, the projection from

$\exp(\mathrm{i}(\theta_j + \omega_j))$ to S has length at least $\dfrac{|S|}{d} > \dfrac{1}{d}$. Then, we can reduce $|S|$ by modifying ω_j, so that the signed projection length is negated while the component orthogonal to S is unchanged. See Figure 6.5 for an illustration.

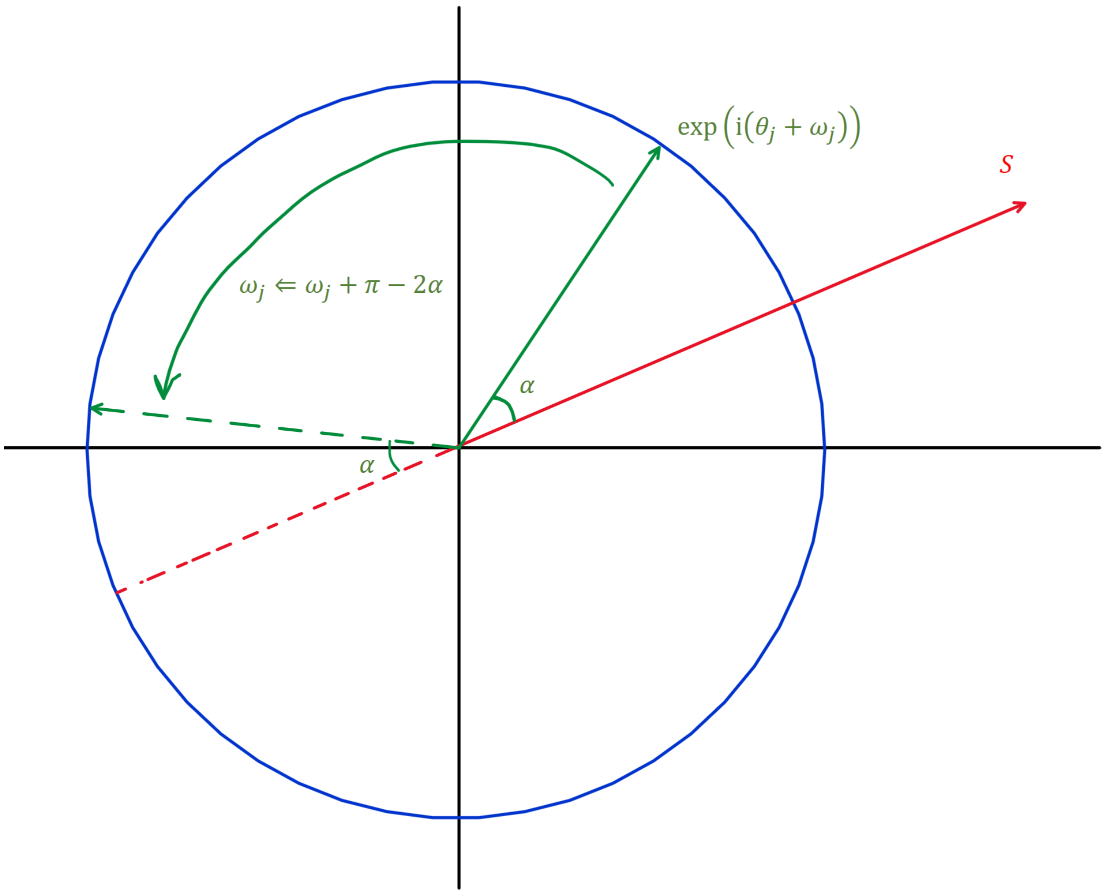

Figure 6.5: Illustration of each while-loop iteration in Algorithm 2.

Let $\beta \coloneqq \dfrac{\pi}{2} - \alpha$, where $\alpha \coloneqq \angle(\exp(\mathrm{i}(\theta_j + \omega_j)), S)$ as on line 5. Each iteration of the while-loop reduces $|S|$ as $||S| - 2\cos\alpha| = ||S| - 2\sin\beta|$ with $\displaystyle\sum_{i=1}^{d} |\omega_i|$ increasing by at most $(\pi - 2\alpha) = 2\beta$. Since $2\sin\beta \in \left(\dfrac{2}{d}, 2\right]$ where the lower bound is from the previous paragraph, either $|S| \geq 2\sin\beta$ and $||S| - 2\sin\beta| = |S| - 2\sin\beta$, or $|S| < 2\sin\beta$, and

$||S| - 2\sin\beta| \le 1$. In the former case, $|S|$ decreases by at least $\frac{2}{d}$ per iteration of the while-loop, and in the latter case, the algorithm terminates immediately. Therefore, Algorithm 2 terminates with $\left|\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i))\right| \le 1$, and as explained below,

$$\sum_{i=1}^{d} |\omega_i| \le \left(\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| + 1\right) \sup_{\beta\in[0,\frac{\pi}{2}]} \left\{\frac{2\beta}{2\sin\beta}\right\} = \frac{\pi}{2}\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| + \frac{\pi}{2}.$$

The first inequality is because S decreases from $\left|\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i))\right|$ down to -1 (it is -1 because $2\sin\beta$ might be larger than S in the last round). In each round, to decrease S by $2\sin(\beta)$, $\sum_{i=1}^{d} |\omega_i|$ increases by at most 2β, and we bound this by the sup.

After the execution of Algorithm 2, $|S| \le 1$, we use Theorem 6.9 to finish the proof. This last step increases $\sum_{i=1}^{d} |\omega_i|$ by at most 14. □

Corollary 6.11. *Suppose we have angles $\theta_1, \cdots, \theta_d \in \mathbb{R}$ such that $\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| > 1$. Then, there exist $\omega_1, \cdots, \omega_d \in \mathbb{R}$ such that $\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i)) = 0$ and $\sum_{i=1}^{d} |\omega_i|^2 < 96\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right|$.*

Proof. Notice in Lemma 6.10, either $\arg(S)$ is unchanged, or in the last iteration, possibly $\arg(S) \leftarrow \pi - \arg(S)$. Whenever an ω_i is changed, either $\exp(\mathrm{i}(\theta_i + \omega_i))$ has a negative projection onto S, or it is in the last iteration. In both cases, ω_i will never be changed again in future iterations of Algorithm 2. Therefore, we can further bound

$$\sum_{i=1}^{d} |\omega_i^2| \le \left(\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| + 1\right) \sup_{\beta\in[0,\frac{\pi}{2}]} \left\{\frac{(2\beta)^2}{2\sin\beta}\right\} = \frac{\pi^2}{2}\left|\sum_{i=1}^{n} \exp(\mathrm{i}\theta_i)\right| + \frac{\pi^2}{2}.$$

Then, we apply Theorem 6.9 with angles $\theta_i + \omega_i$ and obtain angles ω_i' such that

$$\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i + \omega_i')) = 0,$$

and

$$\sum_{i=1}^{d} |\omega_i'| \leq 14\sqrt{\left|\sum_{i=1}^{d} \exp(\mathrm{i}(\theta_i + \omega_i))\right|} \leq 14.$$

We may assume $\omega_i' \in [-\pi, \pi]$, so

$$\sum_{i=1}^{d} |\omega_i'|^2 \leq \pi \sum_{i=1}^{d} |\omega_i'| \leq 14\pi,$$

and

$$\begin{aligned}
\sum_{i=1}^{d} |\omega_i + \omega_i'|^2 &\leq \sum_{i=1}^{d} \left(|\omega_i|^2 + 2|\omega_i||\omega_i'| + |\omega_i'|^2\right) \\
&\leq \frac{\pi^2}{2}\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| + \frac{\pi^2}{2} + 2\sqrt{\left(\frac{\pi^2}{2}\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| + \frac{\pi^2}{2}\right) 14\pi} + 14\pi \\
&\leq \frac{\pi^2}{2}\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| + \frac{\pi^2}{2} + 14\pi + 2\sqrt{14\pi^3}\sqrt{\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right|} \\
&\leq \left(\frac{\pi^2}{2} + \frac{\pi^2}{2} + 14\pi + 2\sqrt{14\pi^3}\right)\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| \\
&< 96\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right|,
\end{aligned}$$

where the second line is by the Cauchy-Schwartz inequality, and the following lines use

$$\left|\sum_{i=1}^{d} \exp(\mathrm{i}\theta_i)\right| > 1.$$

□

6.3 Orthogonalizing two unitary matrices

We combine the algorithms from the previous section to orthogonalize a pair of unitary operators in such a way that one of them is perturbed only slightly.

Theorem 1.5. *Suppose we have $U_1, U_2 \in \mathcal{U}(d)$ for any $d \geq 2$ such that*

$$|\langle U_1, U_2 \rangle| = \left| \frac{1}{d} \operatorname{Tr}\left(U_1^\dagger U_2 \right) \right| \leq \epsilon,$$

then, there exists $U \in \mathcal{U}(d)$ such that $\langle U_1, UU_2 \rangle = 0$ and

$$\|UU_2 - U_2\|_{nhs}^2 = \|U - \mathbb{I}_d\|_{nhs}^2 \leq 196\epsilon = O(\epsilon),$$

where $\|M\|_{nhs} \coloneqq \sqrt{\frac{1}{d} \operatorname{Tr}(M^\dagger M)}$, for any $M \in \mathbb{C}^{d \times d}$.

Proof. We continue from the derivations in section 6.1. Let the spectral decomposition of $U_1^\dagger U_2$ be $\tilde{U} D \tilde{U}^\dagger$ where $\tilde{U}$ is unitary and $D \coloneqq \operatorname{diag}\left(\mathrm{e}^{\mathrm{i}\theta_1}, \mathrm{e}^{\mathrm{i}\theta_2}, \cdots, \mathrm{e}^{\mathrm{i}\theta_d} \right)$ is diagonal. Section 6.1 showed we can find the required $U = U_1 \tilde{U} D' \tilde{U}^\dagger U_1^\dagger$ by finding $D' = \operatorname{diag}\left(\mathrm{e}^{\mathrm{i}\omega_1}, \mathrm{e}^{\mathrm{i}\omega_2}, \cdots, \mathrm{e}^{\mathrm{i}\omega_d} \right)$ such that $\operatorname{Tr}(D'D) = 0$, and $\|U - \mathbb{I}_d\|_{\mathrm{nhs}} = \frac{1}{\sqrt{d}} \|U - \mathbb{I}_d\|_{\mathrm{F}} = \frac{1}{\sqrt{d}} \|D' - \mathbb{I}_d\|_{\mathrm{F}}$.

We are given that $\left| \operatorname{Tr}\left(U_1^\dagger U_2 \right) \right| \leq d\epsilon$. Depending on whether $d\epsilon \leq 1$ or not, there are two cases:

- If $d\epsilon \leq 1$, by Theorem 6.9, there exists D' such that

  $$\operatorname{Tr}(DD') = \sum_{i=1}^{d} e^{\mathrm{i}(\theta_i + \omega_i)} = 0,$$

  and

  $$\sum_{i=1}^{d} |\omega_i| \leq 14 \sqrt{\left| \sum_{i=1}^{d} e^{\mathrm{i}\theta_i} \right|} \leq 14\sqrt{d\epsilon}.$$

  Notice in a sector with central angle $|\omega_i|$ and radius 1, the chord length $\left| \mathrm{e}^{\mathrm{i}\omega_i} - 1 \right|$ is less than the arc length $|\omega_i|$. Thus,

  $$\|D' - \mathbb{I}_d\|_{\mathrm{nhs}}^2 = \frac{1}{d} \sum_{i=1}^{d} \left| \mathrm{e}^{\mathrm{i}\omega_i} - 1 \right|^2 \leq \frac{1}{d} \sum_{i=1}^{d} |\omega_i|^2 \leq \frac{1}{d} \left(\sum_{i=1}^{d} |\omega_i| \right)^2 \leq 196\epsilon,$$

  and the second-last inequality uses the convexity of function $f(x) = x^2$.

- If $d\epsilon > 1$, by Corollary 6.11, there exists D' such that

$$\mathrm{Tr}(DD') = \sum_{i=1}^{d} e^{\mathrm{i}(\theta_i + \omega_i)} = 0,$$

and

$$\sum_{i=1}^{d} |\omega_i|^2 \leq 96 \left| \sum_{i=1}^{d} e^{\mathrm{i}\theta_i} \right| \leq 96 d\epsilon.$$

Then,

$$\|D' - \mathbb{I}_d\|_{\mathrm{nhs}}^2 = \frac{1}{d} \sum_{i=1}^{d} \left| e^{\mathrm{i}\omega_i} - 1 \right|^2 \leq \frac{1}{d} \sum_{i=1}^{d} |\omega_i|^2 \leq 96\epsilon.$$

□

References

[1] Charles H. Bennett and Stephen J. Wiesner. Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states. *Physical Review Letters*, 69(20):2881–2884, November 1992.

[2] Imre Bihari. A generalization of a lemma of bellman and its application to uniqueness problems of differential equations. *Acta Mathematica Academiae Scientiarum Hungaricae*, 7(1):81–94, March 1956.

[3] John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23(15):880–884, October 1969.

[4] Andrea Coladangelo, Alex B. Grilo, Stacey Jeffery, and Thomas Vidick. Verifier-on-a-leash: New schemes for verifiable delegated quantum computation, with quasilinear resources. In *Advances in Cryptology – EUROCRYPT 2019*, pages 247–277. Springer International Publishing, 2019.

[5] Máté Farkas and Jędrzej Kaniewski. Self-testing mutually unbiased bases in the prepare-and-measure scenario. *Physical Review A*, 99(3), March 2019.

[6] Máté Farkas, Jędrzej Kaniewski, and Ashwin Nayak. Mutually unbiased measurements, Hadamard matrices, and Superdense Coding. *IEEE Transactions on Information Theory*, 69(6):3814–3824, June 2023.

[7] Semyon A. Gerschgorin. Uber die abgrenzung der eigenwerte einer matrix. *Izvestija Akademii Nauk SSSR, Serija Matematika*, 7(3):749–754, 1931.

[8] Alexei Gilchrist, Nathan K. Langford, and Michael A. Nielsen. Distance measures to compare real and ideal quantum processes. *Physical Review A*, 71(6), June 2005.

[9] Zhengfeng Ji, Anand Natarajan, Thomas Vidick, John Wright, and Henry Yuen. MIP* = RE. *Commun. ACM*, 64(11):131–138, October 2021.

[10] Phillip Kaye, Raymond Laflamme, and Michele Mosca. *An Introduction to Quantum Computing*. Oxford University Press, Inc., USA, 2007.

[11] Robert S. Kennedy. On the optimum quantum receiver for the M-ary linearly independent pure state problem. Quarterly Progress Report 110, Research Laboratory of Electronics (RLE) at the Massachusetts Institute of Technology (MIT), https://dspace.mit.edu/handle/1721.1/56399, July 1973.

[12] Joseph P. LaSalle. Uniqueness theorems and successive approximations. *The Annals of Mathematics*, 50(3):722, July 1949.

[13] Per-Olov Löwdin. On the nonorthogonality problem. In *Advances in Quantum Chemistry*, pages 185–199. Elsevier, 1970.

[14] Dominic Mayers and Andrew Chi-Chih Yao. Quantum cryptography with imperfect apparatus. In *39th Annual Symposium on Foundations of Computer Science, FOCS '98, November 8-11, 1998, Palo Alto, California, USA*, pages 503–509. IEEE Computer Society, 1998.

[15] Dominic Mayers and Andrew Chi-Chih Yao. Self testing quantum apparatus. *Quantum Info. Comput.*, 4(4):273–286, July 2004.

[16] Carl A. Miller and Yaoyun Shi. Optimal robust self-testing by binary nonlocal xor games. 2013.

[17] Ashwin Nayak and Peter Shor. Bit-commitment-based quantum coin flipping. *Physical Review A*, 67(1), January 2003.

[18] Ashwin Nayak and Henry Yuen. Rigidity of superdense coding. *ACM Transactions on Quantum Computing*, 4(4):1–39, July 2023.

[19] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, USA, 10th edition, 2011.

[20] Alexey E. Rastegin. A lower bound on the relative error of mixed-state cloning and related operations. *Journal of Optics B: Quantum and Semiclassical Optics*, 5(6):S647–S650, October 2003.

[21] Alexey E. Rastegin. Sine distance for quantum states, 2006.

[22] Ivan Šupić and Joseph Bowles. Self-testing of quantum systems: a review. *Quantum*, 4:337, September 2020.

[23] Armin Tavakoli, Massimiliano Smania, Tamás Vértesi, Nicolas Brunner, and Mohamed Bourennane. Self-testing nonprojective quantum measurements in prepare-and-measure experiments. *Science Advances*, 6(16), April 2020.

[24] Umesh Vazirani and Thomas Vidick. Fully device-independent quantum key distribution. *Physical Review Letters*, 113(14), September 2014.

[25] John Watrous. *The Theory of Quantum Information*. Cambridge University Press, April 2018.

[26] Horace P. Yuen, Robert S. Kennedy, and Melvin Lax. Optimum testing of multiple hypotheses in quantum detection theory. *IEEE Transactions on Information Theory*, 21(2):125–134, March 1975.