arXiv:2204.07155v2 [quant-ph] 20 Oct 2022

Tight Bounds for Quantum State Certification with Incoherent Measurements

Sitan Chen* UC Berkeley
Brice Huang† MIT
Jerry Li‡ Microsoft Research
Allen Liu§ MIT

October 21, 2022

Abstract

We consider the problem of quantum state certification, where we are given the description of a mixed state $\sigma \in \mathbb{C}^{d\times d}$, n copies of a mixed state $\rho \in \mathbb{C}^{d\times d}$, and $\varepsilon > 0$, and we are asked to determine whether $\rho = \sigma$ or whether $\|\rho - \sigma\|_1 > \varepsilon$. When σ is the maximally mixed state $\frac{1}{d}I_d$, this is known as mixedness testing. We focus on algorithms which use incoherent measurements, i.e. which only measure one copy of ρ at a time. Unlike those that use entangled, multi-copy measurements, these can be implemented without persistent quantum memory and thus represent a large class of protocols that can be run on current or near-term devices.

For mixedness testing, there is a folklore algorithm which uses incoherent measurements and only needs $O(d^{3/2}/\varepsilon^2)$ copies. The algorithm is non-adaptive, that is, its measurements are fixed ahead of time, and is known to be optimal for non-adaptive algorithms. However, when the algorithm can make arbitrary incoherent measurements, the best known lower bound is only $\Omega(d^{4/3}/\varepsilon^2)$ [BCL20], and it has been an outstanding open problem to close this polynomial gap. In this work:

- We settle the copy complexity of mixedness testing with incoherent measurements and show that $\Omega(d^{3/2}/\varepsilon^2)$ copies are necessary. This fully resolves open questions of [Wri16] and [BCL20].
- We show the instance-optimal bounds for state certification to general σ first derived in [CLO21] for non-adaptive measurements also hold for arbitrary incoherent measurements.

Qualitatively, our results say that adaptivity does not help at all for these problems. Our results are based on new techniques that allow us to reduce the problem to understanding the concentration of certain matrix martingales, which we believe may be of independent interest.

*Email: sitanc@berkeley.edu. This work was supported in part by NSF Award 2103300.

†Email: bmhuang@mit.edu. Supported by an NSF graduate research fellowship, a Siebel scholarship, NSF awards DMS-2022448 and CCF-1940205, and NSF TRIPODS award 1740751.

‡Email: jerrl@microsoft.com

§Email: cliu568@mit.edu. This work was supported in part by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Fellowship

1 Introduction

Quantum mixedness testing, and more generally quantum state certification, are two of the most basic and fundamental tasks in quantum property testing. In quantum state certification, the learner is given n copies of a mixed state $\rho \in \mathbb{C}^{d\times d}$, and an explicit description of a mixed state $\sigma \in \mathbb{C}^{d\times d}$, and the objective is to distinguish with probability at least 0.99 between the case where $\rho = \sigma$ or if it is ε-far from σ in trace distance.[1] Mixedness testing is the special case of state certification where $\sigma = \frac{1}{d}I_d$, i.e., when the target state is the maximally mixed state.

Mixedness testing and state certification are the natural quantum analogues of uniformity testing and identity testing, respectively, two of the most well-studied problems in distribution testing. From a more practical point of view, state certification is also a key subroutine which allows experimentalists to verify the outcomes of their quantum experiments. For instance, if an algorithmist wishes to check that a quantum algorithm with quantum output is correctly outputting the right state, then this is exactly the problem of state certification.

Despite the fundamental nature of the problems, it was not until relatively recently that the copy complexity of state certification and mixedness testing were first understood. The seminal paper of [OW15] first demonstrated that $n = \Theta(d/\varepsilon^2)$ copies were necessary and sufficient to solve mixedness testing. Follow-up work of [BOW19] later demonstrated that $n = O(d/\varepsilon^2)$ is also sufficient for the more general problem of state certification. Combined with the lower bound for mixedness testing, this resolved the copy complexity of state certification, in the worst case over σ.

However, a major downside of the estimators which achieve these copy complexities is that they require heavily entangled measurements over the joint state $\rho^{\otimes n}$. This poses a number of challenges to porting these algorithms into practical settings. First, the descriptions of the measurements are quite large (as the overall joint state is of size $d^n \times d^n$), and cannot be implemented on current (or near-term) quantum devices. Second, the measurements require that all n copies of ρ are simultaneously present. In many realistic settings, such as streaming settings where one copy of ρ is given to the algorithm at a time, this would require that the quantum device be able to store all of these copies in persistent quantum memory. Such a task is also out of reach for current or near-term quantum devices, in essentially any non-trivial regime of the parameters, especially when one considers that d is exponential in the number of qubits in the system!

An appealing class of algorithms which avoids both these issues, and which can be implemented on real world noisy intermediate-scale quantum (NISQ) devices, are algorithms which only rely on *incoherent (a.k.a. unentangled) measurements*. In contrast to general protocols which perform arbitrary measurements on the joint state over all n copies, these algorithms only apply measurements to one copy of ρ at a time, although these measurements can possibly be adaptively chosen based on the (classical) outcomes of the previous measurements. Consequently, these measurements are performed on much smaller states, and moreover, can be performed without any quantum memory.

For these reasons, there has been a considerable amount of attention in recent years devoted to understanding the statistical power of algorithms that only use incoherent measurements, which was also posed as an open problem in Wright's thesis [Wri16]. A recent work of [BCL20] demonstrated that if the measurements are additionally chosen non-adaptively, then $n = \Theta(d^{3/2}/\varepsilon^2)$ copies are necessary and sufficient to solve mixedness testing. They also demonstrated that *any* algorithm using incoherent measurements—even those chosen adaptively—must use at least $n = \Omega(d^{4/3}/\varepsilon^2)$ copies. In other words, there is a polynomial separation between the power of algorithms with and without quantum memory for this problem. Still, this left a gap between the best known upper and lower bounds for mixedness testing with incoherent measurements. This begs the question:

[1] Note that by standard bootstrapping arguments the choice of constant here is arbitrary, and can be any constant larger than 1/2. This only changes the sample complexity by constant factors.

Can we fully characterize the copy complexity of mixedness testing with incoherent measurements?

Closing this gap was posed as an open question in the work of [BCL20].

Underlying this question is another, more qualitative one, regarding the power of adaptivity. Indeed, a recurring theme in a number of different quantum learning settings is that while proving tight lower bounds against adaptive algorithms is quite challenging, the state-of-the-art algorithms almost always tend to be the "obvious" non-adaptive strategies. A very interesting meta-question is understanding for which natural quantum learning problems (if any) adaptivity helps at all for algorithms that use incoherent measurements.

Our first main contribution is to fully resolve this question for mixedness testing: we prove that adaptivity does not improve the sample complexity at all, except possibly up to constant factors.

Theorem 1.1 (Informal, see Theorem 6.1)**.** *The copy complexity of mixedness testing using incoherent measurements is* $n = \Theta(d^{3/2}/\varepsilon^2)$.

By completely pinning down the copy complexity of mixedness testing with incoherent measurements, this answers open questions of [Wri16] and [BCL20]. Qualitatively, our theorem states that adaptivity does not help the copy complexity of this problem whatsoever.

Instance-optimal lower bounds for state certification. We next turn to state certification. Because mixedness testing is a special case of state certification, Theorem 1.1 immediately implies that $n = \Omega(d^{3/2}/\varepsilon^2)$ copies are necessary for state certification, in the worst case over all choices of the reference state σ. This, coupled with a matching upper bound from [CLO21, Lemma 6.2], resolves the copy complexity of state certification with incoherent measurements for worst-case σ.

However, it should be clear that this bound is not the correct bound for all possible σ. For instance, when σ is pure, it is not hard to see that $\Theta(1/\varepsilon^2)$ copies are sufficient and necessary. This raises the natural question: what is the copy complexity of state certification with incoherent measurements, as a function of the reference state σ? This is the quantum analogue of the (classical) distribution testing problem of obtaining instance optimal bounds for identity testing against a known distribution over d elements [ADJ$^+$11, ADJ$^+$12, VV17, DK16, BCG19, JHW18]. In the classical version of the problem, there is a known distribution p over $\{1, \dots, d\}$, and we are given samples from a distribution q. We are asked to distinguish between the case wher $p = q$, and the case when $\|p - q\|_1 > \varepsilon$. A landmark result of [VV17] states that the sample complexity of this question is (essentially) characterized by the $\ell_{2/3}$-quasinorm of p.

In this work, we ask whether or not a similar characterization can be obtained for the quantum version of the question. Prior work of [CLO21] demonstrated such a characterization, but under the caveat that the measurements are chosen non-adaptively. At a high level, they showed that the copy complexity of the problem is governed by the *fidelity* between σ and the maximally mixed state. More precisely, they showed that if $\overline{\sigma}$ and $\underline{\sigma}$ are states given by zeroing out eigenvalues of σ that have total mass at most $\Theta(\varepsilon^2)$ and $\Theta(\varepsilon)$ respectively and normalizing, then the copy complexity with non-adaptive measurements, denoted n, satisfies

$$\widetilde{\Omega}\left(\frac{d \cdot \underline{d}_{\mathsf{eff}}^{1/2}}{\varepsilon^2} \cdot F(\underline{\sigma}, \frac{1}{d} I_d)\right) \leq n \leq \widetilde{O}\left(\frac{d \cdot \overline{d}_{\mathsf{eff}}^{1/2}}{\varepsilon^2} \cdot F(\overline{\sigma}, \frac{1}{d} I_d)\right), \tag{1}$$

where $\underline{d}_{\mathsf{eff}}$ (resp. $\overline{d}_{\mathsf{eff}}$) is the "effective dimension" of the problem, namely, the rank of $\underline{\sigma}$ (resp. $\overline{\sigma}$). In the same work, they also gave lower bounds for arbitrary (possibly adaptive) incoherent measurements, but, like with mixedness testing, these lower bounds were looser and did not match the corresponding upper bound. In light of this, we ask:

Can we give an instance-optimal characterization of the copy complexity of state certification with incoherent measurements?

Our second main contribution is to give such a characterization:

Theorem 1.2 (Informal, see Theorem 8.1)**.** *For any σ, and ε sufficiently small, the copy complexity of state certification w.r.t. σ using incoherent measurements is upper and lower bounded by (1).*

We regard this as strong evidence that, as with mixedness testing, adaptivity does not help for state certification. It is not always a tight bound, as there are states for which the upper and lower bounds in (1) can differ by polynomial factors for some choices of ε, and so this bound can be loose, even in the non-adaptive setting. Still, we conjecture that for all σ, the copy complexity of state certification to σ with incoherent and non-adaptive measurements is the same as that with arbitrary incoherent measurements. Indeed, when ε is sufficiently small compared to the smallest nonzero eigenvalue of σ, our bounds are tight up to logarithmic factors.

Our techniques. We achieve our new lower bounds via a new proof technique which we believe may be of independent interest. As with other lower bounds in this area, we reduce to a "one-versus-many" distinguishing problem. To construct this instance, prior work leveraged the natural quantum analogue of Paninski's famous construction in the lower bound for (classical) uniformity testing [Pan08] – namely, an additive perturbation by a multiple of $UZU^\dagger$, where U is a Haar random matrix and $Z = \mathrm{diag}(1, \ldots, -1, \ldots)$ has equally many $+1$s and -1s.

We instead use a different hard instance based on Gaussian perturbations. While this introduces a number of additional technical challenges, the key advantage of this instance is that the likelihood ratio for this instance has a very clean, self-similar form (see (4)). This allows us to essentially reduce the problem into one of understanding the concentration of a certain matrix martingale defined by the learning process, as well as an auxiliary matrix balancing question. We can then use classical tools from scalar and matrix concentration to demonstrate that the likelihood ratio is close to 1 with high probability over all possible outcomes of the learning algorithm, which yields our desired lower bound. We defer a more detailed explanation of our techniques to Section 3.

Not only does this framework dramatically simplify many of the difficult concentration calculations in prior work such as [BCL20], it also has the conceptual advantage that it never requires a *pointwise* bound on the likelihood ratio. To our knowledge, all prior lower bounds against adaptive algorithms in this literature required some worst-case pointwise bound on the likelihood ratio. For some problems, e.g. shadow tomography [CCHL22], this was already sufficient to prove tight lower bounds. However, for mixedness testing, a worst-case bound cannot be sufficient (as we explain in Section 3), and from a technical perspective, the fact that [BCL20] had to balance between their (much tighter) average case bound on the likelihood ratio and this (fairly large) worst-case bound to control the contribution of certain tail events was why their overall lower bound was loose. Consequently, we believe that this martingale-based technique may also yield tight lower bounds for a number of other problems in the literature.

2 Preliminaries

Throughout, let ρ denote the unknown state, and let $\rho_{\mathsf{mm}} = \frac{1}{d} I_d$ denote the maximally mixed state.

Measurements. We now define the standard measurement formalism, which is the way algorithms are allowed to interact with the unknown quantum state ρ.

Definition 2.1 (Positive operator valued measurement (POVM), see e.g. [NC02])**.** *A positive operator valued measurement $\mathcal{M}$ is a finite collection of psd matrices $\mathcal{M} = \{M_z\}_{z\in\mathcal{Z}}$ satisfying $\sum_z M_z = I_d$. When a state ρ is measured using $\mathcal{M}$, we get a draw from a classical distribution over $\mathcal{Z}$, where we observe z with probability $\mathrm{Tr}(\rho M_z)$. Afterwards, the quantum state is destroyed.*

Incoherent Measurements. Next, we formally define what we mean by an algorithm that uses incoherent measurements. Intuitively, such an algorithm operates as follows: given n copies of ρ, it iteratively measures the i-th copy using a POVM (which could depend on the results of previous measurements), records the outcome, and then repeats this process on the $(i+1)$-th copy. After having performed all n measurements, it must output a decision based on the (classical) sequence of outcomes it has received. More formally, such an algorithm can be represented as a tree:

Definition 2.2 (Tree representation, see e.g. [CCHL22])**.** *Fix an unknown d-dimensional mixed state ρ. A learning algorithm that only uses n incoherent, possibly adaptive, measurements of ρ can be expressed as a rooted tree $\mathcal{T}$ of depth n satisfying the following properties:*

- *Each node is labeled by a string of vectors $\boldsymbol{x} = (x_1, \ldots, x_t)$, where each x_i corresponds to measurement outcome observed in the i-th step.*
- *Each node $\boldsymbol{x}$ is associated with a probability $p^\rho(\boldsymbol{x})$ corresponding to the probability of observing $\boldsymbol{x}$ over the course of the algorithm. The probability for the root is 1.*
- *At each non-leaf node, we measure ρ using a rank-1 POVM $\{\omega_x d \cdot xx^\dagger\}_x$ to obtain classical outcome $x \in \mathbb{S}^{d-1}$. The children of $\boldsymbol{x}$ consist of all strings $\boldsymbol{x}' = (x_1, \ldots, x_t, x)$ for which x is a possible POVM outcome.*
- *If $\boldsymbol{x}' = (x_1, \ldots, x_t, x)$ is a child of $\boldsymbol{x}$, then*

$$p^\rho(\boldsymbol{x}') = p^\rho(\boldsymbol{x}) \cdot \omega_x d \cdot x^\dagger \rho x.$$

- *Every root-to-leaf path is length-n. Note that $\mathcal{T}$ and ρ induce a distribution over the leaves of $\mathcal{T}$.*

We briefly note that in this definition, we assume that the POVMs are always rank-1. It is a standard fact that this is without loss of generality (see e.g. [CCHL22, Lemma 4.8]).

3 Technical Overview

3.1 Mixedness Testing

We begin by describing the proof of our optimal lower bound for mixedness testing. As is standard in this line of work, we first formulate a hard "point-vs-mixture" distinguishing task. Here, we specify some set of states $\{\rho_\alpha\}_\alpha$, and the goal is to distinguish the case where the state ρ is maximally mixed (the "null hypothesis"), and the case where $\rho = \rho_\alpha$, where α is chosen from some distribution $\mathcal{D}$ (the "alternative hypothesis"). Our goal will be to construct such a task so that (1) $\|\rho - \rho_\alpha\|_1 > \varepsilon$ for all α, and (2) for any algorithm that uses incoherent measurements, if p_0 is the distribution over outcomes of the algorithm when run on n copies of the maximally mixed state, and p_α is the distribution over outcomes of the algorithm when run on n copies of ρ_α, then $d_{\mathrm{TV}}(p_0, \mathbb{E}_{\alpha\sim\mathcal{D}}[p_\alpha]) = o(1)$ as long as $n = o(d^{3/2}/\varepsilon^2)$. These two facts together immediately imply our desired lower bound.

Gaussian perturbations. Our first departure from prior work is in the choice of the ensemble of perturbations. All known lower bounds for mixedness testing [BCL20, CCHL21, CLO21, OW15], consider alternate hypotheses of the form $\frac{1}{d}(I_d + \varepsilon UZU^\dagger)$, where $U \in \mathbb{R}^{d\times d}$ is a Haar-random unitary matrix and $Z = \mathrm{diag}(1, \ldots, -1, \ldots)$ has $\frac{d}{2}$ $+1$s and -1s. A drawback of working with these perturbations is that the typical ways of analyzing such distinguishing tasks involve controlling higher-order moments, but the tricky representation-theoretic structure of moments of Haar unitary matrices makes them difficult to work with.

To circumvent this, we work with a Gaussian approximation to the standard Haar-random ensemble: in place of $\frac{1}{d}(I_d + \varepsilon UZU^\dagger)$, we consider the random state $\frac{1}{d}(I_d + \varepsilon M)$, where M is drawn from the *Gaussian orthogonal ensemble* (GOE), suitably shifted to have trace zero (see Definition 5.1). This new alternative hypothesis exhibits comparable tail behavior and fluctuations of the same magnitude as the original, but its moments are much more tractable to analyze and, as we will see, exhibit useful self-similar structure that will be vital to our argument.

Note that strictly speaking, as the distribution over M is supported over all symmetric matrices, with some low probability $\frac{1}{d}(I_d + \varepsilon M)$ may not even be psd, or it may have trace distance $\ll \varepsilon$ from the maximally mixed state. We thus technically need to work with a distribution over M where we condition out these bad events, but it turns out that the impact of this conditioning on our calculations is negligible (see Lemma 6.2 in the proof of Theorem 6.3), and in this overview we will work without conditioning, for simplicity.

Primer on adaptive lower bounds. Having specified the distinguishing task, we now briefly review the usual framework for proving lower bounds against adaptively chosen incoherent measurements. Recall from Definition 2.2 that any learning strategy that uses such measurements can be thought of as specifying a tree, where each internal node corresponds to the transcript of measurement outcomes seen so far, and the edges emanating from that node correspond to the possible outcomes of the POVM that gets chosen to measure the next copy of ρ. At any leaf node, the learner decides based on all the outcomes they have seen along their root-to-leaf path whether the node is maximally mixed or not. As the probabilities for transitioning from any given node to one of its children depend on the unknown state being measured, we can thus think of the null hypothesis and alternative hypothesis as inducing two different distributions p_0 and p_1 over the leaves of the tree. As described above, to show our lower bound for mixedness testing, it suffices to show that for $n = o(d^{3/2}/\varepsilon^2)$, the total variation distance between these distributions satisfies $d_{\mathrm{TV}}(p_0, p_1) = o(1)$.

The main challenge in controlling $d_{\mathrm{TV}}(p_0, p_1)$, and also the key difference from classical distribution testing, is the adaptivity in the measurements. Whereas [BCL20] dealt with this by passing to KL divergence and using chain rule, we will instead work directly with the total variation distance.

Likelihood ratio martingale. In this overview, we will assume for simplicity that every POVM used by the learner consists of rank-1 projectors $yy^\dagger$ to some (adaptively chosen) orthonormal basis.

To bound the total variation distance, we focus on controlling the *likelihood ratio* $L(\boldsymbol{x})$, i.e. the ratio between the probability masses that p_1 and p_0 place on a given leaf $\boldsymbol{x}$. As $d_{\mathrm{TV}}(p_0, p_1) = \mathbb{E}[|L(\boldsymbol{x}) - 1|]$, where the expectation is over $\boldsymbol{x} \sim p_0$, it is enough to show that $L(\boldsymbol{x}) \approx 1$ with high probability over p_0. Henceforth we will thus think of $L(\boldsymbol{x})$ as a random variable where $\boldsymbol{x} \sim p_0$.

Note that for any leaf $\boldsymbol{x} = (x_1, \ldots, x_n)$ specifying a transcript of measurement outcomes corresponding to rank-1 POVM elements $x_1x_1^\dagger, \ldots, x_nx_n^\dagger$, the likelihood ratio between reaching $\boldsymbol{x}$ under the alternative hypothesis versus under the null hypothesis can be expressed as

$$L(\boldsymbol{x}) \triangleq \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \mathop{\mathbb{E}}_{M}\left[\prod_{i=1}^{n}(1 + \varepsilon\, x_i^\dagger M x_i)\right]. \tag{2}$$

We can also extend this to non-leaf nodes $\boldsymbol{x}$: if $\boldsymbol{x} = (x_1, \dots, x_t)$ is a partial transcript for some $t < n$, then $L(\boldsymbol{x}) = \mathbb{E}_M[\prod_{i=1}^t (1 + \varepsilon\, x_i^\dagger M x_i)]$ is simply the ratio between the probability of reaching $\boldsymbol{x}$ after t measurements under the alternative hypothesis versus under the null hypothesis.

Roughly speaking, our strategy will be to track the evolution of the likelihood ratio as t increases. Note that for a fixed node $\boldsymbol{x}$, if $\boldsymbol{x}'$ is the random child node that one transitions to upon measuring another copy of the maximally mixed state, then $\mathbb{E}[L(\boldsymbol{x}')/L(\boldsymbol{x})] = 1$. In other words, the likelihood ratio evolves like a *multiplicative martingale* indexed by t. While this is a basic feature of any likelihood ratio between two sequences of random variables, we are not aware of prior work in quantum learning that exploits this, whereas for us this will be essential to dealing with adaptivity.

We pause to remark that while there have been a number of previous works establishing quantum testing lower bounds by bounding the likelihood ratio [CCHL22, CCHL21, HKP21], in their settings they simply show that the likelihood ratio is bounded for *every leaf*. In contrast, in mixedness testing, such a strategy cannot work, as there can be leaves which are much rarer under the alternative hypothesis than the null hypothesis. For instance, if the algorithm always measures in the standard basis, then a transcript which consists of an equal number of every measurement outcome will be much rarer under the alternative hypothesis than the null.

Recursive structure of L. We now explain how our choice of Gaussian ensemble makes controlling the likelihood ratio martingale particularly convenient. By Isserlis' theorem, one can evaluate (2) explicitly: for (leaf or internal node) $\boldsymbol{x}'$ given by a transcript $x_1, \dots, x_t, x_{t+1}$, we get

$$L(\boldsymbol{x}') = \sum_{k=0}^{\lfloor (t+1)/2 \rfloor} \left(\frac{2\varepsilon^2}{d^2}\right)^k \sum_{\{\{a_i, b_i\}\}} \prod_{i=1}^k (d\langle x_{a_i}, x_{b_i}\rangle^2 - 1), \tag{3}$$

where the latter sum is over all partial matchings of $\{1, \dots, t+1\}$ consisting of k pairs. Now observe that the expression (3) contains a copy of the likelihood ratio for the *parent* of $\boldsymbol{x}'$. If $\boldsymbol{x}$ is the parent corresponding to transcript $x_1, \dots, x_t$, then $L(\boldsymbol{x})$ is precisely the sum of the terms in (3) given by partial matchings which only consist of x_s for $1 \le s \le t$. Moreover, the remaining terms given by partial matchings that contain x_{t+1} also contain likelihood ratio-like terms. Specifically, defining $L(\boldsymbol{x}_{\sim i}) \triangleq \mathbb{E}_M\left[\prod_{j \in [t]: j \neq i} (1 + \varepsilon\, x_j^\dagger M x_j)\right]$,[2] one can verify (Lemma 6.5) that

$$L(\boldsymbol{x}') = L(\boldsymbol{x}) + \frac{2\varepsilon^2}{d^2} \sum_{i=1}^t (d\langle x_i, x_{t+1}\rangle^2 - 1) \cdot L(\boldsymbol{x}_{\sim i}). \tag{4}$$

Now consider the following thought experiment. Imagine for the moment that $L(\boldsymbol{x}_{\sim i}) \approx L(\boldsymbol{x})$ for all $i \in [t]$. Then we could divide by $L(\boldsymbol{x})$ on both sides of (4) to get that

$$\frac{L(\boldsymbol{x}')}{L(\boldsymbol{x})} - 1 \approx \frac{2\varepsilon^2}{d^2} \sum_{i=1}^t (d\langle x_i, x_{t+1}\rangle^2 - 1) = \frac{2\varepsilon^2}{d^2}\, x_{t+1}^\dagger \left(\sum_{i=1}^t (d x_i x_i^\dagger - I_d)\right) x_{t+1}. \tag{5}$$

As $dx_1x_1^\dagger - I_d, dx_2x_2^\dagger - I_d, \cdots$ is a matrix martingale difference sequence, by matrix Freedman [Tro11, Theorem 1.2] we expect the right hand side of (5) to have fluctuations of order roughly $\pm(\varepsilon^2/d^2) \cdot \sqrt{td} = \pm(\varepsilon^2/d^{3/2}) \cdot \sqrt{t}$ (ignoring logarithmic factors). In other words, the likelihood ratio martingale jumps by a multiplicative factor of $1 \pm (\varepsilon^2/d^{3/2}) \cdot \sqrt{t}$ in every step, which means that cumulatively over n steps, it changes by a multiplicative factor of $1 \pm (\varepsilon^2/d^{3/2})n$ with high probability. So if $n = o(d^{3/2}/\varepsilon^2)$, the likelihood ratio is $1 + o(1)$ with high probability over the leaves as desired, and we get the optimal lower bound for mixedness testing.

[2] Note that strictly speaking the transcript $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_t$ does not appear in the tree (unless $i = t$), but this quantity is still well-defined even if it is not a "real" likelihood ratio.

Bootstrapping. This thought experiment is of course inherently circular. Our goal was to show that the likelihood ratio doesn't change very much, but to prove this we assumed that $L(\boldsymbol{x}_{\sim i}) \approx L(\boldsymbol{x})$, i.e. that removing one element from the transcript doesn't change the likelihood ratio very much! Here we outline our approach for resolving this chicken-and-egg problem. The high-level idea is that for $n \le O(d^{3/2}/\varepsilon^2)$, it is actually easy to show that the likelihood ratio can never change by more than a $1 + o(1)$ factor in a single step (see e.g. (23)). For the likelihood ratio martingale argument to work, we need a more refined bound on these multiplicative jumps on the order of $1 \pm (\varepsilon^2/d^{3/2}) \cdot \sqrt{n}$, which we will achieve by recursively bootstrapping the cruder bound– see the proof of Lemma 6.7, which we now sketch.

First, note that the correct version of (5), without approximation, is actually given by

$$\frac{L(\boldsymbol{x}')}{L(\boldsymbol{x})} - 1 = \frac{2\varepsilon^2}{d^2} x_{t+1}^\dagger \left(\sum_{i=1}^t (dx_i x_i^\dagger - I_d) \right) x_{t+1} + x_{t+1}^\dagger \underbrace{\left(\frac{2\varepsilon^2}{d^2} \sum_{i=1}^t (dx_i x_i^\dagger - I_d) \cdot \left(\frac{L(\boldsymbol{x}_{\sim i})}{L(\boldsymbol{x})} - 1 \right) \right)}_{\Delta} x_{t+1}. \quad (6)$$

So the quantity dictating how much the thought experiment deviates from reality is the operator norm of the matrix Δ in (6). Suppose inductively that we have shown that each of the multiplicative jumps $\frac{L(\boldsymbol{x})}{L(\boldsymbol{x}_{\sim i})}$ is bounded by $1 \pm \alpha$ for some $0 < \alpha \ll 1$. Then we can upper bound Δ by

$$\|\Delta\|_{\mathrm{op}} \le O\left(\frac{\varepsilon^2}{d} \cdot \alpha \right) \cdot \sup_{b_1, \dots, b_t \in [-1,1]} \left\| \sum_{i=1}^t b_i (dx_i x_i^\dagger - I_d) \right\|_{\mathrm{op}}. \quad (7)$$

If $\sum_{i=1}^t (dx_i x_i^\dagger - I_d)$ is close to its typical value of $\sqrt{td}$ and $t = \Theta(n)$, then it is not hard to show using a few applications of triangle inequality that the supremum above is upper bounded by $O(t)$ (see Lemma 6.12). In this case, $\|\Delta\| \le O\left(\frac{\varepsilon^2 t}{d^2} \cdot \alpha \right)$, whereas recall that the other term in (6) is of order $(\varepsilon^2/d^{3/2}) \cdot \sqrt{t}$.

The upshot is that we have bootstrapped a bound of $1 \pm \alpha$ on the multiplicative jumps into a better bound on the next multiplicative jump $L(\boldsymbol{x}')/L(\boldsymbol{x})$ which is of order

$$1 \pm \left(\frac{\varepsilon^2}{d^{3/2}} \cdot \sqrt{t} + \frac{\varepsilon^2 t}{d^2} \cdot \alpha \right).$$

In particular, because $t \le n \ll d^2/\varepsilon^2$, our bound has contracted towards the ideal bound of $(\varepsilon^2/d^{3/2}) \cdot \sqrt{t}$ from the thought experiment! Repeating this bootstrapping $O(\log n)$ many rounds and noting that the matrices $\sum_{s \in S} (dx_{i_s} x_{i_s}^\dagger - I_d)$, for $S \subseteq [t], |S| \ge t - O(\log n)$, that arise in recursive applications of the argument above will not be that different from $\sum_{i=1}^t (dx_i x_i^\dagger - I_d)$, we ensure that Δ's contribution to (6) becomes negligible, thus resolving the chicken-and-egg problem.

Log factors. As described, the above would appear to only achieve the optimal bound of $d^{3/2}/\varepsilon^2$ up to log factors. For one, we are using matrix martingale concentration to bound $\sum_{i=1}^t (dx_i x_i^\dagger - I_d)$ and its operator norm thus has fluctuations of order $\sqrt{td \log d}$ rather than $\sqrt{td}$. We also appear to be conditioning on concentration holding for all $t \in [n]$, thus losing another log factor.

To avoid this, instead of bounding the multiplicative jumps pointwise using operator norm, we directly bound the *second moment* of the multiplicative jumps using *expected Frobenius norm*. More precisely, we show that it suffices to control the expected maximum of $\| \sum_{i=1}^t (dx_i x_i^\dagger - I_d) \|_F^2$ across $1 \le t \le n$ (see Lemma 6.8). This can then be bounded without additional log factors using an argument reminiscent of the proof of Doob's L^2 maximal inequality (see Section 6.4).

3.2 State certification

Here we describe how to extend these techniques to the more general setting of state certification with respect to an arbitrary state σ. Without loss of generality we will assume σ is diagonal.

Eigenvalue bucketing. We first describe the hard distinguishing task that we consider. [CLO21] gave a reduction, up to log factors, from showing instance-optimal lower bounds for state certification with respect to arbitrary σ, to showing such bounds when σ takes one of two forms:

(A) σ has eigenvalues that are all within a small multiplicative factor of $1/d$

(B) There are two values $0 \leq a, b \leq 1$ such that each of σ's eigenvalues is within a small multiplicative factor of either a or b.

For completeness, we give a self-contained proof of this reduction in Section 8. At a high level, the idea is that we divide the eigenvalues of σ into logarithmically many buckets where in each bucket, any two eigenvalues are multiplicatively close. Then, the hardest possible distinguishing task one can formulate, up to log factors, is to take the alternative hypothesis to either perturb the submatrix of σ corresponding to a single bucket (this submatrix corresponds to category A above), or to perturb the off-diagonal submatrices of σ corresponding to a pair of buckets (the submatrix of entries from these two buckets corresponds to category B above). The former distinguishing task is sufficient to show optimal lower bounds for states σ like the maximally mixed state, whereas the latter may be harder e.g. for certain approximately low-rank σ.

For σ in category A, the lower bound follows by a simple modification of our analysis for mixedness testing. This proof is presented in Section 6, and includes the proof of the mixedness testing lower bound as a special case. The remaining technical challenge is to prove the lower bound for category B, which we now sketch. This proof is carried out in Section 7.

Off-diagonal perturbations. For simplicity, consider σ of the form $(a \cdot I_{d_1}) \oplus (b \cdot I_{d_2})$ for $a, b > 0$ and $d_1 \geq d_2$. Concretely, the distinguishing task considered in [CLO21] is the following. The null hypothesis is that $\rho = \sigma$, and the alternative hypothesis is that

$$\rho = \begin{pmatrix} a \cdot I_{d_1} & \frac{\varepsilon}{d_2} W \\ \frac{\varepsilon}{d_2} W^\dagger & b \cdot I_{d_2} \end{pmatrix}, \tag{8}$$

where W consists of the first d_2 columns of a Haar-random $d_1 \times d_1$ unitary. Motivated by the Gaussian perturbations used in our proof for mixedness testing, here we consider a Gaussian version of this alternative hypothesis where we instead take W to be a $d_1 \times d_2$ matrix whose entries are independent mean-zero Gaussians with variance $1/d_1$ (see Definition 5.2).

Likelihood ratio pitfalls. To prove this, our goal as before is to show that the likelihood ratio between the distributions p_1, p_0 over leaves of the learning tree induced by the alternative and null hypotheses is close to 1 with high probability with respect to p_0. Here it will be convenient to refer to a transcript $\boldsymbol{x} = (x_1, \ldots, x_t)$ as $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \ldots, (z_t, w_t))$, where $z_i \in \mathbb{C}^{d_1}, w_i \in \mathbb{C}^{d_2}$. We can explicitly compute the likelihood ratio to be

$$L((\boldsymbol{z}, \boldsymbol{w})) = \mathop{\mathbb{E}}_{W} \left[\prod_{i=1}^{n} \left(1 + \frac{2\varepsilon}{d_2} \cdot \frac{z_i^\dagger W w_i}{a\|z_i\|^2 + b\|w_i\|^2} \right) \right],$$

and analogously to (4), we can prove (see Lemma 7.7) that this likelihood ratio has the following nice recursive form. For (leaf or internal node) $(\boldsymbol{z}', \boldsymbol{w}')$ corresponding to the transcript $((z_1, w_1), \ldots,$

(z_{t+1}, w_{t+1})), if $(\boldsymbol{z}, \boldsymbol{w})$ is its parent corresponding to transcript $((z_1, w_1), \ldots, (z_t, w_t))$, then

$$L((\boldsymbol{z}', \boldsymbol{w}')) = L((\boldsymbol{z}, \boldsymbol{w})) + \frac{4\varepsilon^2}{d_1 d_2^2} \sum_{i=1}^{t} \left[\frac{\langle z_i, z_{t+1} \rangle \langle w_i, w_{t+1} \rangle}{(a\|z_i\|^2 + b\|w_i\|^2)(a\|z_{t+1}\|^2 + b\|w_{t+1}\|^2)} \cdot L((\boldsymbol{z}, \boldsymbol{w})_{\sim i}) \right]. \quad (9)$$

The first indication that this distinguishing task could be harder to analyze is the $a\|z\|^2 + b\|w\|^2$ terms that appear in the denominator. For the parameter regimes where we consider this distinguishing task, it turns out that a can be quite small. So any POVM with elements that are aligned with the directions corresponding to the $a \cdot I_{d_1}$ block will lead to measurement outcomes that are rare under the null hypothesis, but not necessarily under the alternative hypothesis.

To see how this issue arises, consider the thought experiment where we imagine $L((\boldsymbol{z}, \boldsymbol{w})_{\sim i}) \approx L((\boldsymbol{z}, \boldsymbol{w}))$ for every i. Then if we divide by $L((\boldsymbol{z}, \boldsymbol{w}))$ on both sides of (9) and define

$$K_t \triangleq \sum_{i=1}^{t} \frac{z_i w_i^\dagger}{a\|z_i\|^2 + b\|w_i\|^2},$$

we get

$$\frac{L((\boldsymbol{z}', \boldsymbol{w}'))}{L((\boldsymbol{z}, \boldsymbol{w}))} - 1 \approx \frac{4\varepsilon^2}{d_1 d^2} \cdot \frac{z_{t+1}^\dagger K_t w_{t+1}}{a\|z_{t+1}\|^2 + b\|w_{t+1}\|^2}. \quad (10)$$

The matrix K_t is the analogue of the $\sum_{i=1}^{t} dx_i x_i^\dagger - I_d$ from mixedness testing. Because

$$\frac{|z_{t+1}^\dagger K_t w_{t+1}|}{a\|z_{t+1}\|^2 + b\|w_{t+1}\|^2} \leq \frac{|z_{t+1}^\dagger K_t w_{t+1}|}{2\sqrt{ab}\|z_{t+1}\|\|w_{t+1}\|} \leq \frac{\|K_t\|_{\mathrm{op}}}{2\sqrt{ab}},$$

we might be tempted to imitate the proof for mixedness testing by bounding $\|K_t\|_{\mathrm{op}}$ using matrix Freedman. Unfortunately this doesn't work: as $a \to 0$, with high probability the operator norm of this matrix is of order at least $\sqrt{t/b}$, so by (10), the multiplicative jumps in the likelihood ratio martingale are of order $1 \pm \frac{\varepsilon^2}{d_1 d_2^2} \cdot \frac{\sqrt{t/b}}{\sqrt{ab}} = 1 \pm \frac{\varepsilon^2}{d_1 d_2^2 \sqrt{ab}} \cdot \sqrt{t}$. So cumulatively over n steps, the likelihood ratio changes by a multiplicative factor of $1 \pm \frac{\varepsilon^2}{d_1 d_2^2 \sqrt{ab}} \cdot n$. This translates to a copy complexity lower bound of $d_1 d_2^2 \sqrt{a} b / \varepsilon^2$. When a and b are both of order $1/d$, this recovers the $d^{3/2}/\varepsilon^2$ lower bound for mixedness testing.[3] But when $a \to 0$, this lower bound becomes vacuous.

From operator to Frobenius. In other words, for this distinguishing task, working with the operator norm is too crude even in the thought experiment! Intuitively the issue is that it yields a *uniform* upper bound on the magnitude of every multiplicative jump, regardless of (z_{t+1}, w_{t+1}). But given that there can be measurement outcomes which are extremely unlikely under the null hypothesis and thus induce rare, huge jumps in the likelihood ratio, it makes more sense to give an upper bound on the magnitude of a *typical* multiplicative jump.

To bound a typical jump, we thus look at the second moment of the jump $\frac{L((\boldsymbol{z}', \boldsymbol{w}'))}{L((\boldsymbol{z}, \boldsymbol{w}))} - 1$ as a random variable in (z_{t+1}, w_{t+1}) under the null hypothesis:

$$\begin{aligned} \mathop{\mathbb{E}}_{(z_{t+1}, w_{t+1})} \left[\left(\frac{z_{t+1}^\dagger K_t w_{t+1}}{a\|z_{t+1}\|^2 + b\|w_{t+1}\|^2} \right)^2 \right] &\leq \sum_{(z_{t+1}, w_{t+1})} \frac{z_{t+1}^\dagger K_t (w_{t+1} w_{t+1}^\dagger) K_t^\dagger z_{t+1}}{b\|w_{t+1}\|^2} \\ &\leq \frac{1}{b} \sum z_{t+1}^\dagger K_t K_t^\dagger z_{t+1} = \frac{1}{b} \|K_t\|_F^2, \end{aligned} \quad (11)$$

[3] The reason we didn't also use this off-diagonal perturbation to prove our mixedness testing lower bound is that this instance is only well-defined for ε sufficiently small; otherwise, the instance (8) is not psd.

where in the second step we used that $w_{t+1}w_{t+1}^\dagger/\|w_{t+1}\|^2 \preceq I_{d_2}$.

It is not hard to show that $\|K_t\|_F^2$ is typically of order td_1d_2 (see Lemma 7.9). So by (11), the typical multiplicative jump in the likelihood ratio martingale is of order $1 \pm \frac{\varepsilon^2}{d_1d_2^2} \cdot \frac{\sqrt{td_1d_2}}{\sqrt{b}} = 1 \pm \frac{\varepsilon^2}{\sqrt{d_1d_2^3b}} \cdot \sqrt{t}$. So cumulatively over n steps, the likelihood ratio changes by a factor of $1 \pm \frac{\varepsilon^2}{\sqrt{d_1d_2^3b}} \cdot n$. This translates to a copy complexity lower bound of $\sqrt{d_1d_2^3b}/\varepsilon^2$. In the parameter regime we care about, $d_2b \geq \Omega(1)$ (see Fact 7.5), so this yields the (optimal) lower bound of $d_2\sqrt{d_1}/\varepsilon^2$.

Bootstrapping. As with our proof for mixedness testing, the above thought experiment is circular. If we no longer pretend that $L((\boldsymbol{z},\boldsymbol{w})_{\sim i}) \approx L((\boldsymbol{z},\boldsymbol{w}))$ for every i, then in place of K_t, the matrix whose Frobenius norm we actually need to bound is

$$H_t \triangleq \sum_{i=1}^t \frac{z_i w_i^\dagger}{a\|z_i\|^2 + b\|w_i\|^2} \cdot \frac{L((\boldsymbol{z},\boldsymbol{w})_{\sim i})}{L((\boldsymbol{z},\boldsymbol{w}))} = K_t + \underbrace{\sum_{i=1}^t \frac{z_i w_i^\dagger}{a\|z_i\|^2 + b\|w_i\|^2} \cdot \left(\frac{L((\boldsymbol{z},\boldsymbol{w})_{\sim i})}{L((\boldsymbol{z},\boldsymbol{w}))} - 1\right)}_{\Delta}, \quad (12)$$

but controlling H_t relies on recursively controlling $\frac{L((\boldsymbol{z},\boldsymbol{w})_{\sim i})}{L((\boldsymbol{z},\boldsymbol{w}))} - 1$. We solve this chicken-and-egg problem by bootstrapping the following crude upper bound. The idea is that for "H_t-like" matrices, the Frobenius norm can always be very loosely upper bounded by $n/\sqrt{ab}$, essentially because the multiplicative jumps in the likelihood ratio are never greater than $O(1)$ (see Lemma 7.14)– we note that the precise polynomial dependence on n in this crude bound is unimportant, as our goal will be to contract this bound by a constant factor in each of $O(\log(n))$ rounds of bootstrapping.

So if we apply the aforementioned *operator norm* bound to control $\frac{L((\boldsymbol{z},\boldsymbol{w})_{\sim i})}{L((\boldsymbol{z},\boldsymbol{w}))} - 1$ and naively upper bound the operator norm of the resulting H_t-like matrix by its Frobenius norm, we get

$$\frac{L((\boldsymbol{z},\boldsymbol{w})_{\sim i})}{L((\boldsymbol{z},\boldsymbol{w}))} - 1 \leq \frac{4\varepsilon^2 n}{d_1d_2^2\sqrt{ab}} \cdot \frac{\|z_i\|\,\|w_i\|}{a\|z_i\|^2 + b\|w_i\|^2}.$$

Substituting this into the right-hand side of (12), we obtain the following analogue of (7):

$$\|\Delta\|_F \leq \frac{4\varepsilon^2 n}{d_1d_2^2\sqrt{ab}} \cdot \sup_{b_1,\ldots,b_t \in [-1,1]} \left\| \sum_{i=1}^t b_i \frac{z_i w_i^\dagger \|z_i\|\|w_i\|}{(a\|z_i\|^2 + b\|w_i\|^2)^2} \right\|_F. \quad (13)$$

As we show in Lemma 7.10, with high probability over $(\boldsymbol{z},\boldsymbol{w})$ this supremum is at most $\frac{1}{10}d_1d_2^2/\varepsilon^2$, so $\|\Delta\|_F \leq n/2\sqrt{ab}$. Before we sketch how to prove this, let us see how to conclude the argument.

Indeed, by plugging the bound on the supremum into (12), we find that we have bootstrapped a crude bound of $n/\sqrt{ab}$ on the Frobenius norm of the "H_t-like" matrices that dictate the preceding multiplicative jumps $\frac{L((\boldsymbol{z},\boldsymbol{w})_{\sim i})}{L((\boldsymbol{z},\boldsymbol{w}))}$ into a better bound on the Frobenius norm of H_t, namely

$$\|H_t\| \leq \|K_t\| + n/2\sqrt{ab}. \quad (14)$$

By repeating this bootstrapping logarithmically many rounds, we can thus shrink the second term in (14) until it is dominated by the contribution from $\|K_t\|$, showing that the above thought experiment is valid.

Supremum bound. Recall that for mixedness testing, we could show that the analogous supremum was bounded as long as $\|\sum_{i=1}^t (dx_ix_i^\dagger - I_d)\|_{\mathrm{op}}$ was (Lemma 6.12). Analogously, one might hope that (13) is bounded as long as $\|K_t\|_F$ is. Unfortunately, this turns out to be false (see Appendix D), essentially because the off-diagonal structure of the distinguishing task makes it possible

for K to be small, in fact zero, even under extremely atypical transcripts (e.g. consider a transcript that repeatedly alternates between a vector (z, w) and the vector $(z, -w)$), whereas the supremum for such transcripts will be extremely large.

This necessitates an entirely different argument for the supremum. The proof involves a careful net argument that is facilitated by a judicious application of Grothendieck's inequality. We defer the details to Section 7.5.

Roadmap. In Section 4 we survey relevant prior work. In Section 5 we provide additional technical preliminaries and formally define the ensembles of perturbations we use. In Section 6, we prove our lower bound for mixedness testing, and in Section 7, we prove our lower bound for the distinguishing task involving "off-diagonal" perturbations that was described in the overview. In Section 8 we state our instance-optimal lower bound for state certification and use the results of Section 6 and 7 to give a simple proof of a slightly weaker version of it. In Appendix A and B we refine our analysis to give a full proof of the instance-optimal bound. In Appendix C we present the deferred proofs that the bad events we condition out when we define our Gaussian perturbations occur with small probability.

4 Related Work

A full literature review on quantum (and classical) testing is out of the scope of this paper. For concision we only discuss some of the more relevant works below.

The questions we consider in this paper fall under the domain of quantum state property testing. See [MdW16] for a more complete survey on property testing of quantum states. In this literature, roughly speaking, there are two settings considered, the *asymptotic regime*, and the *non-asymptotic regime*, the latter of which is the setting we study.

In the former setting, one considers the regime of parameters where $n \to \infty$ and d, ε are held fixed and relatively small, and the goal is to precisely characterize the exponential rate of convergence as a function of n. In this setting, quantum state certification is usually called *quantum state discrimination*, see e.g. [Che00, ANSV08, BC09] and references within. However, since d and ε are fixed, this allows for rates which could depend exponentially on the dimensionality of the problem.

Instead, we consider the "non-asymptotic regime," where the goal is to characterize the statistical rate, as a function of d and ε. Similar work in this regime includes the aforementioned works of [OW15] and [BOW19]. However, as described previously, their algorithms require using fully entangled measurements.

Our work falls into the line of work considering restricted classes of measurements, and specifically, those with without quantum memory. Understanding the power of such algorithms in the context of mixedness testing and, more generally, spectrum testing was posed as an open problem in [Wri16]. Similar questions have also been considered in other settings, such as shadow tomography [Aar18]. However, until recently, lower bounds for algorithms without quantum memory usually only held in the non-adaptive setting, e.g. [HHJ$^+$17, CLO21]. Recent work of [BCL20] demonstrated the first lower bound against general (possibly adaptive) incoherent measurements for such a task. Subsequently, there has been a flurry of work demonstrating similar bounds in a variety of settings [HKP20, ACQ22, HKP21, HBC$^+$21, CCHL22, ALL21, CCHL21, Low21, CZSJ22]. It is an interesting question if our techniques can be extended to also improve any of the lower bounds in these works.

Other restricted models of computation have also been considered in the literature. [Yu19] gives algorithms for various quantum property testing problems using local measurements which

act on each individual qubit, and in an non-adaptive manner. A number of works considers the special case where the measurements are only Pauli matrices [FL11, FGLE12, dSLCP11, AGKE15]. Overall, these classes of measurements seem to be much more restrictive than general non-adaptive measurements. In particular, the copy complexity of tasks such as mixedness testing under these measurements seem to be asymptotically higher than general incoherent measurements.

5 Additional Preliminaries

Notation. Given $z \in \mathbb{R}$, we use z_- to denote $-\min(z, 0)$. We use $\wedge$ and $\vee$ to denote min and max. We use $f \lesssim g$ to denote $f = O(g)$, $f \ll g$ to denote $f = o(g)$, and $f \lll g$ to denote that there exists some absolute constant c for which $f = o(g/\log^c g)$. We will always implicitly assume a sufficiently large system; for example, if $f \gg g$ we will assume where necessary that $f \geq 100g$. We use $f = \widetilde{O}(g)$ (resp. $f = \widetilde{\Omega}(g)$) to denote that there exists some absolute constant c for which $f = O(g \cdot \log^c g)$ (resp. $f = \Omega(g/\log^c g)$).

Given a vector v, we use $\|v\|_p$ to denote its ℓ^p norm; when $p = 2$, we sometimes drop the subscript. Given a matrix M, we use $\|M\|_{\mathrm{op}}$ or $\|M\|$ to denote its operator norm, $\|M\|_1$ to denote its trace norm, and $\|M\|_F$ to denote its Frobenius norm.

For a string $\boldsymbol{x} = (x_1, \ldots, x_n)$, we let $\boldsymbol{x}_{\sim i}$ and $\boldsymbol{x}_{\sim i,j}$ denote the string with the i-th index removed and the string with the i-th and j-th indices removed. For any set $S \subseteq [n]$, we let $\boldsymbol{x}_S$ denote the string restricted to the entries in S.

We will work with the following random matrix ensembles:

Definition 5.1 (Trace-centered Gaussian orthogonal ensemble (GOE))**.** *For $d \in \mathbb{N}$, let $G \sim \mathrm{GOE}(d)$, that is, $G \in \mathbb{R}^{d\times d}$ is symmetric with upper diagonal entries sampled independently from $\mathcal{N}(0, 1/d)$ and diagonal entries sampled independently from $\mathcal{N}(0, 2/d)$.*

Define $M = G - \frac{\mathrm{Tr}(G)}{d} I_d$. We say that M is a trace-centered GOE matrix *and denote its distribution $\mathrm{GOE}^*(d)$. For $U \subseteq \mathbb{R}^{d\times d}$, $\overline{M}$ is a U-*truncated trace-centered GOE matrix *if it is drawn from $\mathrm{GOE}^*(d)$ conditioned on $\overline{M} \in U$. We denote the distribution of $\overline{M}$ by $\mathrm{GOE}^*_U(d)$.*

Definition 5.2 (Truncated Ginibre)**.** *For $d_1, d_2 \in \mathbb{N}$, let $G \sim \mathrm{Gin}(d_1, d_2)$ be the (normalized) $d_1 \times d_2$* Ginibre matrix*, that is, $G \in \mathbb{R}^{d_1\times d_2}$ has i.i.d. entries $\mathcal{N}(0, 1/d_1)$. For $U \subseteq \mathbb{R}^{d_1\times d_2}$, $\overline{G}$ is a U-*truncated $d_1 \times d_2$ Ginibre matrix *if it is drawn from $\mathrm{Gin}(d_1, d_2)$ conditioned on $\overline{G} \in U$. We denote the distribution of $\overline{G}$ by $\mathrm{Gin}_U(d)$.*

Our result for state certification uses the following notion of fidelity.

Definition 5.3 (Fidelity between two quantum states)**.** *The fidelity of quantum states $\rho, \sigma \in \mathbb{C}^{d\times d}$ is $F(\rho, \sigma) = (\mathrm{Tr}\sqrt{\rho^{1/2}\sigma\rho^{1/2}})^2$.*

Our lower bounds are based on Le Cam's two-point method which we briefly review here. The following is an elementary result in binary hypothesis testing:

Fact 5.4 (See e.g. Theorem 4.3 from [Wu17])**.** *Given distributions p_0, p_1 over a domain $\mathcal{S}$, if $d_{\mathrm{TV}}(p_0, p_1) < 1/3$, there is no $\mathcal{A} : \mathcal{S} \to \{0, 1\}$ for which $\Pr_{x\sim p_i}[\mathcal{A}(x) = i] \geq 2/3$ for both $i = 0, 1$.*

Now consider a state distinguishing task of the form

$$H_0 : \rho = \sigma \qquad \text{and} \qquad H_1 : \rho = \sigma_M,$$

where σ_M is a random state sampled from some distribution $\mathcal{D}$ over the set of states satisfying $\|\sigma - \sigma_M\|_1 > \epsilon$. Recall from Definition 2.2 that a learning algorithm that uses n incoherent measurements corresponds to a tree $\mathcal{T}$ of depth n, and $\rho = \sigma$ and $\rho = \sigma_M$ induce distributions p_0

and p_M on the leaves of this tree. We can use Fact 5.4 to reduce proving a copy complexity lower bound for state certification with respect to σ, which is a worst-case guarantee over all possible input states ρ, to bounding $d_{\mathrm{TV}}(p_0, \mathbb{E}_M[p_M])$, which is an average-case bound.

Lemma 5.5 (Le Cam's two-point method, see e.g. Lemma 1 in [Yu97])**.** *If there is a distribution $\mathcal{D}$ over states satisfying $\|\sigma - \sigma_M\|_1 > \epsilon$ for which $d_{\mathrm{TV}}(p_0, \mathbb{E}_M[p_M]) \leq 1/3$ for any tree $\mathcal{T}$ of depth n, then any algorithm $\mathcal{A}$ using incoherent measurements for state certification with respect to σ must make more than n incoherent measurements to achieve success probability at least $2/3$.*

Proof. Suppose to the contrary there existed such an algorithm $\mathcal{A}$ using at most n incoherent measurements, and let p_0 and p_M denote the distributions over the leaves of the tree corresponding to $\mathcal{A}$ when $\rho = \sigma$ and $\rho = \sigma_M$ respectively. Suppose when it succeeds, $\mathcal{A}$ outputs 0 when $\rho = \sigma$ and 1 when $\|\rho - \sigma\|_1 > \epsilon$. Let $p_1 \triangleq \mathbb{E}_{M\sim\mathcal{D}}[p_M]$. Because $\mathcal{A}$ successfully outputs 1 with probability $2/3$ when given as input the state σ_M for any M, $2/3 \leq \mathbb{E}_M[\Pr_{\boldsymbol{x}\sim p_M}[\mathcal{A}(\boldsymbol{x}) = 1]] = \mathbb{E}_{\boldsymbol{x}\sim p_1}[\mathcal{A}(\boldsymbol{x}) = 1]$. Similarly, $2/3 \leq \mathbb{E}_{\boldsymbol{x}\sim p_0}[\mathcal{A}(\boldsymbol{x}) = 0]$. By Fact 5.4, this would contradict the bound on $d_{\mathrm{TV}}(p_0, p_1)$. □

6 Lower Bound for Mixedness Testing

In this section we prove the following theorem, which is the formal version of Theorem 1.1.

Theorem 6.1. *Let $d \gg 1$ and $0 < \varepsilon \leq 1/12$. Any algorithm that uses incoherent measurements which, given n copies of a mixed state $\rho \in \mathbb{C}^{d\times d}$, can distinguish between the case where $\rho = \rho_{\mathsf{mm}}$ and where $\|\rho - \rho_{\mathsf{mm}}\|_1 > \varepsilon$ with probability at least $2/3$, must use at least $n = \Omega(d^{3/2}/\varepsilon^2)$ copies.*

By the upper bound in [BCL20], this is tight up to constant factors. Also note that by standard amplification arguments, the choice of constant in the success probability is arbitrary, and can be taken to be any constant which is strictly larger than $1/2$.

In fact, we will prove a slightly stronger theorem, which will be useful later on for our lower bounds against state certification. Namely, we will show that the same bound holds not just when the null hypothesis is the maximally mixed state, but for any state whose smallest and largest eigenvalues are comparable.

More formally, let $A \in \mathbb{R}^{d\times d}$ be a diagonal matrix with diagonal entries $a_1 \geq \cdots \geq a_d > 0$, satisfying $2a_d \geq a_1$, and $\mathrm{Tr}(A) = d$. We consider the task of distinguishing between the following two alternatives:

$$H_0 : \rho = \frac{1}{d}A \qquad \text{and} \qquad H_1 : \rho = \frac{1}{d}(A + \varepsilon\overline{M}). \tag{15}$$

Here, $\overline{M} \sim \mathrm{GOE}^*_U(d)$ for the U given by Lemma 6.2 below.

Lemma 6.2. *There exists $U \subseteq \mathbb{R}^{d\times d}$ such that if $M \sim \mathrm{GOE}^*(d)$, then $\Pr[M \notin U] \leq \exp(-\Omega(d))$ and on the event $M \in U$, we have $\|M\|_{\mathrm{op}} \leq 3$ and $\|M\|_1 \geq d/12$.*

We defer the proof of this lemma to Appendix C. Our main result for the distinguishing task (15) is the following.

Theorem 6.3. *If $d \gg 1$ and $\varepsilon \leq 1/12$, then any algorithm using incoherent measurements that distinguishes between H_0 and H_1 with success probability at least $2/3$ requires $n = \Omega(d^{3/2}/\varepsilon^2)$ copies.*

Again, by standard amplification arguments, the choice of constant in the success probability is arbitrary, and can be taken to be any constant greater than $1/2$. Note that the bounds in Lemma 6.2 ensure that under H_1, ρ is psd (and thus a valid quantum state) and has trace distance $\Omega(\varepsilon)$ to $\frac{1}{d}A$.

In particular, since any algorithm for mixedness testing must solve this distinguishing problem as well, setting $A = I_d$ into Theorem 6.3 immediately implies Theorem 6.1.

Take any learning tree $\mathcal{T}$ corresponding to an algorithm for this task that uses n incoherent measurements. Recalling the terminology from Definition 2.2, we let p_0 and p_1 denote the distributions over leaves of $\mathcal{T}$ induced by ρ under H_0 and H_1 respectively. In the rest of this section, we assume $n \ll d^{3/2}/\varepsilon^2$ and will prove $d_{\mathrm{TV}}(p_0, p_1) = o(1)$. It is clear that this immediately implies Theorem 6.3.

We let $L^*(\cdot)$ denote the likelihood ratio between p_1 and p_0. That is, for a sequence of vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$, let $L^*(\boldsymbol{x}) \triangleq p_1(\boldsymbol{x})/p_0(\boldsymbol{x})$. Note that

$$L^*(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\overline{M} \sim \mathrm{GOE}^*_U(d)} \left[\prod_{i=1}^n \left(1 + \varepsilon \frac{x_i^\dagger \overline{M} x_i}{x_i^\dagger A x_i} \right) \right].$$

Define similarly

$$L(\boldsymbol{x}) \triangleq \mathop{\mathbb{E}}_{M \sim \mathrm{GOE}^*(d)} \left[\prod_{i=1}^n \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i} \right) \right]. \tag{16}$$

This is an estimate for the likelihood ratio $L^*(\boldsymbol{x})$ where the conditioned Gaussian integral is replaced by a true Gaussian integral. Most of the computations in this section will be done in terms of $L(\boldsymbol{x})$; the proof of Theorem 6.3 below quantifies that $L(\boldsymbol{x})$ is a close approximation of $L^*(\boldsymbol{x})$.

Throughout this section, we will somewhat abuse notation and write $L(\boldsymbol{z})$ for any sequence of unit vectors $\boldsymbol{z} = (z_1, \ldots, z_t)$ of length not necessarily n. This is defined the same way as in (16). We also write $L(\boldsymbol{x}, \boldsymbol{x})$ to denote the value of L on input $(x_1, x_1, x_2, x_2, \ldots, x_n, x_n)$.

The main ingredient in the proof of Theorem 6.3 is the following high-probability bound on L evaluated at the leaves of $\mathcal{T}$.

Proposition 6.4. *There exists a subset S of the leaves of $\mathcal{T}$ such that $\Pr_{p_0}[S] = 1 - o(1)$ and for all $\boldsymbol{x} \in S$, $|L(\boldsymbol{x}) - 1| = o(1)$ and $L(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}}$.*

Let us first prove Theorem 6.3 assuming Proposition 6.4.

Proof of Theorem 6.3. Let U be as in Lemma 6.2. Define

$$\overline{L}(\boldsymbol{x}) = \mathop{\mathbb{E}}_{M \sim \mathrm{GOE}^*(d)} \left[\mathbb{1}\{M \in U\} \prod_{i=1}^n \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i} \right) \right].$$

It is clear that $L^*(\boldsymbol{x}) = \Pr[U]^{-1} \overline{L}(\boldsymbol{x})$. For all $\boldsymbol{x} \in S$, by Cauchy-Schwarz

$$\begin{aligned}
|L(\boldsymbol{x}) - \overline{L}(\boldsymbol{x})| &= \left| \mathop{\mathbb{E}}_{M \sim \mathrm{GOE}^*(d)} \left[\mathbb{1}\{M \notin U\} \prod_{i=1}^n \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i} \right) \right] \right| \\
&\le \Pr[U^c]^{1/2} \mathop{\mathbb{E}}_{M \sim \mathrm{GOE}^*(d)} \left[\prod_{i=1}^n \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i} \right)^2 \right]^{1/2} \\
&= \sqrt{\Pr[U^c] L(\boldsymbol{x}, \boldsymbol{x})} = o(1).
\end{aligned}$$

Here we use that $\Pr[U^c] \le \exp(-\Omega(d))$ and $L(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}}$. Moreover, we have $|L(\boldsymbol{x}) - 1| = o(1)$. Thus, for all $\boldsymbol{x} \in S$, $\overline{L}(\boldsymbol{x}) = 1 + o(1)$ and

$$\begin{aligned}
|L^*(\boldsymbol{x}) - 1| &\le |L^*(\boldsymbol{x}) - \overline{L}(\boldsymbol{x})| + |\overline{L}(\boldsymbol{x}) - 1| \\
&= \frac{\Pr[U^c]}{\Pr[U]} \overline{L}(\boldsymbol{x}) + o(1) = o(1).
\end{aligned}$$

Finally,

$$\begin{aligned} d_{\mathrm{TV}}(p_0, p_1) &= 2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_0}[(L^*(\boldsymbol{x}) - 1)_-] \\ &= 2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_0}[\mathbb{1}\{\boldsymbol{x} \in S\}(L^*(\boldsymbol{x}) - 1)_-] + 2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_0}[\mathbb{1}\{\boldsymbol{x} \notin S\}(L^*(\boldsymbol{x}) - 1)_-] \\ &\le 2 \sup_{\boldsymbol{x}\in S}(L^*(\boldsymbol{x}) - 1)_- + 2 \Pr_{p_0}[S^c] = o(1). \end{aligned}$$

□

6.1 Recursive evaluation of likelihood ratio

Let $\boldsymbol{z} = (z_1, \dots, z_t)$ be a sequence of unit vectors. For $1 \le i \le t$, let $\boldsymbol{z}_{\sim i}$ be the sequence $\boldsymbol{z}$ with z_i omitted. Similarly, for $1 \le i < j \le t$, let $\boldsymbol{z}_{\sim i,j}$ be the sequence $\boldsymbol{z}$ with z_i, z_j omitted. The main result of this subsection is the following recursive formula for $L(\boldsymbol{z})$.

Lemma 6.5. *The function L satisfies*

$$L(\boldsymbol{z}) = L(\boldsymbol{z}_{\sim t}) + \frac{2\varepsilon^2}{d^2} \sum_{i=1}^{t-1} \left[\frac{d\langle z_i, z_t\rangle^2 - 1}{(z_i^\dagger A z_i)(z_t^\dagger A z_t)} L(\boldsymbol{z}_{\sim i,t}) \right].$$

The proof is based on Isserlis' theorem, which we record below. For k even, let $\mathrm{PMat}(k)$ denote the set of perfect matchings of $\{1, \dots, k\}$.

Theorem 6.6 ([Iss18])**.** *Let $g = (g_1, \dots, g_k)$ be a jointly Gaussian vector. If k is odd, then $\mathbb{E}[\prod_{i=1}^k g_i] = 0$. If k is even, then*

$$\mathbb{E}\left[\prod_{i=1}^k g_i\right] = \sum_{\{\{a_1,b_1\},\dots,\{a_{k/2},b_{k/2}\}\}\in \mathrm{PMat}(k)} \prod_{i=1}^{k/2} \mathbb{E}[g_{a_i} g_{b_i}].$$

Proof of Lemma 6.5. For a set $S \subseteq [t]$ with $|S|$ even, let $\mathrm{PMat}(S)$ denote the set of perfect matchings of S. For even $k \le t$, let $\mathrm{Mat}(t, k)$ denote the set of matchings of $[t]$ consisting of $k/2$ pairs. We compute that

$$\begin{aligned} L(\boldsymbol{z}) &= \sum_{S\subseteq[t]} \varepsilon^{|S|} \mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)} \left[\prod_{i\in S} \frac{z_i^\dagger M z_i}{z_i^\dagger A z_i} \right] \qquad \text{(expanding (16))} \\ &= \sum_{\substack{S\subseteq[t] \\ |S| \text{ even}}} \varepsilon^{|S|} \sum_{\{\{a_1,b_1\},\dots,\{a_{|S|/2},b_{|S|/2}\}\}\in \mathrm{PMat}(S)} \prod_{i=1}^{|S|/2} \mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)} \left[\frac{z_{a_i}^\dagger M z_{a_i}}{z_{a_i}^\dagger A z_{a_i}} \cdot \frac{z_{b_i}^\dagger M z_{b_i}}{z_{b_i}^\dagger A z_{b_i}} \right] \qquad \text{(Th. 6.6)} \\ &= \sum_{k=0}^{\lfloor t/2\rfloor} \varepsilon^{2k} \sum_{\{\{a_1,b_1\},\dots,\{a_k,b_k\}\}\in \mathrm{Mat}(t,2k)} \prod_{i=1}^{k} \mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)} \left[\frac{z_{a_i}^\dagger M z_{a_i}}{z_{a_i}^\dagger A z_{a_i}} \cdot \frac{z_{b_i}^\dagger M z_{b_i}}{z_{b_i}^\dagger A z_{b_i}} \right] \\ &= \sum_{k=0}^{\lfloor t/2\rfloor} \left(\frac{2\varepsilon^2}{d^2}\right)^k \sum_{\{\{a_1,b_1\},\dots,\{a_k,b_k\}\}\in \mathrm{Mat}(t,2k)} \prod_{i=1}^{k} \frac{d\langle z_{a_i}, z_{b_i}\rangle^2 - 1}{(z_{a_i}^\dagger A z_{a_i})(z_{b_i}^\dagger A z_{b_i})}. \end{aligned} \tag{17}$$

In the final step we use that for unit vectors $x, y \in \mathbb{C}^d$,

$$\mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)} \Big[(x^\dagger M x)(y^\dagger M y) \Big] = \frac{2}{d^2}(d\langle x, y\rangle^2 - 1),$$

which can be verified by direct computation. The lemma follows by partitioning the summands in (17) based on whether t appears in the matching, and if so which $i \in \{1, \dots, t-1\}$ it is paired with. □

6.2 High probability bound on likelihood ratio at leaves

This subsection gives the main part of the proof of Proposition 6.4. For any sequence of unit vectors $\boldsymbol{z} = (z_1, \dots, z_t)$, define

$$H(\boldsymbol{z}) = \sum_{i=1}^{t} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot \frac{L(\boldsymbol{z}_{\sim i})}{L(\boldsymbol{z})} \qquad \text{and} \qquad K(\boldsymbol{z}) = \sum_{i=1}^{t} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i}.$$

The function H enters our calculations by the following rewriting of Lemma 6.5:

$$\frac{L(\boldsymbol{z})}{L(\boldsymbol{z}_{\sim t})} = 1 + \frac{2\varepsilon^2}{d^2} \cdot \frac{z_t^\dagger H(\boldsymbol{z}_{\sim t}) z_t}{z_t^\dagger A z_t}. \tag{18}$$

If $\boldsymbol{z} = \boldsymbol{x}_{\le t} \triangleq (x_1, \dots, x_t)$ is a prefix of $\boldsymbol{x} \sim p_0$, then $\frac{L(\boldsymbol{z})}{L(\boldsymbol{z}_{\sim t})} = \frac{L(\boldsymbol{x}_{\le t})}{L(\boldsymbol{x}_{\le t-1})}$ is one step in the likelihood ratio martingale. As we will see (proof of Claim 6.10) below, the multiplicative fluctuation of this step is

$$\mathop{\mathbb{E}}_{x_t}\left[\left(\frac{L(\boldsymbol{x}_{\le t})}{L(\boldsymbol{x}_{\le t-1})}\right)^2\right] = 1 + O\left(\frac{\varepsilon^4}{d^5}\right) \|H(\boldsymbol{x}_{\le t-1})\|_F^2 .$$

Thus, an upper bound on $\|H(\boldsymbol{z})\|_F$ over all prefixes $\boldsymbol{z}$ of $\boldsymbol{x}$ controls the fluctuations of the likelihood ratio martingale. Because the matrices output by H are hard to control directly, we will use the function K as a proxy for H. The following lemma quantifies this relationship, showing that if $K(\boldsymbol{z})$ is bounded in Frobenius norm, $H(\boldsymbol{z})$ is bounded at the same scale.

Lemma 6.7. *Suppose $1 \ll \gamma \ll d/(\varepsilon^2 n^{1/2})$. If $\boldsymbol{z} = (z_1, \dots, z_t)$ is a sequence of unit vectors satisfying $t \le n$ and $\|K(\boldsymbol{z})\|_F \le n^{1/2} d\gamma$, then $\|H(\boldsymbol{z})\|_F \le 3n^{1/2} d\gamma$.*

Note that this lemma is a "deterministic" statement about a sequence of vectors. We will prove this in Subsection 6.3 using the bootstrap argument alluded to earlier. The following lemma bounds $K(\boldsymbol{z})$ in Frobenius norm uniformly over all prefixes $\boldsymbol{z}$ of $\boldsymbol{x}$. We will prove this lemma in Subsection 6.4 by mimicking the proof of Doob's L^2 maximal inequality for the matrix valued martingale $K(\boldsymbol{x}_{\le t})$.

Lemma 6.8. *If $\boldsymbol{x} \sim p_0$, then $\mathbb{E}\left[\sup_{1 \le t \le n} \|K(\boldsymbol{x}_{\le t})\|_F^2\right] \lesssim nd^2$.*

We will now prove Proposition 6.4 assuming Lemmas 6.7 and 6.8. We set α, β to be slowly-growing functions such that $1 \ll \alpha \ll \beta \ll d^{3/2}/(\varepsilon^2 n) \wedge d/(\varepsilon^2 n^{1/2})$, and furthermore $\alpha^2 \ll d^{3/2}/(\varepsilon^2 n)$. This is possible because $n \ll d^{3/2}/\varepsilon^2$.

Let $\boldsymbol{x} \sim p_0$. For $1 \le t \le n$, define the filtration $\mathcal{F}_t = \sigma(\boldsymbol{x}_{\le t})$ and the sequences

$$H_t = H(\boldsymbol{x}_{\le t}), \qquad K_t = K(\boldsymbol{x}_{\le t}), \qquad \Phi_t = L(\boldsymbol{x}_{\le t}).$$

Consider the time

$$\tau = \inf\left\{t : \|K_t\|_F > n^{1/2} d\alpha \text{ or } |\Phi_t - 1| > \frac{\varepsilon^2 n}{d^{3/2}}\beta\right\} \cup \{\infty\},$$

which is clearly a stopping time with respect to $\mathcal{F}_t$. Also define the stopped sequence $\Psi_t = \Phi_{t \wedge \tau}$.

Claim 6.9. *With probability $1 - o(1)$, $\|K_t\|_F \le n^{1/2} d\alpha$ for all $1 \le t \le n$.*

Proof. By Lemma 6.8,

$$\Pr\left[\sup_{1\le t\le n}\|K_t\|_F > n^{1/2}d\alpha\right] \le \frac{\mathbb{E}\left[\sup_{1\le t\le n}\|K_t\|_F^2\right]}{nd^2\alpha^2} \lesssim \alpha^{-2} = o(1). \qquad \square$$

Claim 6.10. *With probability* $1-o(1)$, $|\Psi_n - 1| \le \frac{\varepsilon^2 n}{d^{3/2}}\beta$.

Proof. Note that Ψ_t is a multiplicative martingale: if $\tau \le t-1$ then certainly $\mathbb{E}[\frac{\Psi_t}{\Psi_{t-1}}|\mathcal{F}_{t-1}] = 1$, and if $\tau > t-1$, (18) implies

$$\mathbb{E}\left[\frac{\Psi_t}{\Psi_{t-1}}|\mathcal{F}_{t-1}\right] = 1 + \frac{2\varepsilon^2}{d^2}\,\mathbb{E}\left[\frac{x_t^\dagger H_{t-1}x_t}{x_t^\dagger A x_t}|\mathcal{F}_{t-1}\right] = 1,$$

using that

$$\mathbb{E}\left[\frac{x_t^\dagger H_{t-1}x_t}{x_t^\dagger A x_t}|\mathcal{F}_{t-1}\right] = \sum_{x_t}\omega_{x_t}(x_t^\dagger H_{t-1}x_t) = \left\langle H_{t-1}, \sum_{x_t}\omega_{x_t}x_t x_t^\dagger\right\rangle = \langle H_{t-1}, I_d/d\rangle = 0. \tag{19}$$

We next bound the quadratic increment $\mathbb{E}[(\frac{\Psi_t}{\Psi_{t-1}})^2|\mathcal{F}_{t-1}]$. If $\tau \le t-1$ this is 1, and otherwise

$$\mathbb{E}\left[\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2|\mathcal{F}_{t-1}\right] = 1 + \frac{4\varepsilon^2}{d^2}\,\mathbb{E}\left[\frac{x_t^\dagger H_{t-1}x_t}{x_t^\dagger A x_t}|\mathcal{F}_{t-1}\right] + \frac{4\varepsilon^4}{d^4}\,\mathbb{E}\left[\frac{(x_t^\dagger H_{t-1}x_t)^2}{(x_t^\dagger A x_t)^2}|\mathcal{F}_{t-1}\right].$$

The first expectation is zero by (19). To bound the remaining expectation, note that for any unit vector x,

$$x^\dagger A x \ge a_d x^\dagger x = a_d \ge \frac{1}{2}. \tag{20}$$

So,

$$\begin{aligned}\mathbb{E}\left[\frac{(x_t^\dagger H_{t-1}x_t)^2}{(x_t^\dagger A x_t)^2}|\mathcal{F}_{t-1}\right] &\le 2\,\mathbb{E}\left[\frac{(x_t^\dagger H_{t-1}x_t)^2}{x_t^\dagger A x_t}|\mathcal{F}_{t-1}\right] = 2\sum_{x_t}\omega_{x_t}x_t^\dagger H_{t-1}(x_t x_t^\dagger)H_{t-1}x_t \\ &\le 2\sum_{x_t}\omega_{x_t}x_t^\dagger H_{t-1}^2 x_t = 2\left\langle H_{t-1}^2, \sum_{x_t}\omega_{x_t}x_t x_t^\dagger\right\rangle = 2\langle H_{t-1}^2, I_d/d\rangle = \frac{2}{d}\|H_{t-1}\|_F^2.\end{aligned}$$

Moreover, since $\tau > t-1$, $\|K_{t-1}\|_F \le n^{1/2}d\alpha$ and Lemma 6.7 implies $\|H_{t-1}\|_F \le 3n^{1/2}d\alpha$. Thus,

$$\mathbb{E}\left[\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2|\mathcal{F}_{t-1}\right] \le 1 + \frac{8\varepsilon^4}{d^5}\|H_{t-1}\|_F^2 \le 1 + \frac{72\varepsilon^4 n}{d^3}\alpha^2.$$

So, for all $1 \le t \le n$,

$$\mathbb{E}[\Psi_t^2] = \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2|\mathcal{F}_{t-1}\right]\Psi_{t-1}^2\right] \le \left(1 + \frac{72\varepsilon^4 n}{d^3}\alpha^2\right)\mathbb{E}[\Psi_{t-1}^2],$$

and therefore

$$\mathbb{E}[\Psi_t^2] \le \left(1 + \frac{72\varepsilon^4 n}{d^3}\alpha^2\right)^n \le \exp\left(\frac{72\varepsilon^4 n^2}{d^3}\alpha^2\right) \le 2$$

since $\frac{\varepsilon^4 n^2}{d^3}\alpha^2 \ll 1$. Moreover,

$$\begin{aligned}\mathbb{E}[(\Psi_t - 1)^2] &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2 |\mathcal{F}_{t-1}\right]\Psi_{t-1}^2 - 2\,\mathbb{E}\left[\frac{\Psi_t}{\Psi_{t-1}}|\mathcal{F}_{t-1}\right]\Psi_{t-1} + 1\right] \\ &\le \frac{72\varepsilon^4 n}{d^3}\alpha^2\,\mathbb{E}[\Psi_{t-1}^2] + \mathbb{E}[(\Psi_{t-1}-1)^2] \\ &\le \frac{144\varepsilon^4 n}{d^3}\alpha^2 + \mathbb{E}[(\Psi_{t-1}-1)^2],\end{aligned}$$

so by induction

$$\mathbb{E}[(\Psi_n - 1)^2] \le \frac{144\varepsilon^4 n^2}{d^3}\alpha^2.$$

Thus

$$\Pr\left[|\Psi_n - 1| > \frac{\varepsilon^2 n}{d^{3/2}}\beta\right] \le \frac{\mathbb{E}\left[|\Psi_n - 1|^2\right]}{\frac{\varepsilon^4 n^2}{d^3}\beta^2} \le \frac{144\alpha^2}{\beta^2} = o(1).$$

Therefore, $|\Psi_n - 1| \le \frac{\varepsilon^2 n}{d^{3/2}}\beta$ with probability $1 - o(1)$. □

Claim 6.11. *If $\|K_n\|_F \le n^{1/2} d\alpha$, then $L(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}}$.*

Proof. Using the elementary inequality $e^{2z+z^2} \ge (1+z)^2$ and then Cauchy Schwarz, we can write

$$\begin{aligned}L(\boldsymbol{x}, \boldsymbol{x}) &= \mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\prod_{i=1}^n \left(1 + \varepsilon\frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^2\right] \\ &\le \mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\exp\left(\sum_{i=1}^n 2\varepsilon\frac{x_i^\dagger M x_i}{x_i^\dagger A x_i} + \left(\varepsilon\frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^2\right)\right] \\ &\le \sqrt{\mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\exp\left(4\sum_{i=1}^n \varepsilon\frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)\right]\mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\exp\left(2\sum_{i=1}^n \left(\varepsilon\frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^2\right)\right]}. \quad (21)\end{aligned}$$

Now we bound each of the terms in (21). For the first term, we have

$$\mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\exp\left(4\sum_{i=1}^n \varepsilon\frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)\right] = \mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\exp\left(\frac{4\varepsilon}{d}\langle M, K_n\rangle\right)\right], \quad (22)$$

where we used that $\mathrm{Tr}(M) = 0$. As $M = G - \frac{\mathrm{Tr}(G)}{d}I_d$ for $G \sim \mathrm{GOE}(d)$, we have that $\langle M, K_n\rangle = \langle G, K_n\rangle$ is distributed as a Gaussian with variance at most $\frac{2}{d}\|K_n\|_F^2 \le 2nd\alpha^2$. So we can bound (22) by

$$\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,32\varepsilon^2 n\alpha^2/d)}[\exp(g)] = e^{16\varepsilon^2 n\alpha^2/d} \ll e^{\sqrt{d}}$$

as $\alpha^2 \ll d^{3/2}/(\varepsilon^2 n)$ by assumption. Next we bound the second term in the product in (21). Use $\mathsf{vec}(M)$ to denote rearranging M as a vector in $\mathbb{R}^{d^2}$ (done in a consistent way) and use $\otimes$ to denote the Kronecker product of two matrices. Let $Q \in \mathbb{R}^{d^2\times d^2}$ be defined as

$$Q = \sum_{i=1}^n \frac{x_i x_i^\dagger \otimes x_i x_i^\dagger}{(x_i^\dagger A x_i)^2}.$$

We have

$$\mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\exp\left(2\sum_{i=1}^n\left(\varepsilon\frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^2\right)\right] = \mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\exp\left(2\varepsilon^2 \mathsf{vec}(M)^\dagger Q \mathsf{vec}(M)\right)\right].$$

Now note that

$$Q \preceq 2\sum_{i=1}^n \frac{x_i x_i^\dagger \otimes I_d}{x_i^\dagger A x_i} \preceq \left(\frac{2K_n}{d} + \frac{4nI_d}{d}\right)\otimes I_d$$

so we have $\|Q\|_{\mathrm{op}} \le 6n/d$. Also, we have

$$\|Q\|_1 \le 2\sum_{i=1}^n \frac{x_i^\dagger x_i}{x_i^\dagger A x_i} \le 4n$$

Let $\lambda_1,\dots,\lambda_{d^2}$ be the eigenvalues of Q. Next note that we have

$$\begin{aligned}\mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d)}\left[\exp\left(2\varepsilon^2 \mathsf{vec}(M)^\dagger Q \mathsf{vec}(M)\right)\right] &\le \mathop{\mathbb{E}}_{v\sim N(0,I_{d^2})}\left[\exp\left(10\varepsilon^2/d\cdot v^\dagger Q v\right)\right]\\ &\le \prod_{j=1}^{d^2}\mathop{\mathbb{E}}_{g\sim N(0,1)}\left[\exp(10\varepsilon^2/d\cdot g^2\lambda_j)\right]\\ &\le e^{20\varepsilon^2(\lambda_1+\cdots+\lambda_{d^2})/d} \ll e^{\sqrt{d}}.\end{aligned}$$

In the first step above, we used the convexity of the function inside the expectation to replace the distribution over $M\sim \mathrm{GOE}^*(d)$ with another distribution that can be obtained by adding independent, mean-0 noise to M. Afterwards, we used the rotational invariance of $N(0,I_{d^2})$ and then the bound on $\|Q\|_{\mathrm{op}}$ (together with the fact that $\mathbb{E}_g[e^{cx^2}] = (1-2c)^{-1} \le e^{2c}$ for sufficiently small c), and finally the bound on $\|Q\|_1$. Putting everything together, we conclude that $L(\boldsymbol{x},\boldsymbol{x}) \ll e^{\sqrt{d}}$ as desired. □

Proof of Proposition 6.4. Define the event

$$S = \left\{\sup_{1\le t\le n}\|K_t\|_F \le n^{1/2}d\alpha \text{ and } |\Psi_n - 1| \le \frac{\varepsilon^2 n}{d^{3/2}}\beta\right\}.$$

By Claims 6.9 and 6.10, $\Pr_{p_0}[S] = 1-o(1)$. We will show that if S holds, then $\tau=\infty$. Indeed, if $\tau = t < \infty$, then either $\|K_t\|_F > n^{1/2}d\alpha$ or $|\Phi_t - 1| > \frac{\varepsilon^2 n}{d^{3/2}}\beta$ holds. Since $\Psi_n = \Phi_t$, this contradicts S.

So, $\tau=\infty$ on S. This implies $|L(\boldsymbol{x})-1| = |\Phi_n - 1| \le \frac{\varepsilon^2 n}{d^{3/2}}\beta = o(1)$. Moreover $\|K_n\|_F \le n^{1/2}d\alpha$, so by Claim 6.11 we have $L(\boldsymbol{x},\boldsymbol{x}) \ll e^{\sqrt{d}}$. □

6.3 Bounding H in Frobenius norm by bootstrapping

In this subsection, we prove Lemma 6.7. Throughout this subsection, let $\boldsymbol{z} = (z_1,\dots,z_t)$ be a sequence of unit vectors satisfying $t \le n$ and

$$\|K(\boldsymbol{z})\|_F \le n^{1/2}d\gamma$$

for some $1 \ll \gamma \ll d/(\varepsilon^2 n^{1/2})$.

The following lemma bounds a variant of $K(\boldsymbol{z})$ where we multiply each summand by an adversarial $b_i \in [-1,1]$. This will be used to control the discrepancy $H(\boldsymbol{z}) - K(\boldsymbol{z})$ in the bootstrapping argument.

Lemma 6.12. *Uniformly over $b_1, \dots, b_t \in [-1, 1]$, we have*

$$\left\| \sum_{i=1}^{t} b_i \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \right\|_F \le n^{1/2} d\gamma + 2nd^{1/2}.$$

Proof. For any choice of $b_1, \dots, b_t$,

$$\begin{aligned}
\left\| \sum_{i=1}^{t} b_i \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \right\|_F &\le \left\| \sum_{i=1}^{t} b_i \frac{dz_i z_i^\dagger}{z_i^\dagger A z_i} \right\|_F + \left\| \sum_{i=1}^{t} b_i \frac{I_d}{z_i^\dagger A z_i} \right\|_F \\
&\le \left\| \sum_{i=1}^{t} \frac{dz_i z_i^\dagger}{z_i^\dagger A z_i} \right\|_F + \left\| \sum_{i=1}^{t} \frac{I_d}{z_i^\dagger A z_i} \right\|_F \\
&\le \|K(\boldsymbol{z})\|_F + 2 \left\| \sum_{i=1}^{t} \frac{I_d}{z_i^\dagger A z_i} \right\|_F.
\end{aligned}$$

The second inequality holds because the matrices $dz_i z_i^\dagger$ and I_d are both psd. Using (20), we have

$$\left\| \sum_{i=1}^{t} \frac{I_d}{z_i^\dagger A z_i} \right\|_F \le 2td^{1/2} \le 2nd^{1/2}.$$

The result follows by the assumed bound on $\|K(\boldsymbol{z})\|_F$. ∎

For $S \subseteq [t]$, let $\boldsymbol{z}_S = (z_i)_{i \in S}$. Further, let

$$H_S = \sum_{i \in S} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot \frac{L(\boldsymbol{z}_{S \setminus \{i\}})}{L(\boldsymbol{z}_S)} \qquad \text{and} \qquad K_S = \sum_{i \in S} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i}.$$

The following lemma gives a preliminary bound on $\|H_S\|_F$. In the proof of Lemma 6.7, we will use this bound to control $\|H_S\|_F$ for $|S| = t - O(\log n)$, followed by the bootstrap argument over $O(\log n)$ recursive rounds to contract the bound to $O(n^{1/2}d)$.

Lemma 6.13. *For all $S \subseteq [t]$, $\|H_S\|_F \le 2n^{1/2}d\gamma + 4nd^{1/2}$.*

Proof. Note that for any fixed $\overline{M} \in U$, for the U given by Lemma 6.2, and any unit vector z,

$$\varepsilon \left| \frac{z^\dagger \overline{M} z}{z^\dagger A z} \right| \le \frac{1}{12} \cdot \frac{3}{1/2} = \frac{1}{2},$$

so $1 + \varepsilon \frac{z^\dagger \overline{M} z}{z^\dagger A z} \in [1/2, 3/2]$. Thus, for all i, $L(\boldsymbol{z}_S)/L(\boldsymbol{z}_{S \setminus \{i\}}) \in [1/2, 3/2]$, which implies

$$\frac{L(\boldsymbol{z}_{S \setminus \{i\}})}{L(\boldsymbol{z}_S)} \in [2/3, 2]. \tag{23}$$

Lemma 6.12 gives

$$\frac{1}{2} \|H_S\|_F \le n^{1/2} d\gamma + 2nd^{1/2},$$

as desired. ∎

Proof of Lemma 6.7. Let $D = \log \sqrt{n/d}$. If $t < D$, then by equations (20) and (23),

$$\|H(\boldsymbol{z})\|_F \le \sum_{i=1}^{t} \frac{\left\|dz_i z_i^\dagger - I_d\right\|_F}{z_i^\dagger A z_i} \cdot \frac{L(\boldsymbol{z}_{\sim i})}{L(\boldsymbol{z})} \le 4dD \ll n^{1/2} d\gamma$$

as desired. Otherwise $t \ge D$. We will prove by induction on $a \ge 0$ that if $S \subseteq [t]$ satisfies $|S| = t - D + a$, then

$$\|H_S\|_F \le \xi_a \triangleq 2n^{1/2} d\gamma + 4e^{-a} n d^{1/2}.$$

The base case $a = 0$ holds by Lemma 6.13. For the inductive step, assume $a \ge 1$. By the inductive hypothesis and equations (18) and (20), for all $i \in S$

$$\left| \frac{L(\boldsymbol{z}_S)}{L(\boldsymbol{z}_{S\setminus\{i\}})} - 1 \right| \le \frac{2\varepsilon^2}{d^2} \cdot \left\| \frac{H_{S\setminus i}}{z_i^\dagger A z_i} \right\|_{\mathrm{op}} \le \frac{4\varepsilon^2}{d^2} \left\|H_{S\setminus i}\right\|_F \le \frac{4\varepsilon^2}{d^2} \xi_{a-1}.$$

Since this upper bound is $o(1)$, we also have

$$\left| \frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1 \right| \le \frac{5\varepsilon^2}{d^2} \xi_{a-1}.$$

Write $\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1 = \frac{5\varepsilon^2}{d^2}\xi_{a-1} b_i$ for $b_i \in [-1, 1]$. By Lemma 6.12,

$$\begin{aligned}\left\| \sum_{i\in S} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot \left(\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1 \right) \right\|_F &= \frac{5\varepsilon^2}{d^2} \xi_{a-1} \left\| \sum_{i\in S} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot b_i \right\|_F \\ &\le \left(\frac{5\varepsilon^2 n^{1/2}}{d} \gamma + \frac{10\varepsilon^2 n}{d^{3/2}} \right) \xi_{a-1} \le e^{-1} \xi_{a-1},\end{aligned}$$

using the hypotheses $\gamma \ll d/(\varepsilon^2 n^{1/2})$ and $n \ll d^{3/2}/\varepsilon^2$. By the triangle inequality, equation (20), and our choice of D,

$$\|K_S\|_F \le \|K(\boldsymbol{z})\|_F + \sum_{i\in[t]\setminus S} \frac{\left\|dz_i z_i^\dagger - I_d\right\|_F}{z_i^\dagger A z_i} \le n^{1/2} d\gamma + 2dD \le \frac{101}{100} n^{1/2} d\gamma.$$

Hence

$$\begin{aligned}\|H_S\|_F &\le \|K_S\|_F + \left\| \sum_{i\in S} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot \left(\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1 \right) \right\|_F \\ &\le \frac{101}{100} n^{1/2} d\gamma + e^{-1} \xi_{a-1} \le \xi_a,\end{aligned}$$

as $\frac{101}{100} + 2e^{-1} \le 2$. This completes the induction. Finally,

$$\|H(\boldsymbol{z})\|_F = \left\|H_{[t]}\right\|_F \le 2n^{1/2} d\gamma + 4e^{-D} n d^{1/2} = 2n^{1/2} d\gamma + 4n^{1/2} d \le 3n^{1/2} d\gamma. \qquad \square$$

6.4 Uniform Frobenius bound on the $K(\boldsymbol{x}_{\le t})$ matrix martingale

In this subsection, we will prove Lemma 6.8. The proof mimics the proof of Doob's L^2 maximal inequality. Let $\boldsymbol{x} \sim p_0$, recall that $K_t = K(\boldsymbol{x}_{\le t})$, and define $X = \sup_{1 \le t \le n} \|K_t\|_F$.

Lemma 6.14. *We have that* $\mathbb{E}[X^2] \le 4\,\mathbb{E}[\|K_n\|_F^2]$

Proof. We will first upper bound $\Pr[X \ge x]$ for all $x > 0$. Consider the stopping time $\tau = \inf\{t : \|K_t\|_F \ge x\} \cup \{n\}$. Then,

$$\begin{aligned}\Pr[X \ge x] &= \Pr[\|K_\tau\|_F \ge x] \\ &\le x^{-1}\,\mathbb{E}[\|K_\tau\|_F\,\mathbb{1}\{\|K_\tau\|_F \ge x\}] \\ &\le x^{-1}\,\mathbb{E}[\mathbb{E}[\|K_n\|_F\,|\mathcal{F}_\tau]\mathbb{1}\{\|K_\tau\|_F \ge x\}] \\ &= x^{-1}\,\mathbb{E}[\|K_n\|_F\,\mathbb{1}\{X \ge x\}].\end{aligned}$$

The first estimate is by Markov's inequality, and the second is by convexity of the norm $\|\cdot\|_F$. Thus,

$$\begin{aligned}\mathbb{E}[X^2] &= \int_0^\infty \Pr[X^2 \ge x]\;\mathrm{d}x = \int_0^\infty \Pr[X \ge x] 2x\;\mathrm{d}x \le \int_0^\infty 2\,\mathbb{E}[\|K_n\|_F\,\mathbb{1}\{X \ge x\}]\;\mathrm{d}x \\ &= 2\,\mathbb{E}[\|K_n\|_F\,X] \le 2\sqrt{\mathbb{E}\left[\|K_n\|_F^2\right]\mathbb{E}[X^2]}.\end{aligned}$$

Rearranging yields the result. □

Lemma 6.15. *We have that* $\mathbb{E}\left[\|K_n\|_F^2\right] \lesssim nd^2$.

Proof. We can expand

$$\mathbb{E}[\|K_n\|_F^2] = \sum_{i=1}^n \mathbb{E}\left[\left\|\frac{dx_ix_i^\dagger - I_d}{x_i^\dagger A x_i}\right\|_F^2\right] + 2\sum_{1\le i<j\le n} \mathbb{E}\left[\left\langle \frac{dx_ix_i^\dagger - I_d}{x_i^\dagger A x_i}, \frac{dx_jx_j^\dagger - I_d}{x_j^\dagger A x_j}\right\rangle\right]. \tag{24}$$

Since

$$\mathbb{E}\left[\frac{dx_jx_j^\dagger - I_d}{x_j^\dagger A x_j}|\mathcal{F}_{j-1}\right] = \sum_{x_j} \omega_{x_j}(dx_jx_j^\dagger - I_d) = 0,$$

for any $i < j$ we have

$$\mathbb{E}\left[\left\langle \frac{dx_ix_i^\dagger - I_d}{x_i^\dagger A x_i}, \frac{dx_jx_j^\dagger - I_d}{x_j^\dagger A x_j}\right\rangle\right] = \mathbb{E}\left[\left\langle \frac{dx_ix_i^\dagger - I_d}{x_i^\dagger A x_i}, \mathbb{E}\left[\frac{dx_jx_j^\dagger - I_d}{x_j^\dagger A x_j}|\mathcal{F}_{j-1}\right]\right\rangle\right] = 0.$$

The other expectation in (24) can be bounded by (recalling (20))

$$\mathbb{E}\left[\left\|\frac{dx_ix_i^\dagger - I_d}{x_i^\dagger A x_i}\right\|_F^2\right] \le 2\,\mathbb{E}\left[\frac{\langle dx_ix_i^\dagger - I_d, dx_ix_i^\dagger - I_d\rangle}{x_i^\dagger A x_i}\right] = 2d(d-1)\,\mathbb{E}\left[\frac{1}{x_i^\dagger A x_i}\right] = 2d(d-1).$$

Therefore $\mathbb{E}\left[\|K_n\|_F^2\right] \le 2nd(d-1) \lesssim nd^2$. □

Proof of Lemma 6.8. Follows immediately from Lemmas 6.14 and 6.15. □

7 Lower Bound for Off-Diagonal Perturbations

In this section we consider the family of perturbations which correspond to the "off-diagonal" case described in Section 3. More formally, let $d_1 \geq d_2$ and $A \in \mathbb{R}^{d_1 \times d_1}$ and $B \in \mathbb{R}^{d_2 \times d_2}$ be diagonal matrices with diagonal entries $a_1 \geq \cdots \geq a_{d_1} > 0$ and $b_1 \geq \cdots \geq b_{d_2} > 0$ satisfying $2a_{d_1} \geq a_1$, $2b_{d_2} \geq b_1$, and $\mathrm{Tr}(A) + \mathrm{Tr}(B) = 1$. We abbreviate $a_{d_1} = a$, $b_{d_2} = b$. With these settings, we consider the task of distinguishing between the following two alternatives:

$$H_0 : \rho = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \qquad \text{and} \qquad H_1 : \rho = \begin{pmatrix} A & \frac{\varepsilon}{d_2}\overline{G} \\ \frac{\varepsilon}{d_2}\overline{G}^\dagger & B \end{pmatrix}. \tag{25}$$

Here, $\overline{G} \sim \mathrm{Gin}_U(d_1, d_2)$ for the U given by Lemma 7.1 below.

Lemma 7.1. *For $d_1 \geq d_2$, there exists $U \subseteq \mathbb{R}^{d_1 \times d_2}$ such that if $G \sim \mathrm{Gin}(d_1, d_2)$, then $\Pr[G \notin U] \leq \exp(-0.1 d_1)$ and on the event $G \in U$, we have $\|G\|_{\mathrm{op}} \leq 3$ and $\|M\|_1 \geq d_2/3$ for*

$$M = \begin{pmatrix} 0 & G \\ G^\dagger & 0 \end{pmatrix}.$$

We defer the proof of this lemma to Appendix C.

Parameter Settings. We will assume the parameters $a, b, d_1, d_2, \varepsilon$ satisfy the following relations:

$$d_1 \gg 1 \qquad \varepsilon \leq \frac{1}{10^{20}} \frac{d_2\sqrt{ab}}{\log \frac{1}{a}} \qquad d_1\sqrt{a} \leq d_2\sqrt{b} \qquad d_1 \geq d_2$$

Remark 7.2. *For most places, it suffices to use $\varepsilon \leq \frac{1}{10^{20}} d_2\sqrt{ab}$ so we will often drop the $\log(1/a)$ except for the few places where it is actually necessary.*

Our main result for the distinguishing task (25) is the following.

Theorem 7.3. *Under the assumed parameter settings, the copy complexity of distinguishing between H_0 and H_1 with incoherent measurements is $\Omega(d_1^{1/2} d_2/\varepsilon^2)$.*

We first record several elementary consequences of the parameters settings.

Fact 7.4. *Under the parameter settings, $\rho \sim H_1$ is psd with trace distance at least $\varepsilon/3$ to H_0.*

Proof. The trace distance bound is immediate from the properties of U given by Lemma 7.1. To show ρ is psd, note that for any nonzero $x \in \mathbb{C}^{d_1}$, $y \in \mathbb{C}^{d_2}$,

$$\begin{aligned}(x, y)^\dagger \rho (x, y) &= x^\dagger A x + y^\dagger B y + \frac{2\varepsilon}{d_2} x^\dagger \overline{G} y \geq a\|x\|^2 + b\|y\|^2 - \frac{6\varepsilon}{d_2}\|x\|\,\|y\| \\ &\geq a\|x\|^2 + b\|y\|^2 - \frac{6}{10^{20}}\sqrt{ab}\,\|x\|\,\|y\| > 0.\end{aligned}$$

□

Fact 7.5. *Under the parameter settings, $bd_2 \in [1/4, 1]$.*

Proof. Since $d_1 \geq d_2$ and $d_1\sqrt{a} \leq d_2\sqrt{b}$, we have $d_1 a \leq \frac{d_2}{d_1} d_2 b \leq d_2 b$. Thus $1 = \mathrm{Tr}(A) + \mathrm{Tr}(B) \leq 2d_1 a + 2d_2 b \leq 4d_2 b$ and $1 \geq \mathrm{Tr}(B) \geq d_2 b$. □

Take any learning tree $\mathcal{T}$ corresponding to an algorithm for this task that uses $n \ll d_2 d_1^{1/2}/\varepsilon^2$ incoherent measurements. The parameter settings imply that we may further assume, by taking additional superfluous measurements, that $(\log \frac{n}{\sqrt{ab}})^2/(d_1 a) \le n$. Similarly to the previous section, we let p_0 and p_1 denote the distributions over leaves of $\mathcal{T}$ induced by ρ under H_0 and H_1 respectively, and we will show $d_{\mathrm{TV}}(p_0, p_1) \to 0$.

Because of the block structure in (25), we denote leaves of $\mathcal{T}$ by $(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \dots, (x_n, y_n))$. Here, each (x_i, y_i) satisfies $x_i \in \mathbb{C}^{d_1}$, $y_i \in \mathbb{C}^{d_2}$, and $(x_i, y_i) \in \mathbb{C}^{d_1+d_2}$, and corresponds to an outcome from some (adaptively chosen) rank-1 POVM which we write as

$$\{(x,y)(x,y)^\dagger\}_{(x,y)\in\mathcal{P}}.$$

Note that (x, y) are not necessarily unit vectors. We only require that

$$\sum_{(x,y)\in\mathcal{P}} (x,y)(x,y)^\dagger = I_{d_1+d_2}\,.$$

We let $L^*(\cdot)$ denote the likelihood ratio between p_1 and p_0, i.e. $L^*((\boldsymbol{x}, \boldsymbol{y})) \triangleq p_1((\boldsymbol{x}, \boldsymbol{y}))/p_0((\boldsymbol{x}, \boldsymbol{y}))$. Note that

$$L^*((\boldsymbol{x}, \boldsymbol{y})) = \mathop{\mathbb{E}}_{\overline{G}\sim \mathrm{Gin}_U(d_1,d_2)} \left[\prod_{i=1}^n \left(1 + \frac{2\varepsilon}{d_2} \cdot \frac{x_i^\dagger \overline{G} y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)\right]$$

We similarly define the non-truncated estimate

$$L((\boldsymbol{x}, \boldsymbol{y})) = \mathop{\mathbb{E}}_{G\sim \mathrm{Gin}(d_1,d_2)} \left[\prod_{i=1}^n \left(1 + \frac{2\varepsilon}{d_2} \cdot \frac{x_i^\dagger \overline{G} y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)\right]. \tag{26}$$

We will abuse notation and write $L((\boldsymbol{z}, \boldsymbol{w}))$ for any sequence of unit vectors $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \dots, (z_t, w_t))$ of length not necessarily n. This is defined identically to (26). We let $L((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}, \boldsymbol{y}))$ denote the value of L on input $((x_1, y_1), (x_1, y_1), \dots, (x_n, y_n), (x_n, y_n))$.

The following proposition is analogous to Proposition 6.4 and will be the main ingredient in our proof.

Proposition 7.6. *There exists a subset S of the leaves of $\mathcal{T}$ such that $\Pr_{p_0}[S] = 1 - o(1)$ and for all $(\boldsymbol{x}, \boldsymbol{y}) \in S$, $|L((\boldsymbol{x}, \boldsymbol{y})) - 1| = o(1)$ and $L((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}, \boldsymbol{y})) \le e^{0.02\sqrt{d_1 d_2}}$.*

We now prove Theorem 7.3 assuming Proposition 7.6.

Proof of Theorem 7.3. Analogous to the proof of Theorem 6.3 assuming Proposition 6.4. □

7.1 Recursive evaluation of likelihood ratio

Similarly to the previous section, we obtain a recursive expression for L. Let the sequence of unit vectors $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \dots, (z_t, w_t))$ be as above. For $1 \le i \le t$, let $(\boldsymbol{z}, \boldsymbol{w})_{\sim i}$ be this sequence with (z_i, w_i) omitted. Similarly, for $1 \le i < j \le t$, let $(\boldsymbol{z}, \boldsymbol{w})_{\sim i,j}$ be this sequence with (z_i, w_i) and (z_j, w_j) omitted.

Lemma 7.7. *The function L satisfies*

$$L((\boldsymbol{z}, \boldsymbol{w})) = L((\boldsymbol{z}, \boldsymbol{w})_{\sim t}) + \frac{4\varepsilon^2}{d_1 d_2^2} \sum_{i=1}^{t-1} \left[\frac{\langle z_i, z_t\rangle\langle w_i, w_t\rangle}{(z_i^\dagger A z_i + w_i^\dagger A w_i)(z_t^\dagger A z_t + w_t^\dagger A w_t)} \cdot L((\boldsymbol{z}, \boldsymbol{w})_{\sim i,t})\right].$$

Proof. Analogous to Lemma 6.5. The pairwise moments are evaluated by

$$\mathop{\mathbb{E}}_{G\sim \mathrm{Gin}(d_1,d_2)} [(x^\dagger G y)(z^\dagger G w)] = \frac{1}{d_1}\langle x, z\rangle\langle y, w\rangle.$$ □

7.2 High probability bound on likelihood ratio at leaves

This subsection gives the main part of the proof of Proposition 7.6. For the sequence of unit vectors $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \dots, (z_t, w_t))$ as above, define

$$H((\boldsymbol{z}, \boldsymbol{w})) = \sum_{i=1}^{t} \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i} \cdot \frac{L((\boldsymbol{z}, \boldsymbol{w})_{\sim i})}{L((\boldsymbol{z}, \boldsymbol{w}))}.$$

Lemma 7.7 can be rewritten as

$$\frac{L((\boldsymbol{z}, \boldsymbol{w}))}{L((\boldsymbol{z}, \boldsymbol{w})_{\sim t})} = 1 + \frac{4\varepsilon^2}{d_1 d_2^2} \cdot \frac{z_t^\dagger H((\boldsymbol{z}, \boldsymbol{w})_{\sim t}) w_t}{z_t^\dagger A z_t + w_t^\dagger B w_t}. \tag{27}$$

Further define

$$K((\boldsymbol{z}, \boldsymbol{w})) = \sum_{i=1}^{t} \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i} \qquad \text{and} \qquad \kappa((\boldsymbol{z}, \boldsymbol{w})) = \sup_{b_1, \dots, b_t \in [-1,1]} \left\| \sum_{i=1}^{t} b_i \frac{z_i w_i^\dagger \, \|z_i\| \, \|w_i\|}{(z_i^\dagger A z_i + w_i^\dagger B w_i)^2} \right\|_F.$$

As in the previous section, K will be our proxy for H. The condition that the error terms in the bootstrapping argument contract correspond to an upper bound on $\kappa((\boldsymbol{z}, \boldsymbol{w}))$. In contrast to Lemma 6.12, it is no longer true in this setting that boundedness of $K((\boldsymbol{z}, \boldsymbol{w}))$ implies the required bound on $\kappa((\boldsymbol{z}, \boldsymbol{w}))$ (see Appendix D); this will instead be separately proved in Lemma 7.10 below.

The following three lemmas are the analogs of Lemmas 6.7 and 6.8. Lemma 7.8 deterministically controls H given bounds on K and κ, and Lemmas 7.9 and 7.10 give the required high probability bounds on $K((\boldsymbol{x}, \boldsymbol{y}))$ and $\kappa((\boldsymbol{x}, \boldsymbol{y}))$.

Lemma 7.8. *Suppose $\gamma \gg 1$. If $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \dots, (z_t, w_t))$ satisfies $t \le n$, $\|K((\boldsymbol{z}, \boldsymbol{w}))\|_F \le \sqrt{n d_1 d_2}\gamma$ and $\kappa((\boldsymbol{z}, \boldsymbol{w})) \le \frac{1}{10^3} d_1 d_2^2/\varepsilon^2$, then $\|H((\boldsymbol{z}, \boldsymbol{w}))\|_F \le 3\sqrt{n d_1 d_2}\gamma$.*

Lemma 7.9. *For $(\boldsymbol{x}, \boldsymbol{y}) \sim p_0$, let $(\boldsymbol{x}, \boldsymbol{y})_{\le t} = ((x_1, y_1), \dots, (x_t, y_t))$ be the length-t prefix of $(\boldsymbol{x}, \boldsymbol{y})$. Then $\mathbb{E}\left[\sup_{1 \le t \le n} \|K((\boldsymbol{x}, \boldsymbol{y})_{\le t})\|_F^2\right] \lesssim n d_1 d_2$.*

Lemma 7.10. *If $(\boldsymbol{x}, \boldsymbol{y}) \sim p_0$, then $\Pr[\kappa((\boldsymbol{x}, \boldsymbol{y})) > \frac{1}{10^3} d_1 d_2^2/\varepsilon^2] = o(1)$.*

We now prove Proposition 7.6 assuming Lemmas 7.8, 7.9, and 7.10. These lemmas will be proved in Subsections 7.3, 7.4, and 7.5.

Let α, β be slowly-growing functions with $1 \ll \alpha \ll \beta \ll d_1^{1/2} d_2/(\varepsilon^2 n)$ and furthermore $\alpha^2 \ll d_1^{1/2} d_2/(\varepsilon^2 n)$. This is possible because $n \ll d_1^{1/2} d_2/\varepsilon^2$. Let $(\boldsymbol{x}, \boldsymbol{y}) \sim p_0$. For $1 \le t \le n$, define the filtration $\mathcal{F}_t = \sigma((\boldsymbol{x}, \boldsymbol{y})_{\le t})$ and the sequences

$$H_t = H((\boldsymbol{x}, \boldsymbol{y})_{\le t}), \qquad K_t = K((\boldsymbol{x}, \boldsymbol{y})_{\le t}), \qquad \kappa_t = \kappa((\boldsymbol{x}, \boldsymbol{y})_{\le t}), \qquad \Phi_t = L((\boldsymbol{x}, \boldsymbol{y})_{\le t}).$$

Consider the stopping time (with respect to $\mathcal{F}_t$)

$$\tau = \inf \left\{ t : \|K_t\|_F > \sqrt{n d_1 d_2}\alpha \text{ or } \kappa_t > \frac{1}{10^3} d_1 d_2^2 \text{ or } |\Phi_t - 1| > \frac{\varepsilon^2 n}{d_1^{1/2} d_2} \beta \right\} \cup \{\infty\}$$

and stopped sequence $\Psi_t = \Phi_{t \wedge \tau}$.

Claim 7.11. *With probability $1 - o(1)$, $\|K_t\|_F \le \sqrt{n d_1 d_2}\alpha$ for all $1 \le t \le n$.*

Proof. Follows from Lemma 7.9 and Markov's inequality. ☐

Claim 7.12. *With probability* $1-o(1)$, $|\Psi_n - 1| \le \frac{\varepsilon^2 n}{d_1^{1/2} d_2}\beta$.

Proof. This is analogous to Claim 6.10, and we only sketch the differences. Note that Ψ_t is a multiplicative martingale. We will bound the quadratic increment $\mathbb{E}[(\frac{\Psi_t}{\Psi_{t-1}})^2|\mathcal{F}_{t-1}]$. This is 1 if $\tau \le t-1$, and otherwise by (27), (because the linear term expects to 0)

$$\mathbb{E}\left[\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2|\mathcal{F}_{t-1}\right] = 1 + \frac{16\varepsilon^2}{d_1^2 d_2^4}\,\mathbb{E}\left[\frac{(x_t^\dagger H_{t-1} y_t)^2}{(x_t^\dagger A x_t + y_t^\dagger B y_t)^2}|\mathcal{F}_{t-1}\right].$$

This last expectation is bounded by

$$\begin{aligned}\mathbb{E}\left[\frac{(x_t^\dagger H_{t-1} y_t)^2}{(x_t^\dagger A x_t + y_t^\dagger B y_t)^2}|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\frac{x_t^\dagger H_{t-1} y_t y_t^\dagger H_{t-1} x_t}{(x_t^\dagger A x_t + y_t^\dagger B y_t)^2}|\mathcal{F}_{t-1}\right] \le \mathbb{E}\left[\frac{\|y_t\|^2 x_t^\dagger H_{t-1}^2 x_t}{(x_t^\dagger A x_t + y_t^\dagger B y_t)^2}|\mathcal{F}_{t-1}\right] \\ &\le \frac{1}{b}\,\mathbb{E}\left[\frac{x_t^\dagger H_{t-1}^2 x_t}{x_t^\dagger A x_t + y_t^\dagger B y_t}|\mathcal{F}_{t-1}\right] = \frac{\|H_{t-1}\|_F^2}{b} \le 4d_2\,\|H_{t-1}\|_F^2\end{aligned}$$

using Fact 7.5. Since $\tau > t-1$, we have $\|K_{t-1}\|_F \le \sqrt{n d_1 d_2}\alpha$ and $\kappa_t \le \frac{1}{10} d_1 d_2^2$. Thus Lemma 7.8 implies $\|H_{t-1}\|_F \le 3\sqrt{n d_1 d_2}\alpha$, and

$$\mathbb{E}\left[\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2|\mathcal{F}_{t-1}\right] \le 1 + \frac{64\varepsilon^4}{d_1^2 d_2^3}\,\|H_{t-1}\|_F^2 \le 1 + \frac{576\varepsilon^4 n}{d_1 d_2^2}\alpha^2.$$

Analogously to the proof of Claim 6.10, this implies

$$\mathbb{E}[(\Psi_n - 1)^2] \le \frac{2\cdot 576\varepsilon^4 n^2}{d_1 d_2^2}\alpha^2.$$

The result now follows from Markov's inequality. ☐

Claim 7.13. *If* $\|K_n\|_F \le \sqrt{n d_1 d_2}\alpha$ *and* $\kappa_n \le \frac{1}{10^3}\frac{d_1 d_2^2}{\varepsilon^2}$, *then* $L((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x},\boldsymbol{y})) \le e^{0.02\sqrt{d_1 d_2}}$.

Proof. Using the elementary inequality $e^z \ge 1+z$ and then Cauchy-Schwarz, we can upper bound $L((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x},\boldsymbol{y}))$ by

$$\begin{aligned}L((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x},\boldsymbol{y})) &\le \mathop{\mathbb{E}}_{G\sim\mathrm{Gin}(d_1,d_2)}\left[\exp\left(\sum_{i=1}^n \frac{4\varepsilon}{d_2}\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i} + \left(\frac{2\varepsilon}{d_2}\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)^2\right)\right] \\ &\le \sqrt{\mathop{\mathbb{E}}_{G\sim\mathrm{Gin}(d_1,d_2)}\left[\exp\left(\sum_{i=1}^n \frac{8\varepsilon}{d_2}\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)\right]} \\ &\quad\times\sqrt{\mathop{\mathbb{E}}_{G\sim\mathrm{Gin}(d_1,d_2)}\left[\exp\left(\sum_{i=1}^n \frac{8\varepsilon^2}{d_2^2}\left(\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)^2\right)\right]}.\end{aligned} \tag{28}$$

Now, we bound the two terms in the last product in (28) separately. First, we have

$$\mathop{\mathbb{E}}_{G\sim\mathrm{Gin}(d_1,d_2)}\left[\exp\left(\sum_{i=1}^n \frac{8\varepsilon}{d_2}\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)\right] = \mathop{\mathbb{E}}_{G\sim\mathrm{Gin}(d_1,d_2)}\left[\exp\left(\frac{8\varepsilon}{d_2}\langle G, K_n\rangle\right)\right]. \tag{29}$$

Note that $\frac{8\varepsilon}{d_2}\langle G, K_n\rangle$ is distributed as a Gaussian with variance $\frac{64\varepsilon^2}{d_1 d_2^2}\|K_n\|_F^2 \le \frac{64\varepsilon^2 n}{d_2}\alpha^2$. So we can bound (29) by

$$\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,64\varepsilon^2 n\alpha^2/d_2))}[\exp(g)] = e^{64\varepsilon^2 n\alpha^2/d_2} \ll e^{\sqrt{d_1}},$$

as $\alpha^2 \ll d_1^{1/2} d_2/(\varepsilon^2 n)$ by assumption. Next, we bound the second term in the product in (28). Use $\mathsf{vec}(G)$ to denote rearranging G as a vector in $\mathbb{R}^{d_1 d_2}$ (done in a consistent way) and use $\otimes$ to denote the Kronecker product of two matrices. Define $Q \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ as

$$Q = \sum_{i=1}^n \frac{x_i x_i^\dagger \otimes y_i y_i^\dagger}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}\,.$$

Let $X \in \mathbb{C}^{d_1 d_2 \times n}$ be the matrix with columns given by $\frac{\mathsf{vec}(x_i y_i^\dagger)}{\|x_i\|\|y_i\|}$ for $i = 1, 2, \ldots, n$. Let

$$\theta_i = \frac{\|x_i\|^2 \|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}$$

and let $D \in \mathbb{R}^{n\times n}$ be the diagonal matrix whose diagonal entries are θ_i for $i = 1, 2, \ldots, n$. We can write

$$\|Q\|_F^2 = \sum_{i,j} \theta_i\theta_j \left\langle \frac{x_i y_i^\dagger}{\|x_i\|\,\|y_i\|}, \frac{x_j y_j^\dagger}{\|x_j\|\,\|y_j\|}\right\rangle^2 = \langle DX^\dagger X D, X^\dagger X\rangle\,.$$

By Grothendieck's inequality, we can replace the second term $X^\dagger X$ with $\sigma^\dagger\sigma$ for some $\sigma \in [-1,1]^n$ while incurring at most a factor of 2 in the inequality. Thus, we have

$$\|Q\|_F \le \sqrt{2}\max_{\sigma\in[-1,1]^n}\left\|\sum_{i=1}^n \frac{\sigma_i x_i y_i^\dagger \|x_i\|\,\|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}\right\|_F \le \sqrt{2}\kappa_n \le \frac{d_1 d_2^2}{700\varepsilon^2}\,.$$

In particular, we have $\|Q\|_{\mathrm{op}} \le (d_1 d_2^2)/(700\varepsilon^2)$ and $\|Q\|_1 \le d_1^{3/2} d_2^{5/2}/(700\varepsilon^2)$. Let $\lambda_1, \ldots, \lambda_{d_1 d_2}$ be the eigenvalues of Q. Returning to the last term in (28), we can write

$$\begin{aligned}
\mathop{\mathbb{E}}_{G\sim\mathrm{Gin}(d_1,d_2)}\left[\exp\left(\sum_{i=1}^n \frac{8\varepsilon^2}{d_2^2}\left(\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)^2\right)\right] &= \mathop{\mathbb{E}}_{G\sim\mathrm{Gin}(d_1,d_2)}\left[\exp\left(\frac{8\varepsilon^2}{d_2^2}\mathsf{vec}(G)^\dagger Q\,\mathsf{vec}(G)\right)\right] \\
&= \mathop{\mathbb{E}}_{v\sim N(0,I_{d_1 d_2})}\left[\exp\left(\frac{8\varepsilon^2}{d_1 d_2^2} v^\dagger Q v\right)\right] \\
&= \prod_{j=1}^{d_1 d_2}\mathop{\mathbb{E}}_{g\sim N(0,1)}\left[\exp\left(\frac{8\varepsilon^2 g^2}{d_1 d_2^2}\lambda_j\right)\right] \\
&\le e^{10\varepsilon^2(\lambda_1+\cdots+\lambda_{d_1 d_2})/(d_1 d_2^2)} \\
&\le e^{\sqrt{d_1 d_2}/70}\,, \qquad (30)
\end{aligned}$$

where in the fourth step we used our bound on $\|Q\|_{\mathrm{op}}$ together with the fact that $\mathbb{E}_g[e^{cx^2}] = (1-2c)^{-1} \le e^{5c/4}$ for sufficiently small c, and in the last step we used our bound on $\|Q\|_1$.

Putting (28), (29), 7.2, and (30) together, we conclude $L((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x},\boldsymbol{y})) \ll e^{\sqrt{d_1 d_2}}$ as desired. □

Proof of Proposition 7.6. Define the event

$$S = \left\{ \sup_{1\le t\le n} \|K_t\|_F \le \sqrt{nd_1d_2}\alpha \text{ and } \kappa_n \le \frac{1}{10^3}\frac{d_1d_2^2}{\varepsilon^2} \text{ and } \Psi_n \le \frac{\varepsilon^2 n}{d_1^{1/2}d_2}\beta \right\}.$$

By Lemma 7.10 and Claims 7.11 and 7.12, $\Pr_{p_0}[S] = 1 - o(1)$. If S holds, then $\tau = \infty$: indeed, if $\tau = t < \infty$, then one of $\|K_t\|_F > \sqrt{nd_1d_2}\alpha$, $\kappa_t > \frac{1}{10^3}\frac{d_1d_2^2}{\varepsilon^2}$, and $|\Phi_t - 1| > \frac{\varepsilon^2 n}{d_1^{1/2}d_2}\beta$ holds. Since $\Psi_n = \Phi_t$ and $\kappa_n \ge \kappa_t$ (because in the definition of $\kappa((\boldsymbol{x},\boldsymbol{y}))$ we can take $b_{t+1} = \cdots = b_n = 0$) this contradicts S.

So, $\tau = \infty$ on S. This implies $|L((\boldsymbol{x},\boldsymbol{y})) - 1| = |\Phi_n - 1| \le \frac{\varepsilon^2 n}{d_1^{1/2}d_2}\beta = o(1)$. Moreover, $\|K_n\|_F \le \sqrt{nd_1d_2}\alpha$, so by Claim 7.13 we have $L((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x},\boldsymbol{y})) \le e^{0.02\sqrt{d_1d_2}}$. □

7.3 Bounding H in Frobenius norm by bootstrapping

In this section, we will prove Lemma 7.8. Let $(\boldsymbol{z},\boldsymbol{w}) = ((z_1,w_1),\dots,(z_t,w_t))$ be a sequence of unit vectors satisfying $t \le n$, $\|K((\boldsymbol{z},\boldsymbol{w}))\|_F \le \sqrt{nd_1d_2}\alpha$, and $\kappa((\boldsymbol{z},\boldsymbol{w})) \le \frac{1}{10}d_1d_2^2$, where $\gamma \gg 1$. For $S \subseteq [t]$, let $(\boldsymbol{z},\boldsymbol{w})_S = ((z_i,w_i))_{i\in S}$. Let

$$H_S = \sum_{i\in S} \frac{z_iw_i^\dagger}{z_i^\dagger Az_i + w_i^\dagger Bw_i} \cdot \frac{L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i})}{L((\boldsymbol{z},\boldsymbol{w})_S)} \qquad \text{and} \qquad K_S = \sum_{i\in S} \frac{z_iw_i^\dagger}{z_i^\dagger Az_i + w_i^\dagger Bw_i}.$$

Lemma 7.14. *For all $S \subseteq [t]$, $\|H_S\|_F \le n/\sqrt{ab}$.*

Proof. For any fixed $\overline{G} \in U$, for the U given by Lemma 7.1, and any unit vector (z,w),

$$\frac{2\varepsilon}{d_2}\left|\frac{z^\dagger \overline{G} w}{z^\dagger Az + w^\dagger Bw}\right| \le \frac{2\varepsilon}{d_2}\cdot\frac{3\|z\|\|w\|}{a\|z\|^2 + b\|w\|^2} \le \frac{3\varepsilon}{d_2\sqrt{ab}} \le \frac{3}{10^6}.$$

Thus, for all i, $L((\boldsymbol{z},\boldsymbol{w})_S)/L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i}) \in [1/(1+10^{-5}), 1+10^{-5}]$ which implies

$$\frac{L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i})}{L((\boldsymbol{z},\boldsymbol{w})_S)} \in \left[\frac{1}{1+10^{-5}}, 1+10^{-5}\right]. \tag{31}$$

Thus

$$\|H_S\|_F \le \sum_{i\in S}\left\|\frac{z_iw_i^\dagger}{z_i^\dagger Az_i + w_i^\dagger Bw_i}\right\|_F \cdot (1+10^{-5}) \le \frac{n}{2\sqrt{ab}}\cdot(1+10^{-5}) \le \frac{n}{\sqrt{ab}}.$$ □

Proof of Lemma 7.8. Let $D = \log(n/\sqrt{ab})$. If $t < D$, then by equation (31) and the assumption $(\log\frac{n}{\sqrt{ab}})^2/(d_1a) \le n$,

$$\|H((\boldsymbol{z},\boldsymbol{w}))\|_F \le \sum_{i=1}^{t}\left\|\frac{z_iw_i^\dagger}{z_i^\dagger Az_i + w_i^\dagger Bw_i}\right\|_F \cdot (1+10^{-5}) \le \frac{D}{\sqrt{ab}} \le 3\sqrt{nd_1d_2}\gamma.$$

Otherwise $t \ge D$. We will prove by induction on $a \ge 0$ that if $S \subseteq [t]$ satisfies $|S| = t - D + a$, then

$$\|H_S\|_F \le \xi_a \triangleq 2\sqrt{nd_1d_2}\gamma + e^{-a}\frac{n}{\sqrt{ab}}.$$

The base case $a = 0$ holds by Lemma 7.14. For the inductive step, assume $a \geq 1$. By the inductive hypothesis and equation (27),

$$\left| \frac{L((\boldsymbol{z}, \boldsymbol{w})_S)}{L((\boldsymbol{z}, \boldsymbol{w})_{S\setminus i})} - 1 \right| \leq \frac{4\varepsilon^2}{d_1 d_2^2} \cdot \left| \frac{z_i^\dagger H_{S\setminus i} w_i}{z_i^\dagger A z_i + w_i^\dagger B w_i} \right| \leq \frac{4\varepsilon^2 \xi_{a-1}}{d_1 d_2^2} \cdot \frac{\|z_i\| \, \|w_i\|}{z_i^\dagger A z_i + w_i^\dagger B w_i}.$$

Thus,

$$\|H_S\|_F \leq \|K_S\|_F + \left\| \sum_{i \in S} \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i} \cdot \left(\frac{L((\boldsymbol{z}, \boldsymbol{w})_{S\setminus i})}{L((\boldsymbol{z}, \boldsymbol{w})_S)} - 1 \right) \right\|_F$$

These terms are bounded by (in light of (31))

$$\begin{aligned} \|K_S\|_F &\leq \|K((\boldsymbol{z}, \boldsymbol{w}))\|_F + (1 + 10^{-5}) \sum_{i \in [t]\setminus S} \left\| \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i} \right\|_F \\ &\leq \sqrt{n d_1 d_2}\gamma + \frac{D}{\sqrt{ab}} \leq 1.01\sqrt{n d_1 d_2}\gamma \end{aligned}$$

and, for some $b_1, \ldots, b_t \in [-1, 1]$,

$$\begin{aligned} \left\| \sum_{i \in S} \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i} \cdot \left(\frac{L((\boldsymbol{z}, \boldsymbol{w})_{S\setminus i})}{L((\boldsymbol{z}, \boldsymbol{w})_S)} - 1 \right) \right\|_F &= \frac{4\varepsilon^2 \xi_{a-1}}{d_1 d_2^2} \left\| \sum_{i=1}^t b_i \frac{z_i w_i^\dagger \, \|z_i\| \, \|w_i\|}{(z_i^\dagger A z_i + w_i^\dagger B w_i)^2} \right\|_F \\ &\leq \frac{4\varepsilon^2 \kappa((\boldsymbol{z}, \boldsymbol{w}))}{d_1 d_2^2} \xi_{a-1} \leq e^{-1} \xi_{a-1}. \end{aligned}$$

Since $1.01 + \frac{2}{e} \leq 2$, this implies $\|H_S\| \leq \xi_a$. Therefore

$$\|H((\boldsymbol{z}, \boldsymbol{w}))\|_F = \left\|H_{[t]}\right\|_F \leq 2\sqrt{n d_1 d_2}\gamma + e^{-D}\frac{n}{\sqrt{ab}} = 2\sqrt{n d_1 d_2}\gamma + 1 \leq 3\sqrt{n d_1 d_2}\gamma. \qquad \square$$

7.4 Uniform Frobenius bound on $K((\boldsymbol{x}, \boldsymbol{y})_{\leq t})$

In this subsection, we will prove Lemma 7.9. Let $(\boldsymbol{x}, \boldsymbol{y}) \sim p_0$, $K_t = K((\boldsymbol{x}, \boldsymbol{y})_{\leq t})$ and $X = \sup_{1 \leq t \leq n} \|K_t\|_F$.

Lemma 7.15. *We have that* $\mathbb{E}[X^2] \leq 4\,\mathbb{E}[\|K_n\|_F^2]$.

Proof. Analogous to Lemma 6.14. $\square$

Lemma 7.16. *We have that* $\mathbb{E}[\|K_n\|_F^2] \lesssim n d_1 d_2$.

Proof. We expand

$$\mathbb{E}[\|K_n\|_F^2] = \sum_{i=1}^n \mathbb{E}\left[\left\| \frac{x_i y_i^\dagger}{x_i^\dagger A x_i + y_i^\dagger B y_i} \right\|_F^2 \right] + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}\left[\left\langle \frac{x_i y_i^\dagger}{x_i^\dagger A x_i + y_i^\dagger B y_i}, \frac{x_j y_j^\dagger}{x_j^\dagger A x_j + y_j^\dagger B y_j} \right\rangle \right]$$

The cross terms have expectation 0 like in the proof of Lemma 6.15. Now,

$$\mathbb{E}\left[\left\| \frac{x_i y_i^\dagger}{x_i^\dagger A x_i + y_i^\dagger B y_i} \right\|_F^2 \right] = \mathbb{E}\left[\frac{\|x_i\|^2 \, \|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2} \right] \leq \frac{1}{b}\,\mathbb{E}\left[\frac{\|x_i\|^2}{x_i^\dagger A x_i + y_i^\dagger B y_i} \right] = \frac{d_1}{b} \lesssim d_1 d_2$$

in light of Fact 7.5. $\square$

Proof of Lemma 7.9. Follows from Lemmas 7.15 and 7.16. $\square$

7.5 Balancedness of realizations

Finally, it remains to prove Lemma 7.10. Recall that

$$\kappa((\boldsymbol{x},\boldsymbol{y})) = \sup_{b_1,\dots,b_n\in[-1,1]} \left\| \sum_{i=1}^{n} b_i \frac{x_i y_i^\dagger \|x_i\| \|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2} \right\|_F . \tag{32}$$

We will first prove a few preliminary inequalities about the individual terms in the summation above. In particular, since x_i, y_i are the measurements obtained from a POVM, we argue that over the randomness in the ith measurement, the term

$$\frac{x_i y_i^\dagger \|x_i\| \|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}$$

is not too aligned with any given direction. Thus, intuitively, over $i = 1, \dots, n$, the individual terms will be very weakly correlated and this will allow us to bound the signed sum. In the next two claims below, we think of $\mathcal{P}$ as a POVM that we will use to measure our state.

Claim 7.17. *Let $\mathcal{P}$ be a set of vectors (not necessarily unit vectors) in $\mathbb{C}^d$ such that*

$$\sum_{x\in\mathcal{P}} x x^\dagger = I .$$

Then for any coefficients c_x for all $x \in \mathcal{P}$, we have

$$\left\| \sum_{x\in\mathcal{P}} c_x x \right\| \le \sqrt{\sum_{x\in\mathcal{P}} c_x^2} .$$

Proof. For any unit vector v,

$$\left\langle v, \sum_{x\in\mathcal{P}} c_x x \right\rangle \le \sum_{x\in\mathcal{P}} c_x |\langle v, x\rangle| \le \left(\sqrt{\sum_{x\in\mathcal{P}} c_x^2} \right) \left(\sqrt{\sum_{x\in\mathcal{P}} |\langle v, x\rangle|^2} \right) = \sqrt{\sum_{x\in\mathcal{P}} c_x^2}$$

where we used Cauchy-Schwarz and the hypothesis. Because $\|u\| = \max_{\|v\|=1} \langle v, u\rangle$ for all vectors u, we are done. □

Claim 7.18. *Let $\mathcal{P} = \{(x, y)\}$ be a set of vectors where $x \in \mathbb{C}^{d_1}, y \in \mathbb{C}^{d_2}$ such that*

$$\sum_{z=(x,y),z\in\mathcal{P}} z z^\dagger = I .$$

Then for any choice of $c_{x,y} \in [-1, 1]$ for each $(x, y) \in \mathcal{P}$, we have

$$\left\| \sum_{(x,y)\in\mathcal{P}} \frac{c_{x,y} x \|x\| \|y\|^2}{x^\dagger A x + y^\dagger B y} \right\| \le \frac{\sqrt{d_1}}{b} .$$

Proof. For all choices of $c_{x,y} \in [-1, 1]$

$$\left\| \sum_{(x,y)\in\mathcal{P}} \frac{c_{x,y} x \|x\| \|y\|^2}{x^\dagger A x + y^\dagger B y} \right\| \le \frac{1}{b} \left\| \sum_{(x,y)\in\mathcal{P}} c_{x,y} \|x\| x \right\|_F .$$

By Claim 7.17,

$$\left\| \sum_{(x,y)\in\mathcal{P}} c_{x,y} \|x\| \, x \right\|_F \leq \sqrt{\sum_{(x,y)\in\mathcal{P}} c_{x,y}^2 \|x\|^2} \leq \sqrt{d_1}. \qquad \square$$

We can think of the bound in Claim 7.18 as a bound on a single term in (32) over the randomness of the measurement. Now, we will use Claim 7.18 on all of the terms in (32) to bound the signed sum. Of course, the sum in the expression in Claim 7.18 is inside the norm and it is not immediately clear how to use this to reason about the sum in (32). Actually relating the two expressions requires several additional arguments.

Claim 7.19. *Consider measuring the state H_0 with respect to POVMs $\mathcal{P}_1, \ldots, \mathcal{P}_n$ (which may be chosen adaptively). WLOG, the POVMs are rank-1 and can be viewed as sets of vectors such that for all $i \in [n]$*

$$\sum_{z\in\mathcal{P}_i} zz^\dagger = I \,.$$

Let the results of the measurements be $(x_1, y_1), \ldots, (x_n, y_n)$. If $n \leq d_2 d_1^{1/2}/(10^{20}\varepsilon^2)$. Then with probability $1 - e^{-5d_1}$ over the randomness in the measurements, we have the following inequality for any choice of $c_1, \ldots, c_n \in [-1, 1]$:

$$\left\| \sum_{i=1}^n \frac{c_i x_i y_i^\dagger \|x_i\| \|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2} \right\|_F \leq \frac{d_1 d_2}{4 \cdot 10^3 b \varepsilon^2} \,.$$

Proof. For each $i \in [n]$, define

$$\theta_i = \frac{\|x_i\| \|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2} \,.$$

Let Q be the expression inside the Frobenius norm on the LHS of the desired inequality. Define the matrix $X \in \mathbb{C}^{d_1 \times n}$ to have columns given by $x_1, \ldots, x_n$ and the matrix $Y \in \mathbb{C}^{d_2 \times n}$ to have columns given by $y_1, \ldots, y_n$. Let $N, D \in \mathbb{R}^{n\times n}$ be the diagonal matrices whose entries are $\|y_1\|, \ldots, \|y_n\|$ and $c_1\theta_1, \ldots, c_t\theta_n$ respectively. Now we can rewrite

$$\|Q\|_F^2 = \sum_{i,j} c_i c_j \theta_i \theta_j \langle x_i, x_j\rangle \langle y_i, y_j\rangle = \left\langle (XD)^\dagger XD, Y^\dagger Y \right\rangle = \left\langle (XDN)^\dagger (XDN), (YN^{-1})^\dagger (YN^{-1}) \right\rangle \,.$$

Now by Grothendieck's inequality, we can replace the expression $(N^{-1}Y)^\dagger (N^{-1}Y)$ with $\sigma^\dagger\sigma$ for some sign vector $\sigma \in \{-1, 1\}^n$ while incurring at most a factor of 2 loss. Thus,

$$\max_{c_i} \|Q\|_F^2 \leq 2 \max_{c_i} \left\| \sum_{i=1}^n \frac{c_i x_i \|x_i\| \|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2} \right\|^2 \,. \tag{33}$$

Let R denote the quantity inside the norm on the RHS above. Now it suffices to bound $\|R\|$. First, consider any fixed unit vector v. Note that

$$\langle R, v\rangle \leq \sum_{i=1}^n \frac{\|x_i\| \|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2} |\langle x_i, v\rangle| \,.$$

Next, by Claim 7.18, we have that

$$\mathbb{E}\left[\frac{\|x_i\|\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}|\langle x_i, v\rangle|\right] \leq \frac{\sqrt{d_1}}{b}$$

where the randomness is over the ith measurement. Also note that the individual terms

$$\frac{\|x_i\|\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}|\langle x_i, v\rangle|$$

are always bounded in magnitude by $1/(ab)$. Note that the above two observations also imply that

$$\mathbb{E}\left[\left(\frac{\|x_i\|\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}|\langle x_i, v\rangle|\right)^2\right] \leq \frac{\sqrt{d_1}}{ab^2}\,.$$

Thus, by Freedman's inequality, we have

$$\langle R, v\rangle \leq \frac{d_1 d_2}{10^4 b\varepsilon^2}$$

with failure probability at most

$$\exp\left(-\frac{1}{2}\frac{\frac{d_1^2 d_2^2}{10^8 b^2\varepsilon^4}}{\frac{d_1 d_2}{10^{20}\varepsilon^2 ab^2} + \frac{d_1 d_2}{10^4 ab^2\varepsilon^2}}\right) \leq \exp\left(-\frac{d_1 d_2 a}{10^5\varepsilon^2}\right) \leq \exp(-10 d_1)\,,$$

where in the second step we used that $\frac{d_2 a}{\varepsilon^2} \geq \frac{10^{40}}{d_2 b} \geq 10^{40}$. Finally, we can take some 0.1-net of possible choices of v, which we call Γ, and union bound over all elements of Γ. If the desired inequality were false i.e.

$$\|Q\|_F \geq \frac{d_1 d_2}{4\cdot 10^3 b\varepsilon^2}$$

then by (33), we must have

$$\|R\| \geq \frac{d_1 d_2}{8\cdot 10^3 b\varepsilon^2}\,.$$

By the construction of Γ, there must be some $v \in \Gamma$ such that

$$\langle R, v\rangle \geq 0.9\,\|R\| > \frac{d_1 d_2}{10^4 b\varepsilon^2}$$

and thus we are done. □

Proof of Lemma 7.10. Plugging in the parameter settings at the beginning of Section 7, and using Claim 7.19, we get the desired property. □

8 Instance Near-Optimal Lower Bounds

Our main result on general state certification is the following. Recall that for two quantum states σ, ρ, the fidelity F between them is defined to be $F(\sigma,\rho) = \mathrm{Tr}\left(\sqrt{\sigma^{1/2}\rho\sigma^{1/2}}\right)^2$.

Theorem 8.1. *Let $0 < \varepsilon < \widetilde{O}(1/\log\log(d))$. Let $\sigma \in \mathbb{C}^{d\times d}$ be a density matrix. Then any algorithm that uses incoherent measurements which, given n copies of $\rho \in \mathbb{C}^{d\times d}$, can distinguish between the case where $\rho = \sigma$ and $\|\rho - \sigma\|_1 > \varepsilon$ with probability at least $2/3$ must satisfy*

$$n \geq \Omega\left(\frac{d\sqrt{d_{\mathsf{eff}}}}{\varepsilon^2 \mathrm{polylog}(d/\varepsilon)} \cdot F\left(\sigma^*, \rho_{\mathsf{mm}}\right)\right).$$

Here, σ^* is an explicit density matrix given by zeroing out $O(\varepsilon)$ mass from σ and normalizing, and d_{eff} is the rank of σ^* .

As before, the choice of failure probability can be taken to be any constant greater than $1/2$.

In this section we use Theorems 6.3 and 7.3 to give a simple proof of a slightly weaker version of Theorem 8.1 where the construction of σ^* requires removing up to $O(\varepsilon \log(d/\varepsilon))$ mass. The analysis is a simplified version of the analysis from Sections 5.5 and A.3 of [CLO21]. Later, in Appendix A and B we give a full proof of Theorem 8.1, which involves slightly generalizing Theorem 6.3 and carrying out a more delicate version of the analysis below.

As a first step, notice that since we are given an explicit description of σ, by applying an appropriate rotation, we may assume without loss of generality that σ is diagonal. For the remainder of this section, we will let $\sigma_1 \geq \ldots \geq \sigma_d$ be its eigenvalues (equivalently, its diagonal entries in sorted order).

8.1 Bucketing and mass removal

For $j \in \mathbb{Z}_{\geq 0}$, let S_j denote the set of indices $i \in [d]$ for which $2^{-j-1} < \sigma_i \leq 2^{-j}$, and define $d_j \triangleq |S_j|$. Let $\mathcal{J}$ denote the set of j for which $S_j \neq \emptyset$. We will refer to $j \in \mathcal{J}$ as *buckets*. Given $i \in [d]$, let $j(i)$ denote the index of the bucket for which $i \in S_j$.

Let $\mathcal{J}^* \subseteq \mathcal{J}$ denote the buckets j for which $\sum_{i\in S_j} \sigma_i \geq \varepsilon$, and let $S_{\mathsf{light}} \subseteq [d]$ denote all $i \in [d]$ for which $j(i) \in \mathcal{J}^*$. Let σ' denote the unnormalized density matrix given by zeroing out the i-th entry of σ for every $i \in S_{\mathsf{light}}$, and let σ^* denote the density matrix $\sigma'/\mathrm{Tr}(\sigma')$.

Fact 8.2. *$|\mathcal{J}^*| \leq O(\log(d/\varepsilon))$. In particular, σ' is given by removing $O(\varepsilon \log(d/\varepsilon))$ mass from σ.*

Proof. Note that for any $j \in \mathcal{J}^*$, $2^{-j} > \sigma_i \geq \varepsilon/d$ for all $i \in S_j$, so $j < \log_2(\varepsilon/d)$. □

8.2 Helper lemmas

Here we collect some elementary observations that will be useful in our proof of the weaker version of Theorem 8.1. We begin by noting an alternative way of representing fidelity with respect to the maximally mixed state.

Fact 8.3. *Given psd matrix $\sigma \in \mathbb{C}^{d\times d}$, let $\widehat{\sigma} \triangleq \sigma/\mathrm{Tr}(\sigma)$. Then $F(\widehat{\sigma}, \frac{1}{d}I_d) = \frac{1}{d}\|\sigma\|_{1/2} \cdot \mathrm{Tr}(\sigma)^{-2}$.*

Fact 8.4. *Let S be any set of distinct positive integers. Given a collection of numbers $\{d_j\}_{j\in S}$ satisfying $\sum_j d_j 2^{-j} \leq 2$, let p be the vector with d_j entries equal to 2^{-j} for every j. Then $\|p\|_{1/2} \leq |S|^2 \cdot \max_j d_j^2 2^{-j}$.*

Proof. Let $j^* \triangleq \arg\max d_j^2 2^{-j}$. Then $\|p\|_{1/2}^{1/2} = \sum_j d_j 2^{-j/2} \leq |S| \cdot d_{j^*} 2^{-j^*/2}$. □

Our lower bound instances in the proof of the weaker version of Theorem 8.1 will be based on perturbing certain submatrices of σ. We will use the following basic fact to analyze these instances.

Lemma 8.5. *Consider the task of distinguishing between the following alternatives:*

$$H_0 : \rho = \begin{pmatrix} S & 0 \\ 0 & P \end{pmatrix} \qquad \text{and} \qquad H_1 : \rho = \begin{pmatrix} \widetilde{S} & 0 \\ 0 & P \end{pmatrix}$$

where $S \in \mathbb{R}^{d' \times d'}$ and $P \in \mathbb{R}^{(d-d') \times (d-d')}$ are deterministic psd matrices for which $\mathrm{Tr}(S)+\mathrm{Tr}(P) = 1$, and $\widetilde{S} \in \mathbb{R}^{d' \times d'}$ is drawn from some distribution over psd matrices with trace $\mathrm{Tr}(S)$.

Then the copy complexity of this task using incoherent measurements is $\Omega(n/\mathrm{Tr}(S)) \geq \Omega(n)$, where n is the copy complexity of the following distinguishing task:

$$H_0 : \rho = S/\mathrm{Tr}(S) \qquad \text{and} \qquad H_1 : \rho = \widetilde{S}/\mathrm{Tr}(S)$$

using incoherent measurements.

Theorem 6.3 gives a $\Omega(d^{3/2}/\varepsilon^2)$ lower bound for mixedness testing for d larger than some absolute constant. In the following lemma, we complement this with a weaker lower bound that holds for all d, based on the classical lower bound for uniformity testing. For this, consider the task of distinguishing between the two alternatives:

$$H_0 : \rho = \frac{1}{d} A \qquad \text{and} \qquad H_1 : \rho = \frac{1}{d}\big(A + \varepsilon P Z P^\top\big),$$

where $A \in \mathbb{R}^{d \times d}$ is a diagonal matrix with diagonal entries $a_1 \geq \cdots \geq a_d > 0$ and $\mathrm{Tr}(A) = d$, where $Z = \mathrm{diag}(1, \cdots, -1, \cdots)$ if d is even and $Z = \mathrm{diag}(1, \cdots, -1, \cdots, -1, 0)$ otherwise, and where $P \in \mathbb{R}^{d \times d}$ is a random permutation matrix on the first $2\lfloor d/2 \rfloor$ coordinates.

Lemma 8.6. *For all $d > 1$ and $\varepsilon < 1$, the copy of complexity of distinguishing between H_0 and H_1 with incoherent measurements is $\Omega(\sqrt{d}/\varepsilon^2)$.*

Proof. By Lemma 8.5, it suffices to prove the lemma when d is even. As the states under H_0 and H_1 are both diagonal, we can assume without loss of generality that the measurements are all in the standard basis. Let p_0 denote the uniform distribution over $[d]$. Given $S \subset [d]$ of size $d/2$, let p_S denote the discrete distribution over $[d]$ which places mass $\frac{1+\varepsilon}{d}$ on elements in S and mass $\frac{1-\varepsilon}{d}$ on elements in $[d] \backslash S$. Under H_0, if one measures n copies of ρ, the n measurement outcomes are a sample from $p_0^{\otimes n}$. Under H_1, if one measures n copies of ρ, the measurement outcomes are a sample from $\mathbb{E}_S[p_S^{\otimes n}]$ where S is a random subset of $[d]$ of size $d/2$. It is a standard result in distribution testing that distinguishing between $d_{\mathrm{TV}}(p_0^{\otimes n}, \mathbb{E}_S[p_S^{\otimes n}]) = o(1)$ if $n = o(\sqrt{d}/\varepsilon^2)$ (see e.g. the proof of [Wu17, Theorem 24.1] which is based on [Pan08]). □

Finally, we will use the following lower bound to handle a corner case where σ has one especially large eigenvalue.

Lemma 8.7 (Lemma 5.24 from [CLO21])**.** *Let $\varepsilon \leq 1/2$. If $\sigma_1 \geq 3/4$, then state certification to error ε with respect to σ using incoherent measurements is $\Omega(1/\varepsilon^2)$.*

8.3 Proof of weaker variant of Theorem 8.1

We now give a simple proof of a slight weakening of Theorem 8.1 where one removes $O(\varepsilon \log(d/\varepsilon))$ mass from σ, instead of $O(\varepsilon)$. We strengthen this analysis in Appendix B.

Proof. Note that $\mathrm{Tr}(\sigma') \geq 1 - O(\varepsilon \log(d/\varepsilon)) \geq \Omega(1)$, so by Fact 8.3 it suffices to lower bound the copy complexity by

$$\Omega\left(d_{\mathsf{eff}} \left\|\sigma'\right\|_{1/2} / (\varepsilon^2 \log(d/\varepsilon))\right).$$

We proceed by casework depending on whether or not $d_j = 1$ for all $j \in \mathcal{J}^*$.

Case 1. $d_j = 1$ for all $j \in \mathcal{J}^*$. Note that in this case,

$$\left\|\sigma'\right\|_{1/2}^{1/2} = \sum_{j \in \mathcal{J}^*} 2^{-j/2} = O(1).$$

and $\|\sigma^*\|_{1/2} = \Theta(\|\sigma'\|_{1/2})$. As $d_{\mathsf{eff}} = 1$, it thus suffices to show a copy complexity lower bound of $\Omega(1/\varepsilon^2)$ in this case.

If additionally we have $|\mathcal{J}^*| = 1$, then for $\varepsilon \leq \widetilde{O}(1/\log d)$ sufficiently small, the maximum entry of σ is at least $3/4$, so we can apply Lemma 8.7 to obtain a lower bound of $\Omega(1/\varepsilon^2)$ as desired.

Otherwise, let j, j' be the two smallest bucket indices in $\mathcal{J}^*$, and let i, i' be the elements of the singleton sets $S_j, S_{j'}$. If $\varepsilon \leq c 2^{-j/2 - j'/2 - 1}$ for sufficiently small constant $c > 0$, we can invoke [CLO21, Lemma A.4] to conclude a copy complexity lower bound of $\Omega(1/\varepsilon^2)$.[4]

Otherwise, suppose $\varepsilon > c 2^{-j/2 - j'/2 - 1}$. Because $2^{-j} > 2^{-j'}$, we know that $2^{-j'} \leq O(\varepsilon)$. In particular, consider the state σ^{**} given by zeroing out $\sigma_{i'}$ from σ' and normalizing. For this matrix, $d_{\mathsf{eff}} = 1$ and $\|\sigma^{**}\|_{1/2} = O(1)$. Furthermore, because $\varepsilon \leq \widetilde{O}(1/\log(d))$, we have $\sigma_i \geq 3/4$, so we can apply Lemma 8.7 to conclude a lower bound of $\Omega(1/\varepsilon^2)$.

Case 2. $d_j > 1$ for some $j \in \mathcal{J}^*$. In this case, let $j_1 \triangleq \arg\max_{j \in \mathcal{J}^*} d_j$ and $j_2 \triangleq \arg\max_{j \in \mathcal{J}^*} d_j^2 2^{-j}$.

If $\varepsilon \leq c d_{j_2} 2^{-j_1/2 - j_2/2 - 1} / j_1$ for sufficiently small constant $c > 0$, we can apply the lower bound instance in Section 7 to these two buckets. Let P denote the submatrix of σ containing the diagonal entries of σ outside of $S_{j_1} \cup S_{j_2}$. Let Σ_{j_1} and Σ_{j_2} denote the submatrices of σ containing the diagonal entries indexed by S_{j_1} and S_{j_2}.

If $j_1 \neq j_2$, then consider the distinguishing task

$$H_0 : \rho = \begin{pmatrix} \Sigma_{j_1} & 0 & 0 \\ 0 & \Sigma_{j_2} & 0 \\ 0 & 0 & P \end{pmatrix} (= \sigma) \qquad \text{and} \qquad H_1 : \rho = \begin{pmatrix} \Sigma_{j_1} & \frac{\varepsilon}{d_{j_1}^{1/2} d_{j_2}} \overline{G} & 0 \\ \frac{\varepsilon}{d_{j_1}^{1/2} d_{j_2}} \overline{G}^\top & \Sigma_{j_2} & 0 \\ 0 & 0 & P \end{pmatrix},$$

where $\overline{G}$ is a C-truncated $d_{j_1} \times d_{j_2}$ Ginibre matrix. If d_{j_1} is sufficiently large that Theorem 7.3 applies, then by Lemma 8.5 and Theorem 7.3, this has copy complexity at least

$$\Omega\left(\frac{\sqrt{d_{j_1}} \cdot d_{j_2}}{(\varepsilon/(d_{j_1} 2^{-j_1} + d_{j_2} 2^{-j_2}))^2}\right) \geq \Omega(\sqrt{d_{j_1}} \cdot d_{j_2}^2 2^{-j_2} / \varepsilon^2) \geq \Omega(d_{\mathsf{eff}} \left\|\sigma'\right\|_{1/2} / (\varepsilon^2 \mathrm{polylog}(d/\varepsilon))),$$

where in the first step we used that $d_{j_2} 2^{-j_2} \leq 2$, and in the last step we used Fact 8.2 and Fact 8.4. Note that $\|\sigma^*\|_{1/2} = \Theta(\|\sigma'\|_{1/2})$. Otherwise, if $d_{j_1} = O(1)$ and Theorem 7.3 doesn't apply, we can still apply [CLO21, Lemma A.6] which only differs in its suboptimal dependence of $d_{j_1}^{1/3}$ on the parameter d_{j_1}, which does not affect our overall bound as $d_{j_1} = O(1)$ in this case.

[4] Note that [CLO21, Lemma A.4] gives a (suboptimal) lower bound for the distinguishing task in Section 7. The reason we invoke it instead of Theorem 7.3 is that unlike the latter, it holds for the setting $d_j = d_{j'} = 1$ that we consider here.

If $j_1 = j_2$, then because we are in Case 2 we know $d_{j_1} > 1$, so let $\Sigma_{j_1}^{(1)}$ and $\Sigma_{j_1}^{(2)}$ denote an arbitrary partition of Σ_{j_1} into two $\frac{d_1}{2} \times \frac{d_1}{2}$ diagonal submatrices. Consider the distinguishing task

$$H_0 : \rho = \begin{pmatrix} \Sigma_{j_1}^{(1)} & 0 & 0 \\ 0 & \Sigma_{j_1}^{(2)} & 0 \\ 0 & 0 & P \end{pmatrix} (= \sigma) \qquad \text{and} \qquad H_1 : \rho = \begin{pmatrix} \Sigma_{j_1}^{(1)} & \frac{\varepsilon}{d_{j_1}^{1/2} d_{j_2}} \overline{G} & 0 \\ \frac{\varepsilon}{d_{j_1}^{1/2} d_{j_2}} \overline{G}^\top & \Sigma_{j_1}^{(2)} & 0 \\ 0 & 0 & P \end{pmatrix},$$

where $\overline{G}$ is a C-truncated $\frac{d_{j_1}}{2} \times \frac{d_{j_1}}{2}$ Ginibre matrix. If d_{j_1} is sufficiently large that Theorem 7.3 applies, then by Lemma 8.5 and Theorem 7.3, this has copy complexity at least

$$\Omega\left(\frac{\sqrt{d_{j_1}} \cdot d_{j_1}}{(\varepsilon/d_{j_1} 2^{-j_1} + d_{j_2} 2^{-j_2})^2}\right) \geq \Omega(\sqrt{d_{j_1}} \cdot d_{j_1}^2 2^{-j_1}/\varepsilon^2) \geq \Omega(d_{\mathsf{eff}} \left\|\sigma'\right\|_{1/2} / (\varepsilon^2 \mathrm{polylog}(d/\varepsilon))),$$

where in the first step we used that $d_{j_1} 2^{-j_1} \leq 2$, and in the last step we used Fact 8.2 and Fact 8.4. Otherwise, if $d_{j_1} = O(1)$, we can apply [CLO21, Lemma A.4] as above.

It remains to consider the case where $\varepsilon > c 2^{-j_1/2 - j_2/2 - 1}/j_1$. Let $j^* \triangleq \arg\max_{j \in \mathcal{J}^*} d_j 2^{-5j/2}$. We can apply the lower bound instance in Section 6 to bucket j^*. Letting Σ_{j^*} denote the submatrix of σ containing the diagonal entries of S_{j^*} and P denote the submatrix containing the remaining diagonal entries, we consider the distinguishing task

$$H_0 : \rho = \begin{pmatrix} \Sigma_{j^*} & 0 \\ 0 & P \end{pmatrix} (= \sigma) \qquad \text{and} \qquad \rho = \begin{pmatrix} \Sigma_{j^*} + \frac{\varepsilon}{d_{j^*}} \cdot \overline{M} & 0 \\ 0 & P \end{pmatrix},$$

where $\overline{M}$ is a C-truncated trace-centered GOE matrix. By Lemma 8.5 together with either Theorem 6.3 if $d \gg 1$ or Lemma 8.6 if $d = O(1)$, this has copy complexity at least

$$\Omega\left(\frac{d_{j^*}^{3/2}}{(\varepsilon/(d_{j^*} 2^{-j^*}))^2} \cdot \frac{1}{\mathrm{Tr}(\Sigma_{j^*})}\right) = \Omega(d_{j^*}^{5/2} 2^{-j^*}/\varepsilon^2) \geq \Omega(d_{j_1}^{5/2} 2^{-j_1}/\varepsilon^2).$$

To complete the proof of the theorem, it suffices to show that

$$d_{j_1}^{5/2} 2^{-j_1} \mathrm{polylog}(d/\epsilon) \geq \Omega\left(\sqrt{d_{j_1}} d_{j_2}^2 2^{-j_2}\right)$$

Suppose to the contrary. Then we would get

$$d_{j_1}^2 2^{-j_1} \mathrm{polylog}(d/\epsilon) = o\left(d_{j_2}^2 2^{-j_2}\right). \tag{34}$$

But by assumption on ε,

$$c d_{j_2} 2^{-j_1/2 - j_2/2 - 1}/j_1 \leq \varepsilon \leq d_{j_1} 2^{-j_1},$$

where in the last step we used the fact that $d_j 2^{-j} \geq \Omega(\varepsilon)$ for any $j \in \mathcal{J}^*$. Squaring both sides and rearranging, we find that

$$d_{j_2}^2 2^{-j_2} \leq O(d_{j_1}^2 2^{-j_1} j_1^2) \leq O(d_{j_1}^2 2^{-j_1} \log^2(d/\epsilon)),$$

where the last step follows by the fact that $j_1 \leq \log(d/\epsilon)$ because $2^{-j_1} d \geq 2^{-j_1} d_{j_1} \geq \epsilon$, contradicting (34). ∎

Acknowledgments. SC and JL would like to thank Jordan Cotler, Hsin-Yuan Huang, and John Wright for many illuminating discussions on mixedness testing. Part of this work was completed while SC and BH were visiting the Simons Institute for the Theory of Computing. The authors thank Oufkir Aadil for pointing out a bug in the proofs of Claims 6.10 and 7.13 in an earlier version of this manuscript.

References

[Aar18] Scott Aaronson. Shadow tomography of quantum states. In *STOC*, pages 325–338, 2018.

[ACQ22] Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi. Quantum algorithmic measurement. *Nature communications*, 13(1):1–9, 2022.

[ADJ+11] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 47–68. JMLR Workshop and Conference Proceedings, 2011.

[ADJ+12] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In *Conference on Learning Theory*, pages 22–1. JMLR Workshop and Conference Proceedings, 2012.

[AGKE15] Leandro Aolita, Christian Gogolin, Martin Kliesch, and Jens Eisert. Reliable quantum certification of photonic state preparations. *Nature communications*, 6(1):1–8, 2015.

[ALL21] Anurag Anshu, Zeph Landau, and Yunchao Liu. Distributed quantum inner product estimation. *arXiv preprint arXiv:2111.03273*, 2021.

[ANSV08] Koenraad MR Audenaert, Michael Nussbaum, Arleta Szkoła, and Frank Verstraete. Asymptotic error rates in quantum hypothesis testing. *Communications in Mathematical Physics*, 279(1):251–283, 2008.

[BADG01] Gérard Ben Arous, Amir Dembo, and Alice Guionnet. Aging of spherical spin glasses. *Probab. Theory Related Fields*, 120(1):1–67, 2001.

[BC09] Stephen M Barnett and Sarah Croke. Quantum state discrimination. *Advances in Optics and Photonics*, 1(2):238–278, 2009.

[BCG19] Eric Blais, Clément L Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. *ACM Transactions on Computation Theory (TOCT)*, 11(2):1–37, 2019.

[BCL20] Sebastien Bubeck, Sitan Chen, and Jerry Li. Entanglement is necessary for optimal quantum property testing. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 692–703. IEEE, 2020.

[BOW19] Costin Bădescu, Ryan O'Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 503–514, 2019.

[CCHL21] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. A hierarchy for replica quantum advantage. *arXiv preprint arXiv:2111.05874*, 2021.

[CCHL22] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. Exponential separations between learning with and without quantum memory. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 574–585. IEEE, 2022.

[Che00] Anthony Chefles. Quantum state discrimination. *Contemporary Physics*, 41(6):401–424, 2000.

[CLO21] Sitan Chen, Jerry Li, and Ryan O'Donnell. Toward instance-optimal state certification with incoherent measurements. *arXiv preprint arXiv:2102.13098*, 2021.

[CZSJ22] Senrui Chen, Sisi Zhou, Alireza Seif, and Liang Jiang. Quantum advantages for pauli channel estimation. *Physical Review A*, 105(3):032435, 2022.

[DK16] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

[dSLCP11] Marcus P da Silva, Olivier Landon-Cardinal, and David Poulin. Practical characterization of quantum devices without tomography. *Physical Review Letters*, 107(21):210404, 2011.

[FGLE12] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[FL11] Steven T Flammia and Yi-Kai Liu. Direct fidelity estimation from few pauli measurements. *Physical review letters*, 106(23):230501, 2011.

[HBC+21] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, et al. Quantum advantage in learning from experiments. *arXiv preprint arXiv:2112.00778*, 2021.

[HHJ+17] Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Trans. Inf. Theory*, 63(9):5628–5641, 2017.

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

[HKP21] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021.

[Iss18] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1–2):134–139, 1918.

[JHW18] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

[Low21] Angus Lowe. Learning quantum states without entangled measurements. Master's thesis, University of Waterloo, 2021.

[MdW16] Ashley Montanaro and Ronald de Wolf. A survey of quantum property testing. *Theory of Computing*, pages 1–81, 2016.

[NC02] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

[OW15] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 529–538, 2015.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Tro11] Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

[Ver12] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012.

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[Wri16] John Wright. How to learn a quantum state. *Ph. D. dissertation*, 2016.

[Wu17] Yihong Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf, 2017.

[Yu97] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[Yu19] Nengkun Yu. Quantum closeness testing: A streaming algorithm and applications, 2019.

A Multi-Block Distinguishing Task

The proofs from the preceding sections imply a slightly weaker version of Theorem 8.1 where the lower bound involves removing $\varepsilon \log(d/\varepsilon)$ mass from σ. Avoiding the extra log factor requires working with a slightly more involved instance than the one from Section 6 in which diagonal blocks of σ at *all* scales are perturbed.

To that end, here we analyze the following more general distinguishing task.

$$H_0 : \rho = \frac{1}{d}\begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_m \end{pmatrix} \qquad \text{and} \qquad H_1 : \rho = \frac{1}{d}\begin{pmatrix} A_1 + \varepsilon_1 \overline{M}_1 & & \\ & \ddots & \\ & & A_m + \varepsilon_m \overline{M}_m \end{pmatrix}. \tag{35}$$

Here, each $A_\nu \in \mathbb{R}^{d_\nu \times d_\nu}$ is a diagonal matrix. There exist numbers $j_1, \ldots, j_m$ such that each A_ν has diagonal entries in the interval $[d \cdot 2^{-j_\nu}, d \cdot 2^{-j_\nu+1}]$, and $\sum_\nu \mathrm{Tr}(A_\nu) = d$. Furthermore, for every $\nu \in [m]$, $\overline{M}_\nu \sim \mathrm{GOE}^*_{U_\nu}(d_\nu)$ for the events $U_\nu \triangleq U_{d_\nu}$ given by Lemma A.1 below.

We will refer to the set of d_ν row/column indices of ρ which correspond to A_ν as $\mathcal{B}_\nu$. Let $\mathbb{S}_\nu$ denote the set of unit vectors in $\mathbb{C}^{d-1}$ with entries supported on $\mathcal{B}_\nu$.

The following lemma follows easily from the proof of Lemma 6.2.

Lemma A.1. *There is an absolute constant $a^* > 0$ such that for any integer $d' \geq a^*$, there exists $U_{d'} \subseteq \mathbb{R}^{d' \times d'}$ such that if $M \sim \mathrm{GOE}^*(d')$, then $\Pr[M \notin U_{d'}] \leq o(1/m)$ and on the event $M \in U_{d'}$, we have $\|M\|_{\mathrm{op}} \leq 3 + \Theta(\sqrt{\log(m)/d'})$ and $\|M\|_1 \geq d'/12$.*

Our main result for the distinguishing task (35) is the following.

Theorem A.2. *Let $\varepsilon \triangleq \frac{1}{d}\sum_\nu d_\nu \varepsilon_\nu$ and $N \triangleq \frac{1}{m}\min_{\nu\in[m]} \frac{d_\nu^{1/2} d^2}{\varepsilon_\nu^2 2^{j_\nu}}$. For a^* from Lemma A.1, if $N \geq 1/\epsilon$, $d_\nu \geq a^*$ for all ν, and*

$$\varepsilon_\nu \leq d \cdot 2^{-j_\nu}/(12 + \Theta(\sqrt{\log(m)/d_\nu})) \qquad \text{and} \qquad d_\nu/2^{j_\nu} \geq 2\varepsilon/\log(d/\varepsilon) \quad \forall\ \nu \in [m] \tag{36}$$

then the copy complexity of distinguishing between H_0 and H_1 with incoherent measurements with success probability at least $2/3$ is $\widetilde{\Omega}(N)$.

Note that the bounds in Lemma A.1 and the first part of (36) ensure that under H_1, ρ is psd (and thus a valid quantum state) and has trace distance at least $\Omega(\frac{1}{d}\sum_\nu d_\nu \varepsilon_\nu)$ to the null hypothesis.

Block structure of POVMs. Take any learning tree $\mathcal{T}$ corresponding to an algorithm for this task that uses n incoherent measurements. The following lemma shows that we can assume without loss of generality that every POVM respects the block structure in the distinguishing task, that is, it consists of $\omega_x d \cdot xx^\dagger$ for which each $x \in \mathbb{S}_\nu$ for some $\nu \in [m]$.

Lemma A.3. *Given an arbitrary d-dimensional POVM $\{E_x\}$, there is a corresponding rank-1 POVM $\{E'_y\}$ satisfying the following. Let p, p' be the distributions over measurement outcomes from measuring a state ρ with these POVMs respectively. Then:*

- *For every E'_y, there is some ν such that E'_y is zero outside the principal submatrix indexed by $\mathcal{B}_\nu$.*
- *There is an explicit function f mapping outcomes x of the former POVM to outcomes of the latter for which the pushforward of p' under f is p.*

Proof. This immediately follows from [CLO21, Lemma 5.6] and the fact that we can always assume without loss of generality that POVMs are rank-1. □

Given any $x \in \mathbb{S}_\nu$, we use the notation $\nu(x)$ to denote the $\nu \in [m]$ for which $x \in \mathbb{S}_\nu$. Observe that for any $\nu \in [m]$,

$$\sum_{x:\nu(x)=\nu} \omega_x (xx^\dagger - I_{d_\nu}) = 0 \tag{37}$$

The fact that we can assume every POVM respects the block structure will allow our proof to proceed along very similar lines to that of Theorem 6.3, the key distinction being that instead of tracking the overall likelihood ratio, we track one likelihood ratio for each set of coordinates $\mathcal{B}_\nu$.

Recalling the terminology from Definition 2.2, we let p_0 and p_1 denote the distributions over leaves of $\mathcal{T}$ induced by ρ under H_0 and H_1 respectively. In the rest of this section, let ξ be a slowly-growing function satisfying $\xi \gg \log^c(d/\varepsilon)$ for some absolute constant $c > 0$. We assume

$$n \lll \frac{1}{\xi m} \cdot \min_{\nu\in[m]} \frac{d_\nu^{1/2} d^2}{\varepsilon_\nu^2 2^{j_\nu}}. \tag{38}$$

and will prove $d_{\mathrm{TV}}(p_0, p_1) = o(1)$. By the hypothesis in Theorem A.2 that $N \geq 1/\varepsilon$, we may also assume

$$n \geq 1/\varepsilon \tag{39}$$

by adding superfluous measurements to the algorithm. We set α, β to be slowly-growing functions such that $m^{1/2}\xi^{1/2} \ll \alpha \ll \beta \ll \min_\nu \frac{d_\nu^{1/2} d^2}{\varepsilon^2 2^{j_\nu} n}$. Note that these choices are possible by (38) and (39).

We let $L^*(\cdot)$ denote the likelihood ratio between p_1 and p_0. That is, for a sequence of vectors $\boldsymbol{x} = (x_1, \dots, x_n)$, let $L^*(\boldsymbol{x}) \triangleq p_1(\boldsymbol{x})/p_0(\boldsymbol{x})$. For $\overline{M}_1 \sim \mathrm{GOE}^*_{U_1}(d_1), \dots, \overline{M}_m \sim \mathrm{GOE}^*_{U_m}(d_m)$, note that because $\overline{M}_1, \dots, \overline{M}_m$ are independent, $L^*(\boldsymbol{x}) = \prod_{\nu=1}^m L^*_\nu(\boldsymbol{x})$ where

$$L^*_\nu(\boldsymbol{x}) \triangleq \mathop{\mathbb{E}}_{\overline{M}_\nu} \left[\prod_{i \in [n] : \nu(x_i) = \nu} \left(1 + \varepsilon_\nu \frac{x_i^\dagger \overline{M}_\nu x_i}{x_i^\dagger A_\nu x_i} \right) \right].$$

For $M_1 \sim \mathrm{GOE}^*(d_1), \dots, M_m \sim \mathrm{GOE}^*(d_m)$, define similarly $L(\boldsymbol{x}) = \prod_{\nu=1}^m L_\nu(\boldsymbol{x})$ for

$$L_\nu(\boldsymbol{x}) \triangleq \mathop{\mathbb{E}}_{M_\nu} \left[\prod_{i \in [n] : \nu(x_i) = \nu} \left(1 + \varepsilon_\nu \frac{x_i^\dagger \overline{M}_\nu x_i}{x_i^\dagger A_\nu x_i} \right) \right]. \tag{40}$$

Note that $L(\boldsymbol{x})$ (resp. $L_\nu(\boldsymbol{x})$) is an estimate for the likelihood ratio $L^*(\boldsymbol{x})$ (resp. $L^*_\nu(\boldsymbol{x})$) where the conditioned Gaussian integral is replaced by a true Gaussian integral. Most of the computations in this section will be done in terms of L instead of L_ν; the proof of Theorem A.2 below quantifies that $L(\boldsymbol{x})$ is a close approximation of $L^*(\boldsymbol{x})$.

As before, we will somewhat abuse notation and write $L(\boldsymbol{z})$ for any sequence of unit vectors $\boldsymbol{z} = (z_1, \dots, z_t)$ of length not necessarily n, that is, $L(\boldsymbol{z}) = \prod_{\nu=1}^m L_\nu(\boldsymbol{z})$ where $L_\nu(\boldsymbol{z})$ is defined the same way as in (40). We also write $L_\nu(\boldsymbol{x}, \boldsymbol{x})$ to denote the value of L_ν on input $(x_1, x_1, x_2, x_2, \dots, x_n, x_n)$.

The main ingredient in the proof of Theorem A.2 is the following analogue of Proposition 6.4 giving a high-probability bound on each L_ν evaluated at the leaves of $\mathcal{T}$.

Proposition A.4. *There exists a subset S of the leaves of $\mathcal{T}$ such that $\Pr_{p_0}[S] = 1 - o(1)$ and for all $\boldsymbol{x} \in S$, $|L_\nu(\boldsymbol{x}) - 1| = o(1/m)$ and $L_\nu(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d_\nu}}$ for all $\nu \in [m]$.*

Let us first prove Theorem A.2 assuming Proposition A.4.

Proof of Theorem A.2. Let U denote the event that $M_\nu \in U_\nu$ for all $\nu \in [m]$. Define

$$\begin{aligned}
\overline{L}(\boldsymbol{x}) &\triangleq \mathop{\mathbb{E}}_{M_1 \sim \mathrm{GOE}^*(d_1), \dots, M_m \sim \mathrm{GOE}^*(d_m)} \left[\mathbb{1}\{U\} \prod_{i=1}^n \left(1 + \varepsilon_{\nu(x_i)} \frac{x_i^\dagger M_{\nu(x_i)} x_i}{x_i^\dagger A_{\nu(x_i)} x_i} \right) \right] \\
&= \prod_{\nu=1}^m \mathop{\mathbb{E}}_{M_\nu \sim \mathrm{GOE}^*(d_\nu)} \left[\mathbb{1}[U_\nu] \prod_{i \in [n] : \nu(x_i) = \nu} \left(1 + \varepsilon_\nu \frac{x_i^\dagger M_\nu x_i}{x_i^\dagger A_\nu x_i} \right) \right] \triangleq \prod_{\nu=1}^m \overline{L}_\nu(\boldsymbol{x}).
\end{aligned}$$

It is clear that $L^*_\nu(\boldsymbol{x}) = \Pr[U_\nu]^{-1} \overline{L}_\nu(\boldsymbol{x})$. For all $\boldsymbol{x} \in S$ and $\nu \in [m]$, by Cauchy-Schwarz

$$\begin{aligned}
|L_\nu(\boldsymbol{x}) - \overline{L}_\nu(\boldsymbol{x})| &= \left| \mathop{\mathbb{E}}_{M_\nu} \left[\mathbb{1}\{U_\nu^c\} \prod_{i \in [n] : \nu(x_i) = \nu} \left(1 + \varepsilon_\nu \frac{x_i^\dagger M_\nu x_i}{x_i^\dagger A_\nu x_i} \right) \right] \right| \\
&\le \sqrt{\Pr[U^c] L_\nu(\boldsymbol{x}, \boldsymbol{x})} = o(1/m).
\end{aligned}$$

Here we use that $\Pr[U_\nu^c] \le \mathrm{poly}(1/m) \cdot \exp(-\Omega(d_\nu))$ and $L_\nu(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d_\nu}}$. Moreover, we have $|L_\nu(\boldsymbol{x}) - 1| = o(1/m)$. Thus, for all $\boldsymbol{x} \in S$ and $\nu \in [m]$, $\overline{L}_\nu(\boldsymbol{x}) = 1 + o(1/m)$ and

$$\begin{aligned}
|L^*_\nu(\boldsymbol{x}) - 1| &\le |L^*_\nu(\boldsymbol{x}) - \overline{L}_\nu(\boldsymbol{x})| + |\overline{L}_\nu(\boldsymbol{x}) - 1| \\
&= \frac{\Pr[U_\nu^c]}{\Pr[U_\nu]} \overline{L}_\nu(\boldsymbol{x}) + o(1/m) = o(1/m).
\end{aligned}$$

Recalling that $L^*(\boldsymbol{x}) = \prod_{\nu=1}^m L_\nu^*(\boldsymbol{x})$, we conclude that $L^*(\boldsymbol{x}) = 1 + o(1)$. Finally,

$$\begin{aligned} d_{\mathrm{TV}}(p_0, p_1) &= 2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_0}[(L^*(\boldsymbol{x}) - 1)_-] \\ &= 2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_0}[\mathbb{1}\{\boldsymbol{x} \in S\}(L^*(\boldsymbol{x}) - 1)_-] + 2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_0}[\mathbb{1}\{\boldsymbol{x} \notin S\}(L^*(\boldsymbol{x}) - 1)_-] \\ &\le 2 \sup_{\boldsymbol{x}\in S}(L^*(\boldsymbol{x}) - 1)_- + 2 \Pr_{p_0}[S^c] = o(1). \end{aligned}$$

□

A.1 Recursive evaluation of likelihood ratio

Let $\boldsymbol{z} = (z_1, \dots, z_t)$ be a sequence of unit vectors. For $1 \le i \le t$, let $\boldsymbol{z}_{\sim i}$ be the sequence $\boldsymbol{z}$ with z_i omitted. Similarly, for $1 \le i < j \le t$, let $\boldsymbol{z}_{\sim i,j}$ be the sequence $\boldsymbol{z}$ with z_i, z_j omitted. The main result of this subsection is the following recursive formula for $L(\boldsymbol{z})$.

Lemma A.5. *The function L_ν satisfies*

$$L_\nu(\boldsymbol{z}) = L_\nu(\boldsymbol{z}_{\sim t}) + \frac{2\varepsilon_\nu^2}{d_\nu^2} \cdot \mathbb{1}[\nu(z_t) = \nu] \sum_{i<t:\nu(z_i)=\nu} \left[\frac{d_\nu \langle z_i, z_t\rangle^2 - 1}{(z_i^\dagger A_\nu z_i)(z_t^\dagger A_\nu z_t)} L_\nu(\boldsymbol{z}_{\sim i,t}) \right].$$

As with Lemma 6.5, the proof is based on Isserlis' theorem. For k even, recall that $\mathrm{PMat}(k)$ denotes the set of perfect matchings of $\{1, \dots, k\}$.

Proof of Lemma A.5. The case of $\nu(z_t) \ne \nu$ is clear. We now suppose $\nu(z_t) = \nu$. Let $J \subseteq [t]$ denote the indices s for which $\nu(x_s) = \nu$. For a set $S \subseteq J$ with $|S|$ even, let $\mathrm{PMat}(S)$ denote the set of perfect matchings of S. For even k, let $\mathrm{Mat}(J, k)$ denote the set of matchings of S consisting of $k/2$ pairs. We compute that

$$\begin{aligned} L_\nu(\boldsymbol{z}) &= \sum_{S\subseteq J} \varepsilon_\nu^{|S|} \mathop{\mathbb{E}}_{M_\nu\sim \mathrm{GOE}^*(d_\nu)} \left[\prod_{i\in S} \frac{z_i^\dagger M_\nu z_i}{z_i^\dagger A_\nu z_i} \right] \qquad \text{(expanding definition of } L) \\ &= \sum_{\substack{S\subseteq J \\ |S| \text{ even}}} \varepsilon_\nu^{|S|} \sum_{\{\{a_1,b_1\},\dots,\{a_{|S|/2},b_{|S|/2}\}\}\in \mathrm{PMat}(S)} \prod_{i=1}^{|S|/2} \mathop{\mathbb{E}}_{M_\nu\sim \mathrm{GOE}^*(d_\nu)} \left[\frac{z_{a_i}^\dagger M_\nu z_{a_i}}{z_{a_i}^\dagger A_\nu z_{a_i}} \cdot \frac{z_{b_i}^\dagger M_\nu z_{b_i}}{z_{b_i}^\dagger A_\nu z_{b_i}} \right] \text{ (Th. 6.6)} \\ &= \sum_{k=0}^{\lfloor |J|/2 \rfloor} \varepsilon_\nu^{2k} \sum_{\{\{a_1,b_1\},\dots,\{a_k,b_k\}\}\in \mathrm{Mat}(J,2k)} \prod_{i=1}^{k} \mathop{\mathbb{E}}_{M_\nu\sim \mathrm{GOE}^*(d_\nu)} \left[\frac{z_{a_i}^\dagger M_\nu z_{a_i}}{z_{a_i}^\dagger A_\nu z_{a_i}} \cdot \frac{z_{b_i}^\dagger M_\nu z_{b_i}}{z_{b_i}^\dagger A_\nu z_{b_i}} \right] \\ &= \sum_{k=0}^{\lfloor |J|/2 \rfloor} \left(\frac{2\varepsilon_\nu^2}{d^2} \right)^k \sum_{\{\{a_1,b_1\},\dots,\{a_k,b_k\}\}\in \mathrm{Mat}(J,2k)} \prod_{i=1}^{k} \frac{d_\nu \langle z_{a_i}, z_{b_i}\rangle^2 - 1}{(z_{a_i}^\dagger A_\nu z_{a_i})(z_{b_i}^\dagger A_\nu z_{b_i})}. \end{aligned} \tag{41}$$

The lemma follows by partitioning the summands in (41) based on whether t appears in the matching, and if so which $i \in J$ it is paired with. □

A.2 High probability bound on likelihood ratio at leaves

This subsection gives the main part of the proof of Proposition A.4. For any sequence of vectors $\boldsymbol{z} = (z_1, \dots, z_t)$ and $\nu \in [m]$, define

$$H_\nu(\boldsymbol{z}) = \sum_{i\le t:\nu(z_i)=\nu} \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i} \cdot \frac{L(\boldsymbol{z}_{\sim i})}{L(\boldsymbol{z})} \qquad \text{and} \qquad K_\nu(\boldsymbol{z}) = \sum_{i\le t:\nu(z_i)=\nu} \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}.$$

H_ν enters our calculations by the following rewriting of Lemma A.5:

$$\frac{L_\nu(\boldsymbol{z})}{L_\nu(\boldsymbol{z}_{\sim t})} = 1 + \frac{2\varepsilon_\nu^2}{d_\nu^2} \cdot \mathbb{1}[\nu(z_t) = \nu] \cdot \frac{z_t^\dagger H_\nu(\boldsymbol{z}_{\sim t}) z_t}{z_t^\dagger A_\nu z_t}. \tag{42}$$

If $\boldsymbol{z} = \boldsymbol{x}_{\le t} \triangleq (x_1, \dots, x_t)$ is a prefix of $\boldsymbol{x} \sim p_0$, then $\prod_{\nu=1}^m \frac{L_\nu(\boldsymbol{z})}{L_\nu(\boldsymbol{z}_{\sim t})} = \prod_{\nu=1}^m \frac{L_\nu(\boldsymbol{x}_{\le t})}{L_\nu(\boldsymbol{x}_{\le t-1})} = \frac{L(\boldsymbol{x}_{\le t})}{L(\boldsymbol{x}_{\le t-1})}$ is one step in the likelihood ratio martingale. We will control the contribution from each multiplicative martingale L_ν separately. As we will see (proof of Claim A.10) below, the multiplicative fluctuation of any such step is

$$\mathop{\mathbb{E}}_{x_t}\left[\left(\frac{L_\nu(\boldsymbol{x}_{\le t})}{L_\nu(\boldsymbol{x}_{\le t-1})}\right)^2\right] = 1 + \frac{\varepsilon_\nu^4 2^{j_\nu}}{d^2 d_\nu^4} \cdot \|H_\nu(\boldsymbol{x}_{\le t-1})\|_F^2 .$$

Thus, an upper bound on $\|H_\nu(\boldsymbol{z})\|_F$ over all $\nu \in [m]$ and all prefixes $\boldsymbol{z}$ of $\boldsymbol{x}$ controls the fluctuations of the likelihood ratio martingale. Because the matrices output by H_ν are hard to control directly, we will use the function K_ν as a proxy for H_ν. The following analogue of Lemma 6.7 quantifies this relationship, showing that if $K_\nu(\boldsymbol{z})$ is bounded in Frobenius norm, $H_\nu(\boldsymbol{z})$ is bounded at the same scale.

Lemma A.6. *Suppose $\gamma \gg m^{1/2}\xi^{1/2}$. If $\boldsymbol{z} = (z_1, \dots, z_t)$ is a sequence of unit vectors satisfying $t \le n$ and $\|K_\nu(\boldsymbol{z})\|_F \le \left(n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\gamma$, and the number of $s \in [t]$ for which $\nu(z_s) = \nu$ is at most $n \cdot (d_\nu/2^{j_\nu}) \cdot m\xi$, then $\|H_\nu(\boldsymbol{z})\|_F \le C\left(n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\gamma$ for some absolute constant $C > 0$.*

This lemma is a "deterministic" statement about a sequence of vectors. We will prove this in Subsection A.3 using the same bootstrap argument from earlier.

Lemma A.6 requires that for every ν, the number of POVM elements supported on the coordinates $\mathcal{B}_\nu$ is not much greater than its expectation, which we show holds with high probability:

Lemma A.7. *With probability $1 - o(1)$ over $\boldsymbol{x} \sim p_0$, for all $\nu \in [m]$ there are at most $n \cdot (d_\nu/2^{j_\nu}) \cdot m\xi$ indices $s \in [n]$ for which $\nu(x_s) = \nu$.*

Proof. Take any POVM $\{\omega_x d \cdot xx^\top\}$ where for every x there is some ν for which $x \in \mathbb{S}_\nu$. Now fix $\nu \in [m]$ and note that the probability of observing x for which $\nu(x) = \nu$ upon measuring a copy of ρ under the null hypothesis is

$$\sum_{x:\nu(x)=x} \omega_x x^\dagger A_\nu x \le d \cdot 2^{-j_\nu+1} \cdot \sum_{x:\nu(x)=x} \omega_x \|x\|^2 = 2^{-j_\nu+1} d_\nu,$$

where in the last step we used (37). The lemma follows by Markov and a union bound over $\nu \in [m]$. ∎

Finally, Lemma A.6 also requires a bound on $K_\nu(\boldsymbol{z})$. The following analogue of Lemma 6.8 bounds $K_\nu(\boldsymbol{z})$ in Frobenius norm uniformly over all prefixes $\boldsymbol{z}$ of $\boldsymbol{x}$. We will prove this lemma in Subsection A.4.

Lemma A.8. *If $\boldsymbol{x} \sim p_0$, then $\mathbb{E}\left[\sup_{1 \le t \le n} \|K_\nu(\boldsymbol{x}_{\le t})\|_F^2\right] \lesssim n \cdot 2^{j_\nu} d_\nu^3/d^2$.*

We will now prove Proposition A.4 assuming Lemmas A.6 and A.8. Let $\boldsymbol{x} \sim p_0$. For $1 \le t \le n$, define the filtration $\mathcal{F}_t = \sigma(\boldsymbol{x}_{\le t})$ and the sequences

$$H_{\nu,t} = H_\nu(\boldsymbol{x}_{\le t}), \qquad K_{\nu,t} = K_\nu(\boldsymbol{x}_{\le t}), \qquad \Phi_{\nu,t} = L_\nu(\boldsymbol{x}_{\le t}), \qquad \Phi_t = L(\boldsymbol{x}_{\le t}).$$

Consider the times

$$\tau_\nu = \{\infty\} \cup \inf\Bigg\{ t : \|K_{\nu,t}\|_F > (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha \qquad \text{or}$$
$$|s \in [t] : \nu(x_s) = \nu| > n \cdot (d_\nu/2^{j_\nu}) \cdot m\xi \qquad \text{or} \qquad |\Phi_{\nu,t} - 1| > n \cdot \frac{\varepsilon^2 2^{j_\nu}}{d_\nu^{1/2} d^2}\beta \Bigg\}$$

which are clearly stopping times with respect to $\mathcal{F}_t$. Also define the stopped sequences $\Psi_{\nu,t} = \Phi_{\nu,t\wedge\tau_\nu}$.

Claim A.9. *With probability* $1 - o(1)$, $\|K_{\nu,t}\|_F \le n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d$ *for all* $t \in [n]$ *and all* $\nu \in [m]$.

Proof. By Lemma A.8,

$$\Pr\Big[\sup_{1\le t\le n} \|K_{\nu,t}\|_F > n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d\Big] \le \frac{\mathbb{E}\big[\sup_{1\le t\le n} \|K_{\nu,t}\|_F^2\big]}{(n \cdot 2^{j_\nu} d_\nu^3/d^2)\alpha^2} \lesssim \alpha^{-2} = o(1/m). \qquad \square$$

The claim follows by a union bound over ν.

Claim A.10. *With probability* $1 - o(1)$, $|\Psi_{n,\nu} - 1| \le n \cdot \frac{\varepsilon^2 2^{j_\nu}}{d_\nu^{1/2} d^2}\beta$ *for all* $\nu \in [m]$.

Proof. Note that $\Psi_{\nu,t}$ is a multiplicative martingale: if $\tau \le t-1$ then certainly $\mathbb{E}[\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}}|\mathcal{F}_{t-1}] = 1$, and if $\tau > t-1$, (42) implies

$$\mathbb{E}\left[\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}}|\mathcal{F}_{t-1}\right] = 1 + \frac{2\varepsilon^2}{d_\nu^2}\,\mathbb{E}\left[\mathbb{1}[\nu(x_t) = \nu] \cdot \frac{x_t^\dagger H_{\nu,t-1} x_t}{x_t^\dagger A_\nu x_t}|\mathcal{F}_{t-1}\right] = 1,$$

using that

$$\mathbb{E}\left[\mathbb{1}[\nu(x_t) = \nu] \cdot \frac{x_t^\dagger H_{\nu,t-1} x_t}{x_t^\dagger A_\nu x_t}|\mathcal{F}_{t-1}\right] = \sum_{x_t:\nu(x_t)=\nu} \omega_{x_t}(x_t^\dagger H_{\nu,t-1} x_t)$$
$$= \left\langle H_{\nu,t-1}, \sum_{x_t:\nu(x_t)=\nu} \omega_{x_t} x_t x_t^\dagger \right\rangle = \langle H_{\nu,t-1}, I_{d_\nu}/d\rangle = 0. \quad (43)$$

We next bound the quadratic increment $\mathbb{E}[(\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}})^2|\mathcal{F}_{t-1}]$. If $\tau \le t-1$ this is 1, and otherwise it is given by

$$1 + \frac{4\varepsilon^2}{d_\nu^2}\,\mathbb{E}\left[\mathbb{1}[\nu(x_t) = \nu] \cdot \frac{x_t^\dagger H_{\nu,t-1} x_t}{x_t^\dagger A_\nu x_t}|\mathcal{F}_{t-1}\right] + \frac{4\varepsilon^4}{d_\nu^4}\,\mathbb{E}\left[\mathbb{1}[\nu(x_t) = \nu] \cdot \frac{(x_t^\dagger H_{\nu,t-1} x_t)^2}{(x_t^\dagger A_\nu x_t)^2}|\mathcal{F}_{t-1}\right].$$

The first expectation is zero by (43). To bound the remaining expectation, note that for any unit vector $x \in \mathbb{S}_\nu$,

$$x^\dagger A_\nu x \ge d \cdot 2^{-j_\nu}. \quad (44)$$

So,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}[\nu(x_t)=\nu]\cdot\frac{(x_t^\dagger H_{\nu,t-1}x_t)^2}{(x_t^\dagger A_\nu x_t)^2}|\mathcal{F}_{t-1}\right] &\le (2^{j_\nu}/d)\,\mathbb{E}\left[\mathbb{1}[\nu(x_t)=\nu]\cdot\frac{(x_t^\dagger H_{\nu,t-1}x_t)^2}{x_t^\dagger A_\nu x_t}|\mathcal{F}_{t-1}\right]\\ &= (2^{j_\nu}/d)\sum_{x_t:\nu(x_t)=\nu}\omega_{x_t}x_t^\dagger H_{\nu,t-1}(x_tx_t^\dagger)H_{\nu,t-1}x_t\\ &\le (2^{j_\nu}/d)\sum_{x_t:\nu(x_t)=\nu}\omega_{x_t}x_t^\dagger H_{\nu,t-1}^2x_t\\ &= (2^{j_\nu}/d)\left\langle H_{\nu,t-1}^2,\sum_{x_t:\nu(x_t)=\nu}\omega_{x_t}x_tx_t^\dagger\right\rangle\\ &= (2^{j_\nu}/d)\langle H_{\nu,t-1}^2,I_{d_\nu}/d\rangle = \frac{2^{j_\nu}}{d^2}\left\|H_{\nu,t-1}\right\|_F^2.\end{aligned}$$

Moreover, since $\tau > t-1$, $\|K_{\nu,t-1}\|_F \le (n^{1/2}\cdot 2^{j_\nu/2}d_\nu^{3/2}/d)\alpha$ and Lemma A.6 implies $\|H_{\nu,t-1}\|_F \le (Cn^{1/2}\cdot 2^{j_\nu/2}d_\nu^{3/2}/d)\alpha$. Thus,

$$\mathbb{E}\left[\left(\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}}\right)^2|\mathcal{F}_{t-1}\right] \le 1+\frac{4\varepsilon_\nu^4 2^{j_\nu}}{d^2d_\nu^4}\left\|H_{\nu,t-1}\right\|_F^2 \le 1+\frac{36\varepsilon_\nu^4 2^{2j_\nu}\cdot n}{d_\nu d^4}\alpha^2.$$

So, for all $1\le t\le n$,

$$\mathbb{E}[\Psi_{\nu,t}^2] = \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}}\right)^2|\mathcal{F}_{t-1}\right]\Psi_{\nu,t-1}^2\right] \le \left(1+\frac{36\varepsilon_\nu^4 2^{2j_\nu}\cdot n}{d_\nu d^4}\alpha^2\right)\mathbb{E}[\Psi_{\nu,t-1}^2],$$

and therefore

$$\mathbb{E}[\Psi_{\nu,t}^2] \le \left(1+\frac{36\varepsilon_\nu^4 2^{2j_\nu}\cdot n}{d_\nu d^4}\alpha^2\right)^n \le \exp\left(\frac{36\varepsilon_\nu^4 2^{2j_\nu}\cdot n^2}{d_\nu d^4}\alpha^2\right) \le 2$$

since $\frac{\varepsilon_\nu^4 2^{2j_\nu}\cdot n^2}{d_\nu d^4}\alpha^2 \ll 1$ by assumption.

Moreover,

$$\begin{aligned}\mathbb{E}[(\Psi_{\nu,t}-1)^2] &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}}\right)^2|\mathcal{F}_{t-1}\right]\Psi_{\nu,t-1}^2 - 2\,\mathbb{E}\left[\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}}|\mathcal{F}_{t-1}\right]\Psi_{\nu,t-1}+1\right]\\ &\le \frac{36\varepsilon_\nu^4 2^{2j_\nu}\cdot n}{d_\nu d^4}\alpha^2\cdot\mathbb{E}[\Psi_{\nu,t-1}^2]+\mathbb{E}[(\Psi_{\nu,t-1}-1)^2]\\ &\le \frac{72\varepsilon_\nu^4 2^{2j_\nu}\cdot n^2}{d_\nu d^4}\alpha^2+\mathbb{E}[(\Psi_{\nu,t-1}-1)^2],\end{aligned}$$

so by induction

$$\mathbb{E}[(\Psi_{\nu,n}-1)^2] \le \frac{72\varepsilon_\nu^4 2^{2j_\nu}\cdot n^2}{d_\nu d^4}\alpha^2.$$

Thus

$$\Pr\left[|\Psi_{\nu,n}-1| > n\cdot\left(\frac{72\varepsilon_\nu^4 2^{2j_\nu}}{d_\nu d^4}\beta^2\right)^{1/2}\right] \le \frac{72\alpha^2}{\beta^2} = o(1).$$

Therefore, $|\Psi_{\nu,n}-1| \le n\cdot\frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2}d^2}\beta$ with probability $1-o(1)$. □

Claim A.11. *If* $\|K_{\nu,n}\|_F \le (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha$, *then* $L_\nu(\boldsymbol{x},\boldsymbol{x}) \ll e^{\sqrt{d_\nu}}$.

Proof. Using the elementary inequality $e^z \le 1+z$, we can upper bound $L_\nu(\boldsymbol{x},\boldsymbol{x})$ by

$$L_\nu(\boldsymbol{x},\boldsymbol{x}) \le \mathop{\mathbb{E}}_{M\sim \mathrm{GOE}^*(d_\nu)}\left[\exp\left(\left\langle 2\varepsilon_\nu M, \sum_{i\in[n]:\nu(x_i)=\nu} \frac{x_i x_i^\dagger}{x_i^\dagger A_\nu x_i}\right\rangle\right)\right] = \mathop{\mathbb{E}}_{M}\left[\exp\left(\frac{2\varepsilon_\nu}{d_\nu}\langle M, K_{\nu,n}\rangle\right)\right], \quad (45)$$

where in the second step we used that $\mathrm{Tr}(M)=0$. As $M = G - \frac{\mathrm{Tr}(G)}{d} I_{d_\nu}$ for $G\sim \mathrm{GOE}(d_\nu)$, we have that $\langle M, K_{\nu,n}\rangle = \langle G, K_{\nu,n}\rangle$ is distributed as a Gaussian with variance $\frac{2}{d_\nu}\|K_{\nu,n}\|_F^2 \le (2n2^{j_\nu} d_\nu^2/d^2)\alpha^2$. So we can bound (45) by

$$\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,8\varepsilon_\nu^2 n\alpha^2/d)}[\exp(g)] = e^{8\varepsilon_\nu^2 n 2^{j_\nu}\alpha^2/d^2} \ll e^{\sqrt{d_\nu}}$$

where the last step follows by (38). ∎

Proof of Proposition A.4. Define the event

$$S = \left\{ \sup_{1\le t\le n} \|K_{\nu,t}\|_F \le \left(n^{1/2}\cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\alpha \text{ and } |\Psi_{\nu,n}-1| \le n\cdot \frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2} d^2}\beta \ \ \forall\ \nu\in[m]\right\}.$$

By Claims A.9 and A.10, $\Pr_{p_0}[S] = 1-o(1)$. We will show that if S holds, then $\tau=\infty$. Indeed, if $\tau = t < \infty$, then there exists ν such that either $\|K_{\nu,t}\|_F > (n^{1/2}\cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha$ or $|\Phi_{\nu,t}-1| > n\cdot \frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2}d^2}\beta$ holds. Since $\Psi_{\nu,n} = \Phi_{\nu,t}$, this contradicts S.

So, $\tau=\infty$ on S. This implies that for all ν, $|L_\nu(\boldsymbol{x})-1| = |\Phi_{\nu,n}-1| \le n\cdot \frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2}d^2}\beta = o(1/m)$. Moreover $\|K_{\nu,n}\|_F \le (n^{1/2}\cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha$ for all ν, so by Claim A.11 we have $L_\nu(\boldsymbol{x},\boldsymbol{x}) \ll e^{\sqrt{d_\nu}}$ for all ν. ∎

A.3 Bounding H in Frobenius norm by bootstrapping

In this subsection, we prove Lemma A.6. Throughout this subsection, fix some $\nu\in[m]$. To ease notation, we will drop subscripts and refer to K_ν and H_ν simply as K and H. Let $\boldsymbol{z} = (z_1,\dots,z_t)$ be a sequence of unit vectors satisfying $t\le n$ and

$$\|K(\boldsymbol{z})\|_F \le \left(n^{1/2}\cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\gamma$$

for some $\gamma \gg m^{1/2}\xi^{1/2}$. Let $J\subseteq[t]$ denote the set of $s\in[t]$ for which $\nu(z_s)=\nu$. Suppose that

$$|J| \le n\cdot (d_\nu/2^{j_\nu})\cdot m\xi \quad (46)$$

as in Lemma A.7.

The following lemma bounds a variant of $K(\boldsymbol{z})$ where we multiply each summand by an adversarial $b_i\in[-1,1]$. This will be used to control the discrepancy $H(\boldsymbol{z})-K(\boldsymbol{z})$ in the bootstrapping argument.

Lemma A.12. *Uniformly over* $b_1,\dots,b_t\in[-1,1]$, *we have*

$$\left\|\sum_{i\in J} b_i \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}\right\|_F \lesssim n d_\nu^{1/2}\cdot m\xi + (n^{1/2}\cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\gamma$$

Proof. For any choice of $b_1, \dots, b_t$,

$$\begin{aligned}\left\|\sum_{i\in J} b_i \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}\right\|_F &\le \left\|\sum_{i\in J} b_i \frac{d_\nu z_i z_i^\dagger}{z_i^\dagger A_\nu z_i}\right\|_F + \left\|\sum_{i\in J} b_i \frac{I_{d_\nu}}{z_i^\dagger A z_i}\right\|_F \\ &\le \left\|\sum_{i\in J} \frac{d_\nu z_i z_i^\dagger}{z_i^\dagger A_\nu z_i}\right\|_F + \left\|\sum_{i\in J} \frac{I_{d_\nu}}{z_i^\dagger A_\nu z_i}\right\|_F \\ &\le \|K(\boldsymbol{z})\|_F + 2\left\|\sum_{i\in J} \frac{I_{d_\nu}}{z_i^\dagger A_\nu z_i}\right\|_F.\end{aligned}$$

The second inequality holds because the matrices $d_\nu z_i z_i^\dagger$ and I_{d_ν} are both psd. Using (44) and the assume bound on $|J|$ in (46), we have

$$\left\|\sum_{i\in J} \frac{I_{d_\nu}}{z_i^\dagger A_\nu z_i}\right\|_F \le (2^{j_\nu}/d)|J| d_\nu^{1/2} \lesssim n d_\nu^{1/2} \cdot m\xi.$$

The result follows by our assumed bound on $\|K(\boldsymbol{z})\|_F$. ∎

For $S \subseteq J$, let $\boldsymbol{z}_S = (z_i)_{i\in S}$. Further, let

$$H_S = \sum_{i\in S} \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i} \cdot \frac{L_\nu(\boldsymbol{z}_{S\setminus\{i\}})}{L_\nu(\boldsymbol{z}_S)} \qquad \text{and} \qquad K_S = \sum_{i\in S} \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}.$$

The following lemma gives a preliminary bound on $\|H_S\|_F$. In the proof of Lemma A.6, we will use this bound to control $\|H_S\|_F$ for $|S| = t - O(\log n)$, followed by the bootstrap argument over $O(\log n)$ recursive rounds to contract the bound to $O((2^{j_\nu}/d)^{1/2} d_\nu n^{1/2}\gamma)$.

Lemma A.13. *There exists an absolute constant C such that for all $S \subseteq [t]$, $\|H_S\|_F \le C(n d_\nu^{1/2} \cdot m\xi + (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\gamma)$.*

Proof. We take C to be twice the constant hidden by the $\lesssim$ in Lemma A.12. Note that for any fixed $\overline{M}_\nu \in U_\nu$, for the U_ν given by Lemma A.1, and any unit vector $z \in \mathbb{S}_\nu$,

$$\varepsilon \left|\frac{z^\dagger \overline{M}_\nu z}{z^\dagger A_\nu z}\right| \le \frac{1}{12} \cdot \frac{3}{1/2} = \frac{1}{2},$$

so $1 + \varepsilon \frac{z^\dagger \overline{M}_\nu z}{z^\dagger A_\nu z} \in [1/2, 3/2]$. Thus, for all i, $L_\nu(\boldsymbol{z}_S)/L_\nu(\boldsymbol{z}_{S\setminus\{i\}}) \in [1/2, 3/2]$, which implies

$$\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} \in [2/3, 2]. \tag{47}$$

Lemma A.12 gives

$$\frac{1}{2}\|H_S\|_F \le \frac{1}{2} C(n d_\nu^{1/2} \cdot m\xi + (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\gamma).$$

as desired.

∎

Proof of Lemma A.6. Let $\varepsilon^* \triangleq \left(n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\gamma$ and $\varepsilon' \triangleq Cnd_\nu^{1/2} \cdot m\xi$. If $\varepsilon^* \geq \varepsilon'$, then we are already done by Lemma A.13. Otherwise, suppose $\varepsilon^* < \varepsilon'$. and let $D = \log(\varepsilon'/\varepsilon^*)$. If $t < D$, then by equations (44) and (47),

$$\|H(\boldsymbol{z})\|_F \leq \sum_{i=1}^{t} \frac{\left\|d_\nu z_i z_i^\dagger - I_{d_\nu}\right\|_F}{z_i^\dagger A z_i} \cdot \frac{L(\boldsymbol{z}_{\sim i})}{L(\boldsymbol{z})} \leq 2^{j_\nu+1} D d_\nu/d.$$

But note that $2^{j_\nu+1}Dd_\nu/d \ll \varepsilon^*$ provided that $n \ggg 2^{j_\nu}/(d_\nu\gamma^2)$. By (36), $2^{j_\nu}/d_\nu \leq \log(d/\varepsilon)/(2\varepsilon)$, so this holds by (39) and our choice of $\gamma \gg m^{1/2}\xi^{1/2} \geq \mathrm{polylog}(d/\varepsilon)$. So $\|H(\boldsymbol{z})\|_F \ll \varepsilon^*$ when $t < D$.

Now suppose $t \geq D$. By Lemma A.13 and the assumption that $\varepsilon^* \leq \varepsilon'$, $\|H_S\|_F \leq 2\varepsilon'$. We will prove by induction on $a \geq 0$ that if $S \subseteq J$ satisfies $|S\backslash J| = D - a$, then

$$\|H_S\|_F \leq \xi_a \triangleq 2\varepsilon^* + e^{-a} \cdot 2\varepsilon'.$$

The base case $a = 0$ clearly holds. For the inductive step, assume $a \geq 1$. By the inductive hypothesis and equations (42) and (44), for all $i \in S$

$$\left|\frac{L_\nu(\boldsymbol{z}_S)}{L_\nu(\boldsymbol{z}_{S\backslash\{i\}})} - 1\right| \leq \frac{2\varepsilon_\nu^2}{d^2} \cdot \left\|\frac{H_{S\backslash i}}{z_i^\dagger A_\nu z_i}\right\|_{\mathrm{op}} \leq \frac{(2^{j_\nu}/d)\cdot 2\varepsilon_\nu^2}{d^2}\left\|H_{S\backslash i}\right\|_F \leq \frac{(2^{j_\nu}/d)\cdot 2\varepsilon_\nu^2}{d^2}\xi_{a-1}.$$

Since this upper bound is $o(1)$ by (38) and the second part of (36), we also have

$$\left|\frac{L(\boldsymbol{z}_{S\backslash\{i\}})}{L(\boldsymbol{z}_S)} - 1\right| \leq \frac{(2^{j_\nu}/d)\cdot 3\varepsilon_\nu^2}{d^2}\xi_{a-1}.$$

Write $\frac{L_\nu(\boldsymbol{z}_{S\backslash\{i\}})}{L_\nu(\boldsymbol{z}_S)} - 1 = \frac{(2^{j_\nu}/d)\cdot 3\varepsilon_\nu^2}{d^2}\xi_{a-1}b_i$ for $b_i \in [-1,1]$. By Lemma A.12, there is a constant c such that

$$\begin{aligned}\left\|\sum_{i\in S}\frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}\cdot\left(\frac{L_\nu(\boldsymbol{z}_{S\backslash\{i\}})}{L_\nu(\boldsymbol{z}_S)} - 1\right)\right\|_F &= \frac{(2^{j_\nu}/d)\cdot 3\varepsilon_\nu^2}{d^2}\xi_{a-1}\left\|\sum_{i\in S}\frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}\cdot b_i\right\|_F \\ &\leq \frac{(2^{j_\nu}/d)\cdot 3\varepsilon_\nu^2 n d_\nu^{1/2}\cdot m\xi}{d^2}\xi_{a-1} \leq e^{-1}\xi_{a-1},\end{aligned}$$

where in the last step we used that $n \ll \frac{d^2 d_\nu^{1/2}}{\varepsilon_\nu^2 2^{j_\nu}\cdot m\xi}$. By the triangle inequality, equation (44), and our choice of D,

$$\|K_S\|_F \leq \|K(\boldsymbol{z})\|_F + \sum_{i\in J\backslash S}\frac{\left\|d_\nu z_i z_i^\dagger - I_{d_\nu}\right\|_F}{z_i^\dagger A_\nu z_i} \leq \varepsilon^* + 2^{j_\nu} D d_\nu/d \leq \frac{101}{100}\varepsilon^*.$$

Hence

$$\begin{aligned}\|H_S\|_F &\leq \|K_S\|_F + \left\|\sum_{i\in S}\frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i}\cdot\left(\frac{L(\boldsymbol{z}_{S\backslash\{i\}})}{L(\boldsymbol{z}_S)} - 1\right)\right\|_F \\ &\leq \frac{101}{100}\varepsilon^* + e^{-1}\xi_{a-1} \leq \xi_a,\end{aligned}$$

as $\frac{101}{100} + 2e^{-1} \leq 2$. This completes the induction. Finally,

$$\|H(\boldsymbol{z})\|_F = \|H_J\|_F \leq 2\varepsilon^* + e^{-D}2\varepsilon' = 4\varepsilon^*. \qquad \square$$

A.4 Uniform Frobenius bound on the $K_\nu(\boldsymbol{x}_{\le t})$ matrix martingale

In this subsection, we will prove Lemma A.8. Fix any $\nu \in [m]$, let $\boldsymbol{x} \sim p_0$, recall that $K_{\nu,t} = K_\nu(\boldsymbol{x}_{\le t})$. To ease notation, we will drop the subscript ν and refer to this as K_t. Also define $X = \sup_{1\le t\le n} \|K_t\|_F$.

Lemma A.14. *We have that* $\mathbb{E}[X^2] \le 4\,\mathbb{E}[\|K_{\nu,n}\|_F^2]$

Proof. Analogous to Lemma 6.14. □

Lemma A.15. *We have that* $\mathbb{E}\left[\|K_n\|_F^2\right] \lesssim 2^{j_\nu} d_\nu^2 n/d$.

Proof. We can expand

$$\mathbb{E}[\|K_n\|_F^2] = \sum_{i=1}^n \mathbb{E}\left[\mathbb{1}[\nu(x_i)=\nu]\cdot\left\|\frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i}\right\|_F^2\right] + 2\sum_{1\le i<j\le n}\mathbb{E}\left[\left\langle \mathbb{1}[\nu(x_i)=\nu,\nu(x_j)=\nu]\cdot\frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i}, \frac{d_\nu x_j x_j^\dagger - I_{d_\nu}}{x_j^\dagger A_\nu x_j}\right\rangle\right]. \quad (48)$$

By (37),

$$\mathbb{E}\left[\mathbb{1}[\nu(x_j)=\nu]\cdot\frac{d_\nu x_j x_j^\dagger - I_{d_\nu}}{x_j^\dagger A_\nu x_j}|\mathcal{F}_{j-1}\right] = \sum_{x_j:\nu(x_j)=\nu}\omega_{x_j}(d_\nu x_j x_j^\dagger - I_{d_\nu}) = 0,$$

so for any $i<j$ we have

$$\mathbb{E}\left[\mathbb{1}[\nu(x_i)=\nu,\nu(x_j)=\nu]\cdot\left\langle\frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i}, \frac{d_\nu x_j x_j^\dagger - I_{d_\nu}}{x_j^\dagger A_\nu x_j}\right\rangle\right] = \mathbb{E}\left[\mathbb{1}[\nu(x_i)=\nu]\left\langle\frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i}, \mathbb{E}\left[\mathbb{1}[\nu(x_j)=\nu]\cdot\frac{d_\nu x_j x_j^\dagger - I_{d_\nu}}{x_j^\dagger A_\nu x_j}|\mathcal{F}_{j-1}\right]\right\rangle\right] = 0.$$

The other expectation in (48) can be bounded by (recalling (44))

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}[\nu(x_i)=\nu]\cdot\left\|\frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i}\right\|_F^2\right] &\le (2^{j_\nu}/d)\,\mathbb{E}\left[\mathbb{1}[\nu(x_i)=\nu]\cdot\frac{\langle d_\nu x_i x_i^\dagger - I_{d_\nu}, d_\nu x_i x_i^\dagger - I_{d_\nu}\rangle}{x_i^\dagger A_\nu x_i}\right]\\ &= (2^{j_\nu}/d)\cdot d_\nu(d_\nu-1)\,\mathbb{E}\left[\frac{\mathbb{1}[\nu(x_i)=\nu]}{x_i^\dagger A_\nu x_i}\right]\\ &\le (2^{j_\nu}/d)\cdot d_\nu(d_\nu-1)\cdot\sum_{x:\nu(x)=x}\omega_x \le 2^{j_\nu} d_\nu^3/d^2.\end{aligned}$$

Therefore $\mathbb{E}\left[\|K_n\|_F^2\right] \le n\cdot 2^{j_\nu} d_\nu^3/d^2$. □

Proof of Lemma A.8. Follows immediately from Lemmas A.14 and A.15. □

B Refined Bounds for State Certification

In this section we use the lower bound instance from Appendix A to give a refined version of the analysis in Section 8 and prove Theorem 8.1. The steps in this section are essentially already present in [CLO21] (see Sections 5.1, 5.2.2, and 5.5 therein), but we include them for the sake of completeness.

B.1 Bucketing and mass removal

We will use the following bucketing scheme from [CLO21, Definition 5.2].

For $j \in \mathbb{Z}_{\geq 0}$, let S_j denote the set of indices $i \in [d]$ for which $2^{-j-1} < \sigma_i \leq 2^{-j}$, and define $d_j \triangleq |S_j|$. Let $\mathcal{J}$ denote the set of j for which $S_j \neq \emptyset$. We will refer to $j \in \mathcal{J}$ as *buckets*. Given $i \in [d]$, let $j(i)$ denote the index of the bucket for which $i \in S_j$.

Our bounds are based on the following modification of σ given by removing a small fraction of its entries:

Definition B.1. *If b is the largest number for which the b smallest entries of σ sum to at most ε, define $S^1_{\mathsf{light}} \subseteq [d]$ to be the indices of these b smallest entries. Let S^2_{light} denote the set of $i \in [d]$ for which $\sum_{i' \in S_{j(i)}} \sigma_{i'} \leq 2\varepsilon / \log(d/\varepsilon)$. Define $S_{\mathsf{light}} \triangleq S^1_{\mathsf{light}} \cup S^2_{\mathsf{light}}$.*

Without loss of generality, assume that all σ_i are sorted in increasing order based on $\sigma_i / d^2_{j(i)}$. Recall the constant a^* from Lemma A.1 and Theorem A.2. Let $d' \leq d$ denote the largest index for which $\sum_{i \notin S_{\mathsf{light}} : i \leq d'} \sigma_i \leq C_{a^*} \varepsilon$ for a constant C_{a^*} sufficiently lage depending on a^*. Let $S_{\mathsf{tail}} \triangleq \{i : i \notin S_{\mathsf{light}}, i \leq d'\}$.

Let m denote the number of buckets $j \in \mathcal{J}$ for which S_j and S_{light} are disjoint. Let $S_{\mathsf{few}} \subseteq [d]$ denote the set of i belonging to a bucket of size less than a^*, and let $S_{\mathsf{many}} \subseteq [d]$ denote the set of i belonging to a bucket of size at least a^*.

Let σ' denote the matrix given by zeroing out the entries indexed by $S_{\mathsf{tail}} \cup S_{\mathsf{light}}$. Let σ'' denote the matrix by further zeroing out the largest entry of σ'. Let σ^* denote the density matrix $\sigma' / \mathrm{Tr}(\sigma^*)$.

Lastly, define $\mathcal{J}^*$ to be the set of $j \in \mathcal{J}$ for which S_j has nonempty intersection with $S_{\mathsf{many}} \backslash S_{\mathsf{light}}$.

Fact B.2. *We have $m \leq O(\log(d/\varepsilon))$, that is, there are at most $O(\log(d/\varepsilon))$ indices $j \in \mathcal{J}$ for which S_j and S_{light} are disjoint. Furthermore, the total mass of σ in $S_{\mathsf{light}} \cup S_{\mathsf{tail}}$ is $O(\varepsilon)$.*

Proof. This is a slight modification of [CLO21, Fact 5.3]. By definition of S^1_{light}, the $(b+1)$-st smallest entry of σ is at least ε/d. There are thus at most $\log_2(d/\varepsilon)$ buckets containing $[d] \backslash S^1_{\mathsf{light}}$, which concludes the proof of the first part. The second part follows by construction. □

B.2 Tuning the perturbations

The goal of this section will be to tune the perturbations $\{\varepsilon_j\}$ from the lower bound instance in Theorem A.2 in order to show the following:

Lemma B.3. *For $0 < \varepsilon < \widetilde{O}(1/\log\log(d))$, for any mixed state $\sigma \in \mathbb{C}^{d \times d}$, the copy complexity of state certification with respect to σ to error ε is at least $\Omega(1/\epsilon) \vee \widetilde{\Omega}(\|\sigma''\|_{2/5} / (\varepsilon^2 \mathrm{polylog}(d/\varepsilon)))$.*

First we handle a minor corner case. Note that Theorem A.2 can only be applied to the buckets of σ which are of size at least a^*. We now verify that if the Schatten 2/5-norm of σ' is dominated by such buckets, then the $\widetilde{\Omega}(\|\sigma''\|_{2/5} / \varepsilon^2)$ lower bound follows from *classical* lower bounds.

Lemma B.4. *If $\sum_{i \in S_{\mathsf{few}} \backslash (S_{\mathsf{tail}} \cup S_{\mathsf{light}})} \sigma_i^{2/5} \geq \frac{1}{2} \|\sigma''\|^{2/5}_{2/5}$, then state certification with respect to σ using incoherent measurements has copy complexity at least $\Omega(\|\sigma''\|_{2/5} / \varepsilon^2)$.*

For this, we use the following instance-optimal lower bound for classical identity testing:

Theorem B.5 (Theorem 1.1 from [VV17]). *Given a known distribution p and samples from an unknown distribution q, any tester that can distinguish between $q = p$ and $\|p - q\|_1 \geq \varepsilon$ with probability $2/3$ must draw at least $\Omega(1/\varepsilon) \vee \Omega(\left\|p_{-\varepsilon}^{-\max}\right\|_{2/3}/\varepsilon^2)$ samples, where $p_{-\varepsilon}^{-\max}$ denotes the vector given by removing from p the largest element and the smallest elements summing up to at most ε.*

Note that this immediately implies a lower bound for state certification by considering only diagonal mixed states:

Corollary B.6. *State certification with respect to any known mixed state σ to error ε using incoherent measurements requires at least $\Omega(1/\epsilon) \vee \Omega(\left\|\sigma_{-\varepsilon}^{-\max}\right\|_{2/3}/\varepsilon^2)$ samples, where $\sigma_{-\varepsilon}^{-\max}$ denotes the matrix given by projecting out from σ the largest eigenvalue and the smallest eigenvalues summing up to at most ε.*

Proof of Lemma B.4. This is a slight modification of [CLO21, Lemma 5.12]. The idea is that if the hypothesis of the lemma holds, then the spectrum of σ is essentially dominated by eigenvalues in geometric progression, in which case there is no distinction between the 2/5- and 2/3-quasinorms and we can simply apply Corollary B.6.

Formally, Corollary B.6 implies a lower bound of $\Omega(\|\sigma_{\varepsilon}^{-\max}\|_{2/3}/\varepsilon^2)$. We would like to relate $\|\sigma_{\varepsilon}^{-\max}\|_{2/3}$ to

$$\left(\sum_{i \in S_{\mathsf{few}} \setminus (S_{\mathsf{tail}} \cup S_{\mathsf{light}})} \sigma_i^{2/3}\right)^{3/2} \geq a^{*-5/2} \cdot (1 - 2^{-2/5})^{5/2} \cdot \left(\sum_{i \in S_{\mathsf{few}} \setminus S_{\mathsf{tail}}} \sigma_i^{2/5}\right)^{5/2} \geq \Omega(\left\|\sigma''\right\|_{2/5}), \tag{49}$$

where the last step follows by the hypothesis of the lemma and Fact B.2.

Suppose that there is some i for which $d_{j(i)} \leq a^*$ and i is not among the indices removed in the definition of $\sigma_{-\varepsilon}^{-\max}$. Then we can lower bound $\left\|\sigma_{-\varepsilon}^{-\max}\right\|_{2/3}$ by σ_i, which is at least $a^{*-3/2}(1 - 2^{-2/3})^{3/2} = \Omega(1)$ times the left-hand side of (49).

On the other hand, suppose that all i for which $d_{j(i)} \leq a^*$ are removed in the definition of $\sigma_{-\varepsilon}^{-\max}$. As long as $\sigma_{-\varepsilon}^{-\max}$ has some nonzero entry, call it σ_{i^*}, then $\sigma_{i^*} \geq \max_{i \in S_{\mathsf{few}} \setminus (S_{\mathsf{tail}} \cup S_{\mathsf{light}})} \sigma_i$, so we can similarly guarantee that $\left\|\sigma_{-\varepsilon}^{-\max}\right\|_{2/3} \geq \sigma_{i^*}$ is at least $a^{*-3/2}(1 - 2^{-2/3})^{3/2} = \Omega(1)$ times the left-hand side of (49). Otherwise, we note that σ'' is zero as well, in which case we are also done. ∎

It remains to consider the primary case where the hypothesis of Lemma B.4 does not hold, which we can express as

$$\sum_{i \in S_{\mathsf{many}} \setminus (S_{\mathsf{light}} \cup S_{\mathsf{tail}})} \sigma_i^{2/5} > \frac{1}{2} \left\|\sigma''\right\|_{2/5}^{2/5}, \tag{50}$$

and this is the case where we will use Theorem A.2. Because Corollary B.6 already shows that the copy complexity is at least $\Omega(1/\epsilon)$, we will assume henceforth that the lower bound in Theorem A.2 is at least $\Omega(1/\epsilon)$.

First for every $i \in S_{\mathsf{many}} \setminus S_{\mathsf{light}}$, define the perturbations

$$\varepsilon_{j(i)} \triangleq d \cdot \left\{2^{-j(i)-1} \Big/ \left(12 + \Theta\left(\sqrt{\log(m)/d_{j(i)}}\right)\right)\right\} \wedge \left\{\zeta 2^{-2/3(j(i)+1)} d_{j(i)}^{2/3}\right\} \tag{51}$$

for normalizing quantity ζ satisfying

$$\sum_{j\in\mathcal{J}^*} d_j \cdot \left\{2^{-j-1} \wedge \zeta 2^{-2/3(j+1)} d_j^{2/3}\right\} = \varepsilon. \tag{52}$$

Note that this choice ζ ensures that the trace distance between the two states under H_0 and H_1 in Theorem A.2 is $\Omega(\varepsilon)$.

The rest of the proof is devoted to analyzing what Theorem A.2 gives for this choice of $\{\varepsilon_j\}$. The main step is to upper bound the normalizing quantity ζ.

Lemma B.7. $\zeta \le O(\varepsilon) \cdot \left(\sum_{j\in\mathcal{J}^*} 2^{-2j/3} d_j^{5/3}\right)^{-1}$.

To prove this, we will need the following elementary fact.

Fact B.8. *Let $u_1 \le \cdots \le u_m$ be numbers for which there are at most ℓ elements in any interval $[2^{-j-1}, 2^{-j}]$. Let $v_1 \le \cdots v_n$ and let $d_1, \ldots, d_n > 1$ be arbitrary integers. Let $w_1 \le \cdots \le w_{m+n}$ be these numbers $u_1, \ldots, u_m, v_1, \ldots, v_n$ in sorted order. For $i \in [m+n]$, define d_1^* to be 1 if w_i corresponds to some u_j, and d_j if w_i corresponds to some v_j.*

There is an absolute constant C_ℓ depending on ℓ such that the following holds. Let s be the largest index for which $\sum_{i=1}^s w_i d_i^* \le C_\ell \varepsilon$. Let a, b be the largest indices for which u_a, v_b are present among $w_1, \ldots, w_s$ (if none exists, take it to be 0). Then either $b = n$ or $\sum_{i=1}^{b+1} v_i d_i > 2\varepsilon$.

Proof. This is Fact 5.16 from [CLO21] with minor modifications. We may assume $s < m+n$ (otherwise obviously $b = n$). Assume to the contrary that $\sum_{i=1}^{b+1} v_i d_i \le \varepsilon$. We proceed by casework based on whether $w_{s'+1} = u_{a+1}$ or $w_{s'+1} = v_{b+1}$.

If $w_{s'+1} = u_{a+1}$, then

$$C_\ell \varepsilon < \sum_{i=1}^{s+1} w_i d_i^* = \sum_{i=1}^{a+1} u_i + \sum_{i=1}^{b} v_i d_i \le \sum_{i=1}^{a+1} v_{b+1} \cdot 2^{\lceil (1-i)/\ell \rceil} + \sum_{i=1}^{b} v_i d_i \le O_\ell(1)\varepsilon + \sum_{i=1}^{b} v_i d_i,$$

where in the first step we used maximality of s, in the third step we used that $u_{a+1} \le v_{b+1}$ and the assumption on $\{u_i\}$, and in the last step we used that $v_{b+1} \le \sum_{i=1}^{b+1} v_i d_i \le \varepsilon$. From this, if C_ℓ is sufficiently large, then we conclude that $\sum_{i=1}^{b} v_i d_i > 2\varepsilon$, a contradiction. The argument for $w_{s'+1} = v_{b+1}$ is analogous. □

Corollary B.9. *If (50) holds, then $S_{\mathsf{many}} \backslash (S_{\mathsf{light}} \cup S_{\mathsf{tail}})$ is nonempty, and there exists an absolute constant $c > 0$ such that for any $i \in S_{\mathsf{many}} \backslash (S_{\mathsf{light}} \cup S_{\mathsf{tail}})$ in some bucket j, $\zeta \cdot 2^{-2/3(j+1)} d_j^{2/3} \le 2^{-j-1}$.*

Proof. The first part immediately follows from (50). For the second part, take some constant $c \ge 1$ to be optimized later and suppose to the contrary that for some $i^* \in S_{\mathsf{many}} \backslash (S_{\mathsf{light}} \cup S_{\mathsf{tail}})$, lying in some bucket j^*, we have $2^{-j^*-1} < \zeta \cdot 2^{-2/3(j^*+1)} d_{j^*}^{2/3}$, or equivalently $2^{-j^*-1}/d_{j^*}^2 < \zeta^3$. Because in the definition of S_{tail}, we sorted by $\sigma_i/d_{j(i)}^2$, we then also have that $2^{-j(i)-1}/d_{j(i)}^2 < \zeta^3$ for all $i \in S_{\mathsf{tail}}$, or equivalently, $2^{-j(i)-1} < \zeta \cdot 2^{-2/3(j+1)} d_{j(i)}^{2/3}$.

To induce a contradiction, we lower bound the sum on the left-hand side of (52) by the contribution from $j \in \mathcal{J}^*$ for which S_j contains an index i satisfying $i \le i^*$. The above discussion implies that for such j, the corresponding summand in (52) is given by $d_j \cdot 2^{-j(i)-1}$. So the left-hand side of (52) is at least

$$\sum_{i \in S_{\mathsf{many}} \backslash S_{\mathsf{light}} : i \le i^*} \sigma_i > \varepsilon,$$

where in the latter inequality we used Fact B.8 applied to the numbers $\ell \triangleq a^*$, $\{u_i\} \triangleq \{\sigma_i\}_{i \in S_{\mathsf{few}} \setminus S_{\mathsf{light}}}$, $\{v_i\} \triangleq \{\sigma_i / d^2_{j(i)}\}_{i \in S_{\mathsf{many}} \setminus S_{\mathsf{light}}}$, and $\{d_i\} \triangleq \{d^2_{j(i)}\}_{i \in S_{\mathsf{many}} \setminus S_{\mathsf{light}}}$, in light of our definition for S_{tail}. This contradicts (52). ☐

We are finally ready to upper bound the normalizing constant ζ.

Proof of Lemma B.7. By Corollary B.9 and (52),

$$\varepsilon \geq \Omega(\zeta) \cdot \sum_{j \in \mathcal{J}^*} d_j \cdot 2^{-2/3(j+1)} d_j^{2/3} \geq \Omega(\zeta) \sum_{j \in \mathcal{J}^*} 2^{-2j/3} d_j^{5/3}.$$

The claimed bound follows. ☐

We are now ready to complete the proof of Lemma B.3.

Proof of Lemma B.3. As discussed above, because of Lemma B.4 it suffices to consider the case where (50) holds. We will apply Theorem A.2 to the principal submatrix of σ indexed by the indices from buckets in $\mathcal{J}^*$. It suffices to show that the copy complexity in that theorem, when specialized to ε_j from (51), is at least $\widetilde{\Omega}(\|\sigma''\|_{2/5} / (\varepsilon^2 \mathrm{polylog}(d/\varepsilon)))$. Note that we can apply Theorem A.2 to this submatrix because by our definition of S_{light}, the second part of (36) holds, by the first argument of each minimum in (51), the first part of (36) holds, and by the definition of S_{many}, d_j is sufficiently large for every j that appears in this submatrix. Note that our definition of m in Definition B.1 is the same as the parameter m in Theorem A.2. Recall from Fact B.2 that $m \leq O(\log(d/\varepsilon))$.

$$\varepsilon_j \triangleq d \cdot \left\{ 2^{-j-1} \Big/ \left(12 + \Theta\left(\sqrt{\log(m)/d_j} \right) \right) \right\} \wedge \left\{ \zeta 2^{-2/3(j+1)} d_j^{2/3} \right\}$$

First, let us rewrite the lower bound from that theorem as

$$\frac{1}{m} \min_{j \in \mathcal{J}^*} \frac{d_j^{1/2} d^2}{\varepsilon_j^2 2^j} \geq \frac{1}{m} \left(\sum_{j \in \mathcal{J}^*} \frac{\varepsilon_j^4 2^{2j}}{d_j d^4} \right)^{-1/2}$$

Substituting our choice of $\{\varepsilon_j\}$ from (51) and denoting $\alpha_j \triangleq 12 + \Theta(\sqrt{\log(m)/d_j})$, we get

$$\begin{aligned}
&\gtrsim \frac{1}{m} \left(\sum_{j \in \mathcal{J}^*} \frac{2^{-2j}}{\alpha_j^4 d_j} \wedge \zeta^4 2^{-2j/3} d_j^{5/3} \right)^{-1/2} \\
&\geq \frac{1}{m} \left(\sum_{j \in \mathcal{J}^*} \alpha_j^{-1} \zeta^3 2^{-j} d_j \wedge \zeta^4 2^{-2j/3} d_j^{5/3} \right)^{-1/2} \\
&\gtrsim \frac{\zeta^{-3/2}}{m} \left(\sum_{j \in \mathcal{J}^*} d_j \{ \alpha_j^{-1} 2^{-j-1} \wedge \zeta 2^{-2/3(j+1)} d_j^{2/3} \} \right)^{-1/2} \\
&\gtrsim \zeta^{-3/2} \cdot \varepsilon^{-1/2} / m \gtrsim (\varepsilon^{-2}/m) \cdot \left(\sum_{j \in \mathcal{J}^*} 2^{-2j/3} d_j^{5/3} \right)^{3/2} \\
&\geq \max_{j \in \mathcal{J}^*, i \in S_j} \sigma_i d_j^{5/2} / (\varepsilon^2 \log(d/\varepsilon)) \gtrsim \|\sigma''\|_{2/5} / (\varepsilon^2 \mathrm{polylog}(d/\varepsilon)),
\end{aligned}$$

where in the second step we used that the minimum of two nonnegative numbers increases if we replace one of them by a weighted geometric mean of the two, in the fourth step we used (52), in the fifth step we used Lemma B.7, in the penultimate step we used Fact B.2, and in the last step we used (50). □

B.3 Putting everything together

Proof of Theorem 8.1. The proof will be given by modifying a few places in the proof in Section 8. We proceed by the same casework of whether or not $d_j = 1$ for all $j \in \mathcal{J}^*$ (note that our definition of $\mathcal{J}^*$ is slightly different from the one used in Section 8).

First by Fact B.2 we have that $\mathrm{Tr}(\sigma') \geq 1 - O(\varepsilon) \geq \Omega(1)$, so by Fact 8.3 it suffices to lower bound the copy complexity by

$$\Omega\left(d_{\mathsf{eff}} \left\|\sigma'\right\|_{1/2} / (\varepsilon^2 \log^{\Theta(1)}(d/\varepsilon))\right).$$

Case 1. $d_j = 1$ for all $j \in \mathcal{J}^*$. Note that in this case,

$$\left\|\sigma'\right\|_{1/2}^{1/2} = \sum_{j \in \mathcal{J}^*} 2^{-j/2} = O(1)$$

and $\|\sigma^*\|_{1/2} = \Theta(\|\sigma'\|_{1/2})$. As $d_{\mathsf{eff}} = 1$, it thus suffices to show a lower bound of $\Omega(1/\varepsilon^2)$ in this case.

If additionally we have $|\mathcal{J}^*| = 1$, then for ε at most a sufficiently small constant, the maximum entry of σ is at least $3/4$, so we can apply Lemma 8.7 to obtain a lower bound of $\Omega(1/\varepsilon^2)$ as desired.

Otherwise, let j, j' be the two smallest bucket indices in $\mathcal{J}^*$, and let i, i' be the elements of the singleton sets $S_j, S_{j'}$. If $\varepsilon \leq c2^{-j/2-j'/2-1}$ for sufficiently small absolute constant $c > 0$, we can invoke [CLO21, Lemma A.4] to conclude a lower bound of $\Omega(1/\varepsilon^2)$.

Otherwise, suppose $\varepsilon > c2^{-j/2-j'/2-1}$. Because $2^{-j} > 2^{-j'}$, we know that $2^{-j} \leq O(\varepsilon)$. In particular, consider the state σ^{**} given by zeroing out $\sigma_{i'}$ from σ' and normalizing. For this matrix, $d_{\mathsf{eff}} = 1$ and $\|\sigma^{**}\|_{1/2} = O(1)$. Furthermore, because ε is smaller than some absolute constant, we conclude that the nonzero entry of σ^{**} is at least $3/4$, so we can again apply Lemma 8.7 to conclude a lower bound $\Omega(1/\varepsilon^2)$.

Case 2. $d_j > 1$ for some $j \in \mathcal{J}^*$. In this case, let $j_1 \triangleq \arg\max_{j \in \mathcal{J}^*} d_j$ and $j_2 \triangleq \arg\max_{j \in \mathcal{J}^*} d_j^2 2^{-j}$. Note that $d2^{-j_1} \geq d_{j_1} 2^{-j_1} \gtrsim \epsilon / \log(d/\epsilon)$, so

$$j_1 \leq O(\log(d/\epsilon)). \tag{53}$$

If $\varepsilon \leq c d_{j_2} 2^{-j_1/2-j_2/2-1}/j_1$ for sufficiently small constant $c > 0$, then we can invoke the lower bound instance from Section 7. The proof in this case is identical to the corresponding part of the proof in Section 8.

It remains to consider the case of

$$\varepsilon > c d_{j_2} 2^{-j_1/2-j_2/2-1}/j_1. \tag{54}$$

We would like to use the lower bound from Lemma B.3. We would first like to relate $\|\sigma'\|_{2/5}$ to $\|\sigma''\|_{2/5}$ (recall that the difference is that σ'' is defined by removing the largest entry from σ'.

Lemma B.10. *Either $\|\sigma''\|_{2/5} \geq \Omega(\|\sigma'\|_{2/5})$, or the following holds. Let j° be the index maximizing $d_j^{5/2} 2^{-j}$. Then 1) $j^\circ = \min_{j \in \mathcal{J}^*} j$, 2) $d_{j^\circ} = 1$, and 3) $j^\circ = 0$.*

Proof. This is essentially Lemma 5.26 from [CLO21]. We will assume that $\|\sigma''\|_{2/5} = o(\|\sigma'\|_{2/5})$ and show that 1)-3) must hold. Let $i_{\max}$ be the index of the top entry of σ'. Suppose 1) does not hold. Then

$$\frac{\|\sigma'\|_{2/5}^{2/5}}{\|\sigma''\|_{2/5}^{2/5}} \leq \frac{\sigma_{i_{\max}}^{2/5} + \sum_{i \in S_{j^\circ}} \sigma_i^{2/5}}{\sum_{i \in S_{j^\circ}} \sigma_i^{2/5}} \leq 2,$$

where the first inequality follows by the elementary fact that for $a \geq b \geq 0$ and $c \geq 0$, $\frac{a+c}{b+c} \leq \frac{a}{b}$, and the second inequality follows by the definition of j°. This contradicts the assumption that $\|\sigma''\|_{2/5} = o(\|\sigma'\|_{2/5})$.

Next, suppose 1) holds but 2) does not. Then

$$\frac{\|\sigma'\|_{2/5}^{2/5}}{\|\sigma''\|_{2/5}^{2/5}} \leq \frac{\sum_{i \in S_{j^\circ}} \sigma_i^{2/5}}{\sum_{i \in S_{j^\circ} \setminus \{i_{\max}\}} \sigma_i^{2/5}} \leq O(1),$$

where in the first step we again used the above elementary fact and in the second step we used that 2) does not hold. We again get a contradiction.

Finally, suppose 1) and 2) hold, but 3) does not. Because 1) holds and $j^\circ > 0$, this implies that $\|\sigma'\|_{\mathrm{op}} \leq 1/2$. On the other hand, $\|\sigma''\|_{2/5} \geq \|\sigma''\|_1 \geq (1 - O(\varepsilon)) - 1/2 \geq 1/2 - O(\varepsilon)$. So for ε smaller than a sufficiently large constant, we get that $\|\sigma''\|_{2/5} \geq \Omega(\|\sigma'\|_{\mathrm{op}})$, so $\|\sigma''\|_{2/5} \geq \Omega(\|\sigma'\|_{2/5})$, a contradiction. □

Suppose the latter scenario in Lemma B.10 happens, but the former does not. In this case, because $d_{j^\circ} = 1$, we also have that $j^\circ = j_2$, so $1 \geq d_{j^\circ} 2^{-j^\circ} = d_{j^\circ}^2 2^{-j^\circ} = d_{j_c}^2 2^{-j_2}$. Note that this implies that $\|\sigma'\|_{1/2} \leq \log(d/\varepsilon)$. Furthermore, it implies that

$$1 \geq d_{j_2}^2 2^{-j_2} \geq d_{j_1}^2 2^{-j_1} \geq \Omega(d_{j_1}^{3/2} \varepsilon / \log(d/\varepsilon)),$$

where in the second step we used that $j_2 \arg\max_{j \in \mathcal{J}^*} d_j^2 2^{-j}$, and in the last step we used that $d_j 2^{-j} \geq \Omega(\varepsilon / \log(d/\varepsilon))$ by definition of S_{light} and $\mathcal{J}^*$. We conclude that

$$\varepsilon \leq O(d_{j_1}^{-3/2} \log(d/\varepsilon)). \tag{55}$$

But recall that we are assuming that (54) holds, i.e.

$$\varepsilon \gtrsim d_{j_2} \cdot 2^{-j_1/2 - j_2/2} / j_1 = 2^{-j_1/2 - j_2/2} / j_1 \geq \Omega(\varepsilon / (d_{j_1} \log(d/\varepsilon)))^{1/2} / j_1, \tag{56}$$

where the second step is by $d_{j_2} = d_{j^\circ} = 1$ and the last step is by 3) in Lemma B.10 and the fact that $d_j 2^{-j} \geq \Omega(\varepsilon / \log(d/\varepsilon))$ for all $j \in \mathcal{J}^*$. Combining (55) and (56), we conclude that $d_{j_1} \leq \mathrm{polylog}(d/\varepsilon) \cdot j_1 \leq \mathrm{polylog}(d/\varepsilon)$, where in the last step we used (53). But if $d_{j_1} \leq \mathrm{polylog}(d/\varepsilon)$, then $d_{\mathsf{eff}} \leq \mathrm{polylog}(d/\varepsilon)$. Then because we also have $\|\sigma'\|_{1/2} \leq O(\log(d/\varepsilon))$, the claimed lower bound in the theorem would follow from a lower bound of $\Omega(1/\varepsilon^2)$. This then follows in a similar fashion to the analysis from Case 1 above.

Finally, suppose instead that the former scenario in Lemma B.10 happens, in which case Lemma B.3 gives a lower bound of $\Omega(\|\sigma'\|_{2/5} / (\varepsilon^2 \log(d/\varepsilon)))$. Let j° be as defined in Lemma B.10. As $\|\sigma'\|_{2/5} \geq d_{j^\circ}^{5/2} 2^{-j^\circ}$, to complete the proof, it suffices to show that

$$d_{j^\circ}^{5/2} 2^{-j^\circ} \mathrm{polylog}(d/\epsilon) \geq \Omega\left(\sqrt{d_{j_1}} d_{j_2}^2 2^{-j_2}\right). \tag{57}$$

Suppose to the contrary. Then because $d_{j_1}^{5/2}2^{-j_1} \le d_{j^\circ}^{5/2}2^{-j^\circ}$, we would get from the negation of (57) that

$$d_{j_1}^2 2^{-j_1}\mathrm{polylog}(d/\epsilon) = o(d_{j_2}^2 2^{-j_2}). \tag{58}$$

But by (54) and (53),

$$cd_{j_2}2^{-j_1/2-j_2/2-1}/O(\log(d/\epsilon)) \le \varepsilon \le O(d_{j_1}2^{-j_1}\log(d/\varepsilon)),$$

where in the last step we used that $d_{j_1}2^{-j_1} \ge \Omega(\varepsilon \log(d/\varepsilon))$ by definition of S_{light}. Squaring and rearranging, we find that $d_{j_2}^2 2^{-j_2} \le O(d_{j_1}^2 2^{-j_1}\log^2(d/\epsilon))$, contradicting (58). ☐

C Regularity Bounds on GOE and Ginibre Ensembles

Here we provide the proofs of Lemmas 6.2 and 7.1, restated for convenience.

Lemma 6.2. *There exists $U \subseteq \mathbb{R}^{d\times d}$ such that if $M \sim \mathrm{GOE}^*(d)$, then $\Pr[M \notin U] \le \exp(-\Omega(d))$ and on the event $M \in U$, we have $\|M\|_{\mathrm{op}} \le 3$ and $\|M\|_1 \ge d/12$.*

Proof of Lemma 6.2. Let U denote the event that $\|M\|_{\mathrm{op}} \le 3$ and $\|M\|_F^2 \ge d/4$. Let $\lambda_1, \dots, \lambda_d$ denote the eigenvalues of U. On the event U, we have

$$\|M\|_F^2 = \sum_{i=1}^d \lambda_d^2 \le \left(\max_{1\le i\le d}|\lambda_i|\right)\sum_{i=1}^d |\lambda_i| = \|M\|_{\mathrm{op}}\|M\|_1,$$

so $\|M\|_1 \ge \|M\|_F^2 / \|M\|_{\mathrm{op}} \ge d/12$. We will show $\Pr[U^c] \le \exp(-\Omega(d))$. We generate $M = G - \frac{\mathrm{Tr}(G)}{d}I_d$, where $G \sim \mathrm{GOE}(d)$. Note that

$$\begin{aligned}\Pr[\|M\|_{\mathrm{op}} > 3] &\le \Pr[\|G\|_{\mathrm{op}} > 5/2] + \Pr[|\mathrm{Tr}(G)| > d/2] \\ &\le \exp(-\Omega(d)) + \exp(-\Omega(d^2)),\end{aligned}$$

where the first term is bounded by [BADG01, Theorem 6.2] (because $5/2 > 2$) and the second term is bounded by $\mathrm{Tr}(G) \sim \mathcal{N}(0,2)$. Moreover, since

$$\|M\|_F^2 = \|G\|_F^2 - \mathrm{Tr}(G)^2/d,$$

we have

$$\Pr[\|M\|_F^2 < d/4] \le \Pr[\|G\|_F^2 < d/2] + \Pr[|\mathrm{Tr}(G)| > d/2]$$

and the second probability is $\exp(-\Omega(d^2))$ as explained above. To bound the first probability, write $G_{i,i} = \sqrt{\frac{2}{d}}Z_{i,i}$ and for $i<j$, $G_{i,j} = G_{j,i} = \frac{1}{\sqrt{d}}Z_{i,j}$ for i.i.d. $Z_{i,i}, Z_{i,j} \sim \mathcal{N}(0,1)$. Then

$$\|G\|_F^2 = \frac{2}{d}\sum_{1\le i\le j\le d} Z_{i,j}^2.$$

By a standard Chernoff bound, if $X \sim \chi^2(n)$, then $\Pr[X \le (1-\varepsilon)n] \le ((1-\varepsilon)e^{\varepsilon})^{n/2}$. Thus

$$\Pr\left[\|G\|_F^2 \le d/4\right] = \Pr\left[\chi^2(d(d+1)/2) \le d^2/8\right] \le \exp(-\Omega(d^2)).$$ ☐

Lemma 7.1. *For $d_1 \geq d_2$, there exists $U \subseteq \mathbb{R}^{d_1 \times d_2}$ such that if $G \sim \mathrm{Gin}(d_1, d_2)$, then $\Pr[G \notin U] \leq \exp(-0.1d_1)$ and on the event $G \in U$, we have $\|G\|_{\mathrm{op}} \leq 3$ and $\|M\|_1 \geq d_2/3$ for*

$$M = \begin{pmatrix} 0 & G \\ G^\dagger & 0 \end{pmatrix}.$$

Proof of Lemma 7.1. Let U be the event $s_{\max}(G) \leq 3$ and $\|G\|_F^2 \geq d_2/2$, where $s_{\max}$ denotes the largest singular value. On this event, certainly $\|M\|_{\mathrm{op}} \leq 3$ and $\|M\|_F^2 = 2\|G\|_F^2 \geq d_2$. Similarly to the proof of Lemma 6.2, we have $\|M\|_1 \geq \|M\|_F^2 / \|M\|_{\mathrm{op}} \geq d_2/3$. It remains to show $\Pr[U^c] \leq \exp(-0.1d_1)$ By [Ver12, Corollary 5.35], $\Pr[s_{\max}(G) > 3] \leq \exp(-0.11d_1)$. Moreover, $\|G\|_F^2 =_d \frac{1}{d_1}\chi^2(d_1 d_2)$, so similarly to the proof of Lemma 6.2 we have $\Pr[\|G\|_F^2 < d_2/2] \leq \exp(-\Omega(d_1^2))$. □

D Separating K and κ

In this short section we construct an example of a transcript $(\boldsymbol{z}, \boldsymbol{w})$ for which $K((\boldsymbol{z}, \boldsymbol{w})) = 0$, but for which $\kappa((\boldsymbol{z}, \boldsymbol{w})) \gg d_1 d_2^2/\varepsilon^2$. For simplicity, consider $A = a \cdot \mathbb{1}_{d_1}$ and $B = b \cdot \mathbb{1}_{d_2}$.

Consider a unit vector $(z, w) \in \mathbb{S}^{d_1+d_2-1}$ for which $\|z\|^2 = b/(a+b)$ and $\|w\|^2 = a/(a+b)$. Now note that if $(\boldsymbol{z}, \boldsymbol{w}) = ((z, w), (z, -w), (z, w), (z, -w), \ldots)$, then clearly $K((\boldsymbol{z}, \boldsymbol{w})) = 0$. On the other hand, if we take $b_i = (-1)^{i+1}$, we find that

$$\kappa((\boldsymbol{z}, \boldsymbol{w})) \geq \left\| \sum_{i=1}^{t} \frac{z w^\dagger \|z\| \|w\|}{(a\|z\|^2 + b\|w\|^2)^2} \right\|_F = t \cdot \frac{\|z\|^2 \|w\|^2}{(a\|z\|^2 + b\|w\|^2)^2} = \frac{t}{4ab}.$$

Note that for $\varepsilon \asymp d_2\sqrt{ab}$, $d_1 d_2^2/\varepsilon^2 \asymp d_1/(ab)$, so for $t \gg d_1$, $\kappa((\boldsymbol{z}, \boldsymbol{w})) \gg d_1 d_2^2/\varepsilon^2$.