Quantum state certification

Costin Bădescu* Ryan O'Donnell* John Wright†

October 14, 2017

Abstract

We consider the problem of *quantum state certification*, where one is given n copies of an unknown d-dimensional quantum mixed state ρ, and one wants to test whether ρ is equal to some known mixed state σ or else is ϵ-far from σ. The goal is to use notably fewer copies than the $\Omega(d^2)$ needed for full tomography on ρ (i.e., density estimation). We give two robust state certification algorithms: one with respect to fidelity using $n = O(d/\epsilon)$ copies, and one with respect to trace distance using $n = O(d/\epsilon^2)$ copies. The latter algorithm also applies when σ is unknown as well. These copy complexities are optimal up to constant factors.

1 Introduction

A key step in building quantum devices is verifying that they work as intended. Typically, a quantum device is designed with the intent of outputting some known d-dimensional (mixed) state $\sigma \in \mathbb{C}^{d\times d}$, but the possibility of imperfections in the device's construction and noise in the device's operation mean that its actual output state $\rho \in \mathbb{C}^{d\times d}$ is unknown. *Quantum state certification* refers to the problem of testing whether ρ equals σ or is far from σ, given the ability to produce $\rho^{\otimes n}$ (i.e., n copies of ρ). This is the quantum (noncommutative) generalization of the classical statistical problem of testing identity of probability distributions [Can15].

A standard approach for quantum state certification is to first estimate ρ from $\rho^{\otimes n}$ using a *quantum state tomography (estimation)* procedure, then to check that the estimate is close to σ. Given that ρ has d^2-1 real parameters, it is natural that the number of copies needed to estimate it should scale roughly as d^2. This was confirmed in a trio of recent papers [HHJ+16, OW16, OW17]; among other things, those works show that $n = \widetilde{\Theta}(d^2/\epsilon)$ copies of ρ are necessary and sufficient to produce an estimate $\widehat{\rho}$ satisfying the fidelity bound $\mathrm{F}(\rho, \widehat{\rho}) \geq 1 - \epsilon$. (See Section 2.1 for more on prior work, and Section 3.1 for a review of distance measures such as fidelity, trace distance, χ^2-divergence, etc.)

Unfortunately, even small scale quantum systems can have large dimension; for example, a system of q qubits has $d = 2^q$ dimensions. For such systems, the quadratic scaling in d required by full tomography (density estimation) can be prohibitively expensive. For example, a 2005 experiment [HHR+05] designed to produce the entangled 8-particle W-state ($d = 256$) used $n = 656100$ copies to estimate the actually-produced state. (The fidelity to the target state ended up being estimated as .85.)

*Computer Science Department, Carnegie Mellon University. Supported by NSF grant CCF-1618679. {cbadescu,odonnell}@cs.cmu.edu

†Center for Theoretical Physics, Massachusetts Institute of Technology. Supported by NSF grant CCF-6931885. jswright@mit.edu

However for the quantum state certification problem, the goal is not to learn the unknown state $\rho \in \mathbb{C}^{d\times d}$ but merely to test whether it is close to a target σ, or far from it. Learning the entire density matrix might be wasting copies of ρ to gain irrelevant information. As such, it is natural to ask: can we outperform tomography?

1.1 Our results

In this work, we give a unified framework for analyzing the number of copies of ρ needed to estimate polynomial functions of ρ and hence perform various quantum state certification tasks. One of our main results is the following:

Theorem 1.1. *Let $\sigma \in \mathbb{C}^{d\times d}$ be a fixed mixed state, and let $\epsilon > 0$. There is an algorithm that, given $n = O(d/\epsilon)$ copies of ρ, performs a measurement and then reports either "close" or "far". The algorithm has the following guarantee (with high probability[1]): If it reports "close" then we have the fidelity bound $\mathrm{F}(\rho, \sigma) \geq 1 - \epsilon$. If it reports "far" then we have the Bures χ^2-divergence[2] bound $\mathrm{D}_{\chi^2}(\rho \,\|\, \sigma) > .49\epsilon$.*

To put it another way, if $\mathrm{D}_{\chi^2}(\rho \,\|\, \sigma) \leq .49\epsilon$ (in particular, if $\rho = \sigma$) then the algorithm reports "close" and if $\mathrm{F}(\rho, \sigma) < 1 - \epsilon$ then the algorithm reports "far" (whp). We remark that the notions of "close" and "far" in Theorem 1.1 are nearly complementary, since it's known that every pair of states ρ, σ satisfies either $\mathrm{F}(\rho, \sigma) \geq 1 - \epsilon$ or $\mathrm{D}_{\chi^2}(\rho \,\|\, \sigma) > .5\epsilon$.

Theorem 1.1 is stronger than the usual kind of state certification result in that it is *robust*, meaning that the test "accepts" not just if $\rho = \sigma$ but also if ρ is sufficiently close to σ. The simplified (weaker) version would be:

Corollary 1.2. *For a fixed mixed state $\sigma \in \mathbb{C}^{d\times d}$ and $\epsilon > 0$, there is an algorithm that, given $n = O(d/\epsilon)$ copies of ρ, distinguishes (whp) between the cases $\rho = \sigma$ and $\mathrm{F}(\rho, \sigma) < 1 - \epsilon$.*

The stronger version Theorem 1.1 is actually an easy consequence (see Section 6.3) of the following certification procedure for "well-conditioned" states, robust with respect to Bures χ^2-divergence:

Theorem 1.3. *Let $c > 0$ be any small constant. Fix a d-dimensional mixed state σ with smallest eigenvalue at least $c\epsilon^2/d$. Then there is an algorithm that, given $n = O(d/\epsilon^2)$ copies of ρ, (whp) outputs "close" if $\mathrm{D}_{\chi^2}(\rho \,\|\, \sigma) \leq .99\epsilon^2$ and outputs "far" if $\mathrm{D}_{\chi^2}(\rho \,\|\, \sigma) > \epsilon^2$.*

We also obtain a new sample-efficient certification algorithm in the case of *two unknown* states. Here one is given n copies each of mixed states ρ, σ and one wants to distinguish whether $\rho = \sigma$ or ρ is far from σ. Our algorithm here is robust with respect to the Hilbert–Schmidt distance:

Theorem 1.4. *There is an algorithm that, given $n = O(1/\epsilon^2)$ copies each of unknown mixed states $\rho, \sigma \in \mathbb{C}^{d\times d}$, (whp) outputs "close" if $\mathrm{D}_{\mathrm{HS}}(\rho, \sigma) \leq .99\epsilon$ and outputs "far" if $\mathrm{D}_{\mathrm{HS}}(\rho, \sigma) > \epsilon$.*

Of course this result may also be used in the simpler case when σ is a known state (as then the algorithm can simply prepare n copies of σ by itself). We also remark that the sample complexity n has no dependence on d.

Although the Hilbert–Schmidt distance is arguably not too meaningful, operationally, one can use Cauchy–Schwarz to relate it to the very natural trace distance. In this way, Theorem 1.4 immediately yields the following:

[1] Henceforth abbreviated "whp". We may take this to mean probability at least, say, 2/3; however, by standard means this probability can be boosted to $1 - \delta$ at the expense of multiplying n by $O(\log(1/\delta))$.

[2] The Bures χ^2-divergence is reviewed in Section 3.1.

Corollary 1.5. *There is an algorithm that, given $n = O(d/\epsilon^2)$ copies each of unknown mixed states $\rho, \sigma \in \mathbb{C}^{d \times d}$, (whp) distinguishes between the cases $\rho = \sigma$ and $\mathrm{D}_{\mathrm{tr}}(\rho, \sigma) > \epsilon$.*

We stated the above corollary for simplicity, but with slightly more care (see Section 5.4) one also derive from Theorem 1.4 the following much more precise result for trace-distance certification, which has improved sample complexity when one of the states is close to having low rank:

Corollary 1.6. *Assume that one of the two unknown states — say, σ — is close to having rank at most k, in the sense that the sum of its largest k eigenvalues is at least $1 - \delta$. Then there is an algorithm that, given $n = O(k/\epsilon^2)$ copies each of $\rho, \sigma \in \mathbb{C}^{d \times d}$, (whp) distinguishes between the cases $\mathrm{D}_{\mathrm{HS}}(\rho, \sigma) \leq .58\epsilon$ and $\mathrm{D}_{\mathrm{tr}}(\rho, \sigma) > \delta + \epsilon$. (The constant .58 can be anything smaller than $2 - \sqrt{2}$.)*

We note that even the simplest versions of our results — Corollary 1.2 and Corollary 1.5 — have optimal sample complexity (up to a constant), even when σ is promised to be the maximally mixed state $\mathbb{1}/d$. This is a consequence of the following lower bound from [OW15]:

Theorem 1.7 ([OW15])**.** *Given even d and $0 \leq \epsilon \leq 1/2$, let $\sigma = \mathbb{1}/d$ and let $\mathcal{C}_\epsilon$ denote the class of states with eigenvalues $\frac{1+2\epsilon}{d}, \frac{1-2\epsilon}{d}, \frac{1+2\epsilon}{d}, \frac{1-2\epsilon}{d}, \ldots, \frac{1+2\epsilon}{d}, \frac{1-2\epsilon}{d}$. For any $\rho \in \mathcal{C}_\epsilon$, one has*

$$\mathrm{D}_{\mathrm{tr}}(\rho, \sigma) = \epsilon, \quad \mathrm{F}(\rho, \sigma) = 1 - \tfrac{1}{2}\epsilon^2 - O(\epsilon^4), \quad \mathrm{D}_{\mathrm{HS}}(\rho, \sigma) = 2\epsilon/\sqrt{d}, \quad \mathrm{D}_{\chi^2}(\rho \parallel \sigma) = 4\epsilon^2 + O(\epsilon^4).$$

Then any measurement strategy that can distinguish (with probability advantage at least $1/3$) the case $\rho = \sigma$ from the case $\rho \in \mathcal{C}_\epsilon$ using n samples from ρ must have $n > .15d/\epsilon^2$.

Finally, our quantum certification algorithm from Theorem 1.4 is not just copy-efficient, it can be carried out by polynomial-sized (i.e., $\mathrm{poly}(n, d)$-gate) quantum circuits.

1.2 Outline of the remainder of the paper

In Section 2 we review prior work on quantum tomography and state discrimination, as well as some relevant prior work on classical learning and testing of probability distributions. In Section 3.1 we recall various measures of probability distribution distance and quantum state distance that will be important in this work. Sections 3.2 and 3.3 are devoted to background on quantum probability and representation theory. In Section 4, we develop a framework for finding the most efficient (lowest-variance) estimators for symmetric polynomial functions of unknown quantum states. These results are not strictly necessary for our proof of Theorem 1.4 in Section 5; however, they justify that the estimators used therein are optimal. Section 6 contains our proof of Theorems 1.1 and 1.3, as well as a diagonality tester for quantum states. Finally, in Section 7 we give efficient implementations for the algorithm in Theorem 1.4.

2 Prior work on classical and quantum density testing/estimation

In this section we review some results on learning and testing unknown quantum states, and the analogous classical problem of learning and testing unknown probability distributions. As these areas are extremely broad, we cannot completely review all known literature; we will simply give pointers to some of the best known and most relevant results.

2.1 Prior quantum density estimation, testing, and certification

2.1.1 Tomography (density estimation)

Before discussing state certification, we start by reviewing the best known results for the baseline problem of *tomography*; i.e., producing an estimate $\widehat{\rho}$ of an unknown density matrix $\rho \in \mathbb{C}^{d\times d}$, given n copies $\rho^{\otimes n}$, up to error ϵ (whp) for some notion of "distance". We will also let k denote the *rank* of ρ, which is 1 when ρ is a pure state, and in general is at most d. The best results achievable depend on the "figure of merit" — i.e., distance measure — chosen (see Section 3.1 for a review).

In [HHJ+16] it was shown that $n = O(kd/\epsilon) \cdot \log(d/\epsilon)$ copies suffice to obtain infidelity ϵ (i.e., $\mathrm{F}(\rho, \widehat{\rho}) \leq 1-\epsilon$); this also implies that $n = O(kd/\epsilon^2) \cdot \log(d/\epsilon)$ copies suffice to obtain trace distance ϵ (i.e., $\mathrm{D}_{\mathrm{tr}}(\rho, \widehat{\rho}) \leq \epsilon$). Those authors also showed that $n = \Omega(kd/\epsilon^2)/\log(d/k\epsilon)$ copies are necessary, with the log factor being removable in the case $k = d$. Independently, in [OW16] it was shown that $n = O(d/\epsilon^2)$ copies suffice to obtain Hilbert–Schmidt distance ϵ (i.e., $\mathrm{D}_{\mathrm{HS}}(\rho, \widehat{\rho}) \leq \epsilon$); this also implies a copy complexity of $n = O(kd/\epsilon^2)$ for trace distance (slightly better than in [HHJ+16]). More generally, [OW16] showed a kind of "PCA" result: for ρ of any rank, $n = O(kd/\epsilon^2)$ copies suffice to produce an estimate $\widehat{\rho}$ whose trace distance from ρ is at most ϵ more than that of the best rank-k approximator. Finally, a followup work [OW17] gave an alternate proof of the $n = O(kd/\epsilon) \cdot \log(d/\epsilon)$ bound for infidelity, showed also an $n = O(k^2 d/\epsilon)$ bound, and extended these bounds to the PCA case.

2.1.2 Density testing

Tomography results suffer from the inherent issue that $n = \widetilde{\Theta}(d^2)$ copies are needed in the general case (except when the figure of merit is Hilbert–Schmidt distance, but this metric is not considered to be very meaningful, operationally). Thus as mentioned, it is natural to focus on restricted problems like state certification, distance estimation, and other *property testing* problems that can potentially be carried out with $n = O(d)$ or better. Montanaro and de Wolf [MW16] have given an excellent survey on property testing of quantum states; we review a few of the known results here.

A typical quantum property testing problem would involve two disjoint classes $\mathcal{C}_1$, $\mathcal{C}_2$ of d-dimensional quantum states; given n copies of an unknown ρ, promised to be in either $\mathcal{C}_1$ or $\mathcal{C}_2$, the task is to distinguish which is the case (whp) using few copies of ρ. In particular, the *quantum state certification* problem for fixed state $\sigma \in \mathbb{C}^{d\times d}$ is the case when $\mathcal{C}_1 = \{\sigma\}$ and $\mathcal{C}_2 = \{\rho : \mathrm{D}(\rho, \sigma) > \epsilon\}$ for some notion $\mathrm{D}(\cdot, \cdot)$ of distance and some parameter ϵ.

When σ is a *pure* state, it is straightforward to show (see, e.g., [MW16]) that the associated quantum state certification task, with infidelity as the distance measure, can be done using $n = O(1/\epsilon)$ copies (and this implies $n = O(1/\epsilon^2)$ copies suffice for trace distance). Indeed, the same is possible when both ρ *and* σ are unknown pure states, and one is given n copies of each. For practical purposes, it may be useful to have a state certification algorithm for a known pure σ that only uses simple measurements; e.g., Pauli observables. For this problem, it has been shown [FL11, dSLP11, AGKE15] that for σ known and pure, one can solve the certification problem given $n = O(d/\epsilon^2)$ copies of an unknown ρ with infidelity as the distance metric — indeed, with this many copies one can estimate the fidelity $\mathrm{F}(\rho, \sigma)$ to $\pm\epsilon$.

For the state certification problem when σ is mixed (not pure), not much is known except in one case: when $\sigma = \mathbb{1}/d$, the "maximally mixed" state. For this problem, it was shown in [OW15] that $n = \Theta(d/\epsilon^2)$ copies are necessary and sufficient, when the distance measure is trace distance. In fact, for the $n = O(d/\epsilon^2)$ upper bound, [OW15] effectively show that one can estimate the *purity* $\mathrm{tr}(\rho^2)$ of ρ sufficiently well so as to distinguish between purity $1/d$ (achieved by the maximally mixed state) and purity exceeding $1/d + \epsilon^2/d$. Note that the latter case is equivalent to ρ being

$\epsilon/\sqrt{d}$-far from $\mathbb{1}/d$ in Hilbert–Schmidt distance and ϵ^2-far from $\mathbb{1}/d$ in Bures χ^2-divergence. The lower bound was mentioned earlier as Theorem 1.7.

2.1.3 The asymptotic regime for state discrimination

There is a related class of work that we refer to as the "asymptotic regime". Consider the simplest quantum property testing problem, *state discrimination*, in which $\mathcal{C}_1 = \{\sigma_1\}$ and $\mathcal{C}_2 = \{\sigma_2\}$ for two known states $\sigma_1, \sigma_2 \in \mathbb{C}^{d \times d}$. The perspective we take in this paper involves determining the least number of copies n such that one can distinguish $\rho = \sigma_1$ from $\rho = \sigma_2$ with high probability — say, with both "type I" and "type II" errors having probability at most $\delta = 1/3$. One can reduce this δ to any small positive constant at the expense of making n a constant factor larger. We refer to this perspective as the *non-asymptotic regime*, because we do not consider any limiting error rate as $n \to \infty$; rather, we wish to find a concrete upper bound on the n that suffices, depending only on d, the distance between σ_1 and σ_2, and nothing else.

On the other hand, there is substantial work on the *asymptotic regime*, sometimes going under the name *quantum hypothesis testing*, in which the focus is on how exponentially fast the error rate goes to 0 in the limit as $n \to \infty$. Here one might seek the best (smallest) constant R such that, given n copies, one can ensure type I and type II errors have probability at most $(R + o(1))^n$, where the $o(1)$ refers to $n \to \infty$. A downside of such results is that they do not a priori give any information about how large n needs to be before error bounds "kick in"; e.g., the $o(1)$ function might not be less than, say, .1 until n is larger than some uncontrolled function of d (e.g., 2^d) or of some other parameters (e.g., the smallest nonzero eigenvalue of σ_1 or σ_2).

A good survey of the results in the asymptotic regime appears in [ANSV08]; they review known quantum versions of Stein's Lemma and Sanov's Theorem, and prove quantum versions of Chernoff's Bound and the the Hoeffding–Blahut–Csiszár–Longo bound. For example, in the basic hypothesis testing problem described above, they prove that the best rate R is given by $Q_{\min}(\sigma_1 \parallel \sigma_2) = \min_{0 \le s \le 1} \operatorname{tr}(\sigma_1^s \sigma_2^{1-s})$ (a quantity that is within a factor of 2 of the infidelity between σ_1 and σ_2).

2.2 Prior classical density estimation, testing, and certification

For every quantum problem discussed so far, we get a "classical" special case by assuming that all d-dimensional density matrices are diagonal. In this way we obtain basic problems in statistics and property testing: estimation, certification, and identity testing for *probability distributions* $p = (p_1, \dots, p_d)$ on $[d]$. Since our results are partly inspired by these classical analogues, we briefly review some known results here.

2.2.1 Density estimation

The analogue of quantum tomography is *density estimation*: producing an estimate $\widehat{p}$ of an unknown probability distribution p on $[d]$, given n independent samples. For this problem, the most natural algorithm is simply to let $\widehat{p}$ be the empirical distribution of the samples. One can very easily directly calculate that

$$\mathbf{E}[\mathrm{d}_{\ell_2}^2(p, \widehat{p})] = \frac{1}{n}\left(1 - \sum_{i=1}^{d} p_i^2\right) \le \frac{1}{n};$$

hence Markov's inequality implies that $n = O(1/\epsilon^2)$ samples suffice to obtain $\mathrm{d}_{\ell_2}^2(p, \widehat{p}) \le \epsilon$ whp. Cauchy–Schwarz then implies that $n = O(d/\epsilon^2)$ samples suffice to obtain $\mathrm{d}_{\mathrm{TV}}(p, \widehat{p}) \le \epsilon$ with high probability. For the stronger χ^2-divergence, one shouldn't let $\widehat{p}$ be the empirical distribution because then $\mathrm{d}_{\chi^2}(p \parallel \widehat{p}) = \infty$ is possible if $p_i > \widehat{p}_i = 0$ for some i. Instead, standard practice is to take $\widehat{p}$ to

be the "add-one" estimator: $\widehat{p}_i = \frac{\boldsymbol{x}_i+1}{n+d}$, where $\boldsymbol{x}_i \sim \mathrm{Bin}(n, p_i)$ is the number of i's in the sample. Again, one can very easily directly calculate (see, e.g., [KOPS15, Lemma 4]):

$$\mathbf{E}[\mathrm{d}_{\chi^2}(p \parallel \widehat{p})] = \frac{d-1}{n+1} - \frac{n+d}{n+1}\left(1 - \sum_{i=1}^{d}(1-p_i)^{n+1}\right) \leq \frac{d-1}{n+1} \leq \frac{d}{n} \tag{1}$$

and hence $n = O(d/\epsilon)$ samples suffice to obtain $\mathrm{d}_{\chi^2}(p \parallel \widehat{p}) \leq \epsilon$ whp. Thus for natural measures of discrimination like total variation, Hellinger, and χ^2-divergence, $n = O(d)$ samples suffice for density estimation (for constant ϵ). Consequently, for the "distribution certification" problem (known in property testing problems as "identity testing"), the goal is to use $o(d)$ samples.

2.2.2 Identity testing

Three of the main such property testing problems, in increasing order of difficulty, are the following:

0. Testing identity of p to the uniform distribution (which we write as $\mathbb{1}/d$ in this section).
1. Testing identity of p to an arbitrary but known distribution q.
2. Testing identity of two unknown distributions p, q.

Uniformity testing. Historically, property testing researchers considered total variation distance to be the main figure of merit. But beginning with the earliest work of Goldreich and Ron [GR00], it was found that approaching the problems via ℓ_2^2-distance was more expedient. For example, Goldreich and Ron originally showed that with $n = O(\sqrt{d}/\epsilon^2)$ samples, one can (whp) estimate $\|p\|_2^2 = \mathrm{d}_{\ell_2}^2(p, \mathbb{1}/d) + 1/d$ to a multiplicative $1 \pm \epsilon$ factor. A consequence of this (and Cauchy–Schwarz) is that $n = O(\sqrt{d}/\epsilon^4)$ samples suffice to distinguish $p = \mathbb{1}/d$ and $\mathrm{d}_{\mathrm{TV}}(p, \mathbb{1}/d) > \epsilon$. Paninski [Pan08] improved the latter result by a different method to $n = O(\sqrt{d}/\epsilon^2)$ (assuming $\epsilon = \Omega(d^{-1/4})$, a restriction later removed in [VV17]), and showed a matching lower bound.

In fact, a better analysis of Goldreich and Ron's original method yields the optimal result: one simply estimates $\mathrm{d}_{\ell_2}^2(p, \mathbb{1}/d)$ by the natural unbiased estimator (the average number of "collisions" among the n samples, minus $1/d$), computes its variance, and then uses Chebyshev inequality. A little case analysis is needed when applying Chebyshev, which is perhaps why this natural method was not employed until the very recent work of [DGPP16] (for a briefer exposition, see [OW17, Sec. 10]). We will use similar methods in the present work, and the needed version of Chebyshev's inequality is packaged up at the end of this section as Lemma 2.1.

Identity testing to a known distribution. Moving on to Problem 1 above, testing identity of p to an arbitrary known distribution q, Batu et al. [BFF+01] showed that $O(\sqrt{d}\log(d)/\epsilon^2)$ samples from p suffice to distinguish the case $\mathrm{d}_{\mathrm{TV}}(p, q) \ll \epsilon^3/\sqrt{d}\log(d)$ (and in particular, $p = q$) from the case $\mathrm{d}_{\mathrm{TV}}(p, q) > \epsilon$. Valiant and Valiant [VV17] removed the $\log d$ factor from the sample complexity (though without analyzing "robustness"). The analysis in these works showed the importance at looking at "weighted" versions of the ℓ_2^2-distance $\mathrm{d}_{\ell_2}^2(p, q) = \sum_i (p_i - q_i)^2$ in which the ith summand is reweighted by a factor depending on q_i. Indeed, Acharya et al. [ADK15] improved these results by considering an unbiased estimator for the χ^2-divergence of p from q and (implicitly) using a form of Lemma 2.1; they showed that $n = O(\sqrt{d}/\epsilon^2)$ samples from p suffice to distinguish $\mathrm{d}_{\chi^2}(p \parallel q) \leq \epsilon^2/10$ from $\mathrm{d}_{\mathrm{TV}}(p, q) > \epsilon$. Indeed, although it is not stated this way, a close inspection of their proof shows that they actually obtain a robust tester for χ^2-divergence under the assumption that $q_i \geq \Omega(\epsilon^2/d)$ for all i. This observation motivated our result Theorem 1.3. As a not too difficult consequence,

the present authors and others [DKW17] observed that one can upgrade the [ADK15] result from "χ^2-vs.-ℓ_1" to the strictly superior "χ^2-vs.-Hellinger", à la our Theorem 1.1.

On the subject of testing identity of p to a known distribution q, we should mention the line of work on "instance-optimal" results due to Valiant and Valiant [VV17] and Blais et al. [BCG17]. Stating these is slightly technical, but roughly speaking they show that one can distinguish $p = q$ from $\mathrm{d}_{\mathrm{TV}}(p, q) > \epsilon$ using just $n = O(\sqrt{k}/\epsilon^2)$ samples provided the largest k values of q sum to at least $1 - \Theta(\epsilon)$. This can be compared with our Corollary 1.6.

Identity testing with two unknown distributions. Finally, we discussed Problem 2 mentioned above, testing identity of *two* unknown distributions p and q on $[d]$, given n samples from each. This problem was first studied by Batu et al. [BFR+13], who used a natural estimator for $\mathrm{d}^2_{\ell_2}(p, q)$ to show that $n = O(1/\epsilon^4)$ samples suffice to distinguish $\mathrm{d}^2_{\ell_2}(p, q) \leq \epsilon/2$ from $\mathrm{d}^2_{\ell_2}(p, q) > \epsilon$. (This has no dependence on d but a nonoptimal dependence on ϵ; in fact, our Theorem 1.4 improves on this, even in the quantum case.) From this, they were able to derive a total variation tester, using $n = O(d^{2/3} \log(d)/\epsilon^{8/3})$ samples to distinguish $p = q$ from $\mathrm{d}_{\mathrm{TV}}(p, q) > \epsilon$ (in fact, they had a robust condition in place of $p = q$). This was improved by Chan et al. [CDVV14] to an optimal bound of $n = O(\max\{d^{2/3}/\epsilon^{4/3}, \sqrt{n}/\epsilon^2\})$ by means of an estimator resembling the Le Cam (triangular) discrimination. The result was later reproved by Diakonikolas and Kane [DK16], who also obtained a tester for Hellinger distance in the case of unknown p and q with near-optimal sample complexity of $n = \widetilde{O}(\min\{d^{2/3}/\epsilon^{8/3}, d^{3/4}/\epsilon^2\})$ (improving on an $n = \widetilde{O}(d^{2/3}/\epsilon^8)$ bound of Guha et al. [GMV09]). Subsequently, the tilde on the big-Oh was removed by [DKW17], giving the optimal sample complexity for this case. We remark that obtaining an analogous result in the quantum case is an interesting open problem (specifically, obtaining an identity testing algorithm for two unknown states ρ, σ that uses $n = O(d/\epsilon)$ samples to distinguish $\rho = \sigma$ from $\mathrm{F}(\rho, \sigma) < 1 - \epsilon$).

We end this section by stating and proving the useful version of Chebyshev described earlier.

Lemma 2.1. *Let $\boldsymbol{X}^{(n)}$ be a sequence of estimators for a number $\mu \geq 0$, meaning $\mathbf{E}[\boldsymbol{X}^{(n)}] = \mu$ for all n. Suppose we have a variance bound of the form*

$$\mathbf{Var}[\boldsymbol{X}^{(n)}] \leq O\left(\frac{b(\mu)}{n^2} + \frac{v(\mu)}{n}\right), \tag{2}$$

where

$$b(\mu),\ v(\mu),\ \frac{\mu^2}{b(\mu)},\ \frac{\mu^2}{v(\mu)} \text{ are increasing functions of } \mu \geq 0. \tag{3}$$

(The $O(\cdot)$ should hide a universal constant.) Let $\theta > 0$ be a parameter. Then provided

$$n \geq C \max\left\{\sqrt{\frac{b(\theta)}{\theta^2}}, \frac{v(\theta)}{\theta^2}\right\}, \tag{4}$$

one can use $\boldsymbol{X}^{(n)}$ to distinguish (with high probability) whether $\mu \leq .99\theta$ or $\mu > \theta$. Here C is another universal constant. (More generally, to achieve $1 - \gamma$ in place of .99, one should take $\gamma^2\theta^2$ in place of θ^2 in the denominators in (4).)

Proof. We report "$\mu \leq .99\theta$" if $\boldsymbol{X}^{(n)} \leq .995\theta$ and report "$\mu > \theta$" if $\boldsymbol{X}^{(n)} > .995\theta$.

To analyze the correctness, suppose first that $\mu \leq .99\theta$. Then $b(\mu) \leq b(.99\theta) \leq b(\theta)$ (by (3)) and similarly $v(\mu) \leq v(\theta)$. Using these inequalities in (2) and then substituting in (4), we get

$\mathbf{Var}[\boldsymbol{X}^{(n)}] \leq O(\frac{1}{C^2} + \frac{1}{C})\theta^2$. For C sufficiently large this implies $\mathbf{stddev}[\boldsymbol{X}^{(n)}] \leq .001\theta$, say, and then Chebyshev implies $\boldsymbol{X}^{(n)} \leq \mu + .005\theta \leq .995\theta$ with high probability.

On the other hand, suppose that $\mu > \theta$. Then $\frac{b(\theta)}{\theta^2} \geq \frac{b(\mu)}{\mu^2}$ (by (3)) and similarly $\frac{v(\theta)}{\theta^2} \geq \frac{v(\mu)}{\mu^2}$. Using these inequalities in (4) and then substituting into (2), we get $\mathbf{Var}[\boldsymbol{X}^{(n)}] \leq O(\frac{1}{C^2} + \frac{1}{C})\mu^2$. For C sufficiently large this implies $\mathbf{stddev}[\boldsymbol{X}^{(n)}] \leq .001\mu$, say, and then Chebyshev implies $\boldsymbol{X}^{(n)} \geq \mu - .005\mu > .995\theta$ with high probability. □

3 Preliminaries

3.1 Classical and quantum distances and divergences

3.1.1 Distances and divergences for classical probability distributions

There are many distances and divergences used for comparing discrete probability distributions $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$; see, e.g., [GS02, Cro17]. We review some important ones here. All of the distances we review will be *permutation invariant*, meaning they are unchanged if the same permutation $\pi \in \mathfrak{S}_d$ is simultaneously applied to the outcomes of p and q.

Definition 3.1. The *total variation distance* between p and q is

$$\mathrm{d_{TV}}(p, q) = \frac{1}{2}\sum_{i=1}^{d} |p_i - q_i| = \frac{1}{2}\|p - q\|_1.$$

The total variation distance is a metric and has a maximum value of 1, occurring when p and q have disjoint support. It also has an operational meaning: it is the greatest probability with which one can discriminate a draw from p and a draw from q; i.e., $\mathrm{d_{TV}}(p, q) = \max_{A \subseteq [d]} \{|\mathbf{Pr}_p[A] - \mathbf{Pr}_q[A]|\}$.

Definition 3.2. The ℓ_2 *distance* between p and q is

$$\mathrm{d}_{\ell_2}(p, q) = \Big(\sum_{i=1}^{d} (p_i - q_i)^2\Big)^{1/2} = \|p - q\|_2.$$

The ℓ_2 distance is also a metric; nevertheless we more often consider its square, $\mathrm{d}^2_{\ell_2}(p, q)$. As a probability metric, the ℓ_2 distance is somewhat unnatural. For example, it does not satisfy the "data processing inequality", meaning that there is a stochastic operation that *increases* ℓ_2 distance. However it is by far the easiest distance to calculate, as $\mathrm{d}^2_{\ell_2}(p, q)$ is a simple polynomial in p and q; further, it can be related to the total variation distance via $\frac{1}{2}\mathrm{d}_{\ell_2}(p, q) \leq \mathrm{d_{TV}}(p, q) \leq \frac{1}{2}\sqrt{d} \cdot \mathrm{d}_{\ell_2}(p, q)$, using Cauchy–Schwarz.

Definition 3.3. The *Hellinger distance* between p and q is

$$\mathrm{d_H}(p, q) = \Big(\sum_{i=1}^{d} (\sqrt{p_i} - \sqrt{q_i})^2\Big)^{1/2}.$$

Equivalently, its square may be defined as $\mathrm{d}^2_{\mathrm{H}}(p, q) = 2(1 - \mathrm{BC}(p, q))$, where

$$\mathrm{BC}(p, q) = \sum_{i=1}^{d} \sqrt{p_i}\sqrt{q_i}$$

is the *Bhattacharyya coefficient* (or *Hellinger affinity*) of p and q.

The Hellinger distance is also a metric; it has a maximum value of $\sqrt{2}$, occurring when p and q have disjoint support. One of its main advantages comes from the fact that the Bhattacharyya coefficient satisfies the tensorization property $\mathrm{BC}(p \otimes p', q \otimes q') = \mathrm{BC}(p, q) \cdot \mathrm{BC}(p', q')$, where $p \otimes p'$ denotes the product distribution on $[d]^2$ arising from p and p'. We have the following relationship between Hellinger distance and total variation distance: $\frac{1}{2}\mathrm{d}_{\mathrm{H}}^2(p, q) \leq \mathrm{d}_{\mathrm{TV}}(p, q) \leq \mathrm{d}_{\mathrm{H}}(p, q)$. The squared Hellinger distance is also well known to be within a small constant factor of several other popular measures of discrimination, such as the Jensen–Shannon divergence and the Le Cam (triangular) discrimination.

Definition 3.4. The χ^2*-divergence* of p from q is

$$\mathrm{d}_{\chi^2}(p \parallel q) = \sum_{i=1}^{d} \frac{(p_i - q_i)^2}{q_i},$$

which we take to be ∞ if p's support is not a subset of q's support.

Unlike our previous distances, the χ^2-divergence is not a metric since it is not even symmetric with respect to interchanging p and q. (For simplicity, we may still sometimes call it a "distance".) One utility it has is that it bounds the squared Hellinger distance, $\mathrm{d}_{\mathrm{H}}^2(p, q) \leq \mathrm{d}_{\chi^2}(p \parallel q)$, but can be easier to calculate: if q is considered "fixed", then the χ^2-divergence is a simple polynomial in p. Finally, we should mention that the total variation distance, the squared Hellinger distance, and the χ^2-divergence are all "f-divergences", a consequence of which is that they satisfy the data processing inequality [Wu17, Sec. 4]; i.e., none of them increases when the same stochastic operation is applied to p and q.

3.1.2 Distances and divergences for quantum mixed states

There are again many distances and divergences used for comparing two quantum states ρ and σ; see, e.g., [GLN05], [BZ07, Chap. 13], [Aud12] for some surveys. All of the quantum distances we review will be *unitarily invariant*, meaning that $\mathrm{D}(\rho, \sigma) = \mathrm{D}(U\rho U^\dagger, U\sigma U^\dagger)$ for all unitaries U.

Many classical distances have a quantum analogue, and indeed some have *several* quantum analogues. Typically, a quantum distance between ρ and σ reduces to the analogous classical distance between p and q in the case that $\rho = \mathrm{diag}(p)$ and $\sigma = \mathrm{diag}(q)$ are diagonal.

In particular, for every classical f-divergence one can form either the "standard quantum f-divergence" (introduced by Petz) or the "measured quantum f-divergence" — see [HM17]. We will only consider the latter. Given a classical f-divergence $\mathrm{d}_f(\cdot, \cdot)$, one obtains the corresponding measured quantum f-divergence $\mathrm{D}_f(\cdot, \cdot)$ as follows:

$$\mathrm{D}_f(\rho, \sigma) = \sup_{\text{POVMs } \{E_i\}_{i=1}^N} \{\mathrm{d}_f(p_\rho, p_\sigma)\}, \quad \text{where } p_\xi = (\mathrm{tr}(\xi E_1), \ldots, \mathrm{tr}(\xi E_N)). \tag{5}$$

In other words, the quantum divergence is defined as the maximum classical divergence that can be achieved when applying the same POVM to both states. In this section we will encounter the measured quantum f-divergence corresponding to total variation distance, squared Hellinger distance, and χ^2-divergence.

Definition 3.5. The *trace distance* between ρ and σ is

$$\mathrm{D}_{\mathrm{tr}}(\rho, \sigma) = \frac{1}{2}\|p - q\|_1.$$

The trace distance is a metric and it has a maximum value of 1, occurring when ρ and σ have orthogonal support. Helstrom [Hel76] showed that trace distance is the measured version of classical total variation distance in the sense of Equation (5). It therefore equals the maximum probability with which the states ρ and σ can be discriminated by some measurement. It also follows that it satisfies the "quantum data processing inequality". In other words, it can never increase when the same quantum channel (completely positive trace-preserving map) is applied to both ρ and σ.

Definition 3.6. The *Hilbert–Schmidt distance* (or *Frobenius distance*) between ρ and σ is

$$\mathrm{D}_{\mathrm{HS}}(\rho, \sigma) = \|\rho - \sigma\|_{\mathrm{HS}} = \mathrm{tr}\Big(\sum_{i,j=1}^{d} |\rho_{ij} - \sigma_{ij}|^2\Big)^{1/2} = \mathrm{tr}\big((\rho - \sigma)^2\big)^{1/2}.$$

This metric can be seen as analogue of the classical ℓ_2 distance. It is not, however, a direct analogue in the sense of Equation (5); this is related to the fact that it does not satisfy the quantum data processing inequality. Nevertheless, it is useful by virtue of the fact that the squared Hilbert–Schmidt distance, $\mathrm{D}^2_{\mathrm{HS}}(\rho, \sigma) = \mathrm{tr}\big((\rho - \sigma)^2\big)$, is extremely easy to compute, and that it can be related to the trace distance via Cauchy–Schwarz for matrices: $\frac{1}{2}\mathrm{D}_{\mathrm{HS}}(\rho, \sigma) \leq \mathrm{D}_{\mathrm{tr}}(\rho, \sigma) \leq \frac{1}{2}\sqrt{d} \cdot \mathrm{D}_{\mathrm{HS}}(\rho, \sigma)$.

Definition 3.7. The *Bures distance* between ρ and σ is

$$\mathrm{D}_{\mathrm{B}}(\rho, \sigma) = \big(2(1 - \mathrm{F}(\rho, \sigma))\big)^{1/2},$$

where

$$\mathrm{F}(\rho, \sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1$$

is the *fidelity* between ρ and σ. (The quantity $1 - \mathrm{F}(\rho, \sigma)$ is termed the *infidelity*).

The Bures distance is a metric and it has a maximum value of $\sqrt{2}$, occurring when ρ and σ have orthogonal support. The work of Fuchs and Caves [FC95] shows that the (squared) Bures distance is the measured version of classical (squared) Hellinger distance in the sense of Equation (5). It follows that $\frac{1}{2}\mathrm{D}^2_{\mathrm{B}}(p, q) \leq \mathrm{D}_{\mathrm{tr}}(p, q) \leq \mathrm{D}_{\mathrm{B}}(p, q)$. It also follows that the Bures distance satisfies the quantum data processing inequality.

We more often consider the square of the Bures distance, $\mathrm{D}^2_{\mathrm{B}}(\rho, \sigma)$, which is simply twice the infidelity. It is also quite common to consider the squared fidelity, $\mathrm{F}^2(\rho, \sigma)$. The squared fidelity, as shown by Uhlmann [Uhl76], is the maximum overlap between purifications of ρ and σ, where the *overlap* of (mixed) quantum states ρ' and σ' is defined to be $\mathrm{tr}(\rho'\sigma')$.

Note that when ρ and σ are "close", with $\mathrm{F}(\rho, \sigma) = 1 - \epsilon$, we have that $1 - \mathrm{F}^2(\rho, \sigma) \approx 2\epsilon$. Thus there is not much difference if one defines infidelity as $1 - \mathrm{F}(\rho, \sigma)$ or $1 - \mathrm{F}^2(\rho, \sigma)$; these quantities are always within a factor 2 of each other, and also of the squared Bures distance. Also very closely related is the *quantum Hellinger affinity*, $Q_{1/2}(\rho, \sigma) = \mathrm{tr}(\sqrt{\rho}\sqrt{\sigma})$. It satisfies $\mathrm{F}^2(\rho, \sigma) \leq Q_{1/2}(\rho, \sigma) \leq \mathrm{F}(\rho, \sigma)$ and has been used to define a "quantum Hellinger distance" by $\mathrm{D}^2_{\mathrm{H}}(\rho, \sigma) = 2(1 - Q_{1/2}(\rho, \sigma))$; see [ANSV08]. The same bound $\mathrm{F}^2(\rho, \sigma) \leq Q_{\min}(\rho, \sigma) \leq \mathrm{F}(\rho, \sigma)$ also holds [Aud12] for the quantity $Q_{\min}(\rho \parallel \sigma)$ arising in the quantum Chernoff bound mentioned in Section 2.1.3.

Definition 3.8. Assume σ has full rank. The *Bures χ^2-divergence* of ρ from σ is

$$\mathrm{D}_{\chi^2}(\rho \parallel \sigma) = \mathrm{tr}\big((\rho - \sigma) \cdot \Omega_\sigma(\rho - \sigma)\big),$$

where Ω_σ is the linear operator whose inverse is defined by $\Omega_\sigma^{-1}(A) = \frac{1}{2}(\sigma A + A\sigma)$. (There is a simple generalization to the case where σ does not have full rank, so long as ρ's support is a subset

of σ's; we will not need it, however.) In case $\sigma = \mathrm{diag}(\beta_1, \ldots, \beta_d)$, we obtain the following more explicit formula:

$$\mathrm{D}_{\chi^2}(\rho \parallel \sigma) = \sum_{i,j=1}^{d} \frac{2}{\beta_i + \beta_j} |\Delta_{ij}|^2, \quad \text{where } \Delta = \rho - \sigma.$$

The Bures χ^2-divergence is the measured version of the classical χ^2-divergence in the sense of Equation (5), as shown in [BC94, TV15]). As such, it satisfies the quantum data processing inequality, and we can infer from the classical case that $\mathrm{D}_{\mathrm{B}}^2(\rho, \sigma) \leq \mathrm{D}_{\chi^2}(\rho \parallel \sigma)$. Indeed, it is known [TKR+10] that the *quantum relative entropy*, $\mathrm{S}(\rho \parallel \sigma) = \mathrm{tr}(\rho(\log \rho - \log \sigma))$ is sandwiched in between: $\mathrm{D}_{\mathrm{B}}^2(\rho, \sigma) \leq \mathrm{S}(\rho \parallel \sigma) \leq \mathrm{D}_{\chi^2}(\rho \parallel \sigma)$. As in the classical case, such bounds are what makes the Bures χ^2-divergence useful, together with its having a relatively simple formula when σ is considered to be "fixed".

We close this section by commenting that, although we focus on Bures χ^2-divergence, there are many generalizations of χ^2-divergence to the quantum case. For example, the "standard quantum f-divergence" version is $\mathrm{tr}((\rho-\sigma)^2\sigma^{-1})$. More generally, one may consider $\mathrm{tr}((\rho-\sigma)\sigma^{-\alpha}(\rho-\sigma)\sigma^{\alpha-1})$ for any $\alpha \in [0, 1/2]$, and there are further possibilities. See, e.g., [Pet96, TKR+10], wherein it is explained that the Bures χ^2-divergence takes on the smallest value among a wide family of generalizations.

3.2 Quantum probability

Let V be a finite-dimensional vector space over $\mathbb{C}$ and let $\mathrm{End}(V)$ denote the algebra of linear operators on V. An operator $X \in \mathrm{End}(V)$ is self-adjoint or Hermitian if $X^\dagger = X$, where $X^\dagger$ denotes the conjugate-transpose of X; X is positive if there exists an operator $Y \in \mathrm{End}(V)$ such that $X = Y^\dagger Y$. For self-adjoint operators $X, Y \in \mathrm{End}(V)$ we write $X \preceq Y$ provided $Y - X$ is positive. The identity operator is denoted by $\mathbb{1}$, with the dimension of the underlying vector space being inferred from the context.

Definition 3.9. A *quantum state* ϱ is defined to be a positive operator $\varrho \in \mathrm{End}(V)$ with $\mathrm{tr}(\varrho) = 1$.

Definition 3.10. A *positive-operator valued measurement* (POVM) $\mathcal{M}$ consists of a set of positive operators that sum to the identity operator $\mathbb{1}$. When a measurement $\mathcal{M} = \{E_1, \ldots, E_k\}$ is applied to a quantum state ϱ, the *outcome* is $i \in [k]$ with probability $p_i = \mathrm{tr}(\varrho E_i)$.

Definition 3.11. An *observable* $\mathcal{O}$ is a self-adjoint operator $\mathcal{O} \in \mathrm{End}(V)$. It has a unique spectral decomposition $\mathcal{O} = \lambda_1 \Pi_1 + \cdots + \lambda_k \Pi_k$, where the λ_i's are the distinct real eigenvalues of $\mathcal{O}$, and the Π_i's are the orthogonal projections onto the associated eigenspaces. The projections $\{\Pi_i : i \in [k]\}$ form a POVM.

Suppose we perform this POVM on a quantum state $\varrho \in \mathrm{End}(V)$ and then report the eigenvalue λ_i upon receiving outcome i. Then we obtain a discrete real-valued random variable $\boldsymbol{x}$, which takes value λ_i with probability $\mathrm{tr}(\varrho \Pi_i)$ for $i = 1, \ldots, k$.

Fact 3.12. *Given an observable $\mathcal{O}$ and associated real-valued random variable $\boldsymbol{x}$, it holds that $\mathbf{E}[\boldsymbol{x}] = \mathrm{tr}(\rho\mathcal{O})$. It also holds that the observable $\mathcal{O}^2$ is associated to the random variable $\boldsymbol{x}^2$. Thus we can compute $\mathbf{Var}[\boldsymbol{x}]$ as $\mathrm{tr}(\rho\mathcal{O}^2) - \mathrm{tr}(\rho\mathcal{O})^2$.*

In light of these facts, it is reasonable to define the notation $\mathbf{E}_\varrho[\mathcal{O}]$ and $\mathbf{Var}_\varrho[\mathcal{O}]$. In fact, we will extend this notation to all operators, not just self-adjoint ones.

Definition 3.13. The *expectation* of operator $X \in \mathrm{End}(V)$ with respect to state ϱ is defined by

$$\mathop{\mathbf{E}}_{\varrho}[X] = \mathrm{tr}(\varrho X).$$

Since $\mathbf{E}_\varrho[\mathbb{1}] = 1$, $\mathbf{E}_\varrho[X^\dagger] = \overline{\mathbf{E}_\varrho[X]}$, and $\mathbf{E}_\varrho[X^\dagger X] \geq 0$ for all $X \in \mathrm{End}(V)$, the map $\mathbf{E}_\varrho[\,\cdot\,]$ defines a positive linear functional of norm 1 on $\mathrm{End}(V)$. Moreover, $\mathbf{E}_{\varrho\otimes\varrho'}[\,\cdot\,]$ satisfies the following tensorization property: $\mathbf{E}_{\varrho\otimes\varrho'}[\mathcal{O}\otimes\mathcal{O}'] = \mathbf{E}_\varrho[\mathcal{O}]\cdot\mathbf{E}_{\varrho'}[\mathcal{O}']$ for all observables $\mathcal{O}, \mathcal{O}' \in \mathrm{End}(V)$. The following straightforward fact says that $\mathbf{E}_\varrho[\,\cdot\,]$ is also monotone with respect to the Löwner partial order.

Fact 3.14. *If $\mathcal{O}_1, \mathcal{O}_2 \in \mathrm{End}(V)$ are observables, then $\mathcal{O}_1 \preceq \mathcal{O}_2$ if and only if $\mathbf{E}_\varrho[\mathcal{O}_1] \leq \mathbf{E}_\varrho[\mathcal{O}_2]$ for all states $\varrho \in \mathrm{End}(V)$.*

Definition 3.15. The *covariance* of two operators $X_1, X_2 \in \mathrm{End}(V)$ with respect to state ϱ is the sesquilinear form defined by

$$\mathop{\mathbf{Cov}}_\varrho[X_1, X_2] = \mathop{\mathbf{E}}_\varrho[(X_1 - \mu_1\mathbb{1})^\dagger(X_2 - \mu_2\mathbb{1})] = \mathop{\mathbf{E}}_\varrho[X_1^\dagger X_2] - \mu_1^\dagger\mu_2, \quad \text{where } \mu_i = \mathop{\mathbf{E}}_\varrho[X_i].$$

Since $\mathbf{Cov}_\varrho[\mathbb{1}, \,\cdot\,] = \mathbf{Cov}_\varrho[\,\cdot\,, \mathbb{1}] = 0$, it follows that $\mathbf{Cov}_\varrho[\,\cdot\,,\cdot\,]$ is also translation-invariant in each argument; i.e., $\mathbf{Cov}_\varrho[\mathcal{O}_1 + a\mathbb{1}, \mathcal{O}_2 + b\mathbb{1}] = \mathbf{Cov}_\varrho[\mathcal{O}_1, \mathcal{O}_2]$ for all $a, b \in \mathbb{C}$. Furthermore, $\mathbf{Cov}_{\varrho\otimes\varrho'}[\,\cdot\,,\cdot\,]$ satisfies the following tensorization property,

$$\mathop{\mathbf{Cov}}_{\varrho\otimes\varrho'}[X_1 \otimes Y_1, X_2 \otimes Y_2] = \mathop{\mathbf{Cov}}_\varrho[X_1, X_2] \cdot \mathop{\mathbf{Cov}}_{\varrho'}[Y_1, Y_2],$$

for all operators $X_1, X_2, Y_1, Y_2 \in \mathrm{End}(V)$. Hence,

$$\mathop{\mathbf{Cov}}_{\varrho\otimes\varrho'}[X_1 \otimes \mathbb{1}, \mathbb{1} \otimes X_2] = 0. \tag{6}$$

When X_1 and X_2 are observables, the equality above is a quantum analogue of the classical fact that the covariance of independent random variables is zero.

Definition 3.16. The *variance* of operator $X \in \mathrm{End}(V)$ with respect to state ϱ is defined by

$$\mathop{\mathbf{Var}}_\varrho[X] = \mathop{\mathbf{Cov}}_\varrho[X, X].$$

It holds that $\mathbf{Var}_\varrho[X] \geq 0$ for all X, $\mathbf{Var}_\varrho[c\mathcal{O}] = |c|^2 \mathbf{Var}_\varrho[\mathcal{O}]$ for all $c \in \mathbb{C}$, and

$$\mathop{\mathbf{Var}}_\varrho\left[\sum_{i=1}^k X_i\right] = \sum_{i=1}^k \mathop{\mathbf{Var}}_\varrho[X_i] + \sum_{\substack{i,j=1 \\ i\neq j}}^k \mathop{\mathbf{Cov}}_\varrho[X_i, X_j]$$

for all operators $X_1, \ldots, X_k \in \mathrm{End}(V)$.

Remark 3.17. We will ultimately only be concerned about $\mathbf{E}_\varrho$ and $\mathbf{Var}_\varrho$ as applied to observables, since our state certification algorithms will involve measuring according to observables, and then applying Chebyshev's inequality to the reported random variable $\boldsymbol{x}$. Nevertheless, it will be useful in intermediate calculations to allow $\mathbf{E}_\varrho$, $\mathbf{Var}_\varrho$, and $\mathbf{Cov}_\varrho$ to be applied to *all* operators in $\mathrm{End}(V)$, even though there is not an immediate connection to classical probability when non-normal operators are involved.

We end this section with a definition and lemma that will assist us in finding observables with low variance. Let V_1 and V_2 denote finite-dimensional vector spaces over $\mathbb{C}$ and let $\Phi : \mathrm{End}(V_1) \to \mathrm{End}(V_2)$ be a linear map.

Definition 3.18. $\Phi : \mathrm{End}(V_1) \to \mathrm{End}(V_2)$ is *positive* if $\Phi(X) \succeq 0$ for all $X \in \mathrm{End}(V_1)$ with $X \succeq 0$. And, Φ is *unital* if $\Phi(\mathbb{1}) = \mathbb{1}$.

Suppose that $V_1 = V_2 = V$ and Φ is positive and unital. Then the following result holds:

Lemma 3.19. *If* $\mathbf{E}_\varrho \circ \Phi = \mathbf{E}_\varrho$, *then* $\mathbf{Var}_\varrho[\Phi(\mathcal{O})] \leq \mathbf{Var}_\varrho[\mathcal{O}]$ *for all observables* $\mathcal{O} \in \mathrm{End}(V)$.

Proof. Let $\mathcal{O} \in \mathrm{End}(V)$ be an observable. Since $\mathbf{E}_\varrho[\Phi(\mathcal{O})] = \mathbf{E}_\varrho[\mathcal{O}]$, it suffices to show that $\mathbf{E}_\varrho[\Phi(\mathcal{O})^2] \leq \mathbf{E}_\varrho[\mathcal{O}^2]$. Since $\mathcal{O}$ is self-adjoint and Φ is positive and unital, $\Phi(\mathcal{O})^2 \preceq \Phi(\mathcal{O}^2)$, by the Kadison–Schwarz inequality [Kad52]. Hence, by Fact 3.14, $\mathbf{E}_\varrho[\Phi(\mathcal{O})^2] \leq \mathbf{E}_\varrho[\Phi(\mathcal{O}^2)]$. Since $\mathbf{E}_\varrho \circ \Phi = \mathbf{E}_\varrho$, it follows that $\mathbf{E}_\varrho[\Phi(\mathcal{O})^2] \leq \mathbf{E}_\varrho[\mathcal{O}^2]$, as needed. □

Thus, the class of mean-preserving positive unital maps is variance-nonincreasing.

Remark 3.20. Although there are other measurements that can be associated with an observable $\mathcal{O} \in \mathrm{End}(V)$ apart from its spectral decomposition, the variance of the resulting random variables is at least $\mathbf{Var}_\varrho[\mathcal{O}]$. Indeed, suppose $\mathcal{M} = \{ E_1, \ldots, E_k \}$ is a POVM and $x_1, \ldots, x_k$ are real coefficients such that $\mathcal{O} = x_1 E_1 + \cdots + x_k E_k$. Let $\Phi : \mathrm{End}(\mathbb{C}^k) \to \mathrm{End}(\mathbb{C}^k)$ denote the map defined by $\Phi(A) = A_{11} E_1 + \cdots + A_{kk} E_k$ for all $A \in \mathrm{End}(\mathbb{C}^k)$. Since $\mathcal{M}$ is a POVM, the map Φ is positive and unital. Hence, by the Kadison–Schwarz inequality [Kad52],

$$\mathcal{O}^2 = \Phi(\mathrm{diag}(x_1, \ldots, x_k))^2 \preceq \Phi(\mathrm{diag}(x_1, \ldots, x_k)^2) = x_1^2 E_1 + \cdots + x_k^2 E_k$$

and the result now follows from Fact 3.14.

3.3 Representation theory

Let $\mathfrak{S}_n$ denote the symmetric group on the alphabet $[n]$ and let $\mathrm{U}(d)$ denote the group of $d \times d$ unitary matrices.

Definition 3.21. A *partition* λ is a nonincreasing sequence of nonnegative integers of finite support. If $\lambda_1 + \lambda_2 + \cdots = n$, then λ is said to be a partition of n, denoted by $\lambda \vdash n$. The size of the support of λ is called the *length* of the partition and is denoted by $\ell(\lambda)$. The *power sum* symmetric polynomial in d variables $p_\lambda(x_1, \ldots, x_d)$ associated to a partition λ of length k is defined by $p_\lambda = p_{\lambda_1} p_{\lambda_2} \cdots p_{\lambda_k}$, where $p_r(x_1, \ldots, x_d) = x_1^r + \cdots + x_d^r$ for all $r \geq 0$.

The cycle type of a permutation $\pi \in \mathfrak{S}_n$ is denoted by $\mathrm{cyc}(\pi)$. Sorted in nonincreasing order, $\mathrm{cyc}(\pi)$ is a partition of n. Thus, the partitions of n index the conjugacy classes of $\mathfrak{S}_n$.

Definition 3.22. Let $\mathcal{P}$ denote the unitary representation of $\mathfrak{S}_n$ on $(\mathbb{C}^d)^{\otimes n}$ defined by

$$\mathcal{P}(\pi) |x_1\rangle \otimes \cdots \otimes |x_n\rangle = |x_{\pi^{-1}(1)}\rangle \otimes \cdots \otimes |x_{\pi^{-1}(n)}\rangle,$$

for all $|x_1\rangle, \ldots, |x_n\rangle \in \mathbb{C}^d$ and $\pi \in \mathfrak{S}_n$. Furthermore, let Ad_U be the linear map on observables defined by $\mathrm{Ad}_U(X) = (U^{\otimes n}) X (U^{\otimes n})^\dagger$ for all $U \in \mathrm{U}(d)$.

Definition 3.23. The *symmetric group algebra* $\mathbb{C}\mathfrak{S}_n$ is the algebra of functions $f : \mathfrak{S}_n \to \mathbb{C}$. The functions $1_\pi : \mathfrak{S}_n \to \mathbb{C}$ with $\pi \in \mathfrak{S}_n$ form a basis of $\mathbb{C}\mathfrak{S}_n$, where 1_π is defined by

$$1_\pi(\tau) = \begin{cases} 1, & \pi = \tau, \\ 0, & \pi \neq \tau. \end{cases}$$

With a slight abuse of notation, we use π to denote the function 1_π and think of elements of $\mathbb{C}\mathfrak{S}_n$ as linear combinations of permutations $\pi \in \mathfrak{S}_n$. Thus, the product in $\mathbb{C}\mathfrak{S}_n$ is obtained by extending the product in $\mathfrak{S}_n$ to a bilinear map. $\mathbb{C}\mathfrak{S}_n$ also admits a conjugate-linear involution $X \mapsto X^\dagger$ defined by $\pi^\dagger = \pi^{-1}$ for all $\pi \in \mathfrak{S}_n$.

The representation $\mathcal{P}$ of $\mathfrak{S}_n$ extends to a $*$-representation of the $*$-algebra $\mathbb{C}\mathfrak{S}_n$ as follows:

$$X = \sum_{\pi \in \mathfrak{S}_n} a_\pi \pi \mapsto \sum_{\pi \in \mathfrak{S}_n} a_\pi \mathcal{P}(\pi) = \mathcal{P}(X).$$

Since the representation $\mathcal{P}$ is unitary, it follows that $\mathcal{P}(X^\dagger) = \mathcal{P}(X)^\dagger$ for all $X \in \mathbb{C}\mathfrak{S}_n$.

The center of $\mathbb{C}\mathfrak{S}_n$, denoted by $Z(\mathbb{C}\mathfrak{S}_n)$, is the set of elements $X \in \mathbb{C}\mathfrak{S}_n$ with the property that $XY = YX$ for all $Y \in \mathbb{C}\mathfrak{S}_n$. For all partitions $\kappa \vdash n$, let $\mathcal{O}_\kappa \in \mathbb{C}\mathfrak{S}_n$ be defined by

$$\mathcal{O}_\kappa = \operatorname*{avg}_{\substack{\pi \in \mathfrak{S}_n \\ \mathrm{cyc}(\pi) = \kappa}} \{\pi\}.$$

In other words, $\mathcal{O}_\kappa$ is the normalized indicator function of the conjugacy class of permutations of cycle type κ. The following elementary result relates the elements $\mathcal{O}_\kappa$ to the center of $\mathbb{C}\mathfrak{S}_n$.

Proposition 3.24. *$\{\mathcal{O}_\kappa \mid \kappa \vdash n\}$ is a linear basis for $Z(\mathbb{C}\mathfrak{S}_n)$.*

For a proof, see [GW09, Proposition 4.3.7]. Since $\mathcal{O}_\kappa^\dagger = \mathcal{O}_\kappa$ for all $\kappa \vdash n$, it follows that $\{\mathcal{O}_\kappa \mid \kappa \vdash n\}$ is also a basis for the real vector space of self-adjoint elements of $Z(\mathbb{C}\mathfrak{S}_n)$.

4 Efficient quantum estimators

The connection between observables and random variables presented in Section 3.2 allows us to import notions from classical statistics into the quantum setting. In this section, this connection is used to define quantum estimators and introduce the notion of statistical efficiency of a quantum estimator. These notions are used to formulate a structure theorem for efficient quantum estimators in situations where the statistic of interest is unitarily invariant.

As before, let V be a finite-dimensional vector space over $\mathbb{C}$. Let S denote a set of quantum states on V and let $f : S \to \mathbb{R}$ be a statistic on S. The set S serves to restrict an estimation problem to a particular class of quantum states. S will be gradually restricted, as needed, from an arbitrary set of quantum states to a set of multipartite quantum states of the form $\rho^{\otimes n}$ or $\rho^{\otimes m} \otimes \sigma^{\otimes n}$, where ρ and σ are quantum states on $\mathbb{C}^d$.

Definition 4.1. An *estimator* for f is an observable $\mathcal{O} \in \mathrm{End}(V)$ such that $\mathbf{E}_\varrho[\mathcal{O}] = f(\varrho)$ for all $\varrho \in S$. An estimator $\mathcal{O}$ is *efficient* if $\mathbf{Var}_\varrho[\mathcal{O}] \leq \mathbf{Var}_\varrho[\mathcal{O}']$ for all estimators $\mathcal{O}' \in \mathrm{End}(V)$ for f.

Henceforth, fix $V = (\mathbb{C}^d)^{\otimes n}$ and let S denote the set of states of the form $\rho_1 \otimes \cdots \otimes \rho_n$, where $\rho_1, \ldots, \rho_n$ are quantum states on $\mathbb{C}^d$.

Definition 4.2. A statistic $f : S \to \mathbb{R}$ is *unitarily invariant* if $f \circ \mathrm{Ad}_U = f$ for all $U \in \mathrm{U}(d)$. An observable $\mathcal{O} \in \mathrm{End}(V)$ is *unitarily invariant* if $\mathrm{Ad}_U(\mathcal{O}) = \mathcal{O}$ for all $U \in \mathrm{U}(d)$.

Let Φ be the map on observables $\mathcal{O} \in \mathrm{End}(V)$ defined by

$$\Phi(\mathcal{O}) = \int_{\mathrm{U}(d)} \mathrm{Ad}_U(\mathcal{O})\, dU,$$

where dU denotes Haar measure. Note that Φ preserves self-adjointness and, hence, maps observables to observables.

Proposition 4.3. *If $\mathcal{O}$ is an estimator for a unitarily invariant statistic f, then $\Phi(\mathcal{O})$ is also an estimator for f, and $\mathbf{Var}_\varrho[\Phi(\mathcal{O})] \leq \mathbf{Var}_\varrho[\mathcal{O}]$ for all $\varrho \in S$.*

Proof. The map Φ is positive and unital. Since f is unitarily invariant,

$$\mathop{\mathbf{E}}_{\varrho}[\Phi(\mathcal{O})] = \int_{\mathrm{U}(d)} \mathrm{tr}(\mathrm{Ad}_{U^\dagger}(\varrho)\mathcal{O})\, dU = \int_{\mathrm{U}(d)} f(\mathrm{Ad}_{U^\dagger}(\varrho))\, dU = \int_{\mathrm{U}(d)} f(\varrho)\, dU = \mathop{\mathbf{E}}_{\varrho}[\mathcal{O}].$$

Hence, by Lemma 3.19, $\mathbf{Var}_\varrho[\Phi(\mathcal{O})] \leq \mathbf{Var}_\varrho[\mathcal{O}]$. □

The following result relates the image of the map Φ to the symmetric group algebra $\mathbb{C}\mathfrak{S}_n$ and the representation $\mathcal{P}$. It uses the Schur–Weyl duality theorem. For a proof, see e.g. [CS06, Proposition 2.2].

Proposition 4.4. *The map Φ is a projection into $\mathcal{P}(\mathbb{C}\mathfrak{S}_n)$.*

Thus, if $\mathcal{O}$ is an efficient estimator for a unitarily invariant statistic f, then $\Phi(\mathcal{O})$ is also an efficient estimator for f. Hence, the next corollary follows immediately from Proposition 4.4.

Corollary 4.5. *To find an efficient estimator for a unitarily invariant statistic $f : S \to \mathbb{R}$, it suffices to consider estimators of the form $\mathcal{P}(X)$ with $X \in \mathbb{C}\mathfrak{S}_n$.*

In light of Corollary 4.5, we introduce the following notation:

Notation 4.6. Let $\mathbf{E}_\varrho$ be extended to a map on elements $X \in \mathbb{C}\mathfrak{S}_n$ defined by $\mathbf{E}_\varrho[X] = \mathbf{E}_\varrho[\mathcal{P}(X)]$. Thus, $\mathbf{E}_\varrho$, $\mathbf{Cov}_\varrho$, and $\mathbf{Var}_\varrho$ are defined directly on elements of $\mathbb{C}\mathfrak{S}_n$ via the representation $\mathcal{P}$.

If $\gamma = (i_1\ i_2\ \cdots\ i_\ell) \in \mathfrak{S}_n$, let tr_γ be defined by $\mathrm{tr}_\gamma(\varrho) = \mathrm{tr}(\rho_{i_1}\rho_{i_2}\cdots\rho_{i_\ell})$. The following proposition establishes a formula for the expectation $\mathbf{E}_\varrho[\pi]$ of a permutation $\pi \in \mathfrak{S}_n$ with respect to a state $\varrho \in S$. (Caution: π is not in general an observable.)

Proposition 4.7. *Let $\pi \in \mathfrak{S}_n$ be an arbitrary permutation. If $\pi = \gamma_1 \ldots \gamma_k$ is a decomposition of π into disjoint cycles, including cycles of length 1, then*

$$\mathop{\mathbf{E}}_{\varrho}[\pi^{-1}] = \prod_{i=1}^{k} \mathrm{tr}_\gamma(\varrho).$$

Proof. In light of the tensorization property of $\mathbf{E}_\varrho$ and the fact that ϱ is an n-partite quantum state, the problem reduces immediately to the case when π is an n-cycle. Without loss of generality, suppose $\pi = (1\ 2\ \ldots\ n)$. Thus,

$$\begin{aligned}
\mathrm{tr}(\rho_1\rho_2\cdots\rho_n) &= \sum_{v\in[d]^n} \langle v_1|\rho_1|v_2\rangle \cdots \langle v_n|\rho_n|v_1\rangle \\
&= \sum_{v\in[d]^n} \langle v_1|\rho_1|\pi(v)_1\rangle \cdots \langle v_n|\rho_n|\pi(v)_n\rangle \\
&= \sum_{v\in[d]^n} \langle v|\varrho|\pi(v)\rangle \\
&= \mathrm{tr}(\varrho\, \mathcal{P}(\pi^{-1})) \\
&= \mathop{\mathbf{E}}_{\varrho}[\pi^{-1}].
\end{aligned}$$

□

Remark 4.8. In describing the cycle type of a permutation $\pi \in \mathfrak{S}_n$, it is common to omit mentioning 1-cycles. Conveniently, this would have no effect in Proposition 4.7, since $\mathrm{tr}(\rho_i) = 1$ anyway for all i.

Definition 4.9. The group Γ of *permutation invariants* of the set of states S is defined by

$$\Gamma = \{\pi \in \mathfrak{S}_n \mid \forall \varrho \in S,\ \forall X \in \mathbb{C}\mathfrak{S}_n,\ \mathop{\mathbf{E}}_{\varrho}[\pi^{-1} X \pi] = \mathop{\mathbf{E}}_{\varrho}[X]\}.$$

Note that the definition of Γ depends on S. For all $X \in \mathbb{C}\mathfrak{S}_n$, let $X^\Gamma \in \mathbb{C}\mathfrak{S}_n$ be defined by

$$X^\Gamma = \frac{1}{|\Gamma|} \sum_{\pi \in \Gamma} \pi^{-1} X \pi.$$

Thus, $X^\Gamma \tau = \tau X^\Gamma$ for all $\tau \in \Gamma$ and $X \in \mathbb{C}\mathfrak{S}_n$.

Proposition 4.10. *For all self-adjoint elements $\mathcal{O} \in \mathbb{C}\mathfrak{S}_n$, $\mathbf{Var}_\varrho[\mathcal{O}^\Gamma] \leq \mathbf{Var}_\varrho[\mathcal{O}]$.*

Proof. The map $\mathcal{O} \mapsto \mathcal{O}^\Gamma$ is positive and unital. Moreover, $\mathbf{E}_\varrho[\mathcal{O}^\Gamma] = \mathbf{E}_\varrho[\mathcal{O}]$ for all $\mathcal{O} \in \mathbb{C}\mathfrak{S}_n$. Hence, by Lemma 3.19, $\mathbf{Var}_\varrho[\mathcal{O}^\Gamma] \leq \mathbf{Var}_\varrho[\mathcal{O}]$. □

Corollary 4.11. *To find an efficient estimator for a unitarily invariant statistic $f : S \to \mathbb{R}$, it suffices to consider estimators of the form $\mathcal{P}(X)$ with $X \in \mathbb{C}\mathfrak{S}_n$ and $X\tau = \tau X$ for all $\tau \in \Gamma$.*

The group Γ acts on $\mathfrak{S}_n$ by conjugation, viz. $\tau \in \Gamma$ acts on $\mathfrak{S}_n$ by $\pi \mapsto \tau^{-1}\pi\tau$. This action partitions the group $\mathfrak{S}_n$ into disjoint orbits: $\mathfrak{S}_n = O_1 \cup \cdots \cup O_\ell$, where two permutations π_1 and π_2 belong to the same orbit O_i for $i \in [\ell]$ if and only if there exists $\tau \in \Gamma$ such that $\tau^{-1}\pi_1\tau = \pi_2$. It is easy to see that an element $X \in \mathbb{C}\mathfrak{S}_{m+n}$ commutes with all elements of Γ if and only if X is constant on the orbits $O_1, \ldots, O_\ell$ defined by Γ. Let $\phi_i \in \mathbb{C}\mathfrak{S}_{m+n}$ denote the indicator function of the orbit O_i for $i \in [\ell]$. Thus, the set $\{\phi_1, \ldots, \phi_\ell\}$ forms a basis for the elements $X \in \mathbb{C}\mathfrak{S}_n$ that are constant on the orbits $O_1, \ldots, O_\ell$. Therefore, by Corollary 4.11, it holds that:

Proposition 4.12. *To find an efficient estimator for a unitarily invariant statistic $f : S \to \mathbb{R}$, it suffices to consider estimators of the form $\mathcal{P}(X)$ with $X = a_1\phi_1 + \cdots + a_\ell\phi_\ell$, where $a_1, \ldots, a_\ell \in \mathbb{C}$.*

4.0.1 Case: $\varrho = \rho^{\otimes n}$

Let S denote the set of states of the form $\varrho = \rho^{\otimes n}$, where ρ is a quantum state on $\mathbb{C}^d$. Let $\alpha \in \mathbb{R}^d$ denote the spectrum of ρ (taken in some arbitrary order).

When ϱ is a state of the form $\rho^{\otimes n}$, the expectation $\mathbf{E}_\varrho[\pi]$ of $\pi \in \mathfrak{S}_n$ has a particularly simple formula:

Proposition 4.13. *For all $\pi \in \mathfrak{S}_n$ with $\mathrm{cyc}(\pi) = \kappa$, $\mathbf{E}_\varrho[\pi] = p_\kappa(\alpha)$.*

Proof. Let ℓ denote the number of disjoint cycles in the decomposition of π. By Proposition 4.7,

$$\mathop{\mathbf{E}}_{\varrho}[\pi] = \mathrm{tr}(\rho^{\kappa_1}) \cdots \mathrm{tr}(\rho^{\kappa_\ell}) = p_{\kappa_1}(\alpha) \cdots p_{\kappa_\ell}(\alpha) = p_\kappa(\alpha).$$ □

Thus, $\mathbf{E}_\varrho[\pi]$ depends only on the cycle type of π. Since the cycle types of $\pi_1\pi_2$ and $\pi_2\pi_1$ are equal for all $\pi_1, \pi_2 \in \mathfrak{S}_n$, the following result holds:

Proposition 4.14. *For all $X, Y \in \mathbb{C}\mathfrak{S}_n$, $\mathbf{E}_\varrho[XY] = \mathbf{E}_\varrho[YX]$.*

Proof. For all $\pi_1, \pi_2 \in \mathfrak{S}_n$, $\operatorname{cyc}(\pi_1\pi_2) = \operatorname{cyc}(\pi_2\pi_1)$. Hence, by Proposition 4.13, $\mathbf{E}_\varrho[\pi_1\pi_2] = \mathbf{E}_\varrho[\pi_2\pi_1] = p_\kappa(\alpha)$, where $\kappa = \operatorname{cyc}(\pi_1\pi_2)$. It follows by linearity that $\mathbf{E}_\varrho[XY] = \mathbf{E}_\varrho[YX]$ for all $X, Y \in \mathbb{C}\mathfrak{S}_n$. □

Thus, we obtain the following strengthening of Corollary 4.5:

Proposition 4.15. *To find an efficient estimator for a unitarily invariant statistic $f : S \to \mathbb{R}$, it suffices to consider estimators of the form $\mathcal{P}(X)$ with $X \in Z(\mathbb{C}\mathfrak{S}_n)$.*

Proof. By Proposition 4.14, $\Gamma = \mathfrak{S}_n$. The statement follows immediately from Corollary 4.11. □

The expectation $\mathbf{E}_\varrho[X]$ of an estimator $X \in Z(\mathbb{C}\mathfrak{S}_n)$ can be expressed as a linear combination of $p_\kappa(\alpha)$ with $\kappa \vdash n$ where, recall, α is the spectrum of ρ. By Proposition 3.24, the elements $\mathcal{O}_\kappa \in \mathbb{C}\mathfrak{S}_n$ with $\kappa \vdash n$ form real a basis for the real vector space of self-adjoint elements of $Z(\mathbb{C}\mathfrak{S}_n)$. Hence, an estimator $X \in Z(\mathbb{C}\mathfrak{S}_n)$ can be expressed uniquely as a linear combination of the form

$$X = \sum_{\kappa \vdash n} a_\kappa \mathcal{O}_\kappa,$$

where $a_\kappa \in \mathbb{R}$ for all $\kappa \vdash n$. Thus, by Proposition 4.13,

$$\mathop{\mathbf{E}}_{\varrho}[X] = \sum_{\kappa \vdash n} a_\kappa \mathop{\mathbf{E}}_{\varrho}[\mathcal{O}_\kappa] = \sum_{\kappa \vdash n} a_\kappa p_\kappa(\alpha).$$

Moreover, an estimator $X \in Z(\mathbb{C}\mathfrak{S}_n)$ is unique, as the following result shows.

Proposition 4.16. *If $X_1, X_2 \in Z(\mathbb{C}\mathfrak{S}_n)$ are estimators for $f : S \to \mathbb{R}$, then $X_1 = X_2$.*

Proof. Suppose $X_1 = \sum a_\kappa \mathcal{O}_\kappa$ and $X_2 = \sum b_\kappa \mathcal{O}_\kappa$. Since X_1 and X_2 are estimators for $f : S \to \mathbb{R}$, it follows that

$$\sum_{\kappa \vdash n} a_\kappa p_\kappa(\alpha) = \mathop{\mathbf{E}}_{\varrho}[X_1] = \mathop{\mathbf{E}}_{\varrho}[X_2] = \sum_{\kappa \vdash n} b_\kappa p_\kappa(\alpha).$$

Thus, if $h(\alpha)$ is defined by

$$h(\alpha) = \sum_{\kappa \vdash n} (a_\kappa - b_\kappa) p_\kappa(\alpha),$$

then $h(\alpha) = 0$ for all $\alpha \in \mathbb{R}^d_+$ with $\|\alpha\|_1 = 1$. Note that h is a homogeneous polynomial of degree n in α. Hence, if $x \in \mathbb{R}^d_+$ with $\|x\|_1 > 0$, then

$$h(x) = h\left(\|x\|_1 \cdot \frac{x}{\|x\|_1}\right) = \|x\|_1^n \cdot h\left(\frac{x}{\|x\|_1}\right) = 0.$$

Thus, $h(x) = 0$ for all $x \in \mathbb{R}^d_+$. Since h is a polynomial, it follows that $h \equiv 0$. Therefore, $a_\kappa = b_\kappa$ for all $\kappa \vdash n$, so $X_1 = X_2$. □

Therefore, all observables in the center of $\mathbb{C}\mathfrak{S}_n$ are efficient estimators:

Corollary 4.17. *If $X \in Z(\mathbb{C}\mathfrak{S}_n)$ is an estimator for $f : S \to \mathbb{R}$, then X is efficient.*

Proof. The result follows from Proposition 4.15 and Proposition 4.16. □

Example 4.18. By Corollary 4.17, $\mathcal{O}_\kappa$ is an efficient estimator for $f(\varrho) = p_\kappa(\alpha)$. In particular, suppose $\kappa = (k, 1, 1, \dots)$, which we will denote simply as (k) (recalling Remark 4.8). Then $\mathcal{O}_{(k)}$ is an efficient estimator of $f(\varrho) = \operatorname{tr}(\rho^k)$.

4.0.2 Case: $\varrho = \rho^{\otimes m} \otimes \sigma^{\otimes n}$

Let S denote the set of states of the form $\varrho = \rho^{\otimes m} \otimes \sigma^{\otimes n}$, where ρ and σ are quantum states on $\mathbb{C}^d$. Let $\alpha \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$ denote the spectra of ρ and σ, respectively. The group Γ of permutation invariants of S can be described as follows:

Proposition 4.19. *$\Gamma \cong \mathfrak{S}_m \times \mathfrak{S}_n$, where $(\pi_1, \pi_2) \in \Gamma$ embeds in $\mathfrak{S}_{m+n}$ in the natural way, viz. by applying π_1 to $\{1, \ldots, m\}$ and applying π_2 to $\{m, \ldots, m+n\}$.*

Proof. Let Γ be as in the statement of the proposition and let $\tau \in \Gamma$. The conjugation $\pi \mapsto \tau^{-1}\pi\tau$ applies τ to each index in the cycle decomposition of π. Hence, if τ acts as in the statement of the proposition, then, by Proposition 4.7, $\mathbf{E}_\varrho[\pi] = \mathbf{E}_\varrho[\tau^{-1}\pi\tau]$.

Conversely, let $\tau \in \mathfrak{S}_{m+n}$ and suppose there exists an index $i \in \{1, \ldots, m\}$ such that $\tau(i) \in \{m+1, \ldots, m+n\}$. Thus, if $\pi = (1\ i)$, then $\mathbf{E}_\varrho[\pi] = \operatorname{tr}(\rho^2)$ and

$$\mathop{\mathbf{E}}_{\varrho}[\tau^{-1}\pi\tau] = \begin{cases} \operatorname{tr}(\rho\sigma), & \tau(1) \in \{1, \ldots, m\}, \\ \operatorname{tr}(\sigma^2), & \tau(1) \in \{m+1, \ldots, m+n\}. \end{cases}$$

Since ρ and σ are arbitrary quantum states, it follows that $\tau \notin \Gamma$. □

To find an efficient estimator with respect to S, it is sufficient, by Proposition 4.12, to consider functions $X \in \mathbb{C}\mathfrak{S}_{m+n}$ which are constant on the orbits defined by the action of Γ on $\mathfrak{S}_{m+n}$.

Notation 4.20. Since Γ acts on $\mathfrak{S}_{m+n}$ by conjugation, the orbits of Γ refine the conjugacy classes of $\mathfrak{S}_{m+n}$. An orbit of Γ is uniquely determined by a signature consisting of a cycle type and a map that associates each index in the cycle type with either ρ or σ. For instance, the signature $(\rho\,\sigma)$ identifies the orbit of Γ which consists of all transpositions that exchange an index in $\{1, \ldots, m\}$ with an index in $\{m+1, \ldots, m+n\}$. Note that $(\rho\,\sigma) = (\sigma\,\rho)$. Similarly, $(\rho\,\rho\,\sigma)$ denotes the set of 3-cycles with two indices in $\{1, \ldots, m\}$ and one index in $\{m+1, \ldots, m+n\}$.

If $\mathfrak{s}$ is the signature of an orbit of Γ, let $\mathcal{O}_\mathfrak{s} \in \mathbb{C}\mathfrak{S}_{m+n}$ denote the average of all elements in the orbit. For example, $\mathcal{O}_{(\rho\,\sigma)}$ denotes the average of all transpositions in the $(\rho\,\sigma)$ orbit described above.

Example 4.21. By Proposition 4.7, $\mathcal{O}_{(\rho\,\sigma)}$ is an estimator for $f(\varrho) = \operatorname{tr}(\rho\sigma)$.

Moreover, $\mathcal{O}_{(\rho\,\sigma)}$ satisfies the following uniqueness property:

Proposition 4.22. *If $X \in \mathbb{C}\mathfrak{S}_{m+n}$ is an estimator for the statistic $f : S \to \mathbb{R}$ defined by $f(\varrho) = \operatorname{tr}(\rho\sigma)$ and X is of the form presented in Proposition 4.12, then $X = \mathcal{O}_{(\rho\,\sigma)}$.*

Proof. In the case when $\rho = \sigma$, X becomes an estimator for $\operatorname{tr}(\rho^2)$. Then, by Proposition 4.13, $\mathbf{E}_\varrho[X]$ can be expressed as follows:

$$\mathop{\mathbf{E}}_{\varrho}[X] = \sum_{\kappa \vdash m+n} a_\kappa p_\kappa(\alpha),$$

where α is the spectrum of ρ and $a_\kappa \in \mathbb{R}$ for all $\kappa \vdash m+n$. Since $\mathbf{E}_\varrho[X] - p_2(\alpha) = 0$ for all $\alpha \in \mathbb{R}^d$ with $\|\alpha\|_1 = 1$, it follows, as in the proof of Proposition 4.16, that $a_\kappa = 0$ for all $\kappa \vdash m+n$ with $\kappa \neq (2)$ and $a_{(2)} = 1$. Thus, in general, $X = a\mathcal{O}_{(\rho\,\rho)} + b\mathcal{O}_{(\sigma\,\sigma)} + c\mathcal{O}_{(\rho\,\sigma)}$ with $a + b + c = 1$. Since $\mathbf{E}_\varrho[X] = \operatorname{tr}(\rho\sigma)$, it follows that $c = 1$ and $a = b = 0$. □

A similar argument proves the following:

Proposition 4.23. *If $X \in \mathbb{C}\mathfrak{S}_{m+n}$ is an estimator for the statistic $f : S \to \mathbb{R}$ defined by $f(\varrho) = \mathrm{D}^2_{\mathrm{HS}}(\rho, \sigma)$ and X is of the form presented in Proposition 4.12, then $X = \mathcal{O}_{(\rho\,\rho)} + \mathcal{O}_{(\sigma\,\sigma)} - 2\mathcal{O}_{(\rho\,\sigma)}$.*

Thus, the estimators obtained for $\mathrm{tr}(\rho\sigma)$ and $\mathrm{D}^2_{\mathrm{HS}}(\rho, \sigma)$ are efficient:

Corollary 4.24. *$\mathcal{O}_{(\rho\,\sigma)}$ is an efficient estimator for $f(\varrho) = \mathrm{tr}(\rho\sigma)$.*

Corollary 4.25. *$\mathcal{O}_{(\rho\,\rho)} + \mathcal{O}_{(\sigma\,\sigma)} - 2\mathcal{O}_{(\rho\,\sigma)}$ is an efficient estimator for $f(\varrho) = \mathrm{D}^2_{\mathrm{HS}}(\rho, \sigma)$.*

5 Hilbert–Schmidt distance and related estimation

5.1 Purity, and testing identity to the maximally mixed state

Let ρ be a quantum state on $\mathbb{C}^d$, let $\varrho = \rho^{\otimes n}$, and define $f(\varrho) = \mathrm{tr}(\rho^2)$. The quantity $\mathrm{tr}(\rho^2)$ is called the *purity* of ρ. One can also easily compute that the purity is the same as the squared Hilbert–Schmidt distance to the maximally mixed state, up to an additive constant: $\mathrm{D}^2_{\mathrm{HS}}(\rho, \mathbb{1}/d) = \mathrm{tr}(\rho^2) - 1/d$.

By Example 4.18, the observable $\mathcal{O}_{(2)}$ is an efficient estimator for the statistic f. The following result gives an explicit formula for the variance of $\mathcal{O}_{(2)}$.

Lemma 5.1. $\displaystyle\operatorname*{\mathbf{Var}}_{\varrho}[\mathcal{O}_{(2)}] = \frac{1}{\binom{n}{2}}(1 - p_2(\alpha)^2) + \frac{2(n-2)}{\binom{n}{2}}(p_3(\alpha) - p_2(\alpha)^2).$

Proof. We may compute

$$\mathcal{O}_{(2)}^2 = \frac{1}{\binom{n}{2}}\mathbb{1} + \frac{2(n-2)}{\binom{n}{2}}\mathcal{O}_{(3)} + \frac{\binom{n-2}{2}}{\binom{n}{2}}\mathcal{O}_{(2,2)};$$

this follows from the fact that if two transpositions are chosen uniformly at random from $\mathfrak{S}_n$, their product is the identity with probability $\frac{1}{\binom{n}{2}}$, has cycle type (3) with probability $\frac{2(n-2)}{\binom{n}{2}}$, and has cycle type $(2,2)$ with probability $\frac{\binom{n-2}{2}}{\binom{n}{2}}$. Now

$$\operatorname*{\mathbf{E}}_{\varrho}[\mathcal{O}_{(2)}^2] = \frac{1}{\binom{n}{2}} + \frac{2(n-2)}{\binom{n}{2}}p_3(\alpha) + \frac{\binom{n-2}{2}}{\binom{n}{2}}p_{(2,2)}(\alpha) = \frac{1}{\binom{n}{2}} + \frac{2(n-2)}{\binom{n}{2}}p_3(\alpha) + \left(1 - \frac{2(n-2)+1}{\binom{n}{2}}\right)p_2(\alpha)^2,$$

and the lemma follows. □

At this point, we show how to prove our Theorem 1.4 in the special case that σ is known to be the maximally mixed state. (This result was originally proven, in a slightly more opaque way, in [OW15, Theorem 4.1].)

Proposition 5.2. *(Special case of Theorem 1.4.) There is an algorithm that, given $n = O(1/\epsilon^2)$ copies of the state $\rho \in \mathbb{C}^{d\times d}$, (whp) outputs "close" if $\mathrm{D}_{\mathrm{HS}}(\rho, \mathbb{1}/d) \leq .99\epsilon$ and outputs "far" if $\mathrm{D}_{\mathrm{HS}}(\rho, \mathbb{1}/d) > \epsilon$.*

Proof. Since $\mathrm{D}^2_{\mathrm{HS}}(\rho, \mathbb{1}/d) = \mathrm{tr}(\rho^2) - 1/d$, the observable $\mathcal{O}_{(2)} - \mathbb{1}/d$ is an unbiased estimator of $\mathrm{D}^2_{\mathrm{HS}}(\rho, \mathbb{1}/d)$. Let $\alpha \in \mathbb{R}^d$ denote the spectrum of ρ and let $\Delta_i = \alpha_i - 1/d$ for all $i \in [d]$. Thus,

$$p_3(\alpha) - p_2(\alpha)^2 = \frac{p_2(\Delta)}{d} + p_3(\Delta) - p_2(\Delta)^2 \leq p_2(\Delta) = \mathrm{D}^2_{\mathrm{HS}}(\rho, \mathbb{1}/d).$$

Hence, by Lemma 5.1,

$$\operatorname*{\mathbf{Var}}_{\varrho}\left[\mathcal{O}_{(2)} - \frac{\mathbb{1}}{d}\right] = \operatorname*{\mathbf{Var}}_{\varrho}[\mathcal{O}_{(2)}] \leq O\left(\frac{1}{n^2} + \frac{p_2(\Delta)}{n}\right).$$

The result now follows from Lemma 2.1. □

5.2 Linear fidelity

Let ρ and σ be quantum states on $\mathbb{C}^d$, let $\varrho = \rho^{\otimes m} \otimes \sigma^{\otimes n}$, and define $f(\varrho) = \operatorname{tr}(\rho\sigma)$. The quantity $\operatorname{tr}(\rho\sigma)$ is sometimes called the *overlap* or *linear fidelity* between ρ and σ. By Corollary 4.24, $\mathcal{O}_{(\rho\,\sigma)}$ is an efficient estimator for the statistic f. The following result gives an explicit formula for the variance of $\mathcal{O}_{(\rho\,\sigma)}$.

Proposition 5.3. $\operatorname*{\mathbf{Var}}_{\varrho}[\mathcal{O}_{(\rho\,\sigma)}] = \dfrac{1}{mn} + \dfrac{1-m-n}{mn}\operatorname{tr}(\rho\sigma)^2 + \dfrac{1}{n}\left(1-\dfrac{1}{m}\right)\operatorname{tr}(\rho^2\sigma) + \dfrac{1}{m}\left(1-\dfrac{1}{n}\right)\operatorname{tr}(\rho\sigma^2).$

Proof. The result follows straightforwardly from

$$\mathcal{O}_{(\rho\,\sigma)}^2 = \frac{1}{mn}\mathbb{1} + \left(1-\frac{1}{m}\right)\left(1-\frac{1}{n}\right)\mathcal{O}_{(\rho\,\sigma)(\rho\,\sigma)} + \frac{1}{n}\left(1-\frac{1}{m}\right)\mathcal{O}_{(\rho\,\rho\,\sigma)} + \frac{1}{m}\left(1-\frac{1}{n}\right)\mathcal{O}_{(\rho\,\sigma\,\sigma)},$$

which corresponds to the fact that product of two uniformly transpositions of type $(\rho\,\sigma)$ is: the identity probability $\frac{1}{mn}$; of type $(\rho\,\sigma)(\rho\,\sigma)$ with probability $\left(1-\frac{1}{m}\right)\left(1-\frac{1}{n}\right)$; of type $(\rho\,\rho\,\sigma)$ with probability $\frac{1}{n}(1-\frac{1}{m})$; and of type $(\rho\,\sigma\,\sigma)$ with probability $\frac{1}{m}(1-\frac{1}{n})$. □

5.3 Squared Hilbert–Schmidt distance

Let ρ and σ be quantum states on $\mathbb{C}^d$, let $\varrho = \rho^{\otimes m} \otimes \sigma^{\otimes n}$, and define $f(\varrho) = \mathrm{D}_{\mathrm{HS}}^2(\rho,\sigma) = \operatorname{tr}(\rho^2) + \operatorname{tr}(\sigma^2) - 2\operatorname{tr}(\rho\sigma)$. By Corollary 4.25, $\mathcal{O}_{(\rho\,\rho)} + \mathcal{O}_{(\sigma\,\sigma)} - 2\mathcal{O}_{(\rho\,\sigma)}$ is an efficient estimator for the statistic f.

Lemma 5.4. $\operatorname*{\mathbf{Cov}}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}, \mathcal{O}_{(\sigma\,\sigma)}] = 0.$

Proof. Note that $\mathcal{O}_{(\rho\,\rho)} = \mathcal{O}_{(2)} \otimes \mathbb{1}$, where $\mathcal{O}_{(2)}$ is defined on the first m components of the tensor product. Similarly, $\mathcal{O}_{(\sigma\,\sigma)} = \mathbb{1} \otimes \mathcal{O}_{(2)}$, where $\mathcal{O}_{(2)}$ is defined on the last n components of the tensor product. Hence (recalling Equation (6))

$$\operatorname*{\mathbf{Cov}}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}, \mathcal{O}_{(\sigma\,\sigma)}] = \operatorname*{\mathbf{Cov}}_{\varrho}[\mathcal{O}_{(2)} \otimes \mathbb{1}, \mathbb{1} \otimes \mathcal{O}_{(2)}] = 0.$$ □

Lemma 5.5. $\operatorname*{\mathbf{Cov}}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}, \mathcal{O}_{(\rho\,\sigma)}] = \dfrac{2}{m}\big(\operatorname{tr}(\rho^2\sigma) - \operatorname{tr}(\rho^2)\operatorname{tr}(\rho\sigma)\big).$

Proof. A permutation of type $(\rho\,\rho)(\rho\,\sigma)$ or $(\rho\,\rho\,\sigma)$ is uniquely determined by a product of two transpositions of types $(\rho\,\rho)$ and $(\rho\,\sigma)$. Hence,

$$\mathcal{O}_{(\rho\,\rho)}\mathcal{O}_{(\rho\,\sigma)} = \frac{2}{m}\mathcal{O}_{(\rho\,\rho\,\sigma)} + \left(1-\frac{2}{m}\right)\mathcal{O}_{(\rho\,\rho)(\rho\,\sigma)}.$$

Therefore,

$$\begin{aligned}
\operatorname*{\mathbf{Cov}}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}, \mathcal{O}_{(\rho\,\sigma)}] &= \operatorname*{\mathbf{E}}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}\mathcal{O}_{(\rho\,\sigma)}] - \operatorname*{\mathbf{E}}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}]\operatorname*{\mathbf{E}}_{\varrho}[\mathcal{O}_{(\rho\,\sigma)}] \\
&= \frac{2}{m}\operatorname{tr}(\rho^2\sigma) + \left(1-\frac{2}{m}\right)\operatorname{tr}(\rho^2)\operatorname{tr}(\rho\sigma) - \operatorname{tr}(\rho^2)\operatorname{tr}(\rho\sigma) \\
&= \frac{2}{m}\operatorname{tr}(\rho^2\sigma) - \frac{2}{m}\operatorname{tr}(\rho^2)\operatorname{tr}(\rho\sigma).
\end{aligned}$$ □

Proposition 5.6. *When $m = n$,* $\operatorname*{\mathbf{Var}}_{\varrho}[\mathcal{O}_{(\rho\,\rho)} + \mathcal{O}_{(\sigma\,\sigma)} - 2\mathcal{O}_{(\rho\,\sigma)}] = O\left(\dfrac{1}{n^2} + \dfrac{\mathrm{D}_{\mathrm{HS}}^2(\rho,\sigma)}{n}\right).$

Proof. Let $\mathcal{V} = \mathbf{Var}_{\varrho}[\mathcal{O}_{(\rho\,\rho)} + \mathcal{O}_{(\sigma\,\sigma)} - 2\mathcal{O}_{(\rho\,\sigma)}]$. Since $\mathcal{O}_{(\rho\,\rho)}$, $\mathcal{O}_{(\sigma\,\sigma)} \in \mathbb{C}\Gamma$, $\mathcal{O}_{(\rho\,\rho)}$ and $\mathcal{O}_{(\sigma\,\sigma)}$ commute with each other and with $\mathcal{O}_{(\rho\,\sigma)}$. Hence,

$$\mathcal{V} = \mathbf{Var}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}] + \mathbf{Var}_{\varrho}[\mathcal{O}_{(\sigma\,\sigma)}] + 4\,\mathbf{Var}_{\varrho}[\mathcal{O}_{(\rho\,\sigma)}] - 4\,\mathbf{Cov}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}, \mathcal{O}_{(\rho\,\sigma)}] - 4\,\mathbf{Cov}_{\varrho}[\mathcal{O}_{(\sigma\,\sigma)}, \mathcal{O}_{(\rho\,\sigma)}].$$

Using prior results, we have

$$\mathbf{Var}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}] + \mathbf{Var}_{\varrho}[\mathcal{O}_{(\sigma\,\sigma)}] \leq O\left(\frac{1}{n^2}\right) + \frac{4}{n}\big(\operatorname{tr}(\rho^3) + \operatorname{tr}(\sigma^3) - \operatorname{tr}(\rho^2)^2 - \operatorname{tr}(\sigma^2)^2\big),$$

$$\begin{aligned} 4\,\mathbf{Var}_{\varrho}[\mathcal{O}_{(\rho\,\sigma)}] &= \frac{4}{n^2} + \frac{4-8n}{n^2}\operatorname{tr}(\rho\sigma)^2 + \frac{4n-4}{n^2}\operatorname{tr}(\rho^2\sigma) + \frac{4n-4}{n^2}\operatorname{tr}(\rho\sigma^2) \\ &\leq O\left(\frac{1}{n^2}\right) + \frac{4}{n}\big(\operatorname{tr}(\rho^2\sigma) + \operatorname{tr}(\rho\sigma^2) - 2\operatorname{tr}(\rho\sigma)^2\big), \end{aligned}$$

and

$$-4\,\mathbf{Cov}_{\varrho}[\mathcal{O}_{(\rho\,\rho)}, \mathcal{O}_{(\rho\,\sigma)}] - 4\,\mathbf{Cov}_{\varrho}[\mathcal{O}_{(\sigma\,\sigma)}, \mathcal{O}_{(\rho\,\sigma)}] = -\frac{8}{n}\big(\operatorname{tr}(\rho^2\sigma) + \operatorname{tr}(\rho\sigma^2) - \big(\operatorname{tr}(\rho^2) + \operatorname{tr}(\sigma^2)\big)\operatorname{tr}(\rho\sigma)\big).$$

Therefore,

$$\begin{aligned} \mathcal{V} &\leq O\left(\frac{1}{n^2}\right) + \frac{4}{n}\big(\operatorname{tr}(\rho^3) + \operatorname{tr}(\sigma^3) - \operatorname{tr}(\rho^2)^2 - \operatorname{tr}(\sigma^2)^2 + \operatorname{tr}(\rho^2\sigma) + \operatorname{tr}(\rho\sigma^2) - 2\operatorname{tr}(\rho\sigma)^2\big) \\ &\quad - \frac{4}{n}\big(2\operatorname{tr}(\rho^2\sigma) + 2\operatorname{tr}(\rho\sigma^2) - 2\big(\operatorname{tr}(\rho^2) + \operatorname{tr}(\sigma^2)\big)\operatorname{tr}(\rho\sigma)\big) \\ &= O\left(\frac{1}{n^2}\right) + \frac{4}{n}\big(\operatorname{tr}(\rho^3) + \operatorname{tr}(\sigma^3) - \operatorname{tr}(\rho^2)^2 - \operatorname{tr}(\sigma^2)^2 - \operatorname{tr}(\rho^2\sigma) - \operatorname{tr}(\rho\sigma^2) - 2\operatorname{tr}(\rho\sigma)^2\big) \\ &\quad + \frac{4}{n}\big(2\big(\operatorname{tr}(\rho^2) + \operatorname{tr}(\sigma^2)\big)\operatorname{tr}(\rho\sigma)\big) \\ &= O\left(\frac{1}{n^2}\right) + \frac{4}{n}\big(\operatorname{tr}((\rho+\sigma)(\rho-\sigma)^2) - (\operatorname{tr}(\rho^2) - \operatorname{tr}(\rho\sigma))^2 - (\operatorname{tr}(\sigma^2) - \operatorname{tr}(\rho\sigma))^2\big) \\ &\leq O\left(\frac{1}{n^2}\right) + \frac{4}{n}\operatorname{tr}((\rho+\sigma)(\rho-\sigma)^2) \\ &\leq O\left(\frac{1}{n^2}\right) + \frac{4}{n}\|\rho+\sigma\|_{\infty} \cdot \operatorname{tr}\big((\rho-\sigma)^2\big) \\ &\leq O\left(\frac{1}{n^2}\right) + O\left(\frac{1}{n}\right) \cdot \mathrm{D}^2_{\mathrm{HS}}(\rho,\sigma). \end{aligned}$$

□

5.4 Consequences for testing

Theorem 1.4, which uses $O(1/\epsilon^2)$-copies of unknown ρ, σ to distinguish $\mathrm{D}_{\mathrm{HS}}(\rho,\sigma) \leq .99\epsilon$ from $\mathrm{D}_{\mathrm{HS}}(\rho,\sigma) > \epsilon$, is now an immediate consequence of Lemma 2.1 and Proposition 5.6.

In the remainder of this section we give the proof of Corollary 1.6:

Proof. The testing algorithm does not need to know δ, nor which of ρ or σ is δ-close to rank k: it simply applies the robust Hilbert–Schmidt tester Theorem 1.4 with error parameter $c\epsilon/\sqrt{k}$, where $c = \frac{1}{1+1/\sqrt{2}}$. All we need to show is an elementary fact of pure matrix analysis: assuming

$\mathrm{D}_{\mathrm{HS}}(\rho,\sigma) \leq c\epsilon/\sqrt{k}$, it holds that $\mathrm{D}_{\mathrm{tr}}(\rho,\sigma) \leq \delta + \epsilon$. Since the Hilbert–Schmidt and trace distances are symmetric we may assume that it is σ that is close to rank k; and, since these distances are unitarily invariant, we may assume that $\sigma = \mathrm{diag}(\beta_1, \ldots, \beta_d)$, where $\beta_1 + \cdots + \beta_k \geq 1 - \delta$.

Write ρ_A for the top-left $k \times k$ block of ρ, write ρ_B for its bottom-right $(d-k) \times (d-k)$ block, and write ρ_{off} for the "off-diagonal" $d \times d$ matrix given by zeroing out those two blocks. Similarly define σ_A, σ_B, and σ_{off}, so $\sigma_A = \mathrm{diag}(\beta_1, \ldots, \beta_k)$, $\sigma_B = \mathrm{diag}(\beta_{k+1}, \ldots, \beta_d)$, and $\sigma_C = 0$. Now

$$2\mathrm{D}_{\mathrm{tr}}(\rho,\sigma) = \|\rho - \sigma\|_1 \leq \|\rho_A - \sigma_A\|_1 + \|\rho_{\mathrm{off}} - \sigma_{\mathrm{off}}\|_1 + \|\rho_B - \sigma_B\|_1, \tag{7}$$

by the triangle inequality. The matrix $\rho_A - \sigma_A$ of course has rank at most k, and the matrix $\rho_{\mathrm{off}} - \sigma_{\mathrm{off}}$ has rank at most $2k$ (being the sum of a $k \times (d-k)$ matrix and a $(d-k) \times k$ matrix). Thus we use Cauchy–Schwarz to bound the first two terms on the right of (7) by

$$\sqrt{k}\|\rho_A - \sigma_A\|_{\mathrm{HS}} + \sqrt{2k}\|\rho_{\mathrm{off}} - \sigma_{\mathrm{off}}\|_{\mathrm{HS}} \leq \sqrt{k}\mathrm{D}_{\mathrm{HS}}(\rho,\sigma) + \sqrt{2k}\mathrm{D}_{\mathrm{HS}}(\rho,\sigma) \leq (1+\sqrt{2})c\epsilon.$$

Now if we can show

$$\|\rho_B - \sigma_B\|_1 \leq 2\delta + c\epsilon, \tag{8}$$

we will have bounded $2\mathrm{D}_{\mathrm{tr}}(\rho,\sigma)$ by $2\delta + (2+\sqrt{2})c\epsilon = 2\delta + 2\epsilon$, as needed.

To show (8), we begin with the triangle inequality:

$$\|\rho_B - \sigma_B\|_1 \leq \|\rho_B\|_1 + \|\sigma_B\|_1 = \mathrm{tr}(\rho_B) + \mathrm{tr}(\sigma_B) = (1 - \mathrm{tr}(\rho_A)) + (1 - \mathrm{tr}(\sigma_A)),$$

where the first equality used that ρ_B and σ_B are positive, and the second used that ρ and σ have trace 1. Continuing,

$$(1 - \mathrm{tr}(\rho_A)) + (1 - \mathrm{tr}(\sigma_A)) = 2 - 2\,\mathrm{tr}(\sigma_A) + \mathrm{tr}(\sigma_A - \rho_A) \leq 2\delta + \|\sigma_A - \rho_A\|_1 \leq 2\delta + \sqrt{k}\|\sigma_A - \rho_A\|_{\mathrm{HS}},$$

where we used $1 - \mathrm{tr}(\sigma_A) = 1 - (\beta_1 + \cdots + \beta_k) \leq \delta$, and also Cauchy–Schwarz again. Now (8) follows since $\|\sigma_A - \rho_A\|_{\mathrm{HS}} \leq \mathrm{D}_{\mathrm{HS}}(\rho,\sigma) \leq c\epsilon/\sqrt{k}$. □

6 Quantum chi-squared estimation

6.1 A chi-squared observable

In this section, σ will denote a fixed full-rank d-dimensional density matrix, and we will develop a natural unbiased estimator for the Bures χ^2-divergence $\mathrm{tr}((\rho-\sigma)\cdot\Omega_\sigma(\rho-\sigma))$. This formula suggests a natural bilinear form:

Definition 6.1. For matrices $S, T \in \mathbb{C}^{d\times d}$, define the bilinear form

$$\omega_\sigma^{(2)}(S,T) = \mathrm{tr}(S \cdot \Omega_\sigma T).$$

This bilinear form has the following "contraction" property:

Proposition 6.2. *For any $S \in \mathbb{C}^{d\times d}$ it holds that $\omega_\sigma^{(2)}(S,\sigma) = \mathrm{tr}(S) = \omega_\sigma^{(2)}(\sigma,S)$.*

Proof. Both identities are direct from the definition of the Ω_σ: the first uses $\Omega_\sigma\sigma = \mathbb{1}$; the second uses $S = \frac{1}{2}\sigma\cdot\Omega_\sigma S + \frac{1}{2}\Omega_\sigma S\cdot\sigma$. □

It follows that

$$\begin{aligned}\mathrm{D}_{\chi^2}(\rho \parallel \sigma) &= \omega_\sigma^{(2)}(\rho-\sigma, \rho-\sigma) \\ &= \omega_\sigma^{(2)}(\rho,\rho) - \omega_\sigma^{(2)}(\sigma,\rho) - \omega_\sigma^{(2)}(\rho,\sigma) + \omega_\sigma^{(2)}(\sigma,\sigma) = \omega_\sigma^{(2)}(\rho,\rho) - \mathrm{tr}(\rho) - \mathrm{tr}(\rho) + \mathrm{tr}(\sigma),\end{aligned}$$

and from this we arrive at another standard formula for the Bures χ^2-divergence:

Proposition 6.3. *If ρ is a d-dimensional density matrix, then*

$$\mathrm{D}_{\chi^2}(\rho \parallel \sigma) = \omega_\sigma^{(2)}(\rho,\rho) - 1 = \mathrm{tr}(\rho \cdot \Omega_\sigma \rho) - 1.$$

If $\sigma = \mathrm{diag}(\beta_1, \ldots, \beta_d)$, then Ω_σ acts by multiplying the ij-th entry by $\frac{2}{\beta_i+\beta_j} = \mathrm{avg}\{\beta_i, \beta_j\}^{-1}$; thus in this case,

$$\mathrm{D}_{\chi^2}(\rho \parallel \sigma) = \left(\sum_{i,j=1}^{d} \frac{|\rho_{ij}|^2}{\mathrm{avg}\{\beta_i, \beta_j\}} \right) - 1.$$

In light of the above, it is natural to define the following observable.

Definition 6.4. Assume henceforth that $\sigma = \mathrm{diag}(\beta_1, \ldots, \beta_d)$ is diagonal. We define the associated χ^2 *observable*, operating on $(\mathbb{C}^d)^{\otimes 2}$, as follows:

$$\mathcal{X}_\sigma = \sum_{i,j=1}^{d} \frac{|ji\rangle\langle ij|}{\mathrm{avg}\{\beta_i, \beta_j\}}.$$

Evidently, $\mathbf{E}_{\rho\otimes\rho}[\mathcal{X}_\sigma] = \mathrm{D}_{\chi^2}(\rho \parallel \sigma) + 1$.

Definition 6.5. Given distinct $s, t \in [n]$, we write $\mathcal{X}_\sigma^{(s,t)}$ for the operator which acts on $(\mathbb{C}^d)^{\otimes n}$ by applying $\mathcal{X}_\sigma$ to the s-th and the t-th tensor copies of $\mathbb{C}^d$ and acting as the identity on the remaining copies. (The dependence on n in the notation is implicit.)

Observation 6.6. Observe that $\mathcal{X}_\sigma^{(s,t)}$ is rather similar to the observable $\mathcal{P}((s\,t))$; however, when it swaps letters i and j, it picks up a scalar factor of $\frac{2}{\beta_i+\beta_j}$. Thus in comparison with

$$\mathcal{P}((1\,2)) \cdot \mathcal{P}((2\,3)) = \mathcal{P}((1\,2\,3)) = \sum_{i,j,k=1}^{d} |ijk\rangle\langle jki|$$

we have

$$\mathcal{X}_\sigma^{(1,2)} \cdot \mathcal{X}_\sigma^{(2,3)} = \sum_{i,j,k=1}^{d} \frac{|ijk\rangle\langle jki|}{\mathrm{avg}\{\beta_i, \beta_j\} \cdot \mathrm{avg}\{\beta_i, \beta_k\}},$$

the scalar factors in the denominator arising because letters i and k are swapped, and then letters i and j are swapped. As a consequence, rather than the matrix trilinear form mapping (R, S, T) to

$$\mathrm{tr}(\mathcal{P}((1\,2)) \cdot \mathcal{P}((2\,3)) \cdot R \otimes S \otimes T) = \mathrm{tr}(\mathcal{P}((1\,2\,3)) \cdot R \otimes S \otimes T) = \sum_{i,j,k=1}^{d} T_{ij} S_{jk} R_{ki} = \mathrm{tr}(TSR)$$

as in Proposition 4.7, we obtain the trilinear form given in the subsequent definition.

Definition 6.7. For matrices $R, S, T \in \mathbb{C}^{d\times d}$, define the trilinear form

$$\omega_\sigma^{(3)}(R,S,T) = \mathrm{tr}(\mathcal{X}_\sigma^{(1,2)} \cdot \mathcal{X}_\sigma^{(2,3)} \cdot R \otimes S \otimes T) = \sum_{i,j,k=1}^{d} \frac{T_{ij} S_{jk} R_{ki}}{\mathrm{avg}\{\beta_i, \beta_j\} \cdot \mathrm{avg}\{\beta_i, \beta_k\}}.$$

We again get a certain "contraction" property:

Proposition 6.8. *For any $S, T \in \mathbb{C}^{d\times d}$ it holds that $\omega_\sigma^{(3)}(S,T,\sigma) = \omega_\sigma^{(2)}(S,T) = \omega_\sigma^{(3)}(\sigma,S,T)$.*

Proof. We prove the second identity, the first being similar. When we substitute $R = \sigma$ into Definition 6.7 we obtain

$$\omega_\sigma^{(3)}(\sigma,S,T) = \sum_{i,j,k=1}^{d} \frac{T_{ij} S_{jk} \sigma_{ki}}{\mathrm{avg}\{\beta_i, \beta_j\} \cdot \mathrm{avg}\{\beta_i, \beta_k\}}$$

Since σ is diagonal, the summands with $i \neq k$ vanish. When $i = k$ we have $\sigma_{kk} = \beta_k$, which cancels the factor of $\mathrm{avg}\{\beta_i, \beta_k\}$. We are left with

$$\omega_\sigma^{(3)}(\sigma,S,T) = \sum_{j,k=1}^{d} \frac{T_{ij} S_{ji}}{\mathrm{avg}\{\beta_i, \beta_j\}},$$

which is indeed $\omega_\sigma^{(2)}(S,T)$. □

We also observe that unlike

$$P((1\,2))P((1\,2)) = \sum_{i,j=1}^{d} |ij\rangle\langle ij| = \mathbb{1},$$

we have

$$\mathcal{X}_\sigma \mathcal{X}_\sigma = \sum_{i,j=1}^{d} \frac{|ij\rangle\langle ij|}{\mathrm{avg}\{\beta_i, \beta_j\}^2}, \tag{9}$$

a diagonal operator, but not the identity. Finally:

Definition 6.9. For a given $n \geq 2$, we define the *averaged χ^2 observable* on $(\mathbb{C}^d)^{\otimes n}$ to be $\mathcal{O}_{\chi^2} = \mathrm{avg}_{s\neq t}\{\mathcal{X}_\sigma^{(s,t)}\} - \mathbb{1}$, where the average is over all distinct ordered pairs $s, t \in [n]$.

Evidently:

Proposition 6.10. $\mathbf{E}_{\rho^{\otimes n}}[\mathcal{O}_{\chi^2}] = \mathrm{D}_{\chi^2}(\rho \parallel \sigma)$ *and* $\mathbf{Var}_{\rho^{\otimes n}}[\mathcal{O}_{\chi^2}] = \mathbf{Var}_{\rho^{\otimes n}}[\mathrm{avg}_{s\neq t}\{\mathcal{X}_\sigma^{(s,t)}\}]$.

6.2 Analyzing the variance of the average chi-squared observable

The calculation of the variance of the averaged χ^2 observable, $\mathbf{Var}_{\rho^{\otimes n}}[\mathrm{avg}_{s\neq t}\{\mathcal{X}_\sigma^{(s,t)}\}]$, proceeds exactly as does the calculation of the variance of the purity observable in Lemma 5.1. We obtain:

Proposition 6.11. *The averaged χ^2-observable has variance*

$$\frac{1}{\binom{n}{2}}\left(\mathrm{tr}(\mathcal{X}_\sigma^2 \rho^{\otimes 2}) - \omega_\sigma^{(2)}(\rho,\rho)^2\right) + \frac{2(n-2)}{\binom{n}{2}}\left(\omega_\sigma^{(3)}(\rho,\rho,\rho) - \omega_\sigma^{(2)}(\rho,\rho)^2\right).$$

Introducing the shorthand $\Delta = \rho - \sigma$, we analyze the terms in Proposition 6.11.

Proposition 6.12. $\omega_\sigma^{(3)}(\rho,\rho,\rho) - \omega_\sigma^{(2)}(\rho,\rho)^2 = \omega_\sigma^{(3)}(\Delta,\Delta,\Delta) + \omega_\sigma^{(3)}(\Delta,\sigma,\Delta) - \mathrm{D}_{\chi^2}(\rho \parallel \sigma)^2$.

Proof. This is immediate from writing $\rho = \Delta + \sigma$ and using: multilinearity of $\omega_\sigma^{(3)}(\cdot,\cdot,\cdot)$; the contraction properties Propositions 6.2 and 6.8; $\mathrm{tr}(\rho) = \mathrm{tr}(\sigma) = 1$; and, $\mathrm{D}_{\chi^2}(\rho \parallel \sigma) = \omega_\sigma^{(2)}(\Delta,\Delta)$. □

We will ignore the subtracted $\mathrm{D}_{\chi^2}(\rho \parallel \sigma)^2$ and use the following simple bound for $\omega_\sigma^{(3)}(\Delta,\sigma,\Delta)$:

Proposition 6.13. $\omega_\sigma^{(3)}(\Delta,\sigma,\Delta) \le 2\mathrm{D}_{\chi^2}(\rho \parallel \sigma)$.

Proof. Recalling Definition 6.7 and using $\sigma = \mathrm{diag}(\beta_1,\dots,\beta_d)$ we get

$$\omega_\sigma^{(3)}(\Delta,\sigma,\Delta) = \sum_{i,j=1}^d \frac{\Delta_{ij}\beta_j\Delta_{ji}}{\mathrm{avg}\{\beta_i,\beta_j\}^2} \le 2\sum_{i,j=1}^d \frac{|\Delta_{ij}|^2}{\mathrm{avg}\{\beta_i,\beta_j\}} = 2\mathrm{D}_{\chi^2}(\rho \parallel \sigma),$$

where the inequality used $\frac{\beta_j}{\mathrm{avg}\{\beta_i,\beta_j\}} \le 2$. □

We now come to the main term in Proposition 6.12:

Proposition 6.14. *Assume the smallest eigenvalue of σ is at least δ. Then*

$$\omega_\sigma^{(3)}(\Delta,\Delta,\Delta) \le \sqrt{2d/\delta}\cdot \mathrm{D}_{\chi^2}(\rho \parallel \sigma)^{3/2}.$$

Proof. Applying Cauchy–Schwarz to the formula in Definition 6.7 gives

$$\omega_\sigma^{(3)}(\Delta,\Delta,\Delta) \le \sqrt{\sum_{i,j,k=1}^d \frac{|\Delta_{ij}|^2|\Delta_{ki}|^2}{\mathrm{avg}\{\beta_i,\beta_j\}\cdot\mathrm{avg}\{\beta_i,\beta_k\}}}\cdot\sqrt{\sum_{i,j,k=1}^d \frac{|\Delta_{jk}|^2}{\mathrm{avg}\{\beta_i,\beta_j\}\cdot\mathrm{avg}\{\beta_i,\beta_k\}}}.$$

The sum inside the first square-root above is

$$\sum_{i=1}^d \left(\sum_{j=1}^d \frac{|\Delta_{ij}|^2}{\mathrm{avg}\{\beta_i,\beta_j\}}\right)^2 \le \left(\sum_{i,j=1}^d \frac{|\Delta_{ij}|^2}{\mathrm{avg}\{\beta_i,\beta_j\}}\right)^2 = \mathrm{D}_{\chi^2}(\rho \parallel \sigma)^2.$$

For the sum inside the second square-root above, we use the elementary fact that

$$\mathrm{avg}\{\beta_i,\beta_j\}\cdot\mathrm{avg}\{\beta_i,\beta_k\} \ge (\delta/2)\cdot\mathrm{avg}\{\beta_j,\beta_k\}$$

when $\delta \le \beta_i,\beta_j,\beta_k \le 1$. Thus this second sum is at most

$$d\cdot(2/\delta)\cdot\sum_{j,k}\frac{|\Delta_{jk}|^2}{\mathrm{avg}\{\beta_j,\beta_k\}} = (2d/\delta)\cdot\mathrm{D}_{\chi^2}(\rho \parallel \sigma).$$

Combining the two bounds above completes the proof. □

We now analyze the first term in Proposition 6.11, ignoring the subtracted $\omega_\sigma^{(2)}(\rho,\rho)^2$:

Proposition 6.15. *Assume the smallest eigenvalue of σ is at least δ. Then*

$$\mathrm{tr}(\mathcal{X}_\sigma^2\rho^{\otimes 2}) \le 2d^2 + (2d/\delta)\cdot\mathrm{D}_{\chi^2}(\rho \parallel \sigma).$$

Proof. Using $\operatorname{avg}\{\beta_i, \beta_j\} \geq \sqrt{\beta_i \beta_j}$, we may bound $\operatorname{tr}(\mathcal{X}_\sigma^2 \rho^{\otimes 2})$ as

$$\sum_{i,j=1}^{d} \frac{\rho_{ii}\rho_{jj}}{\operatorname{avg}\{\beta_i,\beta_j\}^2} \leq \sum_{i,j=1}^{d} \frac{\rho_{ii}\rho_{jj}}{\beta_i\beta_j} = \left(\sum_{i=1}^{d} \frac{\rho_{ii}}{\beta_i}\right)^2 = \left(d + \sum_{i=1}^{d} \frac{\Delta_{ii}}{\beta_i}\right)^2 \leq 2d^2 + 2\left(\sum_{i=1}^{d} \frac{|\Delta_{ii}|}{\beta_i}\right)^2.$$

Now using $\sqrt{\beta_i} \geq \sqrt{\delta}$ and then Cauchy–Schwarz,

$$\begin{aligned}\left(\sum_{i=1}^{d} \frac{|\Delta_{ii}|}{\beta_i}\right)^2 \leq (1/\delta) \cdot \left(\sum_{i=1}^{d} \frac{|\Delta_{ii}|}{\sqrt{\beta_i}}\right)^2 &\leq (d/\delta) \cdot \sum_{i=1}^{d} \frac{|\Delta_{ii}|^2}{\beta_i} \\ &\leq (d/\delta) \cdot \sum_{i,j=1}^{d} \frac{|\Delta_{ij}|^2}{\operatorname{avg}\{\beta_i,\beta_j\}} = (d/\delta) \cdot \mathrm{D}_{\chi^2}(\rho \parallel \sigma). \qquad \square\end{aligned}$$

Combining all propositions in this section, we have established the following:

Theorem 6.16. *Assume the smallest eigenvalue of σ is at least δ. Then*

$$\operatorname*{\mathbf{Var}}_{\rho^{\otimes n}}[\mathcal{O}_{\chi^2}] \leq \frac{1}{\binom{n}{2}} \cdot \left(2d^2 + (2d/\delta) \cdot \mathrm{D}_{\chi^2}(\rho \parallel \sigma)\right) + \frac{2(n-2)}{\binom{n}{2}} \cdot \left(\sqrt{2d/\delta} \cdot \mathrm{D}_{\chi^2}(\rho \parallel \sigma)^{3/2} + 2\mathrm{D}_{\chi^2}(\rho \parallel \sigma)\right).$$

6.3 Consequences for testing

Assume σ is a fixed known density matrix, and we wish to estimate $\mathrm{D}_{\chi^2}(\rho \parallel \sigma)$ given copies of an unknown density matrix ρ. Since we may first conjugate each copy of ρ by a unitary that diagonalizes σ, we may assume without loss of generality that σ is diagonal. Now the average χ^2 observable is an unbiased estimator for $\mathrm{D}_{\chi^2}(\rho \parallel \sigma)$, and Theorem 6.16 bounds its variance provided σ's eigenvalues are not too small. Then from Lemma 2.1 we immediately obtain Theorem 1.3.

As mentioned, a corollary of Theorem 1.3 is our main Theorem 1.1, a robust "far-in-fidelity vs. close in χ^2-divergence" tester with *no* assumption about σ's eigenvalues. For convenience we restate and prove this theorem in the contrapositive and in terms of the squared Bures distance (which, recall, is exactly half the infidelity and is upper-bounded by the χ^2-divergence):

Corollary 6.17 (Equivalent to Theorem 1.1)**.** *Fix a d-dimensional mixed state σ. Then there is an algorithm that, given $n = O(d/\epsilon)$ copies of ρ, (whp) outputs "close" if $\mathrm{D}_{\chi^2}(\rho \parallel \sigma) \leq .49\epsilon$ and outputs "far" if $\mathrm{D}_{\mathrm{B}}^2(\rho, \sigma) > .5\epsilon$.*

Proof. Let Φ_η denote the depolarizing channel, which maps a state $\nu \in \mathbb{C}^{d \times d}$ to the state $\Phi_\eta(\nu) = (1-\eta)\nu + \eta \mathbb{1}/d$. Define $\rho' = \Phi_{c\epsilon}(\rho)$ and $\sigma' = \Phi_{c\epsilon}(\sigma)$, where $c > 0$ is a small absolute constant to be chosen later.

If $\mathrm{D}_{\chi^2}(\rho \parallel \sigma) \leq .49\epsilon$ then $\mathrm{D}_{\chi^2}(\rho' \parallel \sigma') \leq .49\epsilon$ by the quantum data processing inequality. On the other hand, in case $\mathrm{D}_{\mathrm{B}}^2(\rho, \sigma) > .5\epsilon$,

$$\sqrt{.5\epsilon} < \mathrm{D}_{\mathrm{B}}(\rho, \sigma) \leq \mathrm{D}_{\mathrm{B}}(\rho, \rho') + \mathrm{D}_{\mathrm{B}}(\rho', \sigma') + \mathrm{D}_{\mathrm{B}}(\sigma', \sigma) \tag{10}$$

by the triangle inequality. We can bound the first of these terms by

$$\mathrm{D}_{\mathrm{B}}^2(\rho, \rho') \leq 2\mathrm{D}_{\mathrm{tr}}(\rho, \rho') = \left\|\rho - \rho'\right\|_1 = \|c\epsilon\rho + c\epsilon\mathbb{1}/d\|_1 \leq 2c\epsilon,$$

where at the end we used the triangle inequality and $\|\rho\|_1 = \|\mathbb{1}/d\|_1 = 1$. A similar argument shows that $\mathrm{D}_{\mathrm{B}}^2(\sigma, \sigma') \leq 2c\epsilon$; i.e., $\mathrm{D}_{\mathrm{B}}(\sigma, \sigma') \leq \sqrt{2c\epsilon}$. Now taking c sufficiently small, (10) implies $\mathrm{D}_{\mathrm{B}}(\rho', \sigma') > \sqrt{.495\epsilon}$ and hence $\mathrm{D}_{\chi^2}(\rho' \parallel \sigma') \geq \mathrm{D}_{\mathrm{B}}^2(\rho', \sigma') \geq .495\epsilon$.

In summary, if $\mathrm{D}_{\chi^2}(\rho \,\|\, \sigma) \leq .49\epsilon$ then $\mathrm{D}_{\chi^2}(\rho' \,\|\, \sigma') \leq .49\epsilon$, if $\mathrm{D}^2_{\mathrm{B}}(\rho, \sigma) > .5\epsilon$ then $\mathrm{D}_{\chi^2}(\rho' \,\|\, \sigma') > .495\epsilon$, and all the eigenvalues of σ' are at least $c\epsilon/d$. Thus we can obtain the desired tester by first applying the depolarizing channel $\Phi_{c\epsilon}$ to the n copies of ρ, producing n copies of ρ', and then using the tester from Theorem 1.3 with σ' in place of σ and $.5\epsilon$ in place of ϵ^2. □

We can also use this corollary to test if an unknown state is diagonal:

Theorem 6.18. *Given $n = O(d/\epsilon)$ copies of a d-dimensional mixed state ρ, one can distinguish (whp) the case that ρ is diagonal (in the standard basis) from the case that ρ has infidelity more than ϵ with every diagonal state.*

Proof. Let $p = (\rho_{11}, \ldots, \rho_{dd})$ denote the diagonal of ρ, a probability distribution. We can obtain a sample from p given a copy of ρ simply by measuring ρ in the standard basis. As mentioned near Equation (1), $O(d/\epsilon)$ samples suffice produce an estimate $\widehat{p}$ of p that satisfies $\mathrm{d}_{\chi^2}(p \,\|\, \widehat{p}_{\mathrm{diag}}) \leq .49\epsilon$ (whp). The tester now applies Corollary 6.17 with $\sigma = \mathrm{diag}(\widehat{p})$, using another $O(d/\epsilon)$ samples. If ρ is diagonal, then $\mathrm{D}_{\chi^2}(\rho \,\|\, \sigma) = \mathrm{d}_{\chi^2}(p \,\|\, \widehat{p}) \leq .49\epsilon$ and the tester outputs "close" (whp). If ρ has infidelity more than ϵ with every diagonal state, then in particular $1 - \mathrm{F}(\rho, \sigma) > \epsilon$; i.e., $\mathrm{D}^2_{\mathrm{B}}(\rho, \sigma) > .5\epsilon$, and the tester outputs "far" (whp). □

7 Implementing the observables

In this section, we give efficient algorithms implementing some of our observables. In Section 7.1, we implement the purity observable from Section 5.1, in Section 7.2, we implement the Hilbert–Schmidt observable from Section 5, and in Section 7.3, we implement a different, though related, observable for the Hilbert–Schmidt distance.

Our main tool is *Schur–Weyl duality* from the representation theory of the symmetric and general linear groups. We assume familiarity with representation theory; see [GW09].

Notation 7.1. Given a partition $\lambda \vdash n$, we write SYT_λ for the set of *standard Young tableaus* of shape λ and SSYT^d_λ for the set of *semistandard Young tableaus* of shape λ and alphabet $[d]$.

Notation 7.2. Recall the representations $\mathcal{P}(\pi)$ and $\mathcal{Q}(M)$ of the symmetric and general linear groups, respectively, which act on the vector space $(\mathbb{C}^d)^{\otimes n}$. Because these two commute with each other, $\mathcal{P}(\pi) \cdot \mathcal{Q}(M)$ is a representation of the product group $\mathfrak{S}_n \times \mathrm{GL}(d)$. *Schur–Weyl duality* describes how $(\mathbb{C}^d)^{\otimes n}$ decomposes under this group action:

$$(\mathbb{C}^d)^{\otimes n} \cong \bigoplus_{\substack{\lambda \vdash n \\ \ell(\lambda) \leq d}} \mathrm{Sp}_\lambda \otimes \mathrm{V}^d_\lambda, \tag{11}$$

where Sp_λ and V^d_λ are the irreducible representations of the symmetric and general linear groups, respectively, corresponding to λ. We write $p_\lambda(\pi)$ for the matrix associated with the symmetric group irreducible representation at the permutation $\pi \in \mathfrak{S}_n$.

7.1 Implementing the purity observable

In this section, we describe how to compute the $\mathcal{O}_{(2)}$ observable for estimating the purity, which we used in Section 5.1 to test whether a state is maximally mixed. We begin by deriving the eigendecomposition for *all* $\mathcal{O}_\mu$ observables.

Notation 7.3. Given a partition $\lambda \vdash n$, we write Π_λ for the projector onto the λ-isotypic subspace in Equation (11). If $\ell(\lambda) > d$, then Π_λ is just the all-zeros matrix.

Proposition 7.4. *For any partition $\mu \vdash k \leq n$,*

$$\mathcal{O}_\mu = \sum_\lambda \frac{\chi_\lambda(\mu \cup 1^{n-k})}{\dim(\lambda)} \cdot \Pi_\lambda.$$

Proof. By definition of $\mathcal{O}_\mu$,

$$\mathcal{O}_\mu = \operatorname*{avg}_{\substack{\pi \in \mathfrak{S}_n \\ \mathrm{cyc}(\pi) = \mu}} \{\mathcal{P}(\pi)\} \cong \bigoplus_{\substack{\lambda \vdash n \\ \ell(\lambda) \leq d}} \operatorname*{avg}_{\substack{\pi \in \mathfrak{S}_n \\ \mathrm{cyc}(\pi) = \mu}} \{p_\lambda(\pi)\} \otimes I_{\dim(\mathrm{V}_\lambda^d)} = \bigoplus_{\substack{\lambda \vdash n \\ \ell(\lambda) \leq d}} \frac{\chi_\lambda(\mu \cup 1^{n-k})}{\dim(\lambda)} \left(I_{\dim(\lambda)} \otimes I_{\dim(\mathrm{V}_\lambda^d)} \right),$$

where the last step is by Schur's lemma and the fact that $\mathrm{tr}(p_\lambda(\pi)) = \chi(\mu \cup 1^{n-k})$ if $\mathrm{cyc}(\pi) = \mu$. The right-hand side equals the expression in the proposition, as the $I_{\dim(\lambda)} \otimes I_{\dim(\mathrm{V}_\lambda^d)}$ term just projects into the λ-isotypic subspace. □

Hence, to implement the $\mathcal{O}_\mu$ observable, we measure according to the Π_λ projectors and output $\chi_\lambda(\mu \cup 1^{n-k})/\dim(\lambda)$. As we will see, this can be done efficiently for $\mu = (2)$.

Definition 7.5. *Weak Schur sampling* refers to performing the projective measurement $\{\Pi_\lambda\}_\lambda$ on the space $(\mathbb{C}^d)^{\otimes n}$. It can be implemented in time $\mathrm{poly}(n, d)$; see, for example, [MW16].

Definition 7.6. Given a partition $\lambda \vdash n$, the *second moment estimator* is defined as

$$\mathsf{TN}(\lambda) := \frac{\chi_\lambda(2 \cup 1^{n-2})}{\dim(\lambda)}.$$

In general, computing the characters of the symmetric group is #P-hard [Hep94] (in fact, even deciding whether a character is nonzero is NP-hard [PP17]). However, Frobenius [Fro00] gives an explicit formula for the character ratio $\mathsf{TN}(\lambda)$ (see Ingram [Ing50] for a simple proof of this formula). The following equivalent expression is found, for example, in [IO02]:

$$\mathsf{TN}(\lambda) = \frac{1}{n(n-1)} \sum_{i=1}^{d} \left((\lambda_i - i + \tfrac{1}{2})^2 - (-i + \tfrac{1}{2})^2 \right). \tag{12}$$

As a result, because weak Schur sampling and computing $\mathsf{TN}(\lambda)$ are both efficient operations, we can conclude with the following theorem.

Theorem 7.7. *The $\mathcal{O}_{(2)}$ observable can be computed in time* $\mathrm{poly}(n, d)$.

We note that this is the same algorithm as [OW15] used for testing whether a state is maximally mixed, and it was previously used by [CHW07] to distinguish the maximally mixed state from states which are maximally mixed on a subspace of dimension $d/2$. For a more intuitive view of this algorithm, suppose we perform weak Schur sampling on $\rho^{\otimes n}$, where ρ is a density matrix with sorted eigenvalues $\alpha = (\alpha_1, \ldots, \alpha_d)$. A long line of work [ARS88, KW01, HM02, CM06, OW16, OW17] has shown that the random measurement outcome $\boldsymbol{\lambda}$, when rescaled as $\boldsymbol{\lambda}/n := (\boldsymbol{\lambda}_1/n, \ldots, \boldsymbol{\lambda}_d/n)$, is a good approximation to α. To estimate the purity $p_2(\alpha)$ of α, then, it is natural to output a statistic close to $p_2(\boldsymbol{\lambda}/n)$, and $\mathsf{TN}(\boldsymbol{\lambda})$ is the apparent appropriate statistic.

Remark 7.8. The $\mathcal{O}_\mu$ observables are related to the *central characters*, defined for any $\lambda \vdash n$ and $\mu \vdash k$ as

$$p_\mu^\#(\lambda) = \begin{cases} n^{\downarrow k} \cdot \frac{\chi_\lambda(\mu \cup 1^{n-k})}{\dim(\lambda)} & \text{if } n \geq k, \\ 0 & \text{if } n < k, \end{cases}$$

where $n^{\downarrow k} = n(n-1)\cdots(n-k+1)$. For μ fixed, these are polynomials which are *shifted-symmetric* in the λ_i's, in the sense of [OO98], of which Equation (12) is a special case; see [IO02] for a particularly thorough treatment of these polynomials. Our rule for multiplying the $\mathcal{O}_\mu$'s can be viewed as deriving from the multiplication rule for $p_\mu^\#$ polynomials due to [IK01].

7.2 Implementing the Hilbert–Schmidt observable

In this section, we describe how to compute the $\mathcal{O}_{(\rho\rho)} + \mathcal{O}_{(\sigma\sigma)} - 2\mathcal{O}_{(\rho\sigma)}$ observable for estimating the squared Hilbert–Schmidt distance between ρ and σ, which we used in Section 5 to test whether ρ and σ are equal. There, we considered the general case of states $\rho^{\otimes m} \otimes \sigma^{\otimes n}$, for m possibly not equal to n. For simplicity, we will restrict ourselves to the case $m = n$, though our argument easily extends to the general case. In this section, and this section only, we will write the observable $\mathcal{O}_{(2)} \in \mathbb{C}\mathfrak{S}_k$, for a given integer k, as $\mathcal{O}^k_{(2)}$, so as to make the k dependence explicit. Given this, we can rewrite our Hilbert–Schmidt observable in the following manner.

Proposition 7.9.

$$\mathcal{O}_{(\rho\rho)} + \mathcal{O}_{(\sigma\sigma)} - 2\mathcal{O}_{(\rho\sigma)} = \left(\frac{2n-1}{n}\right) \cdot \mathcal{O}_{(\rho\rho)} + \left(\frac{2n-1}{n}\right) \cdot \mathcal{O}_{(\sigma\sigma)} - \left(\frac{4n-2}{n}\right) \cdot \mathcal{O}^{2n}_{(2)}.$$

Proof. The observable $\mathcal{O}^{2n}_{(2)}$ decomposes as

$$\mathcal{O}^{2n}_{(2)} = \left(\frac{n-1}{4n-2}\right) \cdot \mathcal{O}_{(\rho\rho)} + \left(\frac{n-1}{4n-2}\right) \cdot \mathcal{O}_{(\sigma\sigma)} + \left(\frac{2n}{4n-2}\right) \mathcal{O}_{(\rho\sigma)},$$

where the weights correspond to the probabilities that a random 2-cycle from $\mathfrak{S}_n$ either falls in the first half of $[2n]$, the second half, or falls in both halves. The proposition follows by substitution. □

If we note that $\mathcal{O}_{(\rho\rho)} = \mathcal{O}^n_{(2)} \otimes \mathbb{1}$ and $\mathcal{O}_{(\sigma\sigma)} = \mathbb{1} \otimes \mathcal{O}^n_{(2)}$, where I is the identity matrix acting on $(\mathbb{C}^d)^{\otimes n}$, then by Proposition 7.4 and Definition 7.6, we can rewrite the first two terms in Proposition 7.9 as

$$\left(\frac{2n-1}{n}\right) \cdot \mathcal{O}_{(\rho\rho)} + \left(\frac{2n-1}{n}\right) \cdot \mathcal{O}_{(\sigma\sigma)} = \left(\frac{2n-1}{n}\right) \sum_{\lambda,\mu \vdash n} (\mathsf{TN}(\lambda) + \mathsf{TN}(\mu)) \cdot \Pi_\lambda \otimes \Pi_\mu. \tag{13}$$

We can also rewrite the third term in Proposition 7.9 as

$$\left(\frac{4n-2}{n}\right) \cdot \mathcal{O}^{2n}_{(2)} = \left(\frac{4n-2}{n}\right) \sum_{\nu \vdash 2n} \mathsf{TN}(\nu) \cdot \Pi_\nu. \tag{14}$$

We note that $\mathcal{O}^{2n}_{(2)}$ commutes with $\mathcal{O}_{(\rho\rho)}$ and $\mathcal{O}_{(\sigma\sigma)}$. This is because both of these latter matrices are elements of $\mathbb{C}\mathfrak{S}_{2n}$, and by Proposition 3.24 we know that the center of $\mathbb{C}\mathfrak{S}_{2n}$ contains $\mathcal{O}^{2n}_{(2)}$. By linearity, (13) therefore commutes with (14), and as a result these two matrices are simultaneously diagonalizable, with joint eigenspaces corresponding to the projectors $(\Pi_\lambda \otimes \Pi_\mu)\Pi_\nu = \Pi_\nu(\Pi_\lambda \otimes \Pi_\mu)$. Applying Proposition 7.9, we have that

$$\mathcal{O}_{(\rho\rho)} + \mathcal{O}_{(\sigma\sigma)} - 2\mathcal{O}_{(\rho\sigma)} = \sum_{\substack{\lambda,\mu \vdash n \\ \nu \vdash 2n}} \left(\left(\frac{2n-1}{n}\right) (\mathsf{TN}(\lambda) + \mathsf{TN}(\mu)) - \left(\frac{4n-2}{n}\right) \mathsf{TN}(\nu) \right) \cdot \Pi_\nu(\Pi_\lambda \otimes \Pi_\mu).$$

This equation immediately gives us our algorithm.

Theorem 7.10. *Given $\rho^{\otimes n}$ and $\sigma^{\otimes n}$, the Hilbert–Schmidt observable can be computed as follows:*

- *Perform weak Schur sampling on $\rho^{\otimes n}$ and $\sigma^{\otimes n}$, receiving $\boldsymbol{\mu}, \boldsymbol{\nu} \vdash n$, respectively.*
- *Perform weak Schur sampling on all $2n$ qudits, receiving $\boldsymbol{\lambda} \vdash 2n$.*
- *Output*

$$\binom{2n-1}{n} \cdot \mathsf{TN}(\boldsymbol{\mu}) + \binom{2n-1}{n} \cdot \mathsf{TN}(\boldsymbol{\nu}) - \binom{4n-2}{n} \cdot \mathsf{TN}(\boldsymbol{\lambda}).$$

As noted in Definition 7.5, the Hilbert–Schmidt observable can therefore be computed in time $\mathrm{poly}(n, d)$.

7.3 An alternative Hilbert–Schmidt observable

In the case when the input is $\varrho = \rho^{\otimes n}$ and one already knows σ, one can estimate the squared Hilbert–Schmidt distance between ρ and σ by outputting m copies of σ and measuring the observable from Section 7.2. In this section, we record an alternative observable which performs the same task without first preparing copies of σ.

Definition 7.11. For a word $w \in [d]^n$, its *type* is given by $\tau = (\tau_1, \ldots, \tau_d)$, where τ_i is the number of i's in w, for each $i \in [d]$. Write Types_d^n for the set of types corresponding to words in $[d]^n$; then $(\tau_1, \ldots, \tau_d) \in \mathrm{Types}_d^n$ if and only if each τ_i is a nonnegative integer and $\tau_1 + \cdots + \tau_d = n$. The *$\tau$-subspace* is the span of all vectors $|x\rangle$ of type τ; we write Π_τ for the corresponding projector.

Definition 7.12. Given $\sigma = \mathrm{diag}(\beta)$, for $\beta = (\beta_1, \ldots, \beta_d)$, we define the *inner-product observable*

$$\mathsf{IP} = \sum_{\tau \in \mathrm{Types}_d^n} \frac{\langle \beta, \tau \rangle}{n} \cdot \Pi_\tau.$$

Its name refers to its expectation, $\mathbf{E}_\varrho[\mathsf{IP}] = \mathrm{tr}(\rho\sigma)$. The alternative Hilbert–Schmidt observable is

$$\mathcal{O}_{(2)} + \mathrm{tr}(\sigma^2) \cdot \mathbb{1} - 2 \cdot \mathsf{IP}.$$

By Example 4.18, this has expectation $\mathbf{E}_\varrho[\mathcal{O}_{(2)} + \mathrm{tr}(\sigma^2) \cdot \mathbb{1} - 2 \cdot \mathsf{IP}] = \mathrm{tr}(\rho^2) + \mathrm{tr}(\sigma^2) - 2\,\mathrm{tr}(\rho\sigma) = \mathrm{D}_{\mathrm{HS}}^2(\rho, \sigma)$.

We see that this observable is an unbiased estimator for the squared Hilbert-Schmidt distance. Its variance can be analyzed using the same techniques as for our other observables. Doing so yields a bound that matches the variance of the normal Hilbert-Schmidt observable.

Theorem 7.13. *This observable has variance*

$$\mathop{\mathbf{Var}}_{\varrho}[\mathcal{O}_{(2)} + \mathrm{tr}(\sigma^2) \cdot \mathbb{1} - 2 \cdot \mathsf{IP}] = O\left(\frac{1}{n^2} + \frac{\mathrm{D}_{\mathrm{HS}}^2(\rho, \sigma)}{n}\right).$$

Applying Lemma 2.1, we rederive Theorem 1.4 for the case of known σ: $n = O(1/\epsilon^2)$ copies of ρ are sufficient to distinguish $\mathrm{D}_{\mathrm{HS}}(\rho, \sigma) \leq .99\epsilon$ from $\mathrm{D}_{\mathrm{HS}}(\rho, \sigma) \geq \epsilon$.

To implement this observable, we will need to find a common orthogonal basis for both $\mathcal{O}_{(2)}$ and IP. This is provided by the following definition.

Definition 7.14. Fix a Young diagram $\lambda \vdash n$. The *Young–Yamanouchi basis* of Sp_λ has a vector $|S\rangle$ for each standard Young tableau $S \in \mathrm{SYT}_\lambda$. Similarly, the *Gelfand–Tsetlin basis* of V_λ^d has a vector $|T\rangle$ for each semistandard Young tableau $T \in \mathrm{SSYT}_\lambda^d$. By Notation 7.2, the vectors $|\lambda\rangle \otimes |S\rangle \otimes |T\rangle$, ranging over all $\lambda \vdash n$, $S \in \mathrm{SYT}_\lambda$, and $T \in \mathrm{SSYT}_\lambda^d$, therefore form a basis for the space $(\mathbb{C}^d)^{\otimes n}$. Furthermore, this basis has the following property:

> Write $\tau = (\tau_1, \ldots, \tau_d) \in \mathrm{Types}_d^n$ for the *type* of $|T\rangle$, where τ_i is the number of i's in T, for each $i \in [d]$. Then $|\lambda\rangle \otimes |S\rangle \otimes |T\rangle$ is contained in the τ-subspace of $(\mathbb{C}^d)^{\otimes n}$.

The unitary transformation which maps the standard basis into this basis is known as the *Schur transform*, and by the work of [BCH05, Har05] it can be computed in time $\mathrm{poly}(n, d)$.

Consider the (λ, τ)-*subspace* of $(\mathbb{C}^d)^{\otimes n}$, i.e., the subspace spanned by those vectors of the form $|\lambda\rangle \otimes |S\rangle \otimes |T\rangle$, where T has type τ. Then by Definition 7.14, this is a subspace of both Π_λ and Π_τ and is therefore simultaneously an eigenspace for the $\mathcal{O}_{(2)}$ and IP observables. As a result, writing $\Pi_{\lambda,\tau}$ for the projector onto this subspace, we may decompose our observable as

$$\mathcal{O}_{(2)} + \mathrm{tr}(\sigma^2) \cdot \mathbb{1} - 2 \cdot \mathsf{IP} = \sum_{\lambda,\tau} \left(\mathsf{TN}(\lambda) + \mathrm{tr}(\sigma^2) - 2\frac{\langle \beta, \tau \rangle}{n} \right) \cdot \Pi_{\lambda,\tau}.$$

As we have seen, we can perform the $\Pi_{\lambda,\tau}$ measurement using the Schur transform. We can also compute it by performing the $\{\Pi_\lambda\}_\lambda$ measurement (i.e., weak Schur sampling) followed by the $\{\Pi_\tau\}_\tau$ measurement, using the fact that $\Pi_{\lambda,\tau} = \Pi_\lambda \Pi_\tau = \Pi_\tau \Pi_\lambda$. In conclusion, we derive the following algorithm.

Theorem 7.15. *Given $\rho^{\otimes n}$, the alternative Hilbert–Schmidt observable can be computed as follows:*

- *Measure $\rho^{\otimes n}$ in the Gelfand–Tsetlin basis, receiving a semistandard tableau $\boldsymbol{T}$ of shape $\boldsymbol{\lambda}$ and type $\boldsymbol{\tau}$.*
- *Output $\mathsf{TN}(\boldsymbol{\lambda}) + \mathrm{tr}(\sigma^2) - \langle \beta, \boldsymbol{\tau} \rangle / n$.*

Alternatively, we can receive $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$ by first performing weak Schur sampling and then performing the $\{\Pi_\tau\}_\tau$ projective measurement. By Definition 7.5 and Definition 7.14, both of these algorithms compute the alternative Hilbert-Schmidt observable in time $\mathrm{poly}(n, d)$.

References

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Proceedings of the 29th Annual Conference and Workshop on Neural Information Processing Systems*, pages 3577–3598, 2015. 2.2.2

[AGKE15] Leandro Aolita, Christian Gogolin, Martin Kliesch, and Jens Eisert. Reliable quantum certification of photonic state preparations. *Nature Communications*, 6(8498), 2015. 2.1.2

[ANSV08] Koenraad Audenaert, Michael Nussbaum, Arleta Szkoła, and Frank Verstraete. Asymptotic error rates in quantum hypothesis testing. *Communications in Mathematical Physics*, 279(1):251–283, 2008. 2.1.3, 3.1.2

[ARS88] Robert Alicki, Sławomir Rudnicki, and Sławomir Sadowski. Symmetry properties of product states for the system of N n-level atoms. *Journal of mathematical physics*, 29(5):1158–1162, 1988. 7.1

[Aud12] Koenraad Audenaert. Comparisons between quantum state distinguishability measures. Technical Report 1207.1197, arXiv, 2012. 3.1.2, 3.1.2

[BC94] Samuel Braunstein and Carlton Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439, 1994. 3.1.2

[BCG17] Eric Blais, Clément Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *Proceedings of the 32nd Annual Computational Complexity Conference*, 2017. 2.2.2

[BCH05] Dave Bacon, Isaac Chuang, and Aram Harrow. The quantum Schur transform: I. efficient qudit circuits. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005. 7.14

[BFF+01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001. 2.2.2

[BFR+13] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):Art. 4, 25, 2013. 2.2.2

[BZ07] Ingemar Bengtsson and Karol Życzkowski. *Geometry of Quantum States: an Introduction to Quantum Entanglement*. Cambridge University Press, 2007. 3.1.2

[Can15] Clément Canonne. A survey on distribution testing: Your data is big. But is it blue? Technical Report 63, Electronic Colloquium on Computational Complexity, 2015. 1

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203, 2014. 2.2.2

[CHW07] Andrew Childs, Aram Harrow, and Paweł Wocjan. Weak Fourier-Schur sampling, the hidden subgroup problem, and the quantum collision problem. In *24th Annual Symposium on Theoretical Aspects of Computer Science*, pages 598–609, 2007. 7.1

[CM06] Matthias Christandl and Graeme Mitchison. The spectra of quantum states and the Kronecker coefficients of the symmetric group. *Communications in mathematical physics*, 261(3):789–797, 2006. 7.1

[Cro17] Gavin Crooks. On measures of entropy and information. `http://threeplusone.com/info`, 2017. 3.1.1

[CS06] Benoît Collins and Piotr Śniady. Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, 2006. 4

[DGPP16] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. Technical Report 1611.03579, arXiv, 2016. 2.2.2

[DK16] Ilias Diakonikolas and Daniel Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, pages 685–694, 2016. 2.2.2

[DKW17] Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? Manuscript, 2017. 2.2.2, 2.2.2

[dSLP11] Marcus da Silva, Olivier Landon-Cardinal, and David Poulin. Practical characterization of quantum devices without tomography. *Physical Review Letters*, 107(21):210404, 2011. 2.1.2

[FC95] Christopher Fuchs and Carlton Caves. Mathematical techniques for quantum communication theory. *Open Systems & Information Dynamics*, 3(3):345–356, 1995. 3.1.2

[FL11] Steven Flammia and Yi-Kai Liu. Direct fidelity estimation from few Pauli measurements. *Physical Review Letters*, 106(23):230501, 2011. 2.1.2

[Fro00] Ferdinand Frobenius. Über die charaktere der symmetrischen gruppe. *Sitzungsberichte der Königliche Preussischen Akademie der Wissenschaften zu Berlin*, pages 516–534, 1900. 7.6

[GLN05] Alexei Gilchrist, Nathan Langford, and Michael Nielsen. Distance measures to compare real and ideal quantum processes. *Physical Review A*, 71:062310, Jun 2005. 3.1.2

[GMV09] Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Sublinear estimation of entropy and information distances. *ACM Transactions on Algorithms*, 5(4):Art. 35, 16, 2009. 2.2.2

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000. 2.2.2

[GS02] Alison Gibbs and Francis Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002. 3.1.1

[GW09] Roe Goodman and Nolan Wallach. *Symmetry, representations, and invariants*. Springer, 2009. 3.3, 7

[Har05] Aram Harrow. *Applications of coherent classical communication and the Schur transform to quantum information theory*. PhD thesis, Massachusetts Institute of Technology, 2005. 7.14

[Hel76] Carl Helstrom. *Quantum Detection and Estimation Theory*. Academic Press, 1976. 3.1.2

[Hep94] Charles Hepler. *On the complexity of computing characters of finite groups*. PhD thesis, University of Calgary, 1994. 7.6

[HHJ+16] Jeongwan Haah, Aram Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 913–925, 2016. 1, 2.1.1

[HHR+05] Hartmut Häffner, Wolfgang Hänsel, Christian Roos, Jan Benhelm, Michael Chwalla, Timo Körber, Umakant Rapol, Mark Riebe, Piet Schmidt, Christoph Becher, Otfried Günhe, Wolfgang Dür, and Rainer Blatt. Scalable multiparticle entanglement of trapped ions. *Nature*, 438(7068):643–646, 2005. 1

[HM02] Masahito Hayashi and Keiji Matsumoto. Quantum universal variable-length source coding. *Physical Review A*, 66(2):022311, 2002. 7.1

[HM17] Fumio Hiai and Milán Mosonyi. Different quantum f-divergences and the reversibility of quantum operations. Technical Report 1604.03089, arXiv, 2017. 3.1.2

[IK01] Vladimir Ivanov and Sergei Kerov. The algebra of conjugacy classes in symmetric groups and partial permutations. *Journal of Mathematical Sciences*, 107(5):4212–4230, 2001. 7.8

[Ing50] Richard Ingram. Some characters of the symmetric group. *Proceedings of the American Mathematical Society*, 1(3):358–369, 1950. 7.6

[IO02] Vladimir Ivanov and Grigori Olshanski. Kerov's central limit theorem for the Plancherel measure on Young diagrams. In *Symmetric functions 2001: surveys of developments and perspectives*, pages 93–151. Springer, 2002. 7.6, 7.8

[Kad52] Richard Kadison. A generalized Schwarz inequality and algebraic invariants for operator algebras. *Annals of Mathematics. Second Series*, 56:494–503, 1952. 3.2, 3.20

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of the 28th Annual Conference on Learning Theory*, pages 1066–1100, 2015. 2.2.1

[KW01] Michael Keyl and Reinhard Werner. Estimating the spectrum of a density operator. *Physical Review A*, 64(5):052311, 2001. 7.1

[MW16] Ashley Montanaro and Ronald de Wolf. *A Survey of Quantum Property Testing*. Number 7 in Graduate Surveys. Theory of Computing Library, 2016. 2.1.2, 7.5

[OO98] Andrei Okounkov and Grigori Olshanski. Shifted Schur functions. *St. Petersburg Mathematical Journal*, 9(2):239–300, 1998. 7.8

[OW15] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 529–538, 2015. 1.1, 1.7, 2.1.2, 5.1, 7.1

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 899–912, 2016. 1, 2.1.1, 7.1

[OW17] Ryan O'Donnell and John Wright. Efficient quantum tomography II. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, pages 962–974, 2017. 1, 2.1.1, 2.2.2, 7.1

[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. 2.2.2

[Pet96] Dénes Petz. Monotone metrics on matrix spaces. *Linear algebra and its applications*, 244:81–96, 1996. 3.1.2

[PP17] Igor Pak and Greta Panova. On the complexity of computing Kronecker coefficients. *Computational Complexity*, 26(1):1–36, 2017. 7.6

[TKR+10] Kristan Temme, Michael Kastoryano, Mary Ruskai, Michael Wolf, and Frank Verstraete. The χ^2-divergence and mixing times of quantum Markov processes. *Journal of Mathematical Physics*, 51(12):122201, 2010. 3.1.2

[TV15] Kristan Temme and Frank Verstraete. Quantum chi-squared and goodness of fit testing. *Journal of Mathematical Physics*, 56(1):012202, 18, 2015. 3.1.2

[Uhl76] Armin Uhlmann. The "transition probability" in the state space of a $*$-algebra. *Reports on Mathematical Physics*, 9(2):273–279, 1976. 3.1.2

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. 2.2.2, 2.2.2

[Wu17] Yihong Wu. Lecture notes for ECE598YW: Information-theoretic methods for high-dimensional statistics, 2017. http://www.stat.yale.edu/~yw562/teaching/598/it-stats.pdf. 3.1.1