Adversarial Hypothesis Testing and a Quantum Stein's Lemma for Restricted Measurements

[Extended abstract]*

Fernando G. S. L. Brandão
University College London
fgslbrandao@gmail.com

Aram W. Harrow
MIT
aram@mit.edu

James R. Lee
University of Washington
jrl@cs.washington.edu

Yuval Peres
Microsoft Research, Redmond
peres@microsoft.com

ABSTRACT

Recall the classical hypothesis testing setting with two sets of probability distributions P and Q. One receives either n i.i.d. samples from a distribution $p \in P$ or from a distribution $q \in Q$ and wants to decide from which set the points were sampled. It is known that the optimal exponential rate at which errors decrease can be achieved by a simple maximum-likelihood ratio test which does not depend on p or q, but only on the sets P and Q.

We consider an adaptive generalization of this model where the choice of $p \in P$ and $q \in Q$ can change in each sample in some way that depends arbitrarily on the previous samples. In other words, in the k^{th} round, an adversary, having observed all the previous samples in rounds $1, \ldots, k-1$, chooses $p_k \in P$ and $q_k \in Q$, with the goal of confusing the hypothesis test. We prove that even in this case, the optimal exponential error rate can be achieved by a simple maximum-likelihood test that depends only on P and Q.

We then show that the adversarial model has applications in hypothesis testing for *quantum states* using restricted measurements. For example, it can be used to study the problem of distinguishing entangled states from the set of all separable states using only measurements that can be implemented with local operations and classical communication (LOCC). The basic idea is that in our setup, the deleterious effects of entanglement can be simulated by an adaptive classical adversary.

We prove a quantum Stein's Lemma in this setting: In many circumstances, the optimal hypothesis testing rate is equal to an appropriate notion of quantum relative entropy between two states. In particular, our arguments yield an alternate proof of Li and Winter's recent strengthening of strong subadditivity for von Neumann entropy.

Categories and Subject Descriptors

G.3 [**Mathematics of computing**]: Probability and statistics; H.1.1 [**Information systems**]: Models and principles—*Systems and information theory*

General Terms

Theory

Keywords

quantum information theory, entanglement testing, composite hypothesis testing

1. INTRODUCTION

Asymmetric hypothesis testing is the problem of distinguishing between two sources where one wants to minimize the rate of false positives (type-1 error) subject to a constraint on the rate of false negatives (type-2 error). In the case of n i.i.d. samples from a classical or quantum source, a central result is the Chernoff-Stein Lemma [11, 13, 1] which states that for any constant bound on the type-2 error, the optimal type-1 error decreases at an exponential rate whose exponent is given by the classical (respectively, quantum) relative entropy. Similar results hold even when we generalize the problem so that the sources are described by an unknown parameter and one needs to design a test that works for any choice of the parameter.

Adversarial hypothesis testing. In the first part of this paper (Section 2), we generalize this problem further to allow the parameter to vary adaptively from sample to sample. Since we will allow the parameter to depend arbitrarily on previous samples, this can be thought of as *adversarial* hypothesis testing. That is, we wish to devise a test that can distinguish between samples from two different sets even against an adversary that can choose the distribution in each round based on which samples have previously been observed.

There are some simple cases where it is not hard to see that this additional power cannot help the adversary. For example, suppose we are given a coin with heads probability p

*A full version of this paper is available at http://arxiv.org/abs/1308.6702.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ITCS'14, January 12–14, 2014, Princeton, New Jersey, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2698-8/14/01 ...$15.00.
http://dx.doi.org/10.1145/2554797.2554816.

and wish to distinguish between the cases where $p \in [0, 1/3]$ and where $p \in [2/3, 1]$. It is straightforward to show that this general problem is no harder than simply distinguishing a 1/3-biased coin from a 2/3-biased coin; equivalently, the adversary gains no advantage from the ability to be adaptive. On the other hand, distinguishing between the two settings $p \in \{1/3, 2/3\}$ and $p = 1/2$ is clearly impossible, as the adversary can simply choose with probability 1/2 to flip the 1/3-biased coin, and with probability 1/2 to flip the 2/3-biased coin. The resulting distribution of samples is indistinguishable from the one arising from $p = 1/2$. This stresses the role of *convexity* since even a non-adaptive adversary can simulate a convex combination of distributions by choosing randomly among them.

We will prove in Theorem 2 that this property is sufficient to characterize the optimal error rate for asymmetric hypothesis testing against an adaptive adversary. Specifically, if the two sources vary over convex sets of probability distributions, then the problem is no harder than in the i.i.d. case. Our Theorem 6 also establishes a version of this claim for symmetric hypothesis testing. These two results can be thought of as adversarial versions of the classic Chernoff-Stein Lemma and Chernoff Theorem, respectively.

Quantum hypothesis testing, entanglement, and additivity. One of our main applications for our adversarial Chernoff-Stein Lemma is in quantum hypothesis testing, when the states to be distinguished need not be i.i.d. Indeed, a recurrent challenge in quantum information theory is that even apparently i.i.d. problems can involve complicated entangled states (meaning that they cannot be written as a convex combination of independent states). For example, the quantum capacity of an i.i.d channel requires maximizing over all n-component inputs, and in general it is known that achieving the capacity requires using states that are entangled across channel uses [14, 17]. This phenomenon in quantum information theory—where information-theoretic quantities for n copies of a system are not simply n times the one-copy quantity—is known generally as the "additivity" problem.

A similar additivity problem arises in quantum hypothesis testing when we wish to distinguish many copies of a fixed state against a family of states that include non-i.i.d. states. One important example is the *relative entropy of entanglement* E_R, which is a method of quantifying the entanglement in a state ρ as the minimum of its relative entropy with respect to any separable (i.e. non-entangled) state. Here, ρ is a multipartite state (e.g., shared between systems A, B, C) and separability refers to this partition. However, to establish the asymptotic hypothesis testing rate of ρ against separable states, we need to compare n copies of ρ against states that are separable with respect to our original partition, but not necessarily across the different copies. In our example, $\rho^{\otimes n}$ lives on systems $A_1, B_1, C_1, \ldots, A_n, B_n, C_n$ and we need to compare against states that are separable across the $A_1 \ldots A_n : B_1 \ldots, B_n : C_1 \ldots C_n$ partition, but possibly entangled within the $A_1, \ldots, A_n$ systems (and the $B_1, \ldots, B_n$ and $C_1, \ldots, C_n$ systems). Indeed, such entanglement across copies is known to be necessary to compute the relative entropy of entanglement, since examples exist [39] where $E_R(\rho \otimes \rho) < 2E_R(\rho)$.

Restricted measurements. A further difficulty arises in the quantum setting when we consider restricted families of measurements, such as those arising from locality restrictions. Here, too, the optimal measurement can be entangled across copies. Moreover, since the hypothesis testing problem involves maximizing distinguishability over allowable measurements and minimizing over states, it is possible for entanglement to either increase or decrease the rate.

One particularly relevant example for our work involves distinguishing many copies of a state ρ against a general separable state, using measurements from a class (such as 1-LOCC, defined below) which preserves the set of separable states. This distinguishability scenario was studied extensively in [33, 9, 25, 8]. Though it may initially seem to be an obscure question, it has found applications to understanding the quantum conditional mutual information [9], to channel coding [28], and to classical algorithms for separability testing [10] and the small-set expansion problem [2].

The main result of Section 3 provides quantum versions of the Chernoff-Stein Lemma and Chernoff's theorem for restricted measurements. The main idea is that the deleterious effects of entanglement in this setting are no worse than what could be achieved by an adaptive adversary. Thus quantum analogues follow as a corollary of our classical results. One application of these results is an alternate proof of the improved strong subadditivity inequality of Li and Winter [25].

2. HYPOTHESIS TESTING AGAINST AN ADAPTIVE ADVERSARY

2.1 Asymmetric hypothesis testing

Fix two distributions p and q over a finite domain Ω. Given i.i.d. samples $X_1, X_2, \ldots, X_n$ from a distribution $r \in \{p, q\}$, the goal is to design a test which distinguishes the two possibilities based on the sample. The classical *Chernoff-Stein Lemma* characterizes the optimal exponential rate of error decay achievable in the one-sided error setting.

Consider any acceptance region $A_n \subseteq \Omega^n$ and the corresponding error probabilities $\alpha_n = p^n(A_n^c)$ and $\beta_n = q^n(A_n)$, where we use S^c to denote the complement of a set S. Then for $0 < \varepsilon < 1$, define

$$\beta_n^\varepsilon := \min_{\substack{A_n \subseteq \Omega^n \\ \alpha_n < \varepsilon}} \beta_n \,.$$

We define the optimal error exponent by

$$\mathcal{E}(p, q) := \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{-\log \beta_n^\varepsilon}{n} \,.$$

The following well-known lemma characterizes $\mathcal{E}$ in terms of the relative entropy (see, e.g., Theorem 11.8.3 of [13]).

LEMMA 1 (CHERNOFF-STEIN LEMMA). *Consider any two distributions p and q over a finite domain Ω with $D(p \,\|\, q) < \infty$. Then $\mathcal{E}(p, q) = D(p \,\|\, q)$.*

Here, $D(p \,\|\, q)$ is the *relative entropy,* given by

$$D(p \,\|\, q) := \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)} \,,$$

and we take $D(p \,\|\, q) := \infty$ when there is an $x \in \Omega$ such that $p(x) \neq 0$ but $q(x) = 0$.

The adaptive setting. Suppose now that $P, Q \subseteq \mathbb{R}^\Omega$ are closed, convex sets of probability distributions. An *adaptive*

P-strategy $\hat{p}$ is a collection of functions $\{\hat{p}_k : \Omega^{k-1} \to P : k = 1, 2, \ldots\}$. Let $\mathcal{A}(P)$ denote the set of all adaptive P-strategies. For $x \in \Omega^n$, we denote

$$\hat{p}(x) := \prod_{k=1}^{n} \hat{p}_k(x_1, \ldots, x_{k-1})(x_k) .$$

As before, let $A_n \subseteq \Omega^n$ be an acceptance region, but now we define

$$\alpha_n := \sup_{\hat{p} \in \mathcal{A}(P)} \hat{p}(A_n^c) ,$$

and

$$\beta_n^\varepsilon := \min_{\substack{A_n \subseteq \Omega^n \\ \alpha_n < \varepsilon}} \sup_{\hat{q} \in \mathcal{A}(Q)} \hat{q}(A_n) .$$

We denote the *adversarial one-sided error exponent* by

$$\mathcal{E}_{\mathrm{adv}}(P, Q) := \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{-\log \beta_n^\varepsilon}{n} .$$

Observe that for single distributions $p, q \in \mathbb{R}^\Omega$, we have $\mathcal{E}_{\mathrm{adv}}(\{p\}, \{q\}) = \mathcal{E}(p, q)$.

THEOREM 2 (ADVERSARIAL CHERNOFF-STEIN). *Let Ω be a finite domain. For any closed, convex sets of probability distributions $P, Q \subseteq \mathbb{R}^\Omega$, we have*

$$\mathcal{E}_{\mathrm{adv}}(P, Q) = \min_{p \in P, q \in Q} D(p \,\|\, q) \tag{1}$$

whenever the right-hand side is finite.

Thus in the asymptotic regime, adversarial adaptive hypothesis testing is no harder than the i.i.d. setting. Indeed, the hypothesis test used is a simple Neyman-Pearson test for p, q minimizing the RHS of (1). This result was previously known in the non-adaptive case, where it is sometimes referred to as *composite hypothesis testing* [23].

PROOF. Let $p^* \in P$ and $q^* \in Q$ be the minimizers of $D(p \,\|\, q)$ as p and q vary over P and Q, respectively. We assume they exist and that $0 < D(p^* \,\|\, q^*) < \infty$, else the theorem is vacuously true. By considering non-adaptive strategies that simply play p^* and q^* in each coordinate, one sees that

$$\mathcal{E}_{\mathrm{adv}}(P, Q) \le \mathcal{E}_{\mathrm{adv}}(\{p^*\}, \{q^*\}) = \mathcal{E}(p^*, q^*) = D(p^* \,\|\, q^*) ,$$

where the last equality is Lemma 1. Thus we need only prove that $\mathcal{E}_{\mathrm{adv}}(P, Q) \ge D(p^* \,\|\, q^*)$.

To this end, for $n \in \mathbb{N}$ and $0 < \varepsilon$, we define an acceptance region

$$A_{n,\varepsilon} = \left\{ x \in \Omega^n : \log \frac{p^*(x_1) p^*(x_2) \cdots p^*(x_n)}{q^*(x_1) q^*(x_2) \cdots q^*(x_n)} \ge n(D(p^* \,\|\, q^*) - \varepsilon) \right\} .$$

Our first goal is to argue that, for every adaptive P-strategy $\hat{p}$, and every $\varepsilon > 0$, we have

$$\lim_{n \to \infty} \hat{p}(A_{n,\varepsilon}) = 1 . \tag{2}$$

We will then show that for any adaptive Q-strategy $\hat{q}$, we have

$$\hat{q}(A_{n,\varepsilon}) \le e^{-n(D(p^* \,\|\, q^*) - \varepsilon)} . \tag{3}$$

Once these are proved, letting $\varepsilon \to 0$ yields the desired claim.

Toward proving (2), observe that, for every $\varepsilon > 0$, $\lim_{n \to \infty} (p^*)^n(A_{n,\varepsilon}) = 1$ by the law of large numbers. The following lemma will imply that the same is true for $\hat{p}$.

LEMMA 3. *For any $p \in P$, we have*

$$\sum_{x \in \Omega} p(x) \log \frac{p^*(x)}{q^*(x)} \ge \sum_{x \in \Omega} p^*(x) \log \frac{p^*(x)}{q^*(x)} .$$

PROOF. By Theorem 11.6.1 in [13], we have

$$D(p \,\|\, q^*) \ge D(p \,\|\, p^*) + D(p^* \,\|\, q^*) .$$

Observing that $D(p \,\|\, q^*) - D(p \,\|\, p^*) = \sum_{x \in \Omega} p(x) \log \frac{p^*(x)}{q^*(x)}$, we see that this is precisely the desired inequality. □

Now, for $x \in \Omega$, let $L(x) = \log \frac{p^*(x)}{q^*(x)}$. The preceding lemma states that for any $p \in P$, we have

$$\mathbb{E}_p[L(x)] \ge \mathbb{E}_{p^*}[L(x)] = D(p^* \,\|\, q^*) . \tag{4}$$

Consider a sequence of random variables $\{X_k\}$ distributed according to $\hat{p}$ (in other words, X_k is sampled according to the measure $\hat{p}_k(X_1, X_2, \ldots, X_{k-1})$), and the corresponding martingale difference sequence

$$D_k := L(X_k) - \mathbb{E}[L(X_k) \mid X_1, \ldots, X_{k-1}] .$$

Since the differences are uniformly bounded, Chebyshev's inequality implies that for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(\sum_{k=1}^{n} D_k \ge -\varepsilon n \right) = 1 . \tag{5}$$

On the other hand, (4) implies that for each k, one has $\mathbb{E}[L(X_k) \mid X_1, \ldots, X_{k-1}] \ge D(p^* \,\|\, q^*)$. Combining this with (5) yields

$$\begin{aligned} \lim_{n \to \infty} \hat{p}(A_{n,\varepsilon}) &= \lim_{n \to \infty} \mathbb{P}\left(\sum_{k=1}^{n} L(X_k) \ge n(D(p^* \,\|\, q^*) - \varepsilon) \right) \\ &\ge \lim_{n \to \infty} \mathbb{P}\left(\sum_{k=1}^{n} D_k \ge -\varepsilon n \right) = 1 , \end{aligned}$$

confirming (2).

We now turn to verifying (3).

LEMMA 4. *For any $q \in Q$, we have*

$$\sum_{x \in \Omega} q(x) \frac{p^*(x)}{q^*(x)} \le 1 .$$

PROOF. For $\lambda \in [0, 1]$, write $q_\lambda = \lambda q + (1 - \lambda) q^*$. Since q^* is the minimizer of $D(p^* \,\|\, q)$ for q in the convex set Q, we know that the derivative of $D(p^* \,\|\, q_\lambda)$ at $\lambda = 0$ is non-negative.

Calculate

$$\begin{aligned} \frac{d}{d\lambda} D(p^* \,\|\, q_\lambda) &= \sum_{x \in \Omega} p^*(x) \frac{d}{d\lambda} \log \frac{p^*(x)}{q_\lambda(x)} \\ &= -\sum_{x \in \Omega} p^*(x) \frac{d}{d\lambda} \log \left(\frac{\lambda q(x) + (1 - \lambda) q^*(x)}{p^*(x)} \right) \\ &= -\sum_{x \in \Omega} p^*(x) \frac{q(x) - q^*(x)}{\lambda q(x) + (1 - \lambda) q^*(x)} . \end{aligned}$$

Using the fact that the derivative is non-negative at $\lambda = 0$ yields

$$\sum_{x \in \Omega} \frac{p^*(x) q^*(x)}{q^*(x)} \ge \sum_{i=1}^{k} \frac{p^*(x) q(x)}{q^*(x)} ,$$

but the left-hand side is equal to 1, yielding the desired result. □

With the preceding lemma in hand, we finish the proof of (3). Fix some adaptive Q-strategy $\hat{q}$. By Markov's inequality,

$$\hat{q}(A_{n,\varepsilon}) \le e^{-n(D(p^* \,\|\, q^*)-\varepsilon)}\, \mathbb{E}_{\hat{q}}\left[\frac{p^*(x_1)\cdots p^*(x_n)}{q^*(x_1)\cdots q^*(x_n)}\right]. \tag{6}$$

We now use the fact that, by Lemma 4, the sequence of likelihood ratios $\prod_{i=1}^n \frac{p^*(x_i)}{q^*(x_i)}$ is a supermartingale with respect to $\hat{q}$. In particular,

$$\begin{aligned}
&\mathbb{E}_{\hat{q}}\left[\frac{p^*(x_1)\cdots p^*(x_n)}{q^*(x_1)\cdots q^*(x_n)}\right] \\
&\quad= \mathbb{E}_{\hat{q}}\left[\frac{p^*(x_1)\cdots p^*(x_{n-1})}{q^*(x_1)\cdots q^*(x_{n-1})}\mathbb{E}_{\hat{q}_n(x_1,x_2,\ldots,x_{n-1})}\frac{p^*(x)}{q^*(x)}\right] \\
&\quad\le \mathbb{E}_{\hat{q}}\left[\frac{p^*(x_1)\cdots p^*(x_{n-1})}{q^*(x_1)\cdots q^*(x_{n-1})}\right] \\
&\quad\le \cdots \\
&\quad\le 1\,,
\end{aligned}$$

where in the third line we have applied Lemma 4 to the distribution $\hat{q}_n(x_1, x_2, \ldots, x_{n-1}) \in Q$, and then we have continued by induction. Combining this with (6) completes our verification of (3) and hence our proof of the theorem.

2.2 Chernoff information and symmetric hypothesis testing

Suppose again that we have two distributions p and q over a finite domain Ω. We also have n i.i.d. samples $X_1, X_2, \ldots, X_n$ from a distribution $r \in \{p, q\}$, and a Bayesian hypothesis: The samples come from p with probability π_p and from q with probability π_q. Consider a test $T_n \subseteq \Omega^n$. If $(X_1, X_2, \ldots, X_n) \in T_n$, we declare that the sample came from p.

Our goal is to minimize the expected error

$$\delta_n(T_n) := \pi_p\, p^n(T_n^c) + \pi_q\, q^n(T_n)\,.$$

In this case, the best achievable error exponent is

$$\gamma(p,q) := \lim_{n\to\infty} -\frac{1}{n} \min_{T_n \subseteq \Omega^n} \log \delta_n(T_n)\,.$$

Observe that the constants π_p and π_q do not affect $\gamma(p, q)$.

For $\lambda \in [0, 1]$, let us define

$$\Gamma^\lambda(p,q) := -\log \sum_{x\in\Omega} p(x)^\lambda q(x)^{1-\lambda}\,,$$

and for p and q distinct, let $\lambda(p, q)$ be the value of $\lambda \in [0, 1]$ that maximizes $\Gamma^\lambda(p, q)$. Finally, put $\Gamma^*(p, q) := \Gamma^{\lambda(p,q)}(p, q)$. We have the following characterization due to Chernoff (see, e.g., Theorem 11.9.1 of [13]).

THEOREM 5. *For any distributions p and q, one has*

$$\gamma(p,q) = \Gamma^*(p,q) = D(r \,\|\, q) = D(p \,\|\, r)\,,$$

where r is the distribution given by

$$r(x) := \frac{p(x)^{\lambda(p,q)} q(x)^{1-\lambda(p,q)}}{\sum_{y\in\Omega} p(y)^{\lambda(p,q)} q(y)^{1-\lambda(p,q)}}$$

We will prove a corresponding theorem in the adaptive setting. To this end consider again two closed, convex sets of distributions $P, Q \subseteq \mathbb{R}^\Omega$. Define the *adversarial two-sided error exponent*

$$\gamma_{\mathrm{adv}}(P,Q) := \lim_{n\to\infty} -\frac{1}{n} \min_{T_n\subseteq\Omega^n} \max_{\hat{p},\hat{q}} \log\left(\hat{p}(T_n^c) + \hat{q}(T_n)\right)$$

where the maximum is over all adaptive P-strategies $\hat{p}$ and adaptive Q-strategies $\hat{q}$.

THEOREM 6 (ADVERSARIAL CHERNOFF THEOREM). *For any finite domain Ω and closed, convex sets of distributions $P, Q \subseteq \mathbb{R}^\Omega$, we have*

$$\gamma_{\mathrm{adv}}(P,Q) = \min_{p\in P, q\in Q} \Gamma^*(p,q)\,.$$

PROOF. Assume P and Q are disjoint, since otherwise both $\gamma_{\mathrm{adv}}(P, Q)$ and $\Gamma^*(P, Q)$ are trivially equal to zero. Let $p^* \in P$ and $q^* \in Q$ denote the minimizers of $\Gamma^*(p, q)$ and put $\lambda^* = \lambda(p, q)$. First, we have

$$\gamma_{\mathrm{adv}}(P,Q) \le \gamma_{\mathrm{adv}}(\{p^*\},\{q^*\}) = \gamma(p^*,q^*) = \Gamma^*(p^*,q^*)\,,$$

where the latter equality is given by Theorem 6. Thus we are left to prove $\gamma_{\mathrm{adv}}(P, Q) \ge \Gamma^*(p^*, q^*)$.

To this end, for $n \in \mathbb{N}$, define

$$T_n := \left\{ x \in \Omega_n : \prod_{i=1}^n p^*(x_i) \ge \prod_{i=1}^n q^*(x_i) \right\}.$$

Fix also an adaptive P-strategy $\hat{p}$ and an adaptive Q-strategy $\hat{q}$. We will show that

$$\lim_{n\to\infty} \frac{-\log(\hat{p}(T_n^c) + \hat{q}(T_n))}{n} \le \Gamma^*(p^*,q^*)\,. \tag{7}$$

We will need to employ the following easy variant of the "envelope theorem."

LEMMA 7. *Consider a differentiable function $f : [0, 1]^2 \to \mathbb{R}$. Define $V(t) = \inf_{\lambda\in[0,1]} f(\lambda, t)$ and suppose that for every $t \in [0, 1]$, there is a unique $\lambda^*(t) \in (0, 1)$ such that $V(t) = f(\lambda^*(t), t)$. If λ^* is differentiable at $t \in [0, 1]$, then $V'(t) = f_2(\lambda^*(t), t)$ where f_2 is the partial derivative of f with respect to its second argument.*

PROOF. Let f_1 denote the partial derivative of f with respect to its first argument. Writing $V(t) = f(\lambda^*(t), t)$ and applying the chain rule yields

$$V'(t) = f_2(\lambda^*(t), t) + f_1(\lambda^*(t), t)\frac{d}{dt}\lambda^*(t).$$

The second term is zero because $f_1(\lambda^*(t), t) = 0$ by optimality of $\lambda^*(t)$. □

REMARK 8. *Observe that if $f(\lambda, t)$ has $\frac{\partial^2}{\partial\lambda^2} f(\lambda, t) > 0$ for some $t \in [0, 1]$, then $\lambda^*(t)$ is the unique solution of $\frac{\partial}{\partial\lambda} f(\lambda, t) = 0$ and is differentiable by the implicit function theorem. Note that the assumptions of Lemma 7 can be relaxed considerably; see, e.g., [29, Ch. 3].*

This allows us to prove the following.

LEMMA 9. *For any distribution $q \in Q$, one has*

$$\sum_{x\in\Omega} q(x)\frac{p^*(x)^{\lambda^*}}{q^*(x)^{\lambda^*}} \le \sum_{x\in\Omega} q^*(x)\frac{p^*(x)^{\lambda^*}}{q^*(x)^{\lambda^*}}\,.$$

PROOF. For $t \in [0,1]$, define a distribution $q_t := tq + (1-t)q^* \in Q$. Moreover, define a function $f : [0,1]^2 \to \mathbb{R}$ by

$$f(\lambda, t) = \sum_{x \in \Omega} q_t(x)^{1-\lambda} p^*(x)^{\lambda} .$$

Observe that since $q_t \neq p^*$ for any t, we have

$$\frac{\partial^2}{\partial \lambda^2} f(\lambda, t) = \sum_x q_t(x)^{1-\lambda} p^*(x)^{\lambda} \left(\ln \left(\frac{p^*(x)}{q_t(x)} \right) \right)^2 > 0$$

for all $t \in [0,1]$ and $\lambda \in (0,1)$. Moreover, for fixed t, the minimum of $f(\lambda, t)$ is achieved for some $\lambda \in (0,1)$.

Let f_2 be the partial derivative of f in its second argument; then one computes:

$$f_2(\lambda, t) = \sum_{x \in \Omega} (q(x) - q^*(x))(1-\lambda) q_t(x)^{-\lambda} p^*(x)^{\lambda} .$$

If we let $V(t) = \min_{\lambda \in (0,1)} f(\lambda, t)$, then optimality of q^* implies $V'(0) \leq 0$. But now Lemma 7 (in conjunction with Remark 8) yields

$$\begin{aligned} 0 \geq V'(0) &= f_2(\lambda^*, 0) \\ &= \sum_{x \in \Omega} (q(x) - q^*(x))(1-\lambda^*) q^*(x)^{-\lambda^*} p^*(x)^{\lambda^*} . \end{aligned}$$

Rearranging yields the desired claim. □

The preceding lemma shows that the sequence $\prod_{i=1}^n \frac{p^*(x_i)^{\lambda^*}}{q^*(x_i)^{\lambda^*}}$ is a supermartingale with respect to $\hat{q}$. Thus we can write

$$\begin{aligned} &\mathbb{E}_{\hat{q}} \left[\prod_{i=1}^n \frac{p^*(x_i)^{\lambda^*}}{q^*(x_i)^{\lambda^*}} \right] \\ &= \mathbb{E}_{\hat{q}} \left[\prod_{i=1}^{n-1} \frac{p^*(x_i)^{\lambda^*}}{q^*(x_i)^{\lambda^*}} \mathbb{E}_{\hat{q}_n(x_1,\ldots,x_{n-1})} \frac{p^*(x_n)^{\lambda^*}}{q^*(x_n)^{\lambda^*}} \right] \\ &\leq e^{-\Gamma^*(p^*,q^*)} \mathbb{E}_{\hat{q}} \left[\prod_{i=1}^{n-1} \frac{p^*(x_i)^{\lambda^*}}{q^*(x_i)^{\lambda^*}} \right] \\ &\leq \cdots \\ &\leq e^{-n\Gamma^*(p^*,q^*)} , \end{aligned}$$

where in the third line we have used Lemma 9 along with the fact that $q = \hat{q}_n(x_1, \ldots, x_{n-1}) \in Q$, and then we have continued by induction.

By Markov's inequality, this implies $\hat{q}(T_n) \leq e^{-n\Gamma^*(p^*,q^*)}$. By the symmetry of the preceding argument with respect to P and Q, the same bound of $\hat{p}(T_n^c) \leq e^{-n\Gamma^*(p^*,q^*)}$ holds for $\hat{p}$. Combining these yields $\gamma_{\mathrm{adv}}(P,Q) \geq \Gamma^*(p^*,q^*)$, completing the proof.

3. DISTINGUISHING QUANTUM STATES WITH RESTRICTED MEASUREMENTS

A central problem in quantum information is to distinguish between a pair of quantum states ρ and σ. Necessary background and definitions for the reader unfamiliar with quantum information theory can be found in Appendix A. As usual, there is a tradeoff between errors of type 1 and 2, i.e., mistaking ρ for σ and vice versa. The quantum Neyman-Pearson lemma states that the optimal tradeoff curve between errors of type 1 and 2 is achieved by choosing

$$\mathcal{M} = \{\theta\rho - \sigma \geq 0\},$$

for some $\theta \geq 0$, where $\{X \geq 0\}$ denotes the projector onto the eigenvectors of X with nonnegative eigenvalue. The estimation strategy is then to perform the measurement $\{\mathcal{M}, I - \mathcal{M}\}$ and guess ρ upon obtaining outcome $\mathcal{M}$ or σ upon obtaining outcome $I - \mathcal{M}$.

One well-known case is when ρ and σ have prior probabilities p and $1-p$, respectively, and we wish to minimize the total probability of error. In this case the optimal $\mathcal{M}$ is given by $\mathcal{M} = \{p\rho - (1-p)\sigma \geq 0\}$, and the probability of error is $\frac{1 - \|p\rho - (1-p)\sigma\|_1}{2}$, where $\|\cdot\|_1$ denotes the Schatten 1-norm. The familiar *trace distance* $\frac{1}{2}\|\rho - \sigma\|_1$ corresponds to the case $p = 1/2$.

We modify this basic problem of state distinguishability in three (simultaneous) ways:

1. We consider only measurements $\mathcal{M}$ from some restricted class M.
2. We allow ρ, σ to be drawn adversarially from some sets R, S, respectively.
3. We consider the asymptotic limit in which M, R, S are replaced by families $\mathbf{M} = (M_1, M_2, \ldots), \mathbf{R} = (R_1, R_2, \ldots), \mathbf{S} = (S_1, S_2, \ldots)$ with M_n, R_n, S_n describing measurements and states on $V^{\otimes n}$. Our goal is then, for each n, to find a measurement $\mathcal{M} \in M_n$ that will effectively distinguish any state $\rho \in R_n$ from any state $\sigma \in S_n$.

These changes render the problem a good deal more abstract, and introduce a large number of new parameters. Thus, it may be helpful to keep in mind a prototypical example that was one of the motivations for this work. For some fixed bipartite state ρ over $A \otimes B$, let R_n be the singleton set $\{\rho^{\otimes n}\}$, and let $S_n := \mathrm{Sep}(A^{\otimes n} : B^{\otimes n})$. This corresponds to studying the asymptotic distinguishability of many copies of ρ from a separable state on the same number of systems. For this special case, we introduce the notation $\boldsymbol{\rho} := (\{\rho\}, \{\rho^{\otimes 2}\}, \ldots)$ and $\mathbf{Sep}(A : B) := (\mathrm{Sep}(A : B), \mathrm{Sep}(A^{\otimes 2} : B^{\otimes 2}), \ldots)$. Where the context is understood, we will often omit the reference to A, B and simply write Sep or **Sep**. Finally, we will consider a restricted class of measurements **M**, such as the class of 1-LOCC measurements (as discussed in [33, 9, 25, 8]).

3.1 Background on restricted quantum measurements

We begin by introducing notation, describing known results on restricted-measurement distinguishability, and presenting a few small new results to help clean up the landscape. In Section 3.2, we describe our restricted-measurement version of the quantum Stein's Lemma, and in Section 3.3 we give an application to quantum conditional mutual information.

3.1.1 *Quantum Stein's Lemma*

Let us first introduce some definitions analogous to the classical setup discussed in Section 2. We replace our finite domain Ω with a finite-dimensional vector space V, and denote the set of density matrices over V by $\mathcal{D}(V)$. Often we will be interested in the case where V is a composite system, e.g., a bipartite space $A \otimes B$. If ρ, σ are density matrices on V, then the *relative entropy of ρ with respect to σ* is

$$D(\rho \,\|\, \sigma) := \mathrm{tr}\left(\rho(\log \rho - \log \sigma)\right). \tag{8}$$

If $\ker(\sigma) \not\subseteq \ker(\rho)$, we take $D(\rho \,\|\, \sigma) := \infty$.

Following the classical case, we define an *acceptance operator* $\mathcal{M}_n$ (analogous to the acceptance region T_n) to be an operator on $V^{\otimes n}$ satisfying $0 \leq \mathcal{M}_n \leq I$ (i.e., a POVM element), with corresponding error probabilities $\alpha_n = \operatorname{tr}\left((I - \mathcal{M}_n)\rho^{\otimes n}\right)$ and $\beta_n := \operatorname{tr}\left(\mathcal{M}_n \sigma^{\otimes n}\right)$. Again we can define $\beta_n^\varepsilon := \min\{\beta_n : \alpha_n < \varepsilon\}$ and

$$E(\rho, \sigma) := \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{-\log \beta_n^\varepsilon}{n} \tag{9}$$

Hiai and Petz [19] proved the following quantum analogue of Lemma 1:

$$D(\rho \,\|\, \sigma) = E(\rho, \sigma). \tag{10}$$

See also [5, 24] for elegant and elementary proofs. The "strong converse" of (10) was proved by Ogawa and Nagaoka [31], and can be thought of as showing that (10) holds when the limit of $\varepsilon \to 0$ in (9) is replaced by any fixed $\varepsilon \in (0, 1)$.

3.1.2 Asymptotic composite hypothesis testing

An important generalization of hypothesis testing is when ρ and σ are chosen from sets $R, S \subseteq \mathcal{D}(V)$, respectively, and we need to design our test with knowledge only of R and S. This problem is known as *composite hypothesis testing* and is closely related to the classical Sanov's theorem. In [4, 18], it is proved that the best error exponent when R is general and S is the singleton set $S = \{\sigma\}$ is given by

$$D(R \,\|\, \sigma) := \min_{\rho \in R} D(\rho \,\|\, \sigma)\,. \tag{11}$$

One case of particular interest to quantum information is when $\rho \in \mathcal{D}(A \otimes B)$ and S is the set of separable states on $A \otimes B$, i.e., $S = \operatorname{Sep}(A : B)$. The quantity $D(\rho \,\|\, \operatorname{Sep}) := D(\rho \,\|\, \operatorname{Sep}(A : B))$ is known as the *relative entropy of entanglement* [38] and has been widely studied as an entanglement measure (see, e.g., Table I in [9]); note that it is usually written as $E_R(\rho)$.

One challenge in working with the relative entropy of entanglement is that $D(\rho^{\otimes n} \,\|\, \operatorname{Sep})$ will not in general be equal to $n \cdot D(\rho \,\|\, \operatorname{Sep})$, reflecting the fact that $\operatorname{Sep}(A^{\otimes n} : B^{\otimes n})$ is larger than the convex hull of $\{\sigma_1 \otimes \cdots \otimes \sigma_n : \sigma_1, \ldots, \sigma_n \in \operatorname{Sep}(A : B)\}$. Intuitively, $\operatorname{Sep}(A^{\otimes n} : B^{\otimes n})$ can be thought of as the set of states on the $2n$ systems $A_1 \ldots A_n B_1 \ldots B_n$ which are separable across the $A_1 \ldots A_n : B_1 \ldots B_n$ cut, but may be entangled arbitrarily among the A systems and among the B systems. This is an example of the quantum-information phenomenon known as the *additivity* problem (see, e.g., [41, 36]).

Definition 1. *Let* $\mathbf{R} = (R_1, R_2, \ldots)$, $\mathbf{S} = (S_1, S_2, \ldots)$, *with* $R_n, S_n \subseteq \mathcal{D}(V^{\otimes n})$. *Then the* asymptotic relative entropy *of* $\mathbf{R}$ *with respect to* $\mathbf{S}$ *is*

$$D(\mathbf{R} \,\|\, \mathbf{S}) := \lim_{n \to \infty} \inf_{\substack{\rho \in R_n \\ \sigma \in S_n}} \frac{D(\rho \,\|\, \sigma)}{n}. \tag{12}$$

We further define

$$\alpha_n(\mathcal{M}) := \sup_{\rho \in R_n} \operatorname{tr}\left((I - \mathcal{M})\rho\right) \tag{13}$$

$$\beta_n(\mathcal{M}) := \sup_{\sigma \in S_n} \operatorname{tr}\left(\mathcal{M}\sigma\right) \tag{14}$$

$$\beta_n^\varepsilon := \inf\{\beta_n(\mathcal{M}) : \alpha_n(\mathcal{M}) < \varepsilon\} \tag{15}$$

$$E(\mathbf{R}, \mathbf{S}) := \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{-\log \beta_n^\varepsilon}{n} \tag{16}$$

Note that the limits of Eq. (12) (resp. Eq. (16)) may not exist, in which case we leave $D(\mathbf{R} \,\|\, \mathbf{S})$ (resp. $E(\mathbf{R}, \mathbf{S})$) undefined. See [6] for a discussion of replacing the lim with lim inf or lim sup.

An important special case of Eq. (12) is the *regularized relative entropy of entanglement* [37], which is defined to be $\lim_{n \to \infty} \frac{1}{n} D(\rho^{\otimes n} \,\|\, \operatorname{Sep})$, and is normally denoted $E_R^\infty(\rho)$. In our notation this quantity is given by

$$D(\boldsymbol{\rho} \,\|\, \mathbf{Sep}), \tag{17}$$

In terms of Definition 1, the result of [4, 18] can be expressed as

$$D(\mathbf{R} \,\|\, \mathbf{S}) = E(\mathbf{R}, \mathbf{S}), \tag{18}$$

whenever $\mathbf{R}, \mathbf{S}$ are of the form $R_n = \{\rho^{\otimes n} : \rho \in R_1\}$ and $S_n = \{\sigma^{\otimes n}\}$, for some set R_1 and some state σ. We call results of the form (18) "quantum Stein's Lemmas," because, like the classical Chernoff-Stein Lemma, they give an equality between a relative entropy and an error exponent for hypothesis testing.

A quantum Stein's Lemma has also been proven in the case when $\mathbf{R} = \boldsymbol{\rho}$ for a fixed state ρ and $\mathbf{S}$ is a family of sets. In this case, (18) is proved in [7] in the case where $\mathbf{S}$ is a *self-consistent* family of states, meaning that:

1. Each S_n is convex and closed.
2. There exists a full-rank state σ such that each S_n contains $\sigma^{\otimes n}$.
3. For each $\sigma \in S_n$, $\operatorname{tr}_n \sigma \in S_{n-1}$.
4. If $\sigma_n \in S_n, \sigma_m \in S_m$ then $\sigma_n \otimes \sigma_m \in S_{n+m}$.
5. S_n is closed under permutation.

Some important cases of self-consistent families of states are $\mathbf{Sep}$ (defined in Section 3.1.1), PPT (defined in Appendix A, although it will not be used in this paper) and $\boldsymbol{\sigma}$ for any full-rank state σ.

3.1.3 Hypothesis testing with restricted measurements

We now introduce the problem of quantum hypothesis testing with restricted measurements. In general, two-outcome measurements on $V^{\otimes n}$ have the form $\{\mathcal{M}, I - \mathcal{M}\}$ where $0 \leq \mathcal{M} \leq I$. However, it is often useful to consider smaller classes of measurements, such as those that two parties can perform with local operations and classical communication (LOCC).

Definition 2. *Let* $\mathbf{R} = (R_1, R_2, \ldots)$, $\mathbf{S} = (S_1, S_2, \ldots)$, *with* $R_n, S_n \subseteq \mathcal{D}(V^{\otimes n})$, *and* $\mathbf{M} = (M_1, M_2, \ldots)$, *with* M_n *a set of measurements on* $\mathcal{D}(V^{\otimes n})$. *Then the* asymptotic relative entropy *of* $\mathbf{R}$ *with respect to* $\mathbf{S}$ *under measurements* $\mathbf{M}$ *is*

$$D_{\mathbf{M}}(\mathbf{R} \,\|\, \mathbf{S}) := \lim_{n \to \infty} \sup_{\mathcal{M} \in M_n} \inf_{\substack{\rho \in R_n \\ \sigma \in S_n}} \frac{D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma))}{n}. \tag{19}$$

We further define

$$\alpha_n(\mathcal{M}) := \sup_{\rho \in R_n} \operatorname{tr}\left((I - \mathcal{M})\rho\right) \tag{20}$$

$$\beta_n(\mathcal{M}) := \sup_{\sigma \in S_n} \operatorname{tr}\left(\mathcal{M}\sigma\right) \tag{21}$$

$$\beta_n^\varepsilon(\mathbf{M}) := \inf_{\mathcal{M} \in \mathbf{M}} \{\beta_n(\mathcal{M}) : \alpha(\mathcal{M}) < \varepsilon\} \tag{22}$$

$$E_{\mathbf{M}}(\mathbf{R}, \mathbf{S}) := \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{-\log \beta_n^\varepsilon}{n} \tag{23}$$

As before, the quantities (19) and (23) are left undefined when the corresponding limit does not exist.

Following our notation for families of states, we use boldface (e.g. $\mathbf{M}$) to denote families of measurements. In particular, we define $\mathrm{SEP}(A : B)$ to denote separable measurements on $A : B$ (i.e. M of the form $\sum_i X_i \otimes Y_i$ with $X_i, Y_i \geq 0$) and denote the corresponding family by

$$\mathbf{SEP}(A : B) = (\mathrm{SEP}(A : B), \mathrm{SEP}(A^{\otimes 2} : B^{\otimes 2}), \ldots).$$

Again we will often write SEP or **SEP** where the systems A, B are clear from context. Note that $\mathrm{Sep}(A : B)$ and $\mathrm{SEP}(A : B)$ both refer to sets of matrices that can be written as $\sum_i X_i \otimes Y_i$ with $X_i, Y_i \geq 0$; the difference is that Sep refers to density matrices (i.e. matrices with trace one) and SEP to measurement outcomes (i.e. matrices with operator norm ≤ 1).

Another important class of measurements is ALL, which is simply the set of all valid quantum measurements: i.e. $\mathsf{ALL} = \{0 \leq M \leq I\}$. The corresponding family is denoted ALL.

One further definition we will need (following [33], but with different notation) is the idea of a *compatible pair*.

Definition 3. *For $\mathbf{M}$ is a collection of measurements and $\mathbf{S}$ is a collection of states, we say that $(\mathbf{M}, \mathbf{S})$ are a compatible pair if applying a measurement in $\mathbf{M}$ to a state in $\mathbf{S}$ and conditioning on any outcome leaves a residual state that is still in $\mathbf{S}$. In other words, for any positive integers n, k, applying a measurement in M_k to S_{n+k} and conditioning on any outcome leaves a state that is still in S_n.*

For example $(\mathbf{SEP}, \mathbf{Sep})$ is a compatible pair. Our main results (in Sections 3.2 and 3.4) involve compatible pairs, and we also discuss previously known results about compatible pairs in Section 3.1.5.

3.1.4 Relations between distinguishability measures

Finally, we state some known and new results that relate the different versions of $D, E, D_{\mathbf{M}}, E_{\mathbf{M}}$. The following statement is a consequence of the minimax theorem.

Lemma 10. *Let R, S and M be closed convex sets. Then*

$$\max_{\mathcal{M} \in M} \min_{\substack{\rho \in R \\ \sigma \in S}} D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma)) = \min_{\substack{\rho \in R \\ \sigma \in S}} \max_{\mathcal{M} \in M} D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma)) \tag{24}$$

The notation $\mathcal{M}(rho)$ refers to the probability distribution of measurement outcomes resulting from applying $\mathcal{M}$ to ρ.

Proof. The proof is an application of Sion's minimax theorem [35]. The function $f(\mathcal{M}, (\rho, \sigma)) := D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma))$ is jointly convex in ρ and σ [27]. However, to apply minimax we also need that it is quasi-concave in the measurement $\mathcal{M}$. In order to do so, we linearize the function in the measurement by maximizing over the set of probability measures on M instead. Let $\mathcal{P}(M)$ be the set of probability measures over M. We have,

$$\begin{aligned}
&\max_{\mathcal{M} \in M} \min_{\substack{\rho \in R \\ \sigma \in S}} D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma)) \\
&\quad = \max_{\mu \in \mathcal{P}(M)} \min_{\substack{\rho \in R \\ \sigma \in S}} \mathbb{E}_{\mathcal{M} \sim \mu} D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma)) \\
&\quad = \min_{\substack{\rho \in R \\ \sigma \in S}} \max_{\mu \in \mathcal{P}(M)} \mathbb{E}_{\mathcal{M} \sim \mu} D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma)) \\
&\quad = \min_{\substack{\rho \in R \\ \sigma \in S}} \max_{\mathcal{M} \in M} D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma)),
\end{aligned} \tag{25}$$

where the second equality follows from Sion's minimax theorem applied to the function $g(\mu, (\rho, \sigma)) := \mathbb{E}_{\mathcal{M} \sim \mu} D(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma))$, which is linear in μ and convex in (ρ, σ). □

Known facts: The following relations between the quantities have been derived previously.

$$E(\boldsymbol{\rho}, \boldsymbol{\sigma}) = D(\rho \,\|\, \sigma) \tag{26}$$

quantum Stein's Lemma [19]

$$D(\{\rho\} \,\|\, S_1) \geq D(\boldsymbol{\rho} \,\|\, \mathbf{S}\}) \tag{27}$$

for $\mathbf{S}$ satisfying self-consistency property (4)

$$D(\mathbf{R} \,\|\, \mathbf{S}) \geq D_{\mathbf{M}}(\mathbf{R} \,\|\, \mathbf{S}) \tag{28}$$

from monotonicity of relative entropy

$$E(\boldsymbol{\rho}, \mathbf{S}) = D(\boldsymbol{\rho} \,\|\, \mathbf{S}) \tag{29}$$

for S a self-consistent class [7]

We can, in fact, relate D_{ALL}, D, E using

$$D_{\mathsf{ALL}}(\cdot \,\|\, \mathbf{S}) \overset{(41)}{\geq} E(\cdot, S) \overset{(29)}{=} D(\cdot \,\|\, \mathbf{S}) \overset{(28)}{\geq} D_{\mathsf{ALL}}(\cdot \,\|\, \mathbf{S}) \tag{30}$$

3.1.5 Superadditivity

When we consider families of states and measurements, it is not *a priori* clear whether the distinguishability per system should increase or decrease with the number of systems. We say that a quantity $f(\rho)$ is *subadditive* if $f(\rho_{XY}) \leq f(\rho_X) + f(\rho_Y)$ (e.g., entropy) and *superadditive* if $f(\rho_{XY}) \geq f(\rho_X) + f(\rho_Y)$ (e.g., most entanglement measures). A function f is weakly subadditive (resp. superadditive) if $f(\rho^{\otimes n})$ is $\leq n f(\rho)$ (resp., $\geq n f(\rho)$). If a function is both (weakly) subadditive and superadditive then we say it is (weakly) *additive*.

One of the main results known so far about relative entropy with restricted measurements is due to Piani [33], who used these measures to prove a superadditivity inequality:

$$D(\rho_{XY} \,\|\, S_2) \geq D_{\mathbf{M}}(\rho_X \,\|\, S_1) + D(\rho_Y \,\|\, S_1) \tag{31}$$

for compatible $(\mathbf{M}, S)$ [33]

$$D(\boldsymbol{\rho} \,\|\, \mathbf{S}) \geq D_{\mathbf{M}}(\rho \,\|\, S_1) \tag{32}$$

as a corollary of (31) [33]

In fact, Piani's result can easily be improved to show that $D_{\mathbf{M}}(\mathbf{R} \,\|\, \mathbf{S})$ is superadditive whenever $(\mathbf{M}, \mathbf{R})$ and $(\mathbf{M}, \mathbf{S})$ are compatible pairs.

Lemma 11. *Let $(\mathbf{M}, \mathbf{R})$ and $(\mathbf{M}, \mathbf{S})$ be compatible pairs. Then for all ρ_{XY}*

$$D_{M_2}(\rho_{XY} \,\|\, S_2) \geq D_{M_1}(\rho_X \,\|\, S_1) + D_{M_1}(\rho_Y \,\|\, S_1). \tag{33}$$

Moreover,

$$D_{\mathbf{M}}(\mathbf{R} \,\|\, \mathbf{S}) \geq D_{M_1}(R_1 \,\|\, S_1), \tag{34}$$

PROOF. The argument is a direct adaptation of the proof of Theorem 1 in [33].

Let $\rho_{XY}, \sigma_{XY} \in S_2$ be states and $\mathcal{M}_X, \mathcal{M}_Y \in M_1$ optimal measurements for $D_{M_1}(\rho_X \,\|\, S_1)$ and $D_{M_1}(\rho_Y \,\|\, S_1)$, respectively. Let k be the number of outcomes of $\mathcal{M}_X$, and let $e_1, \ldots, e_k$ be an orthonormal basis of $\mathbb{C}^k$. Then:

$$\begin{aligned}
&D\left((\mathcal{M}_X \otimes \mathcal{M}_Y)(\rho_{XY}) \,\|\, (\mathcal{M}_X \otimes \mathcal{M}_Y)(\sigma_{XY})\right) \\
&= D\left(\sum_{i=1}^k p_i(\rho_X) e_i e_i^* \otimes \mathcal{M}_Y(\rho_Y^i) \,\|\, \sum_{i=1}^k p_i(\sigma_X) e_i e_i^* \otimes \mathcal{M}_Y(\sigma_Y^i)\right) \\
&= D\left(p_i(\rho_X) \,\|\, p_i(\sigma_X)\right) + \sum_{i=1}^k p_i(\rho_X) D\left(\mathcal{M}_Y(\rho_Y^i) \,\|\, \mathcal{M}_Y(\sigma_Y^i)\right) \\
&\geq D\left(\mathcal{M}_X(\rho_X) \,\|\, \mathcal{M}_X(\sigma_X)\right) \\
&\quad + D\left(\sum_{i=1}^k p_i(\rho_X) \mathcal{M}_Y(\rho_Y^i) \,\|\, \sum_{i=1}^k p_i(\rho_X) \mathcal{M}_Y(\sigma_Y^i)\right) \\
&= D\left(\mathcal{M}_X(\rho_X) \,\|\, \mathcal{M}_X(\sigma_X)\right) \\
&\qquad + D\left(\mathcal{M}_Y(\rho_Y) \,\|\, \mathcal{M}_Y\left(\sum_{i=1}^k p_i(\sigma_X) \sigma_Y^i\right)\right),
\end{aligned} \tag{35}$$

where the first inequality follows from Proposition 1 of [33], the second by definition with $p_i(\rho_X) = \mathrm{tr}(M_X^i \rho_X)$ and $\rho_Y^i = \mathrm{tr}_X(M_X^i \otimes I_Y \rho_{XY})/p_i(\rho_X)$, the third from Lemma 1 of [33], and the fourth from property 2 of Proposition 1 of [33].

Since $(\mathbf{M}, \mathbf{R})$ and $(\mathbf{M}, \mathbf{S})$ are compatible, we can lower bound the last term of (35) by $D_{M_1}(\rho_X \,\|\, S_1) + D_{M_1}(\rho_Y \,\|\, S_1)$, from which (33) follows. (34), in turn, is a direct consequence of (33). □

The preceding lemma says that $D_{\mathbf{M}}(\cdot \,\|\, \mathbf{S})$ is superadditive for compatible pairs $(\mathbf{M}, \mathbf{S})$. The compatibility requirement here is essential. The pair $(\mathsf{ALL}, \mathbf{Sep})$ is not compatible, and here $D(\cdot \,\|\, \mathbf{S})$ is known to be strictly subadditive (i.e. not superadditive) in some cases [39].

On the other hand, $D_{\mathbf{M}}(\cdot \,\|\, \mathbf{S})$ can be strictly superadditive (i.e., not subadditive). Let us consider the simple situation in which $R_n = \{\rho^{\otimes n}\}$ and $S_n = \{\sigma^{\otimes n}\}$. It is a consequence of the quantum Stein's Lemma (10) that

$$D(\rho \,\|\, \sigma) = \lim_{n\to\infty} \frac{1}{n} D_{\mathsf{ALL}}(\rho^{\otimes n} \,\|\, \sigma^{\otimes n}).$$

Thus, any example in which

$$\max_{M \in \mathsf{ALL}} D(M(\rho) \,\|\, M(\sigma)) < D(\rho \,\|\, \sigma) \tag{36}$$

will yield an example in which $D_{\mathbf{M}}(\cdot \,\|\, \mathbf{S})$ is strictly superadditive. In fact, Lemma 1 of [32] states that (36) holds whenever $D(\rho \,\|\, \sigma)$ is finite and $\rho\sigma \neq \sigma\rho$. Thus superadditivity is a generic property of $D_{\mathbf{M}}(\cdot \,\|\, \cdot)$.

3.2 A quantum Stein's Lemma for restricted measurements

The following theorem can be thought of as a Quantum Stein's Lemma for restricted measurements.

THEOREM 12. *For any compatible pairs* $(\mathbf{M}, \mathbf{R})$ *and* $(\mathbf{M}, \mathbf{S})$,

$$D_{\mathbf{M}}(\mathbf{R} \,\|\, \mathbf{S}) = E_{\mathbf{M}}(\mathbf{R}, \mathbf{S}). \tag{37}$$

PROOF. For any positive integer k, let $E_k := D_{M_k}(R_k \,\|\, S_k)$, and choose some $\mathcal{M}_k \in M_k$ achieving the maximum in $D_{M_k}(R_k \,\|\, S_k)$. Let $P := \mathcal{M}_k(R_k)$ and $Q := \mathcal{M}_k(S_k)$. By our choice of $\mathcal{M}$, we have

$$D(p \,\|\, q) \geq E_k \qquad \forall p \in P, q \in Q. \tag{38}$$

Given a state $\rho \in \mathcal{D}(V^{\otimes nk})$, we apply $\mathcal{M}_k$ to each block of k systems, obtaining outcomes $x_1, \ldots, x_n$. Then since $(\mathbf{M}, \mathbf{R})$ and $(\mathbf{M}, \mathbf{S})$ are compatible pairs, the distribution of each x_i, conditioned on any possible value of $x_1, \ldots, x_{i-1}$, is an element of P (if $\rho \in R_{nk}$) or Q (if $\rho \in S_{nk}$). Thus, according to Theorem 2, there is an acceptance region that achieves the rate E_k. Thus

$$E_{\mathbf{M}}(\mathbf{R}, \mathbf{S}) \geq E_k. \tag{39}$$

Since (39) holds for any k, we obtain

$$E_{\mathbf{M}}(\mathbf{R}, \mathbf{S}) \geq D_{\mathbf{M}}(\mathbf{R} \| \mathbf{S}). \tag{40}$$

The reverse inequality can be obtained by the following standard argument: Let $\mathcal{M}_n \in M_n$ be an optimal sequence of measurements in $D_{\mathbf{M}}(\mathbf{R} \,\|\, \mathbf{S})$ and $\rho_n \in R_n$, and $\sigma_n \in S_n$ optimal sequences of states in $E_{\mathbf{M}}(\mathbf{R}, \mathbf{S})$. (Here "optimal" is in the sense of Lemma 10; i.e. $\mathcal{M}_n$ achieves the maximum on the LHS of (24) and ρ_n, σ_n achieve the minimum on the RHS of (24).) Then by monotonicity of relative entropy (see, e.g., [31]),

$$\begin{aligned}
D_{\mathbf{M}}(\mathbf{R} \,\|\, \mathbf{S}) &\geq \lim_{n\to\infty} \frac{D(\mathcal{M}_n(\rho_n) \,\|\, \mathcal{M}_n(\sigma_n))}{n} \\
&\geq -\lim_{n\to\infty} \frac{(1 - \alpha_n(\mathcal{M})) \log \beta_n(\mathcal{M})}{n} \\
&\geq E_{\mathbf{M}}(\mathbf{R}, \mathbf{S}).
\end{aligned} \tag{41}$$

□

This is analogous to the result in [7], which established $E(\boldsymbol{\rho}, \mathbf{S}) = D_{\mathsf{ALL}}(\boldsymbol{\rho} \,\|\, \mathbf{S})$ for self-consistent sets of states $\mathbf{S}$, but incomparable because in general $(\mathsf{ALL}, \mathbf{S})$ will not be a compatible pair.

3.3 Stronger Subadditivity of Quantum Entropy

We now present an application of Theorem 12 to a strengthening of the celebrated strong subadditivity inequality of Lieb and Ruskai for the quantum entropy [26], which can be written as

$$I(A : B \mid C)_\rho \geq 0 \tag{42}$$

where

$$\begin{aligned}
I(A : B \mid C)_\rho &:= H(AC)_\rho + H(BC)_\rho - H(ABC)_\rho - H(C)_\rho \\
&:= H(\rho_{AC}) + H(\rho_{BC}) - H(\rho_{ABC}) - H(\rho_C)
\end{aligned}$$

denotes the conditional mutual information of a state ρ_{ABC}. In what follows we will often omit the subscript ρ when the state is understood. See Appendix A for additional discussion.

In [9], the following lower bound was shown for any state ρ_{ABC}:

$$\begin{aligned}
&I(A : B \mid C) \geq \\
&D_{\mathsf{ALL}}(\boldsymbol{\rho}_{ABC} \,\|\, \mathbf{Sep}(A : BC)) - D_{\mathsf{ALL}}(\boldsymbol{\rho}_{AC} \,\|\, \mathbf{Sep}(A : C))
\end{aligned}$$

Moreover the following inequality was shown

$$D_{\mathsf{ALL}}(\boldsymbol{\rho}_{ABC} \,\|\, \mathbf{Sep}(A:BC)) - D_{\mathsf{ALL}}(\boldsymbol{\rho}_{AC} \,\|\, \mathbf{Sep}(A:C)) \geq E_{\text{1-LOCC}}(\boldsymbol{\rho}, \mathbf{Sep}(A:B)), \quad (43)$$

with 1-LOCC the class of all measurements that can be implemented by quantum local operations and classical communication from Bob to Alice (see Appendix A for the precise definition). This implies that the conditional mutual information is lower bounded by $E_{\text{1-LOCC}}(\boldsymbol{\rho}, \mathbf{Sep}(A:B))$.

In [25] the following strengthening of (43) was obtained:

$$D_{\mathsf{ALL}}(\boldsymbol{\rho}_{ABC} \,\|\, \mathbf{Sep}(A:BC)) \geq D_{\mathsf{ALL}}(\boldsymbol{\rho}_{AC} \,\|\, \mathbf{Sep}(A:C)) + D_{\text{1-LOCC}}(\boldsymbol{\rho}_{AB} \,\|\, \mathbf{Sep}(A:B)), \quad (44)$$

which implies

$$I(A:B \mid C) \geq D_{\text{1-LOCC}}(\boldsymbol{\rho}_{AB} \,\|\, \mathbf{Sep}(A:B)). \quad (45)$$

Theorem 12 shows that (44) is equivalent to (43) and so it can be used in conjunction with [9] to give an alternative proof of (45).

3.4 Symmetric hypothesis testing with restricted measurements

Our main result on symmetric hypothesis testing against an adaptive adversary (Theorem 6) implies a corresponding result for symmetric quantum hypothesis testing. For quantum states ρ, σ, define

$$\Gamma^*(\rho,\sigma) := \max_{0\leq\lambda\leq 1} \Gamma^\lambda(\rho,\sigma) := \max_{0\leq\lambda\leq 1} -\log \operatorname{tr}(\rho^\lambda \sigma^{1-\lambda})$$

$$\Gamma^*_{\mathbf{M}}(\mathbf{R},\mathbf{S}) := \lim_{n\to\infty} \sup_{\mathcal{M}\in M_n} \inf_{\substack{\rho\in R_n \\ \sigma\in S_n}} \frac{\Gamma^*(\mathcal{M}(\rho) \,\|\, \mathcal{M}(\sigma))}{n}$$

$$\gamma_{\mathbf{M}}(\mathbf{R},\mathbf{S}) := \lim_{n\to\infty} \sup_{\mathcal{M}\in M_n} \inf_{\substack{\rho\in R_n \\ \sigma\in S_n}} -\frac{1}{n} \log \operatorname{tr}(\mathcal{M}\sigma + (I-\mathcal{M})\rho)$$

A quantum analogue of the Chernoff Theorem was proven in [1] and in our notation can be expressed as

$$\gamma_{\mathsf{ALL}}(\boldsymbol{\rho},\boldsymbol{\sigma}) = \Gamma^*(\rho,\sigma).$$

Using the same idea behind the proof of Theorem 12, one can prove a restricted-measurement quantum Chernoff Theorem.

THEOREM 13. *If* $(\mathbf{M},\mathbf{R})$ *and* $(\mathbf{M},\mathbf{S})$ *are compatible pairs, then*

$$\gamma_{\mathbf{M}}(\mathbf{R},\mathbf{S}) = \Gamma^*_{\mathbf{M}}(\mathbf{R},\mathbf{S}).$$

The proof is essentially the same as that of Theorem 12 with the adversarial Chernoff-Stein's Lemma replaced by the adversarial Chernoff Theorem (Theorem 6). We omit the details.

3.5 Open questions

Having established a quantum Stein's Lemma for restricted measurements, we would like to know if a strong converse can also be proven, or more generally if we can calculate the error exponent for type-2 error when type-1 error is required to be $< \varepsilon$ for some fixed $\varepsilon \in (0,1)$. The difficulty is that $D_{\mathbf{M}}(\cdot \,\|\, \mathbf{S}) > D_{M_1}(\cdot \,\|\, S_1)$ in general, and we would need to control the rate of convergence as a function of n in the lim used to define $D_{\mathbf{M}}(\cdot \,\|\, \mathbf{S})$.

Like many information-theoretic quantities, $D(\boldsymbol{\rho} \,\|\, \mathbf{Sep})$ and $D_{\mathbf{M}}(\boldsymbol{\rho} \,\|\, \mathbf{Sep})$ (for various natural choices of $\mathbf{M}$) are operationally interesting, but are hard in practice to compute. We would like to know the complexity of estimating them (which is a variant of the usual question about the hardness of testing separability, cf. [16, 10]) and whether good relaxations exist (cf. [3]).

Finally, a major application of restricted-measurement distinguishability is to the related questions of k-extendable states[1], tripartite states with low conditional mutual information (i.e. "approximate Markov states", cf. [20]), and the quality of approximations achieved by the sum-of-squares hierarchy (cf. [2]). A few of the more prominent open questions here are:

- If $I(A:B \mid E)_\rho \leq \varepsilon$ then does there exist an "approximate recovery" map $T : E \to E \otimes B$ such that $(\mathrm{id} \otimes T)\rho_{AE} \approx_\delta \rho_{ABE}$, with $\delta \to 0$ as $\varepsilon \to 0$? (Here we use A, B, E both to denote quantum subsystems and the corresponding vector spaces.) This conjecture is due to Andreas Winter, and would imply an approximate equivalence between k-extendability and low conditional mutual information.

- How large can $D_{\mathbf{M}}(\rho \,\|\, \mathrm{Sep})$ be when ρ is k-extendable and $\mathbf{M}$ is the class of separable measurements? Sharp bounds are known [10] when $\mathbf{M}$ = 1-LOCC, and if they could be extended to separable measurements it would have implications for quantum Merlin-Arthur games with multiple Merlins [16] as well as for classical optimization algorithms.

- The ability of semidefinite programming hierarchies to estimate small-set expansion can be understood in terms of a restricted-measurement distinguishability problem [2]. A major open question is whether small-set expansion on graphs of size n can be well-approximated by $O(\log n)$ levels of these hierarchies, which would imply a quasipolynomial-time algorithm for the problem. Can tools from quantum information shed further light here?

APPENDIX

A. APPENDIX: BACKGROUND ON QUANTUM INFORMATION

This appendix contains a very brief review of the quantum formalism and notation used in this paper. For a much more detailed introduction to quantum information theory, see [40], or for an overview of the field of quantum computing and quantum information more generally see [30, 22].

Density matrices. The quantum analogue of a probability distribution over $[d] = \{1,\ldots,d\}$ is called a *density matrix*, or simply a *state*. Density matrices must be positive semi-definite and have trace one. These conditions are analogous to the requirement that probabilities must be

[1] A bipartite state ρ_{AB} is said to be k-extendable if there exists a state $\tilde{\rho}_{AB_1\ldots B_k}$ such that $\tilde{\rho}_{AB_i} = \rho_{AB}$ for each i. The idea of k-extendability was introduced in [34, 15], where it was proved that for any fixed dimension of A and/or B, the set of k-extendable states approaches the set of separable states. However, the rate of convergence is an open question.

nonnegative and normalized; indeed diagonal density matrices correspond exactly to probability distributions. If A is a complex vector space, then define $\mathcal{D}(A)$ to be the set of density matrices on A, meaning the set of operators on A that are positive semi-definite and have trace one. Let $\mathcal{L}(A, B)$ denote the set of linear operators from A to B, and let $\mathcal{L}(A) := \mathcal{L}(A, A)$.

Tensor product. To describe composite quantum systems, we use the tensor product. The tensor product of a vector $x \in \mathbb{C}^{d_1}$ and a vector $y \in \mathbb{C}^{d_2}$ is denoted $x \otimes y$ and has entries that run over all $x_{i_1} y_{i_2}$ for $i_1 \in [d_1], i_2 \in [d_2]$. Similarly, if X and Y are matrices, then their tensor product $X \otimes Y$ has matrix elements $(X \otimes Y)_{(i_1,i_2),(j_1,j_2)} = X_{i_1,j_1} Y_{i_2,j_2}$. For vector spaces A, B, we let $A \otimes B$ denote the span of $\{a \otimes b : a \in A, b \in B\}$. Note that $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2} \cong \mathbb{C}^{d_1 d_2}$. Finally, in each case we use the tensor power notation $X^{\otimes n}$ to stand for

$$\overbrace{X \otimes X \otimes \cdots \otimes X}^{n \text{ times}}.$$

Product and separable states. The tensor product is used to combine quantum states in the same way that independent classical probability distributions are combined to form a joint distribution. Indeed, if p, q are probability distributions of independent random variables, then $p \otimes q$ denotes the joint distribution. Similarly, if ρ and σ are density matrices, then $\rho \otimes \sigma$ denotes the state of a system that is in a so-called *product state*. The convex hull of the set of product states is called the set of *separable states*. We write $\mathrm{Sep}(A : B)$ to indicate the split along which we demand that the states be separable, e.g.

$$\mathrm{Sep}(A : B) = \mathrm{conv}\{\alpha \otimes \beta : \alpha \in \mathcal{D}(A), \beta \in \mathcal{D}(\beta)\}. \quad (46)$$

Although the set $\mathrm{Sep}(A : B)$ is convex, it is not easy to work with. For example, computational hardness results are known for the weak membership problem. Instead, it is sometimes more convenient to consider the relaxation PPT, which denotes the set of states with Positive Partial Transpose. The partial transpose operator Γ (meant to resemble the right half of the T that usually denotes transpose) acts linearly on $\mathcal{L}(A \otimes B)$ by mapping $X \otimes Y$ to $X \otimes Y^T$; equivalently we can write it as $\mathrm{id}_A \otimes T_B$, where id_A is the identity operator on $\mathcal{L}(A)$ and T_B is the transpose operator on $\mathcal{L}(B)$. We define $\mathrm{PPT}(A : B) = \{\rho \in \mathcal{D}(A \otimes B) : \rho^\Gamma \in \mathcal{D}(A : B)\}$. This set is easier to work with because it has a semidefinite-programming characterization. Moreover, it is straightforward to show that $\mathrm{Sep}(A : B) \subset \mathrm{PPT}(A : B)$. However, in general this inclusion is strict, and as the dimensions of A, B grow large, PPT can be an arbitrarily bad approximation for Sep [3].

Partial trace. Another concept from probability theory that we will need to generalize is the idea of a marginal distribution. Say we have a density matrix $\rho_{AB} \in \mathcal{D}(A \otimes B)$. The subscript emphasizes the systems which ρ describes, which are analogous to the random variables corresponding to a probability distribution. To obtain the state on only the A system, we apply the *partial trace* operator $\mathrm{tr}_B := \mathrm{id}_A \otimes \mathrm{tr}_B$ to ρ_{AB}. The action of the partial trace is often denoted by writing only the subscripts, as in

$$\rho_A := \mathrm{tr}_B \rho_{AB} \qquad \text{and} \qquad \rho_B := \mathrm{tr}_A \rho_{AB}. \quad (47)$$

(This notation generalizes; e.g. if $\rho \in \mathcal{D}(A \otimes B \otimes C)$, then $\rho_B = \mathrm{tr}_{AC} \rho_{ABC} = \mathrm{tr}_A \mathrm{tr}_C \rho_{ABC}$, etc.) Concretely, $(\rho_A)_{i,i'} = \sum_j (\rho_{AB})_{(i,j),(i',j)}$ and $(\rho_B)_{j,j'} = \sum_i (\rho_{AB})_{(i,j),(i,j')}$. We see that if ρ is diagonal then this coincides with the idea of a marginal distribution from classical probability theory.

Measurements. Although technically all of physics is described by quantum mechanics, it is often convenient to make a distinction between quantum information, which is often carried in very small systems such as single atoms or single photons, and classical information, which is carried in macroscopic systems, such as a bit in a classical RAM. The bridge from quantum state to probability distribution is given by a *measurement* (also sometimes called a POVM), which formally is a collection of matrices $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ satisfying $\mathcal{M}_i \geq 0$ for each i (meaning each $\mathcal{M}_i$ is positive semi-definite) and $\mathcal{M}_1 + \cdots + \mathcal{M}_k = I$. Performing the measurement $\mathcal{M}$ on state ρ yields outcome i with probability $\mathrm{tr}[\rho \mathcal{M}_i]$. Thus we can interpret $\mathcal{M}$ as a linear map from $\mathcal{L}(V)$ to $\mathbb{R}^k$, with the psd and normalization conditions serving to guarantee that $\mathcal{M}$ maps $\mathcal{D}(V)$ to valid probability distributions.

Measurements on multipartite states. For our purposes, we will consider a quantum state to be destroyed after it is measured. However, if we have a quantum state on multiple systems, such as $A \otimes B$, and we measure only system A, then we will still have a quantum state on system B. In this case, the probability of obtaining outcome i is $\mathbb{P}[i] = \mathrm{tr}[\mathcal{M}_i \rho_A]$ and the residual state in this case is

$$\frac{\mathrm{tr}_A[(\mathcal{M}_i \otimes I)\rho_{AB}]}{\mathbb{P}[i]}. \quad (48)$$

Since $\sum_i \mathcal{M}_i = I$, we can verify that if we average over all measurement outcomes, then system B is left in the state ρ_B, independent of the choice of measurement. This is an important feature of quantum mechanics; despite the possibility of entanglement, there is no way for Alice (who controls system A) to signal to Bob (who controls system B) through her choice of measurement.

Restricted classes of measurements. Consider a bipartite system $A \otimes B$, with systems A, B held by Alice and Bob respectively. Performing a general measurement on $A \otimes B$ may require that Alice and Bob exchange quantum messages, so it is often more practical for them to consider only measurements that they can perform using Local Operations and Classical Communication (LOCC). Although such restricted measurements were initially introduced to model these practical restrictions, they have since arisen in settings such as [9, 25] for completely different reasons. The class LOCC is difficult to work with and is cumbersome to even properly define—see [12] for a discussion—so we will often work with various restrictions or relaxations of it. A restriction which is interesting in its own right is the class 1-LOCC, which corresponds to Alice performing a measurement locally and sending the outcome to Bob. We say that $\mathcal{M} \in$ 1-LOCC if $\mathcal{M} = \{\mathcal{M}_{i,j}\}$ with $\mathcal{M}_{i,j} = X_i \otimes Y_{i,j}$, each $X_i, Y_{i,j} \geq 0$, $\sum_i X_i = I$ and for each i, $\sum_j Y_{i,j} = I$. On the other hand, a useful relaxation is the set SEP, for which each $\mathcal{M}_i$ should have the form $\mathcal{M}_i = \sum_j X_{i,j} \otimes Y_{i,j}$ with each $X_{i,j}, Y_{i,j} \geq 0$. An even further relaxation is PPT for which we demand only that each $\mathcal{M}_i^\Gamma \geq 0$ (apart from the

usual conditions that $\sum_i \mathcal{M}_i = I$ and each $\mathcal{M}_i \geq 0$). Finally we use ALL to denote the set of all measurements. Summarizing, we have

$$\text{1-LOCC} \subset \text{LOCC} \subset \text{SEP} \subset \text{PPT} \subset \text{ALL}.$$

In each case, we consider measurements with any finite number of outcomes, so these classes are technically not compact.

Entanglement swapping. An important concept in our work (building on [33]) is that of compatible pairs of families of measurements and states. We say that a measurement outcome $\mathcal{M}_i$ is compatible with a family of states $\mathbf{S}$ if for each n and each $\rho \in S_n$, applying $\mathcal{M}_i$ to the first system leaves a residual state (defined by (48)) that is in S_{n-1}. A family of measurements $\mathbf{M}$ is compatible with $\mathbf{S}$ if each outcome of each measurement in $\mathbf{M}$ is compatible with $\mathbf{S}$. If $\mathbf{S} = \text{Sep}$, then 1-LOCC, LOCC, SEP are all compatible with $\mathbf{S}$. If $\mathbf{S} = \text{PPT}$ then the set of compatible measurements includes PPT. However, it is easy to construct examples of incompatible pairs. Let $e_1, \ldots, e_d$ be an orthonormal basis for $\mathbb{C}^d$ and define $\Psi = \frac{1}{d}\sum_{i,j\in[d]} e_i \otimes e_j \otimes e_i \otimes e_j$. Observe that Ψ has entanglement between systems 1:3 and systems 2:4, but is product across the 13:24 cut. Now consider a measurement acting on systems 12. One can calculate that

$$\text{tr}_{12}[(\mathcal{M}_i \otimes I)\Psi\Psi^*] = \frac{\mathcal{M}_i^T}{d}. \tag{49}$$

Thus, if $\mathcal{M}_i^T$ is proportional to an entangled state, then the measurement can create entanglement on the previous unentangled states 3,4 that were not measured. This phenomenon—in which we start with $A_1 : A_2$ and $B_1 : B_2$ entanglement, measure A_1B_1 and end with $A_2 : B_2$ entanglement—is called entanglement swapping [21] and is one of the main new difficulties encountered in attempting to perform hypothesis testing with respect to classes such as Sep.

Entropy. The classical (Shannon) entropy of a distribution p is given by $H(p) = -\sum_i p_i \log(p_i)$. The quantum analogue is called the von Neumann entropy, and is given by $H(\rho) = -\text{tr}[\rho \log \rho]$. Observe that $H(\rho)$ is the Shannon entropy of the eigenvalues of ρ, and coincides with the Shannon entropy when we consider probability distributions to be diagonal density matrices. If ρ_{ABC} is a multipartite state, then we let $H(A)_\rho := H(\rho_A), H(AB)_\rho = H(\rho_{AB})$, etc. When ρ is understood, we may write simply $H(A), H(AB), \ldots$. Analogous to the classical mutual information, conditional entropy, etc. we can define

$$H(A \mid B) := H(AB) - H(B) \tag{50}$$

$$I(A : B) := H(A) + H(B) - H(AB) \tag{51}$$

$$I(A : B \mid C) := H(AC) + H(BC) - H(ABC) - H(C), \tag{52}$$

in each case with an implicit dependence on some state ρ. Finally, the quantum relative entropy is $D(\rho \,\|\, \sigma) := \text{tr}[\rho(\log \rho - \log \sigma)]$. Many of these quantities behave similarly to their classical analogues, but a number of new subtleties emerge; see Chapter 11 of [40] or Chapter 11 of [30] for more information.

Acknowledgments

We are grateful to Keiji Matsumoto for helpful conversations about hypothesis testing, and AWH and FGSLB also thank the Mittag-Leffler Institute for their hospitality while some of this work was done. FB was funded by EPSRC. AWH was funded by NSF grant CCF-1111382 and ARO contract W911NF-12-1-0486. JRL was supported by NSF grants CCF-1217256 and CCF-0905626.

B. REFERENCES

[1] K. Audenaert, J. Calsamiglia, L. Masanes, R. Munoz-Tapia, A. Acin, E. Bagan, and F. Verstraete. Discriminating states: The quantum Chernoff bound. *Phys. Rev. Lett.*, 98, 2007, arXiv:quant-ph/0610027.

[2] B. Barak, F. G. Brandão, A. W. Harrow, J. Kelner, D. Steurer, and Y. Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the 44th symposium on Theory of Computing*, STOC '12, pages 307–326, 2012, arXiv:1205.4484.

[3] S. Beigi and P. W. Shor. Approximating the set of separable states using the positive partial transpose test. *J. Math. Phys.*, 51(4):042202, 2010, arXiv:0902.1806.

[4] I. Bjelaković, J.-D. Deuschel, T. Krüger, R. Seiler, R. Siegmund-Schultze, and A. Szkoła. A quantum version of Sanov's theorem. *Commun. Math. Phys.*, 260(3):659–671, 2005, arXiv:quant-ph/0412157.

[5] I. Bjelaković and R. Siegmund-Schultze. Quantum Stein's lemma revisited, inequalities for quantum entropies, and a concavity theorem of Lieb, 2012, arXiv:quant-ph/0307170.

[6] G. Bowen and N. Datta. Beyond i.i.d. in quantum information theory, 2006, arXiv:quant-ph/0604013.

[7] F. G. Brandão and M. B. Plenio. A generalization of quantum Stein's lemma. *Commun. Math. Phys.*, 295:791, 2010, arXiv:0904.0281.

[8] F. G. S. L. Brandão and A. W. Harrow. Quantum de Finetti theorems under local measurements with applications, 2012, arXiv:1210.6367.

[9] F. G. S. L. Brandão, M. Christandl, and J. Yard. Faithful squashed entanglement. *Commun. Math. Phys.*, 306(3):805–830, 2011, arXiv:1010.1750.

[10] F. G. S. L. Brandão, M. Christandl, and J. Yard. A quasipolynomial-time algorithm for the quantum separability problem. pages 343–351, NY, USA, 2011. ACM New York, arXiv:1011.2751.

[11] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[12] E. Chitambar, D. Leung, L. Mancinska, M. Ozols, and A. Winter. Everything you always wanted to know about LOCC (but were afraid to ask), 2012, arXiv:1210.4583.

[13] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Series in Telecommunication. John Wiley and Sons, New York, 1991.

[14] D. P. DiVincenzo, P. W. Shor, and J. A. Smolin. Quantum channel capacity of very noisy channels. *Phys. Rev. A*, 57:830, 1998, arXiv:quant-ph/9706061.

[15] A. C. Doherty, P. A. Parrilo, and F. M. Spedalieri. Complete family of separability criteria. *Phys. Rev. A*, 69:022308, Feb 2004, arXiv:quant-ph/0308032.

[16] A. W. Harrow and A. Montanaro. An efficient test for product states, with applications to quantum Merlin-Arthur games. In *Proc. 51st Symp. on FOCS*, pages 633–642, 2010, arXiv:1001.0017.

[17] M. B. Hastings. A counterexample to additivity of minimum output entropy. *Nature Physics*, 5, 2009, arXiv:0809.3972.

[18] M. Hayashi. Optimal sequence of quantum measurements in the sense of Stein's lemma in quantum hypothesis testing. *J. Phys. A*, 35(50):10759–10773, 2002, arXiv:quant-ph/0208020.

[19] F. Hiai and D. Petz. The proper formula for relative entropy and its asymptotics in quantum probability. *Commun. Math. Phys.*, 143(1):99–114, 1991.

[20] B. Ibinson, N. Linden, and A. Winter. Robustness of quantum markov chains. *Commun. Math. Phys.*, 277(2):289–304, 2008, arXiv:quant-ph/0611057.

[21] M. Żukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert. "Event-ready-detectors" Bell experiment via entanglement swapping. *Phys. Rev. Lett.*, 71:4287–4290, Dec 1993.

[22] A. Y. Kitaev, A. H. Shen, and M. N. Vyalyi. *Classical and Quantum Computation*, volume 47 of *Graduate Studies in Mathematics*. AMS, 2002.

[23] E. Levitan and N. Merhav. A competitive Neyman-Pearson approach to universal hypothesis testing with applications. *Information Theory, IEEE Transactions on*, 48(8):2215–2229, 2002.

[24] K. Li. Second order asymptotics for quantum hypothesis testing, 2012, arXiv:1208.1400.

[25] K. Li and A. Winter. Relative entropy and squashed entanglement, 2012, arXiv:1210.3181.

[26] E. Lieb and M. Ruskai. Proof of the strong subadditivity of quantum-mechanical entropy. *J. Math. Phys.*, 14(12):1938, 1973.

[27] G. Lindblad. Expectations and entropy inequalities for finite quantum systems. *Comm. Math. Phys.*, 39:111–119, 1974.

[28] W. Matthews and S. Wehner. Finite blocklength converse bounds for quantum channels, 2012, arXiv:1210.4722.

[29] P. Milgrom. *Putting Auction Theory to Work*. Cambridge University Press, 2004.

[30] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, New York, 2000.

[31] T. Ogawa and H. Nagaoka. Strong converse and Stein's lemma in quantum hypothesis testing. *Information Theory, IEEE Transactions on*, 46(7):2428–2433, 2000, arXiv:quant-ph/9906090.

[32] D. Petz. Monotonicity of quantum relative entropy revisited. *Rev. Math. Phys.*, 15(01):79–91, 2003, arXiv:quant-ph/0209053.

[33] M. Piani. Relative entropy of entanglement and restricted measurements. *Phys. Rev. Lett.*, 103:160504, Oct 2009, arXiv:0904.2705.

[34] G. A. Raggio and R. F. Werner. Quantum statistical mechanics of general mean field systems. *Helv. Phys. Acta*, 62:980–1003, 1989.

[35] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958.

[36] G. Smith. Quantum channel capacities. In *Information Theory Workshop (ITW), 2010 IEEE*, pages 1–5, 2010, arXiv:1007.2855.

[37] V. Vedral and M. B. Plenio. Entanglement measures and purification procedures. *Phys. Rev. A*, 57:1619–1633, Mar 1998, arXiv:quant-ph/9707035.

[38] V. Vedral, M. B. Plenio, M. A. Rippin, and P. L. Knight. Quantifying entanglement. *Phys. Rev. Lett.*, 78:2275–2279, Mar 1997, arXiv:quant-ph/9702027.

[39] K. G. H. Vollbrecht and R. F. Werner. Entanglement measures under symmetry. *Phys. Rev. A*, 64:062307, Nov 2001, arXiv:quant-ph/0010095.

[40] M. M. Wilde. *Quantum Information Theory*. Cambridge University Press, 2013, arXiv:1106.1445.

[41] M. M. Wolf, T. S. Cubitt, and D. Perez-Garcia. Are problems in quantum information theory (un)decidable?, 2011, arXiv:1111.5425.