We study the non-asymptotic fundamental limits for transmitting classical information over memoryless quantum channels, i.e. we investigate the amount of classical information that can be transmitted when the channel is used a finite number of times and a constant average error is permissible. We show that, in the absence of entanglement between different channel uses, the non-asymptotic fundamental limit admits a Gaussian approximation that illustrates the speed at which the rate of optimal codes converges to the Holevo capacity as the blocklength tends to infinity. The behavior is governed by a new channel parameter, called channel dispersion, for which we provide a geometrical interpretation.

I. INTRODUCTION

One of the landmark achievements in quantum information theory is the establishing of the coding theorem for sending classical information across a noisy quantum channel by Holevo [18], and independently by Schumacher-Westmoreland [32]—the so-called HSW theorem. For a classical-quantum (c-q) channel, in which the encoder is restricted to prepare product states by picking them from a finite (and fixed) ensemble for each communication round, the HSW theorem can be formally stated as follows: Let $M^*(W^n, \varepsilon)$ denote the maximum size of a length-$n$ block code for the c-q $W$ with average error probability $\varepsilon \in (0, 1)$. Then, the HSW theorem, together with the weak converse established by Holevo [19] in the 1970s (the Holevo bound), asserts that

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \log M^*(W^n, \varepsilon) = C(W)$$

where $C(W) = \max_P I(P, W)$ is the Holevo capacity of the channel and $I(P, W)$ is the mutual information between the classical channel input $X \leftarrow P$ and the quantum system at the output of $W$. We define all quantities precisely in the following. Let us emphasize that products of c-q channels do not allow entangled inputs—their capacity is thus additive and Hastings’ counterexamples [13] are not applicable.

The converse part of HSW theorem was strengthened significantly by Ogawa-Nagaoka [24] and Winter [41] who proved the strong converse for discrete memoryless c-q channels, namely

$$\lim_{n \to \infty} \frac{1}{n} \log M^*(W^n, \varepsilon) = C(W), \quad \text{for all } \varepsilon \in (0, 1).$$

In the work by Ogawa-Nagaoka [24], the strong converse was proved using ideas from Arimoto’s strong converse proof [1] for classical channels, which itself was based on techniques to prove Gallager’s random coding error exponent [9]. Hence, Ogawa-Nagaoka [24] proof also applies to c-q channels whose inputs are not necessarily discrete. Winter’s strong converse proof [41], on the other hand, is based on the method of types [3] which is a powerful tool developed in classical information theory for discrete memoryless systems. We also mention the work by Hayashi-Nagaoka [15] in
which a necessary and sufficient condition was provided for the strong converse property to hold for general (not only memoryless) c-q channels.

We are interested in investigating the behavior of \( \log M^*(W^n, \varepsilon) \). This quantity characterizes the fundamental backoff from the Holevo capacity for finite block lengths \( n \) as a function of the error tolerance \( \varepsilon \) and is thus of immediate practical relevance when designing communication systems. In particular, we want to approximate \( \log M^*(W^n, \varepsilon) \) for large but finite \( n \). Note that Winter [41] in fact showed for discrete memoryless c-q channels that
\[
\log M^*(W^n, \varepsilon) = nC(W) + O(\sqrt{n}), \quad \text{for all } \varepsilon \in (0, 1).
\]

Our present work refines the \( O(\sqrt{n}) \) term by identifying the implied constant in this remainder term as a function of \( \varepsilon \) and a new channel parameter called the dispersion of the channel. The resulting Gaussian approximation generalizes results for classical channels that go back to Strassen’s work in the 1962 [35]. In this seminal work, he showed for most well-behaved discrete memoryless channels \( W : X \to Y \) that
\[
\log M^*(W^n, \varepsilon) = nC(W) + \sqrt{nV_\varepsilon(W)} \Phi^{-1}(\varepsilon) + O(\log n),
\]
where \( C(W) \) is the Shannon capacity, \( \Phi \) the cumulative normal Gaussian distribution, and \( V_\varepsilon(W) \) is another fundamental property of the channel known as the \( \varepsilon \)-channel dispersion, a term coined by Polyanskiy et al. [29]. Refinements to and extensions of the expansion of \( \log M^*(W^n, \varepsilon) \) were pursued by Hayashi [14], Polyanskiy et al. [29] and the present authors [37].

There are at least three main contributions in this paper, which we detail in Section III.

1. We go significantly beyond the discrete memoryless c-q setting and refine the asymptotic expansion of \( \log M^*(W^n, \varepsilon) \) for memoryless c-q whose input ensemble is neither discrete nor otherwise structured. In fact our only requirement is that the image of the channel is comprised of quantum states on a finite-dimensional Hilbert space. We then prove a quantum analogue of Strassen’s [35] refinement to the Shannon capacity in (1). This result is presented in Theorem 2 and discussed in Section III.B.

2. It is a well-known fact that the capacity of a classical or c-q channel can be represented geometrically as the divergence radius of the channel image. In this paper, in the course of proving our main result, and especially the converse part, we leverage on this fact heavily and refine the geometric interpretation in Section III.A.

We develop a one-shot converse bound on \( M^*(W, \varepsilon) \) in terms of the geometry of the image of the channel by employing a non-asymptotic quantity known as the \( \varepsilon \)-hypothesis testing divergence radius. This is a one-shot analogue of the divergence radius that is used to characterize the channel capacity. We find that such an approach allows to to shift our attention from the input to the output space already in the non-asymptotic regime. Indeed, all the necessary calculations to yield the Gaussian approximation are done in the output space, thus allowing the input space to be arbitrary.

This approach of working solely on the output space by leveraging on a non-asymptotic divergence radius to find the converse of the Gaussian approximation is new (to the best of the authors’ knowledge) and does not have a classical analogue.

3. Because of the generality that is being afforded in our setup, several auxiliary technical results have to be developed either by modifying arguments from the literature or proving them from scratch. These results may be of independent interest in other contexts. We mention some examples here.
First, we develop several alternative representations of the divergence radius that turn out to be amenable for computations involved in both the direct part and converse parts. Second, in the course of proving the direct part, we also show, by appealing to Caratheodory’s theorem, that it suffices to choose a finite input ensemble in order to achieve the Gaussian approximation. Third, for the converse part, to deal with ensembles of “bad” states that are not close to Holevo capacity-achieving, we construct an appropriate $\gamma$-net whose size can be controlled appropriately and whose elements serve to approximate those ensembles of “bad” states. Finally, we also prove several useful continuity properties of quantum information quantities. These allow us to establish that the third-order term in the Strassen-type asymptotic expansion in (1) for cq channels with discrete support is $O(\log n)$ instead of the weaker $o(\sqrt{n})$.

**II. PRELIMINARIES**

We consider the real vector space of self-adjoint (Hermitian) operators on a finite-dimensional inner product (Hilbert) space. We denote the space of self-adjoint operators by $\mathcal{H}$ and keep it fixed throughout to ease notation. For $A, B \in \mathcal{H}$, we write $A \geq B$ iff $A - B$ is positive semi-definite. Moreover, we denote by $\{A > B\}$ and $\{A \geq B\}$ the projects onto the positive and non-negative subspaces of $A - B$, respectively. We write $A \gg B$ to denote the fact that the kernel of $A$ is contained in the kernel of $B$. Let $\lambda_{\text{min}}(A)$ denote the minimum eigenvalue of $A$. We equip $\mathcal{H}$ with a metric, the trace distance $\delta_{\text{tr}}(A, B) := \frac{1}{2} \text{tr}|A - B|$, where tr denotes the trace. The identity operator is denoted by $id$. The set of quantum states is given by $S := \{\rho \in \mathcal{H} | \rho \geq 0 \land \text{tr}(\rho) = 1\}$. Clearly, $(S, \delta_{\text{tr}})$ is a (sequentially) compact metric space.

For any closed (and thus compact) subset $S_0 \subseteq S$, we denote by $\mathcal{P}(S_0)$ the set of probability measures on $(S_0, \Sigma_0)$, where $\Sigma_0$ is the Borel $\sigma$-algebra on $(S_0, \delta_{\text{tr}})$. Since $(S_0, \delta_{\text{tr}})$ is a compact metric space, $(\mathcal{P}(S_0), \delta_{\text{wc}})$ is a compact metric space, where $\delta_{\text{wc}}$ denotes the Prohorov metric [27, Sec. 6 and Thm. 6.4]. We will not use $\delta_{\text{wc}}$ explicitly but simply note that convergence in $\delta_{\text{wc}}$ is equivalent to weak convergence of probability measures. As such, any function of the form

$$\mathcal{P}(S_0) \rightarrow \mathbb{R}, \quad \mathbb{P} \mapsto \int_{S_0} d\mathbb{P}(\rho) f(\rho)$$

is continuous if $f$ is bounded and continuous.\(^1\) If $S_0$ is discrete, we abuse notation and also use $\mathcal{P}(S_0)$ to denote the set of discrete probability distributions on $S_0$. We then use $P \in \mathcal{P}(S_0)$ to denote its elements. We often use the abbreviations $\rho^{(\mathbb{P})}$ and $\rho^{(P)}$ to denote the averaged states

$$\rho^{(\mathbb{P})} := \int_{S_0} d\mathbb{P}(\rho) \rho \quad \text{and} \quad \rho^{(P)} := \sum_{\rho \in S_0} P(\rho) \rho.$$

For any $n \in \mathbb{N}$, we also consider the $n$-fold products of the underlying inner-product space and denote the associated set of self-adjoint operators and states with $\mathcal{H}^n$ and $S^n$, respectively. For any $S_0 \subseteq S$, we denote by $S_0^{\otimes n} \subseteq S^n$ the set of $n$-tuples of states in $S_0$, represented as a product state $\bigotimes_{i=1}^n \rho_i$, where $\rho_i \in S_0$. We have $S_0^{\otimes n} \subseteq S^n$ with equality only if $n = 1$ or $d = 1$.

We employ the cumulative distribution function of the standard normal distribution

$$\Phi(a) := \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} x^2 \right) dx$$

and define its inverse as $\Phi^{-1}(\varepsilon) := \sup\{a \in \mathbb{R} | \Phi(a) \leq \varepsilon\}$, which evaluates to the usual inverse for $0 < \varepsilon < 1$ and extends to take values $\pm \infty$ outside that range.

\(^1\) Note that in order for a sequence $\{\mathbb{P}_n\}_{n \in \mathbb{N}}$ to convergence weakly to $\mathbb{P}$ we must have $\lim_{n \to \infty} g(\mathbb{P}_n) = g(\mathbb{P})$. 

A. Codes and $\varepsilon$-Error Capacity

We consider general cq channels, i.e. arbitrary functions $W: \mathcal{X} \rightarrow \mathcal{S}$, where $\mathcal{X}$ is an arbitrary set. A special case of this is a quantum channel, namely a completely positive trace-preserving (CPTP) map $W: \mathcal{S}' \rightarrow \mathcal{S}$, where $\mathcal{S}'$ denotes a set of quantum states. We denote the image of the channel by

$$\text{im}(W) := \{\rho \in \mathcal{S} | \exists x \in \mathcal{X} : \rho = W(x)\},$$

and its closure by $\overline{\text{im}(W)}$. Without loss of generality, we may assume that $\text{im}(W)$ has full support on the underlying Hilbert space, i.e. every vector (of the underlying Hilbert space) is supported by at least one element in $\text{im}(W)$. Thus, we will usually set $d = |\text{supp}(\text{im}(W))|$.

A code $C$ for $W$ is defined by the triple $\{M, e, D\}$, where $M$ is a (discrete) set of messages, $e: M \rightarrow \mathcal{X}$ an encoding function and $D = \{Q_m\}_{m \in M}$ is a positive operator valued measure (POVM).\footnote{A POVM in this context is a set of operators $\{Q_m\}_{m \in M}$ satisfying $Q_m \geq 0$ for all $m \in M$ and $\sum_{m \in M} Q_m = \text{id}$.} We write $|C| = |M|$ for the cardinality of the message set. We define the average error probability of a code $C$ for the channel $W$ as

$$p_{\text{err}}(C, W) := \text{Pr}[M \neq M'] = 1 - \frac{1}{|M|} \sum_{m \in M} \text{tr}(W(e(m))Q^m)$$

where the distribution over messages $P_M$ is assumed to be uniform on $M$,

$$M \xrightarrow{e} X \xrightarrow{W} W(X) \xrightarrow{D} M'$$

forms a Markov chain, $W(X)$ denotes the (random) output of the channel, and $M'$ thus denotes the output of the decoder.

To characterize the non-asymptotic fundamental limit of data transmission over a single use of the channel, we define the maximum volume of a codebook for $W$ with average error $\varepsilon$ as

$$M^*(W, \varepsilon) := \max \{m \in \mathbb{N} | \exists C : |C| = m \land p_{\text{err}}(C, W) \leq \varepsilon\}.$$  

We are also interested in the $\varepsilon$-error capacity for $n \geq 1$ uses of a memoryless channel. The $n$-fold i.i.d. repetition of the channel, $W^n: \mathcal{X}^n \rightarrow \mathcal{S}^\otimes n$, takes as input a vector $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ and maps it to $W(x_1) \otimes W(x_2) \otimes \ldots \otimes W(x_n) \in \mathcal{S}^\otimes n$. In particular, our model thus does not allow for entangled channel outputs. The fundamental limit of data transmission over $n$ uses of the channel is consequently given by $M^*(W^n, \varepsilon)$.

B. Basic Information Quantities

The following basic quantities are of interest here. For any $\rho \in \mathcal{S}$, we employ the von Neumann entropy $H(\rho) := -\text{tr}(\rho \log \rho)$. Moreover, for positive semi-definite $\sigma$ satisfying $\sigma \gg \rho$, the relative entropy [17, 38] and the relative entropy variance [20, 36] are respectively defined as

$$D(\rho||\sigma) := \text{tr}(\rho(\log \rho - \log \sigma)) \quad \text{and}$$

$$V(\rho||\sigma) := \text{tr}(\rho(\log \rho - \log \sigma - D(\rho||\sigma) \cdot \text{id})^2).$$

As usual, we implicitly use the convention $0 \log^k 0 \equiv 0$ for all $k \in \mathbb{N}$. 
Classically, for two distributions $P, Q \in \mathcal{P}(\mathcal{X})$, the Kullback-Leibler divergence $D(P \parallel Q)$ is the expectation value of the log-likelihood ratio $\log \left( \frac{P(X)}{Q(X)} \right)$ where $X \leftarrow P$, and $V(P \parallel Q)$ is the corresponding variance. The above definition of $V(\rho \parallel \sigma)$ is thus a natural non-commutative generalization of the classical concept, with its operational meaning firmly established in [20, 36].

We summarize some properties of the above quantities, which we will employ later.

1. $\rho \mapsto H(\rho)$ is strictly concave (cf., e.g., Lemma 20) and continuous.
2. $(\rho, \sigma) \mapsto D(\rho \parallel \sigma)$ is jointly convex and lower semi-continuous. In fact, it is continuous except when it diverges to infinity, i.e. when $\sigma \not\gg \rho$.
3. $D(\rho \parallel \sigma)$ is positive definite, i.e. $D(\rho \parallel \sigma) \geq 0$ with equality iff $\rho = \sigma$.
4. $(\rho, \sigma) \mapsto V(\rho \parallel \sigma)$ is continuous except when $\sigma \not\gg \rho$.

III. MAIN RESULTS

A. The Divergence Radius of a Set of Quantum States

It is a well-known fact that the capacity of a classical or classical-quantum channel can be represented geometrically as the divergence radius of the channel image. (For the quantum case, see, e.g. [26] and [33].) Here, we take a complementary approach and investigate the divergence radius of subsets of the set of quantum states. If such a set is the image of a channel, our analysis allows us to construct capacity achieving ensembles by just looking at the channel image. Furthermore, this viewpoint leads to a natural quantum generalization of the concept of channel dispersion. Thus, somewhat surprisingly, we will see that not only the capacity but also the finite blocklength behavior of channels is governed by the geometry of the channel image.

1. Divergence Radius

Let us start by investigating the divergence radius of arbitrary closed subsets of the set of quantum states on a finite-dimensional Hilbert space.

Definition 1. Let $S_0 \subseteq S$ be closed. The divergence radius of $S_0$ (in $S$) is defined as

$$\chi(S_0) := \inf_{\sigma \in S} \sup_{\rho \in S_0} D(\rho \parallel \sigma).$$

We show the following properties of the divergence radius.

Theorem 1. Let $S_0 \subseteq S$ be closed. We find the following:

1. The divergence center, defined as $\sigma^*(S_0) := \arg \min_{\sigma \in S} \left\{ \sup_{\rho \in S_0} D(\rho \parallel \sigma) \right\}$, exists and is unique. Moreover, $\sigma^*(S_0) \gg \rho$ for all $\rho \in S_0$.
2. Define the set of peripheral points of $S_0$, i.e.

$$\Gamma(S_0) := \arg \max_{\rho \in S_0} D(\rho \parallel \sigma^*(S_0)).$$

Then, $D(\rho \parallel \sigma^*(S_0)) \leq \chi(S_0)$ for all $\rho \in S_0$ with equality iff $\rho \in \Gamma(S_0)$.
3. We have $\sigma^*(S_0) \in \text{conv}(\Gamma(S_0))$. 


FIG. 1. Example of a set $S_\circ$ with divergence center $\sigma^* = \sigma^*(S_\circ)$ and a partly continuous and disconnected set of peripheral points, $\Gamma = \Gamma(S_\circ)$. The set $\Gamma$ must lie on the boundary of $S_\circ$ due to the quasi-convexity of $\rho \mapsto D(\rho\|\sigma)$ (cf. Lemma 7). As seen in Theorem 1, the center $\sigma^*$ lies in the convex hull of $\Gamma(S_\circ)$ consistent with the Euclidean intuition.

4. The divergence radius has the following alternative representation:

$$\chi(S_\circ) = \sup_{P \in \mathcal{P}(S_\circ)} \int dP(\rho) D\left(\rho \big\| \int dP(\rho)\rho\right).$$

(4)

5. The set of probability measures that achieve the supremum is given by the peripheral decompositions of the divergence center, namely the compact convex set

$$\Pi(S_\circ) := \left\{ P \in \mathcal{P}(\Gamma(S_\circ)) \big| \int dP(\rho)\rho = \sigma^*(S_\circ) \right\}.$$  

(5)

Moreover, $\Pi(S_\circ)$ contains a discrete measure with support on at most $d^2$ points in $\Gamma(S_\circ)$.

The proof of this theorem is presented in Section IV.

Remark 1. Uniqueness of $\sigma^*(S_\circ)$ was also claimed by Ohya, Petz and Watanabe [26, Lem. 3.4] in a related context. However, they argue that this directly follows from the “fact that the relative entropy functional is strictly convex in the second variable”. We submit that more care has to be taken to establish uniqueness. Notably, the functional $\sigma \mapsto D(\rho\|\sigma)$ is only strictly convex if $\rho > 0$ is positive definite and trivial counterexamples can be constructed otherwise. It is then unclear how to apply this property directly to the situation at hand.

Remark 2. Property 3 is of particular importance for our argument and has not been shown before. A weaker property, namely $\sigma^*(S_\circ) \in \text{conv}(S_\circ)$ was already pointed out in [26, Lem. 3.4]. However, our stronger Property 3 implies that $\sigma^*(S_\circ)$ can be written as a convex combination of states in $\Gamma(S_\circ)$, i.e. $\sigma^*(S_\circ) = (P)$ for some $P \in \mathcal{P}(\Gamma(S_\circ))$. If $S_\circ$ is the image of a quantum channel $W$, we write $W^{-1}(\Gamma(S_\circ))$ to denote any pre-image of $\Gamma(S_\circ)$. Then, the tuple $\{P, W^{-1}(\Gamma(S_\circ))\}$ corresponds to an optimal ensemble of input states, i.e. an ensemble that achieves the maximum mutual information. In particular, $\Pi(S_\circ)$ as defined in (5) is non-empty.

We illustrate this result in Figure 1.

2. Peripheral Information Variance

The above observations allow us to define the minimal and maximal peripheral information variance of $S_\circ$ in terms of the information variance of peripheral decompositions of the divergence center. To do so, we consider measures $P \in \Pi(S_\circ)$ and optimize the conditional information variance:
Definition 2. Let \( \mathcal{S}_o \subseteq \mathcal{S} \) be closed and \( \Pi(\mathcal{S}_o) \) defined in (5). Then, the minimal and maximal peripheral information variance of \( \mathcal{S}_o \) (in \( \mathcal{S} \)) are respectively defined as

\[
\begin{align*}
v_{\min}(\mathcal{S}_o) &:= \inf_{P \in \Pi(\mathcal{S}_o)} \int dP(\rho) V(\rho \Vert \sigma^*(\mathcal{S}_o)), \quad \text{and} \quad (6) \\
v_{\max}(\mathcal{S}_o) &:= \sup_{P \in \Pi(\mathcal{S}_o)} \int dP(\rho) V(\rho \Vert \sigma^*(\mathcal{S}_o)). \quad (7)
\end{align*}
\]

It is evident from the compactness of \( \Pi(\mathcal{S}_o) \) that the infimum and supremum are achieved. We will see later (cf. Lemma 10) that the optima are in fact achieved by discrete measures with support on at most \( d^2 + 1 \) points.

B. Gaussian Approximation for the Channel Capacity

1. General Classical-Quantum Channels

Our main result is the evaluation of the second order asymptotics of general cq channels. For this purpose, we consider a channel \( W : \mathcal{X} \to \mathcal{S} \), where \( \mathcal{X} \) is an arbitrary set.\(^3\)

Theorem 2. Let \( \varepsilon \in (0, 1) \) and \( W \) be a cq channel. Setting \( \mathcal{S}_o = \overline{\text{im}(W)} \), we find

\[
\log M^*(W^n, \varepsilon) = n C(W) + \sqrt{n V_\varepsilon(W)} \Phi^{-1}(\varepsilon) + K(n, \mathcal{S}_o, \varepsilon),
\]

where

\[
C(W) = \chi(\mathcal{S}_o) \quad \text{and} \quad V_\varepsilon(W) = v_\varepsilon(\mathcal{S}_o) := \begin{cases} v_{\min}(\mathcal{S}_o) & \text{if } 0 < \varepsilon \leq \frac{1}{2} \\ v_{\max}(\mathcal{S}_o) & \text{if } \frac{1}{2} < \varepsilon < 1. \end{cases}
\]

We have \( K(n, \mathcal{S}_o, \varepsilon) = o(\sqrt{n}) \) for all channels. Moreover, if \( \mathcal{S}_o \) is discrete and \( v_\varepsilon(\mathcal{S}_o) > 0 \), we have \( K(n, \mathcal{S}_o, \varepsilon) = O(\log n) \).

Remark 3. The \( \varepsilon \)-channel dispersion is an operational quantity defined as \([29, \text{Eq. (221)}]\)

\[
V_\varepsilon(W) := \limsup_{n \to \infty} \frac{1}{n} \left( \frac{nC(W) - \log M^*(W^n, \varepsilon)}{\Phi^{-1}(\varepsilon)} \right)^2.
\]

Our results imply that it equals \( v_\varepsilon(\mathcal{S}_o) \), the minimal or maximal peripheral variance of the channel image, depending on the value of \( \varepsilon \).

Remark 4. Traditionally, classical-quantum channels are studied for the case when \( \mathcal{X} \) is discrete. In our framework, this corresponds to a discrete set \( \mathcal{S}_o = \{ W(x) \mid x \in \mathcal{X} \} \).

Remark 5. Some restrictions on \( \mathcal{S}_o \) are necessary in order to show that \( K(n, \mathcal{S}_o, \varepsilon) = O(\log n) \). Indeed, there exists a class of classical discrete memoryless channels, so-called exotic channels \([29, \text{p. 2231 and App. H}]\), for which \( v_\varepsilon(\mathcal{S}_o) = 0 \) and \( K(n, \mathcal{S}_o, \varepsilon) = \Theta(n^{1/3}) \) hold \([28, \text{Thm. 51}]\).

We sketch the main ideas of our proof in the following.

\(^3\) In particular, this set is not assumed to be discrete, finite or have any topological structure.
Proof Summary: The direct part, shown in Section V B, is derived employing a one-shot bound due to Wang and Renner that relates $M^*(W^n, \varepsilon)$ with the $\varepsilon$-hypothesis-testing divergence, $D_{h}(\cdot\|\cdot)$, defined in Definition 3 in Sec. V A. The bound is valid for classical-quantum channels with finite input alphabets and the asymptotics are derived in this setting based upon the second order asymptotics of the hypothesis testing divergence evaluated on i.i.d. states established in [20] and [36]. Finally, Caratheodory’s theorem shows that it is possible to achieve the second order asymptotics with finite alphabets (of size depending on the dimension of the output space).

For the converse, shown in Section V C, we derive a bound that directly relates the one-shot fundamental limit with the $\varepsilon$-hypothesis testing divergence radius of the channel image. The latter is defined as $\chi_{h}(\mathcal{S}) = \inf_{\sigma \in \mathcal{S}} \sup_{\rho \in \mathcal{S}} D_{h}(\rho\|\sigma)$ and is a one-shot analogue of the divergence radius. We establishes that

$$\log M^*(W, \varepsilon) \lesssim \chi_{h}(\mathcal{S}), \quad (8)$$

independently of the input alphabet supported by the channel. This allows to treat the remaining evaluation as a problem on the output space—depending only on the geometry of the channel image, $\mathcal{S}_o$. In particular, we find

$$\chi_{h}(\mathcal{S}o^n) = \inf_{\sigma_o \in \mathcal{S}_o} \sup_{\rho^n \in \mathcal{S}o^n} D_{h}(\rho^n\|\sigma^n) \lesssim n \chi(\mathcal{S}_o) + \sqrt{n \psi(\mathcal{S}_o)} \Phi^{-1}(\varepsilon) + K(n, \mathcal{S}_o, \varepsilon).$$

To evaluate these asymptotics for a suitable choice of $\sigma^n$ we extend the Gaussian approximation of [36] to non-identical product distributions. Moreover, we show that these bounds hold uniformly in all sequences $\rho^n = \bigotimes_{i=1}^n \rho_i \in \mathcal{S}_o^o$ that appear in the supremum above. This is particularly challenging because we have to treat separately sequences for which the average relative entropy variance is small, and hence the convergence to the Gaussian approximation is too slow.4 To tackle this, we employ a net on $\mathcal{S}_o$ and in particular do not refer to constant composition codes and type-counting arguments, which are cornerstones of the second order analysis for discrete memoryless channels in the classical setting. Our novel proof thus departs from the usual treatment, which in particular allows us to consider general input alphabets.

2. On the Classical Capacity of Quantum Channels

Theorem 2 has a neat corollary when we consider the classical capacity of quantum channels. Let $\mathcal{E}$ be a quantum channel from an arbitrary set of states to $\mathcal{S}$. Clearly, the set $\mathcal{X}$ could itself be a set of quantum states, and thus such channels naturally fit into our framework.5 Hence, the one-shot classical capacity of a quantum channel is certainly also upper bounded by $\chi_{h}(\text{im}(\mathcal{E}))$ as in (8). However, to treat the $n$-fold memoryless repetition of the channel, it now does not suffice to simply evaluate the $\varepsilon$-hypothesis testing divergence radius for $\text{im}(\mathcal{E})^o$, as the channel image can be enlarged in the presence of non-product input states, i.e. $\text{im}(\mathcal{E}^o) \neq \text{im}(\mathcal{E})^o$.

Interestingly, our results show that questions related to the additivity of the channel capacity in the non-asymptotic regime (and in the asymptotic regime, as is well-known) can be phrased in terms of the geometry of the channel image.

If the input of the channel is restricted to separable states then indeed $\text{im}(\mathcal{E}^o) = \text{conv}(\text{im}(\mathcal{E})^o)$ and Theorem 2 applies rather directly. The only missing observation is that $\chi_{h}(\text{conv}(\mathcal{S}_o)) \leq \chi_{h}(\mathcal{S}_o)$ for all $\mathcal{S}_o \subseteq \mathcal{S}$, which is an immediate consequence of the quasi-convexity of $\rho \mapsto D_{h}(\rho\|\sigma)$, proved in Lemma 7. We thus arrive at the following corollary.

4 For a classical analogue, recall that the convergence speed in the Berry Esseen theorem is inversely proportional to $\sigma^4$, where $\sigma^4$ is the average variance of a sequence of non-i.i.d. random variables.

5 Whether the states in $\mathcal{X}$ are modeled as density operators on a (finite-dimensional or not) Hilbert space or states of a C* algebra is irrelevant.
Corollary 3. Let $\varepsilon \in (0, 1)$, let $\mathcal{E}$ be a quantum channel and set $S_0 = \text{im}(\mathcal{E})$. Let $M_{\text{sep}}^* (\mathcal{E}^n, \varepsilon)$ denote the maximum volume of a codebook for classical information transmission over $\mathcal{E}$ with average error $\varepsilon$ when the channel is restricted to separable input states. Then,

$$\log M_{\text{sep}}^* (\mathcal{E}^n, \varepsilon) = n \chi(S_0) + \sqrt{n v_{\text{min}}(S_0)} \Phi^{-1} (\varepsilon) + o(\sqrt{n}).$$

(9)

Example 1. Qubit Pauli channels are symmetric under reflection at the center of the Bloch sphere. As such, $\sigma^*(S_0) = \frac{1}{2} \text{id}$ and it is furthermore easy to verify that any capacity achieving ensemble (of minimal size) is commutative. Hence, the capacity and dispersion of a Pauli channel equal those of a (classical) binary symmetric channel (see, e.g., [29, Thm. 52]).

Example 2. The amplitude damping channel with magnitude $\gamma$ is given as

$$\mathcal{E}_{\text{ad}}^{\gamma} : \rho \mapsto \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{1 - \gamma} & 0 \\ 0 & \sqrt{1 - \gamma} & 0 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{\gamma}}{\sqrt{1 - \gamma}} \end{pmatrix} \rho \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{1 - \gamma} & 0 \\ 0 & \sqrt{1 - \gamma} & 0 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{\gamma}}{\sqrt{1 - \gamma}} \end{pmatrix}.$$

Its channel image, $S_0^\gamma = \text{im}(\mathcal{E}_{\text{ad}}^{\gamma})$, is displayed in Figure 2(a). In Fig. 2(b), the channel capacity and dispersion are evaluated numerically for different values of $\gamma$. The Gaussian approximation, i.e. the first two terms on the right-hand side of (9) are plotted as a function of $n$ in Figure 2(c).
It was already noted in [33, Fig. 1] that it is necessary to consider non-orthogonal input states to achieve $\chi(S^n_0)$—in particular, $\mathcal{E}_{ad}^\gamma(|0\rangle\langle 0|) \notin \Gamma(S^n_0)$ for general $\gamma \in (0, 1)$.

This naturally leaves many open questions. Most intriguingly, it was recently shown that the for entanglement-breaking and Hadamard channels, we have [40]

$$\log M^*(\mathcal{E}^n, \varepsilon) = n \chi(S_0) + O(\sqrt{n})$$

Thus, one could conjecture that a Gaussian approximation of the form (9) also holds for such channels (without assuming separable input states). In particular, it would be interesting to see if the second order term is again given by the peripheral information variance. The proof of the strong converse in [40] relies on the additivity of a suitable Rényi divergence radius [22, 40] of the channel image. However, it appears challenging to deduce a second order expansion of the $\varepsilon$-hypothesis testing divergence radius from these results.

IV. PROOFS: QUANTUM DIVERGENCE RADIUS

This section contains various lemmas which, combined, establish Theorem 1. Recall that $\mathcal{S}$ denotes the set of quantum states on a Hilbert space of dimension $d$, and $\mathcal{S}_0 \subseteq \mathcal{S}$ is an arbitrary closed subset of $\mathcal{S}$, and thus also compact.

We will later show that the divergence center $\sigma^*(\mathcal{S}_0)$, as defined in Theorem 1, is indeed a singleton, but at his point we have to be satisfied with the following statement.

**Lemma 1.** The set $\sigma^*(\mathcal{S}_0)$ is nonempty, convex and $\sigma \in \sigma^*(\mathcal{S}_0)$ implies $\sigma \gg \rho$ for all $\rho \in \mathcal{S}_0$.

**Proof.** The function $f : \sigma \mapsto \sup_{\rho \in \mathcal{S}_0} D(\rho||\sigma)$ is finite only if $\sigma \gg \rho$ for all $\rho \in \mathcal{S}_0$. Moreover, since $\mathcal{S}_0$ is compact and $f$ convex, the set of minima contains at least one element and is convex. \qed

In analogy to Theorem 1, we define the set of extremal points in $\mathcal{S}_0$ corresponding to the center $\sigma \in \sigma^*(\mathcal{S}_0)$ as $\Gamma_\sigma(\mathcal{S}_0) := \arg\max_{\rho \in \mathcal{S}_0} D(\rho||\sigma)$.

**Proposition 2.** For every $\sigma \in \sigma^*(\mathcal{S}_0)$, we have $\sigma \in \text{conv}(\Gamma_\sigma(\mathcal{S}_0))$.

**Proof.** Let us fix $\sigma \in \sigma^*(\mathcal{S}_0)$ to simplify notation. We define

$$\Theta^\nu := \{ \rho \in \mathcal{S}_0 \mid D(\rho||\sigma) \geq \chi(S_0) - \nu \},$$

and its complement $\bar{\Theta}^\nu := \mathcal{S}_0 \setminus \Theta^\nu$ for any $\nu \geq 0$. We first observe that $\Theta^\nu \subseteq \mathcal{S}_0$ is closed since $D(\cdot||\sigma)$ is continuous and $\mathcal{S}_0$ is closed itself. Thus, both $\Theta^\nu$ and $\text{conv}(\Theta^\nu)$ are compact. Moreover, we clearly have $\bigcap_{\nu > 0} \Theta^\nu = \Theta^0 = \Gamma_\sigma(\mathcal{S}_0)$.

For the sake of contradiction, let us now assume that $\sigma \notin \text{conv}(\Theta^\nu)$ for some fixed $\nu > 0$. We employ the following lemma.

**Lemma 3.** [26, Lem. 3.3] Let $\mathcal{S}_0 \subseteq \mathcal{S}$ be compact convex and let $\sigma \in \mathcal{S}$. Then, $\tau := \arg\min_{\tau \in \mathcal{S}_0} D(\tau||\sigma)$ is unique. Moreover, for all $\rho \in \mathcal{S}_0$, we have

$$D(\rho||\sigma) \geq D(\rho||\tau) + D(\tau||\sigma).$$

This establishes that there exists a unique state $\tau \in \text{conv}(\Theta^\nu)$ that minimizes $D(\tau||\sigma)$. Furthermore, $D(\rho||\sigma) > D(\rho||\tau)$ for all $\rho \in \Theta^\nu$. Consequently, using the parametrization $\tau^\lambda := \lambda \tau + (1 - \lambda)\sigma$ and the convexity of $D(\rho||\cdot)$, we find

$$D(\rho||\tau^\lambda) \leq \lambda D(\rho||\tau) + (1 - \lambda)D(\rho||\sigma) < D(\rho||\sigma) \quad \forall \lambda \in (0, 1).$$
Hence, $D(\rho \| \tau^\lambda) < D(\rho \| \sigma)$ for all $\rho \in \Theta^\nu$ for all $\lambda \in (0, 1)$.

Furthermore, recall that $D(\rho \| \sigma)$ is bounded away from $\chi(S_0)$ for all $\rho \in \bar{\Theta}^\nu$ by definition. Due to the continuity of $D(\rho \| \cdot)$, we thus find that for sufficiently small $\lambda > 0$,

$$D(\rho \| \tau^\lambda) < \chi(S_0) \quad \forall \rho \in S_0.$$ 

Thus, $\sigma \notin \text{conv}(\Theta^\nu)$ leads to a contradiction with the fact that $\sigma \in \sigma^*(S_0)$. Hence, we established that $\sigma \in \text{conv}(\Theta^\nu)$ and since this holds for all $\nu > 0$, we find $\sigma \in \bigcap_{\nu > 0} \text{conv}(\Theta^\nu)$. The statement then follows by the following lemma proven in Appendix A.

**Lemma 4.** Let $\Theta_1 \supseteq \Theta_2 \supseteq \ldots$ be a sequence of (sequentially) compact sets in a finite-dimensional vector space. Then,

$$\bigcap_{n \in \mathbb{N}} \text{conv}(\Theta_n) = \text{conv}(\Theta_\infty) \quad \text{whenever} \quad \Theta_\infty := \bigcap_{n \in \mathbb{N}} \Theta_n \neq \emptyset.$$ 

This establishes that $\bigcap_{\nu > 0} \text{conv}(\Theta^\nu) = \text{conv}(\Theta^0)$ and concludes the proof.

The fact that $\sigma \in \sigma^*(S_0) \implies \sigma \in \text{conv}(\Gamma_{\sigma}(S_0))$, first established here, is crucial since it allows the following construction:

Due to Caratheodory’s theorem, we may decompose $\sigma$ into a convex combination of (at most $d^2$) peripheral states, namely we may write

$$\sigma = \sum_{\rho \in \mathcal{X}_0} P(\rho) \rho, \quad \text{where} \quad \mathcal{X}_0 \subseteq \Gamma_{\sigma}(S_0), \quad |\mathcal{X}_0| \leq d^2 \quad \text{and} \quad P \in \mathcal{P}(\mathcal{X}_0). \quad (10)$$

Using this decomposition and the fact that $D(\rho \| \sigma) = \chi(S_0)$ for all $\rho \in \mathcal{X}_0$, we find

$$\chi(S_0) = \sum_{\rho \in \mathcal{X}_0} P(\rho) D(\rho \| \sigma) = H(\sigma) - \sum_{\rho \in \mathcal{X}_0} P(\rho) H(\rho). \quad (11)$$

The uniqueness of $\sigma^*(S_0)$ now follows from a standard argument (see, e.g., [10, Sec. 4.5]) and using the strict concavity of $H$.

**Lemma 5.** The center $\sigma^*(S_0)$ is unique.

*Proof.* Assume for the sake of contradiction that $\sigma_0, \sigma_1 \in \sigma^*(S_0)$ with $\sigma_0 \neq \sigma_1$. Consequently, $\sigma_\lambda := \lambda \sigma_1 + (1 - \lambda) \sigma_0$ is in $\sigma^*(S_0)$ for all $\lambda \in [0, 1]$. Following (10), we may write

$$\sigma_\lambda = \sum_{\rho \in \mathcal{X}_0} P_\lambda(\rho) \rho, \quad \text{where} \quad |\mathcal{X}_0| \leq \Gamma_{\sigma_0}(S_0) \cup \Gamma_{\sigma_1}(S_0), \quad |\mathcal{X}_0| \leq 2d^2.$$ 

and $P_\lambda(\rho) = \lambda P_1(\rho) + (1 - \lambda) P_0(\rho)$ for $P_\lambda \in \mathcal{P}(\mathcal{X}_0^\prime)$ . Then, due to (11), we have

$$\chi(S_0) = H(\sigma_0) - \sum_{\rho \in \mathcal{X}_0} P_0(\rho) H(\rho) = H(\sigma_1) - \sum_{\rho \in \mathcal{X}_0} P_1(\rho) H(\rho).$$

Hence, using the strict concavity of $H(\cdot)$, we find

$$\chi(S_0) = \lambda H(\sigma_1) + (1 - \lambda) H(\sigma_0) - \sum_{\rho \in \mathcal{X}_0} P_\lambda(\rho) H(\rho)$$

$$< H(\sigma_\lambda) - \sum_{\rho \in \mathcal{X}_0} P_\lambda(\rho) H(\rho)$$

$$= \sum_{\rho \in \mathcal{X}_0} P_\lambda(\rho) D(\rho \| \sigma_\lambda).$$

Finally, the fact that $D(\rho \| \sigma_\lambda) \leq \sup_{\rho \in S_0} D(\rho \| \sigma_\lambda) = \chi(S_0)$ since $\sigma_\lambda \in \sigma^*(S_0)$ yields the desired contradiction. \qed
The previous lemma justifies writing $\Gamma(\mathcal{S}_\sigma)$ in Theorem 1, i.e. $\Gamma_\sigma(\mathcal{S}_\sigma)$ does not depend on $\sigma$. We will thus drop the subscript $\sigma$ in $\Gamma_\sigma$ hereafter hereafter.

For any $\mathbb{P} \in \mathcal{P}(\mathcal{S})$ and $\sigma \in \mathcal{S}$, let us introduce the notation

$$I(\mathbb{P}|\sigma) := \int_0^\infty d\mathbb{P}(\rho) I(\rho|\sigma) \quad \text{and} \quad I(\mathbb{P}) := I\left(\mathbb{P}||\rho^{(\mathbb{P})}\right)$$

(12)

in analogy with the conditional mutual information.

**Lemma 6.** We have $\chi(\mathcal{S}_\sigma) = \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S}_\sigma)} I(\mathbb{P})$. The supremum is achieved by a discrete measure with support on at most $d^2$ points in $\Gamma(\mathcal{S}_\sigma)$.

**Proof.** First, note that for every $\mathbb{P} \in \mathcal{P}(\mathcal{S})$ we have $I(\mathbb{P}) = \min_{\sigma \in \mathcal{S}} I(\mathbb{P}|\sigma)$ due to the positive-definiteness of $D(\cdot||\cdot)$. Now, Sion’s minimax theorem [34] yields

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S}_\sigma)} \min_{\sigma \in \mathcal{S}} I(\mathbb{P}|\sigma) = \min_{\sigma \in \mathcal{S}} \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S}_\sigma)} I(\mathbb{P}|\sigma)$$

(13)

Indeed, it is easy to verify that the $I(\mathbb{P}|\sigma)$ is convex in $\sigma$ and linear in $\mathbb{P}$. Moreover, $\mathcal{S}$ is compact convex and $\mathcal{P}(\mathcal{S}_\sigma)$ is convex, as required. Finally, the supremum over distributions on the right-hand side of (13) can be replaced by a supremum over Dirac measures on $\mathcal{S}_\sigma$ without loss of generality. This establishes

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S}_\sigma)} \int_0^\infty d\mathbb{P}(\rho) I(\rho|\sigma^{(\mathbb{P})}) = \min_{\sigma \in \mathcal{S}} D(\rho||\sigma).$$

The second statement follows immediately due to the construction given in Eq. (10) and (11). \qed

We are now ready to summarize the proof of Theorem 1.

**Proof of Theorem 1.** Property 1 follows from Lemmas 1 and 5. Property 2 is a trivial consequence of Property 1 and the definition of $\Gamma$. Property 3 is implied by Proposition 2 whereas Property 4 is established in Lemma 6. Finally, Property 5 is established as follows:

Clearly, every $\mathbb{P} \in \Pi(\mathcal{S}_\sigma)$ achieves the supremum in (4), $\chi(\mathcal{S}_\sigma)$, by definition of $\Gamma(\mathcal{S}_\sigma)$. Let us assume that there exists a distribution $\mathbb{P} \in \mathcal{P}$ that achieves $\chi(\mathcal{S}_\sigma)$. Then, $\rho^{(\mathcal{P})} = \sigma^*(\mathcal{S}_\sigma)$ by the argument in Lemma 5. Moreover, $\mathbb{P}[\Gamma(\mathcal{S}_\sigma)] = 1$ is necessary due to the definition of $\Gamma$. \qed

V. PROOFS: GAUSSIAN APPROXIMATION FOR THE CHANNEL CAPACITY

A. Properties of the Hypothesis-Testing Divergence

In order to express the one-shot results, we introduce the $\varepsilon$-hypothesis-testing divergence [39].

**Definition 3.** Let $\varepsilon \in (0,1)$ and $\rho,\sigma \in \mathcal{S}$. The $\varepsilon$-hypothesis-testing divergence is defined as

$$D_h^\varepsilon(\rho||\sigma) := -\log \frac{\beta_{1-\varepsilon}(\rho||\sigma)}{1-\varepsilon} , \quad \beta_{1-\varepsilon}(\rho||\sigma) := \min_{\frac{\sigma \in \mathcal{Q} \subset \mathcal{S}}{\operatorname{tr}(Q) \geq 1-\varepsilon}} \operatorname{tr}(Q\sigma).$$

Note that $\beta_{1-\varepsilon}$ is the smallest type-II error of a hypothesis test between $\rho$ and $\sigma$ with type-I error at most $\varepsilon$. The $\varepsilon$-hypothesis testing divergence satisfies the following basic properties, which we summarize here for later reference.

**Lemma 7.** Let $\varepsilon \in (0,1)$, let $\mathcal{S}_o, \mathcal{S}_o' \subset \mathcal{S}$ be discrete sets, and let $\mathbb{P} \in \mathcal{P}(\mathcal{S}_o)$, $Q \in \mathcal{P}(\mathcal{S}_o')$. Define $\rho = \sum_{\tau \in \mathcal{S}_o} P(\tau) \tau$ and $\sigma = \sum_{\omega \in \mathcal{S}_o'} Q(\omega) \omega$. Then $D_h^\varepsilon(\rho||\sigma)$ satisfies the following properties:
1. $D_h^e(\rho\|\sigma) \geq 0$ with equality if and only if $\rho = \sigma$. (cf. [5, Prop. 3.2])
2. For any CPTP map $\mathcal{M}$ we have $D_h^e(\rho\|\sigma) \geq D_h^e(\mathcal{M}(\rho)\|\mathcal{M}(\sigma))$. (cf. [39, Lem. 1])
3. $D_h^e(\rho\|\sigma) \leq \min_{\omega \in S_\varepsilon} \{ D_h^e(\rho\|\omega) + \log \frac{1}{Q(\omega)} \}$.
4. $D_h^e(\rho\|\sigma) \leq \max_{\tau \in S_\varepsilon} D_h^e(\tau\|\sigma)$.

The last inequality shows that $\rho \mapsto D_h^e(\rho\|\sigma)$ is quasi-convex. The last two inequalities can be verified by a close inspection of Definition 3 and we omit the proof.

Some of the main ingredients of our proof are the following non-asymptotic bounds on the $\varepsilon$-hypothesis testing divergence evaluated for product states. Before we state the bounds, recall that $I(\mathbb{P}\|\sigma) = \int d\mathbb{P}(\rho) D(\rho\|\sigma)$ and define $V(\mathbb{P}\|\sigma) := \int d\mathbb{P}(\rho) V(\rho\|\sigma)$ analogously for any $\mathbb{P} \in \mathcal{P}(S)$ and $\sigma \in \mathcal{S}$. Moreover, given a sequence of states $\rho^n = \bigotimes_{i=1}^n \rho_i$, we denote by $P_{\rho^n}(\rho) := \frac{1}{n} \sum_{i=1}^n \{ \rho = \rho_i \}$ the empirical distribution of $\rho^n$.

**Proposition 8.** Let $\varepsilon \in (0, 1)$, $S_\varepsilon \subseteq \mathcal{S}$ and $\lambda_0 > 0$. Let $\{\varepsilon_n\}_{n=1}^\infty$ be any sequence satisfying $|\varepsilon_n - \varepsilon| \leq 1/\sqrt{n}$ for all $n$ and set $\varepsilon^* := \min\{\varepsilon, 1 - \varepsilon\}$. Then, there exist constants $N_1(\varepsilon, S_\varepsilon, \lambda_0)$ and $K_1(\varepsilon, S_\varepsilon, \lambda_0)$ and $L_1$ such that the following holds. For every $n \geq N_1$, every $\sigma \in \mathcal{S}$ with $\lambda_{\min}(\sigma) \geq \lambda_0$ and every sequence $\rho^n = \bigotimes_{i=1}^n \rho_i$, $\rho_i \in S_\varepsilon$, we have

$$
\left| D_h^e(\rho^n\|\sigma^{\otimes n}) - nI(\rho^n\|\sigma) \right| \leq \sqrt{\frac{nV(P_{\rho^n}\|\sigma)}{\varepsilon^*}} + L_1 \log n \leq K_1 \sqrt{n}. \tag{14}
$$

Further let $\xi > 0$ and fix $\sigma \in \mathcal{S}$ with $\lambda_{\min}(\sigma) > 0$. Then, there exist constants $N_2(\varepsilon, S_\varepsilon, \sigma, \xi)$ and $L_2$ such that the following holds. For every $n \geq N_2$ and every sequence $\rho^n = \bigotimes_{i=1}^n \rho_i$, $\rho_i \in S_\varepsilon$ satisfying $V(P_{\rho^n}\|\sigma) \geq \xi$, we have

$$
\left| D_h^e(\rho^n\|\sigma^{\otimes n}) - nI(\rho^n\|\sigma) - \sqrt{\frac{nV(P_{\rho^n}\|\sigma)}{\Phi^{-1}(\varepsilon)}} \right| \leq L_2 \log n. \tag{15}
$$

Finally, if $\sigma = \rho^{(P_{\rho^n})}$ in (15), then the statement holds for $n \geq N_3(\varepsilon, S_\varepsilon, \xi)$ independent of $\sigma$.

In the asymptotic limit where $n \to \infty$, all inequalities imply the seminal quantum Stein’s lemma [17] and its strong converse [25] when the sequence is chosen i.i.d. The proof is based on the techniques of [20, 36] and presented in Appendix B. It is crucial for our application that $L_1, L_2, K_1, N_1, N_2$ and $N_3$ are uniform over $\sigma$ and sequences $\rho^n$ satisfying the constraints. This is nontrivial and requires arguments beyond those in [20, 36] which only treat the i.i.d. case.\footnote{For this reason we also do not rely on the infamous $O(\cdot)$ notation here, which tends to hide such subtleties.}

### B. Direct Part

We base our result on the following straight-forward generalization of the one-shot bounds by Hayashi and Nagaoka [15] and Wang and Renner [39] (see also [4, 6, 31] for recent one-shot bounds for c-q channels).\footnote{To compare with [39, Thm. 2], simply note that we may restrict our channel to a discrete classical-quantum channel bijectively mapping from an arbitrary index set to element in $\mathcal{X}_\varepsilon$. The direct sum notation reveals the classical quantum structure of the underlying state. Finally, the constant $c$ in [39] can be optimized over.}

**Proposition 9.** [39, Thm. 2] Let $\varepsilon \in (0, 1)$, $\eta \in (0, \varepsilon)$, and let $\mathcal{X}_\varepsilon \subseteq \text{im}(\mathcal{W})$ be discrete. Then,

$$
\log M^*(\mathcal{W}, \varepsilon) \geq \sup_{P \in \mathcal{P}(\mathcal{X}_\varepsilon)} D_h^{\varepsilon - \eta} \left( \bigoplus_{\rho \in \mathcal{X}_\varepsilon} P(\rho) \otimes \rho \right) \otimes \left( \bigoplus_{\rho \in \mathcal{X}_\varepsilon} P(\rho) \right) - \log \frac{4\varepsilon(1 - \varepsilon + \eta)}{\eta^2}.
$$
We can include the closure of $\text{im}(\mathcal{W})$ due to the continuity of the above expression when the set $\mathcal{X}_o$ is varied by replacing an element with one that is close in $(\mathcal{S}, d)$. Thus, our bound reads

$$M^*(\mathcal{W}, \varepsilon) \geq \sup_{\mathcal{X}_o \subseteq \overline{\text{im}}(\mathcal{W})} \sup_{P \in \mathcal{P}(\mathcal{X}_o)} D_h^\varepsilon - \eta \left( \omega(P) \left\| \tau(P) \otimes \rho(P) \right\| \right) - \log \frac{4\varepsilon(1 - \varepsilon + \eta)}{\eta^2},$$  

(16)

where $\mathcal{X}_o$ is discrete and we introduced the shorthands $\omega(P) := \bigoplus_{\rho \in \mathcal{X}_o} P(\rho) \otimes \rho$ and $\tau(P) := \bigoplus_{\rho \in \mathcal{X}_o} P(\rho)$. The similarity of the above expression with the asymptotic expression in (4) is evident once (4) is specialized to the discrete case as well.

The restriction to finite subsets of $\mathcal{S}_o$ may appear problematic on first sight, however, an elementary application of Caratheodory’s theorem reveals that the minimal and maximal variance are achieved by a discrete measure.

Lemma 10. There exist discrete probability measures with support on at most $d^2 + 1$ points in $\Gamma(\mathcal{S}_o)$ that achieve the infimum and supremum in (6) and (7), respectively.

Proof. Note first that $\Pi(\mathcal{S}_o)$ is compact and the optima are thus achieved. Now, we are simply looking for a probability measure $P \in \mathcal{P}(\Gamma(\mathcal{S}_o))$ that satisfies the linear constraints

$$\int d\mathbb{P}(\rho) \rho = \sigma^*(\mathcal{S}_o) \quad \text{and} \quad \int d\mathbb{P}(\rho) V(\rho\|\sigma^*(\mathcal{S}_o)) = v_\varepsilon(\mathcal{S}_o).$$

These constitute $d^2 - 1$ real constraints for the first equality and one additional constraint for the second one. Since $\Gamma(\mathcal{S}_o)$ is not connected in general, Caratheodory’s theorem (see, e.g., [7, Thm. 18]) establishes that there exists a discrete $P \in \Pi(\mathcal{S}_o)$ with support on at most $d^2 + 1$ states that satisfies the constraints. \qed

Let us then proceed to prove the lower bound in Theorem 2.

Theorem 4. Let $\varepsilon \in (0, 1)$ and let $\mathcal{W}$ be a channel. Set $\mathcal{S}_o := \overline{\text{im}(\mathcal{W})}$. We have

$$\log M^*(\mathcal{W}^n, \varepsilon) \geq n \chi(\mathcal{S}_o) + \sqrt{n v_\varepsilon(\mathcal{S}_o)} \Phi^{-1}(\varepsilon) + O(\log n).$$

Proof. First, let us apply (16) to the $n$-fold repetition of the channel $\mathcal{W}$. Fixing any discrete set $\mathcal{X}_o \subseteq \mathcal{S}_o$ and $P \in \mathcal{P}(\mathcal{X}_o)$, we first confirm that

$$\log M^*(\mathcal{W}^n, \varepsilon) \geq D_h^{r - \eta} \left( (\omega(P))^{\otimes n} \left\| (\tau(P) \otimes \rho(P))^{\otimes n} \right\| \right) - \log \frac{4\varepsilon(1 - \varepsilon + \eta)}{\eta^2}$$

(17)

Note that we applied (16) using the set $\mathcal{X}_o^{\otimes n} \subseteq \overline{\text{im}(\mathcal{W})^{\otimes n}} = \overline{\text{im}(\mathcal{W})^{\otimes n}}$ and the $n$-fold product distribution $P^{\otimes n}$. By Lemma 10 there exists a discrete probability measure (let it be our choice of $P$) with support on $\Gamma(\mathcal{S}_o)$ (let the support set be our choice of $\mathcal{X}_o$) such that

$$\rho(P) = \sigma, \quad \sum_{\rho \in \mathcal{X}_o} P(\rho) D(\rho\|\sigma) = \chi(\mathcal{S}_o), \quad \text{and} \quad \sum_{\rho \in \mathcal{X}_o} P(\rho) V(\rho\|\sigma) = v_\varepsilon(\mathcal{S}_o),$$

(18)

where we set $\sigma = \sigma^*(\mathcal{S}_o)$. Now, is easy to verify that

$$\sum_{\rho \in \mathcal{X}_o} P(\rho) D(\rho\|\rho(P)) = D\left( \bigoplus_{\rho \in \mathcal{X}_o} P(\rho) \rho \bigg| \bigoplus_{\rho \in \mathcal{X}_o} P(\rho) \otimes \rho(P) \right),$$

(19)

and, the following simple generalization of [29, Lm. 62] proved in Appendix D holds.
Lemma 11. For any discrete probability measure $P \in \Pi(S_\circ)$, we have

$$\sum_{\rho \in X_\circ} P(\rho) V(\rho \| \rho^{(P)}) = V\left( \bigoplus_{\rho \in X_\circ} P(\rho) \rho \bigoplus_{\rho \in X_\circ} P(\rho) \rho^{(P)} \right).$$

As such, we are left to evaluate the asymptotics of $D_{h}^{\varepsilon-\eta}$ for identical product states. Let us set $\varepsilon_n := \varepsilon - \eta$ with $\eta = 1/\sqrt{n}$. First, consider the case where $v_\varepsilon(S_\circ) > 0$. Proposition 8, i.e. (15) specialized to i.i.d., establishes that

$$D_{h}^{\varepsilon-\eta}\left((\omega^{(P)}) \otimes \rho^{(P)}) \otimes \rho^{(P)}\right) \geq nD(\omega^{(P)} \| \rho^{(P)} + \sqrt{nV(\omega^{(P)} \| \rho^{(P)}) \Phi^{-1}(\varepsilon) - L_2 \log n}
$$

= $n\chi(S_\circ) + \sqrt{n v_\varepsilon(S_\circ) \Phi^{-1}(\varepsilon) - L_2 \log n}$.

for all $n \geq N_2$ and some constants $L_2$ and $N_2(\varepsilon, S_\circ)$. In the last step we employed (19), Lemma 11 and (18). Moreover, the last summand in (17) is of the form $O(\log n)$ and we are done.

The proof for the case $v_\varepsilon(S_\circ) = 0$ proceeds similarly but employs the first bound in Eq. (14) in Proposition 8 instead. This yields

$$D_{h}^{\varepsilon-\eta}\left((\omega^{(P)}) \otimes \rho^{(P)} \otimes \rho^{(P)}\right) \geq nD(\omega^{(P)} \| \rho^{(P)} - L_1 \log n.$$ 

for all $n \geq N_1$ and some constants $L_1$ and $N_1(\varepsilon - \eta, S_\circ)$. The rest then proceeds analogously to the discussion above.

\[\Box\]

C. Converse Part

1. One-Shot Bounds and Geometry of Channel Image

We define a one-shot analogy of the divergence radius in Definition 1.

Definition 4. Let $\varepsilon \in (0, 1)$ and $S_\circ \subseteq S$. The $\varepsilon$-hypothesis-testing divergence radius is defined as

$$\chi^{\varepsilon}(S_\circ) := \inf_{\sigma \in S} \sup_{\rho \in S_\circ} D_{h}^{\varepsilon}(\rho \| \sigma).$$

This quantity, evaluated on the channel image, constitutes a tight upper bound on $M^*(\mathcal{W}, \varepsilon)$ for general channels, as the following shows. The bound is similar to a recent bound by Matthews and Whener [21] in terms of the $\varepsilon$-hypothesis testing divergence, but our result differs in that we find a bound that only depends on the image of the channel.

Proposition 12. Let $\varepsilon \in (0, 1)$ and let $\mathcal{W}$ be a channel. For any $\mu \in (0, 1 - \varepsilon)$, we have

$$\log M^*(\mathcal{W}, \varepsilon) \leq \chi^{\varepsilon+\mu}(\mathcal{W}) \log \frac{\varepsilon + \mu}{\mu(1 - \varepsilon - \mu)}.$$

Proof. Let $\mathcal{C} = \{\mathcal{M}, e, D\}$ be a code with $p_{\text{err}}(\mathcal{C}, \mathcal{W}) \leq \varepsilon$ given by codewords $x_m = e(m) \in X$ and a decoder $\mathcal{D} = \{Q_m\}_{m \in \mathcal{M}}$. By assumption, we thus have $\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \text{tr}(Q_m \mathcal{W}(x_m)) \geq 1 - \varepsilon$. For an arbitrary but fixed $\sigma \in S$, we define the set

$$\mathcal{K} := \{m \in \mathcal{M} \big| \text{tr}(Q_m \mathcal{W}(x_m)) \geq 1 - \varepsilon - \mu\}, \text{ and } m^* := \arg \min_{m \in \mathcal{K}} \text{tr}(Q_m \sigma).$$
By definition of this set, we have
\[
1 - \varepsilon \leq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \text{tr}(Q_n \mathcal{W}(x_m)) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{K}} \text{tr}(Q_m \mathcal{W}(x_m)) + \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M} \setminus \mathcal{K}} \text{tr}(Q_m \mathcal{W}(x_m)) < |\mathcal{K}| + \frac{|\mathcal{M} - |\mathcal{K}|}{|\mathcal{M}|} (1 - \varepsilon - \mu).
\]
Hence, $|\mathcal{K}| > |\mathcal{M}| \frac{\mu}{\varepsilon + \mu}$. Moreover, we have
\[
1 = \text{tr}(\sigma) = \sum_{m \in \mathcal{M}} \text{tr}(Q_{x_m} \sigma) \geq |\mathcal{K}| \text{tr}(Q_{x_m}^* \sigma) > |\mathcal{M}| \frac{\mu}{\varepsilon + \mu} \text{tr}(Q_{x_m}^* \sigma).
\]
By definition of the $\varepsilon$-hypothesis testing divergence we find
\[
D_h^\varepsilon + \mu(\mathcal{W}(x_m^*)) \| (\sigma) \geq -\log \frac{\text{tr}(Q_{x_m}^* \sigma)}{1 - \varepsilon - \mu} > \log |\mathcal{M}| - \log \frac{\varepsilon + \mu}{\mu(1 - \varepsilon - \mu)}.
\]
Thus, in particular we have
\[
\sup_{\rho \in \text{im}(\mathcal{W})} D_h^\varepsilon + \mu(\rho) \| (\sigma) > \log |\mathcal{M}| - \log \frac{\varepsilon + \mu}{\mu(1 - \varepsilon - \mu)}.
\]
Finally, Eq. (20) follows by observing that the above bound holds for all $\sigma \in \mathcal{S}$.

2. Asymptotics of the $\varepsilon$-Hypothesis-Testing Divergence Radius: First Order

As a warmup, we use our techniques to provide a simple proof of the strong converse property of general classical-quantum channels. The strong converse is evidently a corollary of Proposition 12 and the following result.\footnote{To verify this, apply Proposition 12 for the $n$-fold repetition of the channel, $\mathcal{W}^n$ with image $\mathcal{S}_o^n$, and choose $\mu(n) = 1/\sqrt{n}$ such that $\varepsilon_n = \varepsilon + \mu$ in Proposition 13.}

**Proposition 13.** Let $\varepsilon \in (0, 1)$ and $\mathcal{S}_o \subseteq \mathcal{S}$ closed. Let $\{\varepsilon_n\}_{n=1}^\infty$ be any sequence satisfying $|\varepsilon_n - \varepsilon| \leq 1/\sqrt{n}$ for all $n$. Then,
\[
\chi_h^\varepsilon_n(\mathcal{S}_o^n) \leq n \chi(\mathcal{S}_o) + O(\sqrt{n}).
\]

Note that Winter [41] and Ogawa-Nagaoka [24] first showed the strong converse for classical-quantum channels. The latter proof also applies for the generality we consider here, whereas the former assumes a discrete input alphabet (or, in our context, that $\mathcal{S}_o$ is a discrete set).

**Proof.** By definition of the $\varepsilon$-hypothesis testing divergence radius, we have
\[
\chi_h^\varepsilon_n(\mathcal{S}_o^n) \leq \sup_{\sigma \in \mathcal{S}_o^n} D_h^n(\rho^n \| \sigma^\otimes n),
\] (21)
where we chose an $n$-fold product of the divergence center, $\sigma = \sigma^\otimes(\mathcal{S}_o) \in \mathcal{S}$, as the output state. The states $\rho^n$ are of the form $\rho^n = \bigotimes_{i=1}^n \rho_i$. For a fixed and arbitrary $\rho^n$, we define the set $\mathcal{S}_o^n := \{\rho_i\}_{i=1}^n \subseteq \mathcal{S}_o$ and the empirical distribution $P_{\rho^n} \in \mathcal{P}(\mathcal{S}_o^n)$ given by $P_{\rho^n}(\rho) = 1/n \sum_{i=1}^n 1\{\rho = \rho_i\}$.

We then use (14) in Proposition 8 to assert that
\[
D_h^n(\rho^n \| \sigma^\otimes n) \leq nI(P_{\rho^n} \| \sigma) + K_1 \sqrt{n}
\] (22)
for sufficiently large $n \geq N_1$. Here, we used that $\lambda_{\min}(\sigma) > 0$ and recall that $I(P \| \sigma)$ is defined as $I(P \| \sigma) = \int dP(\rho) D(\rho \| \sigma)$. Therefore, Theorem 1 ensures that $D(\rho \| \sigma) \leq \chi(\mathcal{S}_o)$ for all $\rho \in \mathcal{S}_o$ and we have established that
\[
\chi_h^\varepsilon_n(\mathcal{S}_o^n) \leq \sup_{\rho^n \in \mathcal{S}_o^n} nI(P_{\rho^n} \| \sigma) + K_1 \sqrt{n} \leq n \chi(\mathcal{S}_o) + K_1 \sqrt{n}.
\]
This section provides the universal upper bound in Theorem 2.

**Theorem 5.** Let $\epsilon \in (0, 1)$ and let $\mathcal{W}$ be a channel. Set $\mathcal{S}_o := \overline{\text{im}(\mathcal{W})}$. We have
\[
\log M^*(\mathcal{W}^n, \epsilon) \leq n \chi(\mathcal{S}_o) + \sqrt{n} v_\epsilon(\mathcal{S}_o) \Phi^{-1}(\epsilon) + o(\sqrt{n}).
\]
In view of Proposition 12, we therefore want to find a second order upper bound on
\[
\chi^\epsilon_h(\mathcal{S}_o^{\otimes n}) = \min_{\sigma^n \in \mathcal{B}_n} \sup_{\rho^n \in \mathcal{S}_o^{\otimes n}} D_h^\epsilon(\rho^n || \sigma^n).
\]

The proof of the strong converse relies on choosing $\sigma^n$ as the $n$-fold product of the divergence center and then taking advantage of the fact that $D(\rho || \sigma) \leq \chi(\mathcal{S}_o)$ for all $\rho \in \mathcal{S}_o$. This will not be sufficient if we want to pin down the exact second order term proportional to $\sqrt{n}.$

Before we commence, we thus introduce an appropriate choice of auxiliary state $\sigma^n$. To construct it, we require the following auxiliary result whose proof is provided in Appendix C. This establishes that there exists a $\gamma$-net on $\mathcal{S}_o$ whose cardinality can be bounded appropriately.

**Lemma 14.** For every $\gamma \in (0, 1)$, there exists a set of states $\mathcal{G}^\gamma \subseteq \mathcal{S}$ of size
\[
|\mathcal{G}^\gamma| \leq \left(\frac{5}{\gamma}\right)^{2d^2} \left(\frac{2d}{\gamma} + 2\right)^{d-1}
\]
such that, for every $\rho \in \mathcal{S}$, there exists a state $\tau \in \mathcal{G}^\gamma$ satisfying the following:
\[
\frac{1}{2} \|\rho - \tau\|_1 \leq \gamma, \quad D(\rho || \tau) \leq \gamma \cdot 4(2d + 1), \quad \text{and} \quad \lambda_{\min}(\tau) \geq \frac{\gamma}{2d + \gamma}.
\]

Now, for a $\gamma$ to be specified below, we choose the output state $\sigma^n \in \mathcal{S}_o$ as follows:
\[
\sigma^n := \frac{1}{2} \sigma^{\otimes n} + \frac{1}{2 |\mathcal{G}^\gamma|} \sum_{\tau \in \mathcal{G}^\gamma} \tau^{\otimes n}, \quad \text{where} \quad \sigma = \sigma(\mathcal{S}_o).
\]  

Note that $\sigma^n$ is normalized and is, in fact, a convex combination of the $n$-fold tensor product of the divergence center and the $n$-fold tensor product of the elements of the net, of which there are only finitely many. With this choice of $\sigma^n$ we bound $D_h^\epsilon(\rho^n || \sigma^n)$ in the following.

We will also need to treat different types of state sequences separately. We keep $\mathcal{S}_o$ fixed for the following to simplify notation. Let us define $\Omega^\nu_1, \Omega^\nu_2 \subseteq \mathcal{S}_o^{\otimes n}$ for some $0 < \nu \leq 1$, which describe sets of state sequences of length $n$ that are close to achieving the first order optimum. (We omit the dependence on $n$ in our notation here.) The first set ensures that the states are close to $\Gamma(\mathcal{S}_o)$, and is defined as
\[
\Omega^\nu_1 := \left\{ \rho^n \in \mathcal{S}_o^{\otimes n} \left| \frac{1}{n} \sum_{i=1}^n \min_{\tau \in \Gamma(\mathcal{S}_o)} \frac{1}{2} \|\rho_i - \tau\|_1 \leq \nu \right\} =: \Delta(\rho^n, \Gamma(\mathcal{S}_o)).
\]

---

9 To see why this is so, consider a sequence of states $\rho^n = \bigotimes_{i=1}^n \rho_i$ with $\rho_i \in \Gamma(\mathcal{S}_o)$. Then, following the notation in the proof of Proposition 13, we realize that $D_n = \chi(\mathcal{S}_o)$. However, since $\frac{1}{n} \sum_{i=1}^n \rho_i \neq \sigma(\mathcal{S}_o)$ in general, the empirical distribution $P_{\rho^n}$ can be arbitrarily far from $\Pi(\mathcal{S}_o)$. Thus, we cannot hope to bound $V_n$ in terms of $v_\epsilon$. 

---
The interesting, close to capacity achieving sequences are those that are in \( \Omega_2^{\nu} \). The second set ensures that the average state is close to the divergence center, and is defined as

\[
\Omega_2^{\nu} := \left\{ \rho^n \in S_0 \mid \frac{1}{n} \left\| \frac{1}{n} \sum_{i=1}^{n} \rho_i - \sigma^*(S_0) \right\|_1 \leq \nu \right\}.
\]

The interesting, close to capacity achieving sequences are those that are in \( \Omega_2^{\nu} \).

The following results constitute the main technical contribution of this paper. We first deal with sequences that are far from optimal in the sense prescribed above.

**Proposition 15.** Let \( \varepsilon \in (0, 1) \), \( \nu > 0 \) and \( S_0 \subseteq S \). Let \( \{\varepsilon_n\}_{n=1}^{\infty} \) be any sequence satisfying \( |\varepsilon_n - \varepsilon| \leq 1/\sqrt{n} \) for all \( n \). Then, there exists constants \( N_0(\varepsilon, S_0, \nu) \) and \( \gamma_0(S_0, \nu) \) such that, for all \( n \geq N_0 \) and all \( \rho^n \notin \Omega_1^{\nu} \cap \Omega_2^{\nu} \), we have

\[
D_h^{\varepsilon_n}(\rho^n, \|\sigma^n\|) \leq n \chi(S_0) + \sqrt{n} v_\varepsilon(S_0) \Phi^{-1}(\varepsilon),
\]

where \( \sigma^n \) is defined as in (23) for a fixed \( \gamma = \gamma_0 \).

**Proof.** The technique for bounding \( D_h^{\varepsilon_n}(\rho^n, \|\sigma^n\|) \) differs depending on the state sequence \( \rho^n \). We consider two cases: (a) \( \rho^n \notin \Omega_1^{\nu} \) and (b) \( \rho^n \notin \Omega_2^{\nu} \) in the following subsections.

(a) Sequences \( \rho^n \notin \Omega_1^{\nu} \): Applying Property 3 of Lemma 7 to \( D_h^{\varepsilon_n}(\rho^n, \|\sigma^n\|) \) with our choice of \( \sigma^n \) in (23) and picking out the divergence center \( \sigma^{\otimes n} \) yields an upper bound of the form

\[
D_h^{\varepsilon_n}(\rho^n, \|\sigma^n\|) \leq C D_h^{\varepsilon_n}(\rho^n, \|\sigma^{\otimes n}\|) + \log(2).
\]

Furthermore, as in the proof of Proposition 13, we employ (14) in Proposition 8 to obtain

\[
D_h^{\varepsilon_n}(\rho^n, \|\sigma^{\otimes n}\|) \leq \sum_{i=1}^{n} D(\rho_i, \|\sigma\|) + K_1 \sqrt{n},
\]

for all \( n \geq N_1 \). Now, we define \( \hat{\chi}_1^\nu := \sup_{\rho \in S_0: \Delta(\rho, \Gamma) > \frac{\nu}{2}} D(\rho, \|\sigma\|) < \chi(S_0) \) and employ the following auxiliary lemma which is shown in Appendix D.

**Lemma 16.** Let \( \rho^n \in S_0^{\otimes n} \) be fixed and let \( \nu \in (0, 1) \). If \( \rho^n \notin \Omega_1^{\nu} \), then there exists a set \( \Xi' \subseteq [n] \) of cardinality \( |\Xi'| > n \nu^2 \) such that, for all \( i \in \Xi' \), we have \( \Delta(\rho_i, \Gamma) > \frac{\nu}{2} \).

This leads us to bound

\[
D_h^{\varepsilon_n}(\rho^n, \|\sigma^n\|) \leq \sum_{i \in \Xi'} \hat{\chi}_1^\nu + \sum_{i \notin \Xi'} \chi(S_0) + K_1 \sqrt{n} \leq n \chi(S_0) - n(\chi(S_0) - \hat{\chi}_1^\nu) + K_1 \sqrt{n}.
\]

In particular, we have \( D_h^{\varepsilon_n}(\rho^n, \|\sigma^n\|) \leq n \chi(S_0) + \sqrt{n} v_\varepsilon(S_0) \Phi^{-1}(\varepsilon) \) for sufficiently large \( n \geq N \), where \( N \) is appropriately chosen.

(b) Sequences \( \rho^n \notin \Omega_2^{\nu} \): For these sequences, we extract the state \( \tau^{\otimes n} \) from the convex combination that defines \( \sigma^n \) in (23), where \( \tau \) is the state closest (in the relative entropy sense) to the average output state \( \tilde{\rho} = \rho(P_{\rho^n}) = \frac{1}{n} \sum_{i=1}^{n} \rho_i \) in \( S' \) and the constant \( \gamma > 0 \) is to be chosen later. In other words, \( \tau \in \arg\min_{\tau \in S'} D(\tilde{\rho}, \|\tau\|) \). Thus, by Property 3 of Lemma 7, we have

\[
D_h^{\varepsilon_n}(\rho^n, \|\sigma^n\|) \leq D_h^{\varepsilon_n}(\rho^n, \|\tau^{\otimes n}\|) + \log |S'|.
\]

Then, by using (14) in Proposition 8 we find for all \( \rho^n \notin \Omega_2^{\nu} \) that

\[
D_h^{\varepsilon_n}(\rho^n, \|\tau^{\otimes n}\|) \leq \sum_{i=1}^{n} D(\rho_i, \|\tau\|) + K_1' \sqrt{n}.
\]
for \( n \geq N'_1 \). Here, we take advantage of the fact that the minimum eigenvalue of \( \tau \) satisfies 
\[ \lambda_{\min}(\tau) \geq \frac{\gamma}{2d+1} \] 
such that the constants \( K'_1 \) and \( N'_1 \) can be chosen uniformly for all \( \tau \in \mathcal{G}^\gamma \).

We continue to bound 
\[ D_h^n(\rho^n\|\sigma^n) \leq \sum_{i=1}^n D(\rho_i\|\tilde{\rho}) + \sum_{i=1}^n \text{tr}(\rho_i(\log \tilde{\rho} - \log \rho_i)) + K'_1 \sqrt{n} \]
\[ = \sum_{i=1}^n D(\rho_i\|\tilde{\rho}) + nD(\tilde{\rho}\|\tau) + K'_1 \sqrt{n} \]
\[ \leq nI(P_{\rho^n}\|\rho(\rho^n)) + n \cdot 4\gamma(2d+1) + K'_1 \sqrt{n}, \]
where the second inequality follows from the properties of the \( \gamma \)-net stated in Lemma 14 and on the last line we introduced the empirical distribution of \( \rho^n \), defined as 
\[ P_{\rho^n}(\rho) = \frac{1}{n} \sum_{i=1}^n 1\{\rho = \rho_i\}. \]

Then, by Theorem 1 and the definition of \( \Pi(\mathcal{S}_o) \) and \( \nu \in (0,1) \), we know that 
\[ \bar{\chi}_2 := \sup \left\{ I(P\|\rho(P)) \middle| P \in \mathcal{P}(\mathcal{S}_o) : \frac{1}{2} \left\| \rho(P) - \sigma^*(\mathcal{S}_o) \right\|_1 > \nu \right\} < \chi(\mathcal{S}_o). \]
Summarizing the above, we have 
\[ D_h^n(\rho^n\|\sigma^n) \leq n\chi(\mathcal{S}_o) - n(\chi(\mathcal{S}_o) - \bar{\chi}_2 - 4\gamma(2d+1)) + K'_1 \sqrt{n} + \log |\mathcal{G}^\gamma|. \]

By choosing \( \gamma = \gamma_0(\nu,\mathcal{S}_o) \) small enough such that \( \chi(\mathcal{S}_o) - \bar{\chi}_2 - 4\gamma(2d+1) > 0 \), we find that 
\[ D_h^n(\rho^n\|\sigma^n) \leq n\chi(\mathcal{S}_o) + \sqrt{n\bar{v}_c(\mathcal{S}_o)}\Phi^{-1}(\varepsilon) \]
for sufficiently large \( n \geq N'_1 \), appropriately chosen.

We conclude by observing that the statement of the proposition holds for \( n \geq \max\{N, N'_1\} \). \( \square \)

Theorem 5 is now a corollary of Proposition 12 and the following result.

**Proposition 17.** Let \( \varepsilon \in (0,1) \) and \( \mathcal{S}_o \subseteq \mathcal{S} \). Let \( \{\varepsilon_n\}_{n=1}^\infty \) be any sequence satisfying \( |\varepsilon_n - \varepsilon| \leq 1/\sqrt{n} \) for all \( n \). Then,
\[ \chi_{\mathcal{S}_o}^{\varepsilon_n}(S^\otimes n) \leq n\chi(\mathcal{S}_o) + \sqrt{n\bar{v}_c(\mathcal{S}_o)}\Phi^{-1}(\varepsilon) + o(\sqrt{n}). \]

**Proof.** For any \( \nu \in (0,1) \), we first invoke Proposition 15 to certify that 
\[ \sup_{\rho^n \not\in \Omega^\varepsilon_{\mathcal{S}_o} \cap \Omega^\nu_{\mathcal{S}_o}} D_h^n(\rho^n\|\sigma^n) \leq n\chi(\mathcal{S}_o) + \sqrt{n\bar{v}_c(\mathcal{S}_o)}\Phi^{-1}(\varepsilon) \]  
(24)
for \( n \geq N_0(\varepsilon, \nu, \mathcal{S}_o) \) sufficiently large. We are left to deal with sequences \( \rho^n \in \Omega^\varepsilon_{\mathcal{S}_o} \cap \Omega^\nu_{\mathcal{S}_o} \). Define the set of sequences \( \rho^n \) with empirical distribution \( P_{\rho^n} \) resulting in a \( \xi \)-positive relative entropy variance as 
\[ \Omega^\xi_{\mathcal{S}_o} := \{\rho^n \in S^\otimes n : V(P_{\rho^n}\|\sigma) \geq \xi\}, \]
(25)
where \( \xi > 0 \) is a constant to be chosen later.

For \( \rho^n \not\in \Omega^\xi_{\mathcal{S}_o} \), we again pick out \( \sigma^\otimes n \) from (23) to find 
\[ D_h^n(\rho^n\|\sigma^n) \leq D_h^n(\rho^n\|\sigma^\otimes n) + \log 2. \]
Then, we employ (14) in Proposition 8 to obtain 
\[ D_h^n(\rho^n\|\sigma^\otimes n) \leq nI(P_{\rho^n}\|\sigma) + \sqrt{nV(P_{\rho^n}\|\sigma)}\Phi^{-1}(\varepsilon) + L_1 \log n \leq n\chi(\mathcal{S}_o) + \sqrt{n\bar{v}_c(\mathcal{S}_o)}\Phi^{-1}(\varepsilon) + K_1 \log n. \]  
(26)
For sequences $\rho^n \in \Omega^n_3$, by the Berry-Esseen-type bound (15) in Proposition 8, we have

$$D_h^n(\rho^n||\sigma^\otimes n) \leq nI(P_{\rho^n}|\sigma) + \sqrt{nV(P_{\rho^n}|\sigma)\Phi^{-1}(\varepsilon)} + L_2 \log n$$

$$\leq n\chi(S_o) + \sqrt{n v^\nu(S_o) \Phi^{-1}(\varepsilon)} + L_2 \log n,$$

(27)

where we define $v^\nu(S_o)$ similarly to $v_\varepsilon(S_o) = v^0_\varepsilon(S_o)$ as

$$v^\nu(S_o) := \begin{cases} \inf_{P \in \Pi^\nu} V(P||\sigma) & \text{if } 0 < \varepsilon \leq \frac{1}{2} \\ \sup_{P \in \Pi^\nu} V(P||\sigma) & \text{if } \frac{1}{2} < \varepsilon < 1 \end{cases}$$

where we employed the set $\Pi^\nu \subseteq \mathcal{P}(S_o)$ of probability measures close to $\Pi(S_o)$, given as

$$\Pi^\nu := \left\{ P \in \mathcal{P}(S_o) \left| \int dP(\rho) \Delta(\rho, \Gamma) \leq \nu \wedge \frac{1}{2} \|\rho^{(P)} - \sigma(S_o)\|_1 \leq \nu \right. \right\}.$$

Clearly, the empirical distribution of a sequence $\rho^n$ is in $\Pi^\nu$ if and only if $\rho^n \in \Omega^\nu_1 \cup \Omega^\nu_2$. The sets $\Pi^\nu$ are compact. Moreover, we may write $\nu > 0$ and $\Pi^\nu$ to recover the definition in (5).

Now, we will choose the parameters $\xi$ and $\nu$ differently depending on some properties of $S_o$. Let us first consider two cases for which $v_\varepsilon(S_o) > 0$.

1. $v_{\min}(S_o) > 0$. In this case, the constant $\xi > 0$ is chosen to be $\xi = \frac{v_{\min}(S_o)}{2} > 0$. Now, for all $\nu$ sufficiently small we have $\inf_{P \in \Pi^\nu} V(P||\sigma) > \xi$ so that $\Omega^\nu_1 \cap \Omega^\nu_2 \setminus \Omega^\nu_3$ is empty. Thus, combining (24) and (27), we find

$$\chi^\varepsilon_h(S_o^\otimes n) \leq \sup_{\rho^n \in S_o^\otimes n} D_h^n(\rho^n||\sigma^\otimes n) \leq n\chi(S_o) + \sqrt{n v^\nu(S_o) \Phi^{-1}(\varepsilon)} + O(\log n)$$

(28)

2. $\varepsilon > \frac{1}{2}$ and $v_{\max}(S_o) > v_{\min}(S_o) = 0$. Here we note that $\Phi^{-1}(\varepsilon) > 0$ and $v_\varepsilon(S_o) > 0$. Thus, we may choose $\xi > 0$ sufficiently small so that

$$\sqrt{\frac{\xi}{\varepsilon}} \leq \sqrt{v_\varepsilon(S_o) \Phi^{-1}(\varepsilon)} \leq \sqrt{v^\nu(S_o) \Phi^{-1}(\varepsilon)}$$

for any $\nu > 0$. The bounds (24), (26) and (27) can then be summarized and (28) holds.

The bounds for cases 1 and 2 can be restated as follows. For all $\nu > 0$ sufficiently small, we have

$$\limsup_{n \to \infty} \frac{\chi^\varepsilon_h(S_o^\otimes n) - n\chi(S_o)}{\sqrt{n}} \leq v^\nu \Phi^{-1}(\varepsilon)$$

Since $\nu > 0$ is arbitrary small we take $\nu \searrow 0$. Then, it remains to show that $\lim_{\nu \to 0} v^\nu(S_o) = v_\varepsilon(S_o)$. This is a consequence of the following lemma (proved in Appendix D).

**Lemma 18.** Let $\Theta_1 \supseteq \Theta_2 \supseteq \ldots$ be a sequence of compact sets in a metric space and let $f: \Theta_1 \to \mathbb{R}$ be continuous and bounded. Then,

$$\lim_{n \to \infty} \inf_{x \in \Theta_n} f(x) = \inf_{x \in \Theta_\infty} f(x) \quad \text{whenever} \quad \Theta_\infty := \bigcap_{n \in \mathbb{N}} \Theta_n \neq \emptyset.$$

(29)

This establishes that $\chi^\varepsilon_h(S_o^\otimes n) \leq n\chi(S_o) + \sqrt{n v^\nu \Phi^{-1}(\varepsilon)} + o(\sqrt{n})$, as desired.

Let us now turn our attention to the cases for which $v_\varepsilon(S_o) = 0$. 
3. \( v_{\max}(\mathcal{S}_o) = v_{\min}(\mathcal{S}_o) = 0 \). Here, we note that for any \( \xi > 0 \) there exists a \( \nu > 0 \) such that 
\[ \sup_{\mathcal{P} \in \Pi^*} V(\mathcal{P}|\sigma) < \xi \] 
and, thus, the set \( \Omega_1^r \cap \Omega_2^r \cap \Omega_3^r \) is empty. Hence, the bounds (24) and (26) can be combined to yield 
\[ \chi_h^{\xi_n}(\mathcal{S}_o^{\otimes n}) \leq \sup_{\rho^n \in \mathcal{B}^{\otimes n}} D_h^{\xi_n}(\rho^n||\sigma^n) \leq n \chi(\mathcal{S}_o) + \sqrt{\frac{n\xi}{\epsilon^*}} + O(\log n). \quad (30) \]

4. \( \varepsilon \leq \frac{1}{2} \) and \( v_{\max}(\mathcal{S}_o) > v_{\min}(\mathcal{S}_o) = 0 \). Here, any choice of \( \xi > 0 \) enforces that 
\[ \sqrt{\frac{\xi}{\epsilon^*}} \geq 0 = \sqrt{v_{\varepsilon}(\mathcal{S}_o)} \Phi^{-1}(\varepsilon). \]

Thus, the bounds (24), (26) and (27) together establish that (30) holds.

Again, let us restate the bounds for cases 3 and 4 as follows. For all \( \xi > 0 \), we have 
\[ \limsup_{n \to \infty} \frac{\chi_h^{\xi_n}(\mathcal{S}_o^{\otimes n}) - n \chi(\mathcal{S}_o)}{\sqrt{n}} \leq \sqrt{\frac{\xi}{\epsilon^*}}. \]
Since \( \xi > 0 \) is arbitrary, we may take \( \xi \searrow 0 \) and deduce that 
\[ \chi_h^{\xi_n}(\mathcal{S}_o^{\otimes n}) \leq n \chi(\mathcal{S}_o) + o(\sqrt{n}). \]
This concurs with the Gaussian approximation since \( v_{\varepsilon}(\mathcal{S}_o) \) is zero and concludes the proof. \( \square \)

4. Asymptotics of the \( \varepsilon \)-Hypothesis-Testing Divergence Radius: Beyond Second Order

In this section we want to improve the upper bound of \( o(\sqrt{n}) \) in Theorem 5 to \( O(\log n) \) for the important special case where \( \mathcal{S}_o \) is a discrete set. To simplify the exposition here, we further assume that \( v_{\min}(\mathcal{S}_o) > 0 \). Comparing with the proof of Proposition 17, it is however easy to see that this condition can be relaxed to \( v_{\varepsilon}(\mathcal{S}_o) > 0 \).

**Proposition 19.** Let \( \varepsilon \in (0, 1) \) and \( \mathcal{S}_o \subseteq \mathcal{S} \) be discrete and \( v_{\min}(\mathcal{S}_o) > 0 \). Let \( \{\varepsilon_n\}_{n=1}^\infty \) be any sequence satisfying \( |\varepsilon_n - \varepsilon| \leq 1/\sqrt{n} \) for all \( n \). Then,
\[ \chi_h^{\varepsilon_n}(\mathcal{S}_o^{\otimes n}) \leq n \chi(\mathcal{S}_o) + \sqrt{n v_{\varepsilon}(\mathcal{S}_o)} \Phi^{-1}(\varepsilon) + O(\log n). \]

**Proof.** For any \( n \), consider all sequences \( \rho^n = \bigotimes_{i=1}^n \rho_i \) with \( \rho_i \in \mathcal{S}_o \). The method of types [3] reveals that \( P_{\rho^n} \) is in a set \( \mathcal{P}_n(\mathcal{S}_o) \subseteq \mathcal{P}(\mathcal{S}_o) \) with cardinality satisfying \( \log |\mathcal{P}_n(\mathcal{S}_o)| = O(\log n) \).

We use this for a further refinement of our state \( \sigma^n \) as follows:
\[ \sigma^n := \frac{1}{3} \sigma^{\otimes n} + \frac{1}{3} \chi^{\otimes n} + \frac{1}{3} \chi^{\otimes n} \sum_{r \in \mathcal{R}} \frac{1}{3 |\mathcal{P}_n(\mathcal{S}_o)|} \sum_{P \in \mathcal{P}_n(\mathcal{S}_o)} (P^{(P)})^{\otimes n}. \quad (31) \]

Clearly, Proposition 15 still applies with this definition, and for any \( \nu \in (0, 1) \) we find that
\[ \sup_{\rho^n \in \Omega_1^r \cap \Omega_2^r} D_h^{\xi_n}(\rho^n||\sigma^n) \leq n \chi(\mathcal{S}_o) + \sqrt{n v_{\varepsilon}(\mathcal{S}_o)} \Phi^{-1}(\varepsilon) \]
Now, observe that due to our condition on the channel, we have \( v_{\min}(\mathcal{S}_o) > 0 \). Thus, \( V(P) = \sum P(\rho) V(\rho||\rho^{(P)}) \) evaluated for \( P \in \Pi \) is lower bounded by \( v_{\min}(\mathcal{S}_o) \). Moreover, by continuity,
inf_{P \in \Pi^v} V(P) > v_{\min}(S_0)/2 is \nu is chosen sufficiently small. Thus, we in particular have that 
\( V(P^{\rho^n}) > v_{\min}/2 \) for all \( \rho^n \in \Omega^n_1 \cap \Omega^n_2 \). For such a sequence \( \rho^n \), we apply Proposition 8 to find 
\[
D_h^n(\rho^n || \sigma^n) \leq D_h^\epsilon(\rho^n \mid (\rho(P))^{\otimes n}) + \log |\mathcal{P}_n(S_0)| \leq nI(P^{\rho^n}) + \sqrt{nV(P^{\rho^n})\Phi^{-1}(\epsilon)} + \log |\mathcal{P}_n(S_0)| + L_3 \log n
\]
for \( n \geq N_3 \). Thus, we immediately find 
\[
\chi^{\epsilon,n}_h(S_0^{\otimes n}) \leq \sup_{P \in \Pi^v} \left( nI(P) + \sqrt{nV(P)\Phi^{-1}(\epsilon)} \right) + O(\log n)
\]
and it only remains to show that the supremum is achieved in \( \Pi \), without too much loss. As Polyanskiy, Poor and Verdú discuss in [29, App. J], we indeed have 
\[
\min \{ \lambda \in \mathbb{R} \mid \lambda \gg 0 \} \text{ achieving}
\]
\[
\sup_{P \in \Pi^v} \left( nI(P) + \sqrt{nV(P)\Phi^{-1}(\epsilon)} \right) \leq \lambda \sum_{v \in S_0} v(\rho) = 0 \text{ such that } P + v \notin \Pi.\]
This is equivalent to the condition 
\[
\frac{d^2}{d\alpha^2}H\left( \rho^{(P)} + \alpha \Delta^{(v)} \right)_{\alpha=0} > 0, \quad \text{where } \Delta^{(v)} := \sum_{\rho \in S_0} v(\rho),
\]
which is satisfied due to Lemma 20 below.

\textbf{Lemma 20.} Let \( \rho \in S \) and let \( \Delta \in \mathfrak{H} \) with \( \text{tr}(\Delta) = 0, \Delta \neq 0 \) and \( \Delta \ll \rho \). Then, we have 
\[
\frac{d^2}{d\lambda^2}H(\rho + \lambda \Delta)_{\lambda=0} < 0. \quad \text{In particular, } \rho \mapsto H(\rho) \text{ is strictly concave.}
\]

Note that that strict negativity of the second derivative is a stronger property than strict concavity, which it implies.\[11\]

We are grateful to David Reeb for allowing us to present a proof based on his ideas here [30].

\textbf{Proof.} We define \( \rho_\lambda := \rho + \lambda \Delta \). First, we note that 
\[
\frac{d}{d\lambda} \rho_\lambda^{-1} = -\rho_\lambda^{-1} \left( \frac{d}{d\lambda} \rho_\lambda \right) \rho_\lambda^{-1} = -\rho_\lambda^{-1} \Delta \rho_\lambda^{-1}
\]
by applying the product rule to \( \frac{d}{d\lambda} (\rho_\lambda^{-1} \rho_\lambda) \). Since \( \Delta \ll \rho \), we can restrict to the subspace \( \{ \rho > 0 \} \) without loss of generality. There, for \( \lambda \) small enough such that \( \rho_\lambda > 0 \), we use the integral representation 
\[
\log \rho_\lambda = \int_0^\infty ds \left( 1 - s \right)^{-1} \text{id} - (\rho_\lambda + s \text{id})^{-1}
\]
which directly follows from its scalar analogue. As such, it is easy to compute 
\[
\frac{d}{d\lambda} \log \rho_\lambda = \int_0^\infty ds - \frac{d}{d\lambda} (\rho_\lambda + s \text{id})^{-1} = \int_0^\infty ds (\rho_\lambda + s \text{id})^{-1} \Delta (\rho_\lambda + s \text{id})^{-1}.
\]
Since \( \frac{d}{d\lambda} \text{tr} \left( f(\rho_\lambda) \right) = \text{tr} \left( f'(\lambda) \frac{d}{d\lambda} \rho_\lambda \right) \) we thus find that 
\[
\frac{d}{d\lambda} H(\rho_\lambda) = -\text{tr}(\Delta \log \rho_\lambda) \quad \text{and}
\]
\[
\frac{d^2}{d\lambda^2} H(\rho_\lambda) = -\int_0^\infty ds \text{tr} \left( \Delta (\rho_\lambda + s \text{id})^{-1} \Delta (\rho_\lambda + s \text{id})^{-1} \right).
\]
For all \( s \geq 0 \) we have \( \text{tr} \left( (\sqrt{\Delta} (\rho + s \text{id})^{-1} \sqrt{\Delta})^2 \right) > 0 \) whenever \( \sqrt{\Delta} (\rho + s \text{id})^{-1} \sqrt{\Delta} \neq 0 \), which is evident since \( \rho \gg \Delta \neq 0 \). Hence, the desired inequality holds.\[12\]

\[10\] Note that \( \frac{d}{d\alpha} I(P + \alpha v)_{|\alpha=0} = 0 \) on \( \Pi \) by definition as \( I(P) \) is maximized on \( \Pi \).

\[11\] This is revealed, for example, by the behavior of the function \( t \mapsto -t^4 \) at \( t = 0 \).
Acknowledgements: We thank Andreas Winter for pointing us to [33] and Mark Wilde for discussions and comments on a previous version of this manuscript. MT thanks David Reeb, Milán Mosonyi and especially Corsin Pfister for many insightful discussions, and the Isaac Newton Institute (Cambridge) for its hospitality while part of this work was completed. MT acknowledges funding by the Ministry of Education (MOE) and National Research Foundation Singapore, as well as MOE Tier 3 Grant “Random numbers from quantum processes” (MOE2012-T3-1-009). VT work is supported by a startup grant from the Faculty of Engineering, NUS.

Appendix A: Proof of Lemma 4

Proof. The inclusion $\supseteq$ is obvious because by the monotonicity of the convex hull operator and the fact that $\Theta_n \supseteq \Theta_\infty$ for any $n \in \mathbb{N}$, we have $\text{conv}(\Theta_n) \supseteq \text{conv}(\Theta_\infty)$.

It remains to prove the inclusion $\subseteq$. Let

$$\rho \in \bigcap_{n \in \mathbb{N}} \text{conv}(\Theta_n).$$

This means that for every $n \in \mathbb{N}$, $\rho$ can be written as $\rho = \sum_{j=1}^{\ell} \alpha_j \rho_j$ where $\rho_j \in \Theta_n$ for each $j = 1, \ldots, \ell$ and $(\alpha_1, \ldots, \alpha_{\ell})$ is a probability distribution. Note that $\ell$ is finite and does not depend on $n$ due to Caratheodory’s theorem since $\Theta_n$ for each $n$ are subsets of the same finite-dimensional vector space.

Consider the sequence $\{\rho_{1n}\}_{n \in \mathbb{N}} \subset \Theta_1$, i.e., $j = 1$. Since $\Theta_1$ is compact, there must exists a convergent subsequence, say indexed by $n_k[1]$, i.e., the sequence $\{\rho_{1n_k[1]}\}_{k \in \mathbb{N}}$ is convergent and

$$\lim_{k \to \infty} \rho_{1n_k[1]} = \rho_1$$

where $\rho_1 \in \Theta_\infty$ since $\Theta_n$ decrease to $\Theta_\infty$. Now, consider the sequence $\{\rho_{2n_k[1]}\}_{k \in \mathbb{N}}$. By the same argument, we may extract a subsequence of $n_k[1]$ indexed by $n_k[2]$ for which

$$\lim_{k \to \infty} \rho_{2n_k[2]} = \rho_2, \quad \text{and} \quad \rho_2 \in \Theta_\infty.$$ 

Continue extracting subsequences until we reach $\ell$. Now consider the subsequence indexed by $n_k := n_k[\ell]$. Clearly, $\rho$ can be written also as

$$\rho = \sum_{j=1}^{\ell} \alpha_j \rho_{jm_k}.$$  \hfill (A2)

By construction, each $\rho_{jm_k}$ converges to $\rho_j \in \Theta_\infty$ when we take $k \to \infty$. So by representation of $\rho$ in (A2), and the arbitrariness of $k$, we have that $\rho$ is a convex combination of elements from $\Theta_\infty$, i.e., $\rho \in \text{conv}(\Theta_\infty)$ as desired.

Appendix B: Background and Proof of Proposition 8

1. Nussbaum-Skoła Distributions

For the proof we leverage on a hierarchy of information measures in quantum information that was introduced in [36]. To apply these results, let us first review the following concept.
For any two quantum states $\rho, \sigma \in S$, we define their (classical) Nussbaum-Skoła distributions $P^{\rho,\sigma}, Q^{\rho,\sigma} \in \mathcal{P}([d] \times [d])$ via the relations [23]

$$P^{\rho,\sigma}(a,b) = r_{a}\langle \phi_a | \psi_b \rangle^2 \quad \text{and} \quad Q^{\rho,\sigma}(a,b) = s_{b}\langle \phi_a | \psi_b \rangle^2,$$

where $\rho = \sum_{a} r_{a}|\phi_a\rangle \langle \phi_a |$ and $\sigma = \sum_{b} s_{b}|\psi_b\rangle \langle \psi_b |$. We summarize some properties of the Nussbaum-Skoła distributions that will turn out to be of great use in the sequel (these were already pointed out in [36]). First, it is easy to verify by substitution that

$$D(\rho || \sigma) = D(P^{\rho,\sigma} || Q^{\rho,\sigma}) \quad \text{and} \quad V(\rho || \sigma) = V(P^{\rho,\sigma} || Q^{\rho,\sigma}). \quad (B1)$$

Second, for product states $\rho_1 \otimes \rho_2$ and $\sigma_1 \otimes \sigma_2$, we have

$$P^{\rho_1 \otimes \rho_2,\sigma_1 \otimes \sigma_2} = P^{\rho_1,\sigma_1} \otimes P^{\rho_2,\sigma_2}, \quad \text{and} \quad Q^{\rho_1 \otimes \rho_2,\sigma_1 \otimes \sigma_2} = Q^{\rho_1,\sigma_1} \otimes Q^{\rho_2,\sigma_2}. \quad (B2)$$

Third, the condition $\sigma \gg \rho$ holds if and only if $Q^{\rho,\sigma} \gg P^{\rho,\sigma}$. Now, let

$$\Xi(\sigma) := 2 \left[ \log \frac{\lambda_{\max}(\sigma)}{\lambda_{\min}(\sigma)} \right],$$

where $\lambda_{\max}(\sigma)$ and $\tilde{\lambda}_{\min}(\sigma)$ denote the largest and smallest nonzero eigenvalues of $\sigma$, respectively.

**Lemma 21.** [36, Thm. 14] Let $\rho, \sigma \in S$ and $\sigma \gg \rho$. Then, for $0 < \delta < \min\{\varepsilon, \frac{1-\varepsilon}{\delta}\}$,

$$D^{\varepsilon}_h(\rho || \sigma) \leq D^{\varepsilon+4\delta}_s(P^{\rho,\sigma} || Q^{\rho,\sigma}) + \log \Xi(\sigma) + 4 \log \frac{1}{\delta} + F_1(\varepsilon, \delta), \quad \text{and}$$

$$D^{\varepsilon}_h(\rho || \sigma) \geq D^{\varepsilon-\delta}_s(P^{\rho,\sigma} || Q^{\rho,\sigma}) - \log \Xi(\sigma) - \log \frac{1}{\delta} - F_2(\varepsilon),$$

where $F_1(\varepsilon, \delta) := \log \frac{(1-\varepsilon)(\varepsilon+3\delta)}{1-(\varepsilon+3\delta)}$ and $F_2(\varepsilon) := \log \frac{1}{1-\varepsilon}$.

Here, the (classical) information spectrum divergence (in the spirit of Verdú and Han [11, 12]) for two probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$ (where $\mathcal{X}$ is a discrete set) is given by

$$D^{s}_s(P || Q) := \sup \left\{ R \in \mathbb{R} \left| \Pr_{X \sim P} \left[ \log \frac{P(X)}{Q(X)} \leq R \right] \leq \varepsilon \right\} \right. \quad (B3)$$

In the following, we also need the absolute third moment of the log-likelihood ratio between $P$ and $Q$, given as\footnote{It is not evident how a non-commutative version of this quantity should be defined directly; however, the commutative case is sufficient for our work.}

$$T(P || Q) := \sum_{x \in \mathcal{X}} P(x) \left| \log \frac{P(x)}{Q(x)} - D(P || Q) \right|^3 \quad \text{and} \quad T(\rho || \sigma) := T(P^{\rho,\sigma} || Q^{\rho,\sigma}).$$

2. Non-Asymptotic Bounds on the $\varepsilon$-Hypothesis-Testing Divergence

It is immediate that the probability appearing in the definition of the information spectrum divergence evaluated for product distributions is subject to the central limit theorem if the variance of $\log \frac{P}{Q}$ is bounded away from zero.
Lemma 22. Let \( n \geq 1 \), \( \{\rho_i\}_{i=1}^n \), for \( \rho_i \in \mathcal{S} \) a set of states and let \( \sigma \in \mathcal{S} \) such that \( \sigma \gg \rho_i \) for all \( i \in [n] \). Moreover, let \( \varepsilon \in (0, 1) \) and \( \delta < \min\{\varepsilon, \frac{1}{4}\} \). Define

\[
D_n := \frac{1}{n} \sum_{i=1}^{n} D(\rho_i \| \sigma), \quad V_n := \frac{1}{n} \sum_{i=1}^{n} V(\rho_i \| \sigma), \quad T_n := \frac{1}{n} \sum_{i=1}^{n} T(\rho_i \| \sigma).
\]

Then, the following Chebyshev-type inequalities hold:

\[
D_n^\varepsilon \left( \bigotimes_{i=1}^{n} \rho_i \| \sigma^{\otimes n} \right) \leq nD_n + \sqrt{\frac{nV_n}{1 - \varepsilon - 4\delta}} + \log \left( n\Xi(\sigma) \right) + 4 \log \frac{1}{\delta} + F_1(\varepsilon, \delta),
\]

\[
D_n^\varepsilon \left( \bigotimes_{i=1}^{n} \rho_i \| \sigma^{\otimes n} \right) \geq nD_n - \sqrt{\frac{nV_n}{\varepsilon - \delta}} - \log \left( n\Xi(\sigma) \right) - \log \frac{1}{\delta} - F_2(\varepsilon).
\]

Moreover, if \( V_n > 0 \), then the following Berry-Esseen type bounds holds:

\[
D_n^\varepsilon \left( \bigotimes_{i=1}^{n} \rho_i \| \sigma^{\otimes n} \right) \leq nD_n + \sqrt{nV_n} \Phi^{-1} \left( \varepsilon + 4\delta + \frac{6T_n}{\sqrt{nV_n^3}} \right) + \log \left( n\Xi(\sigma) \right) + 4 \log \frac{1}{\delta} + F_1(\varepsilon, \delta),
\]

\[
D_n^\varepsilon \left( \bigotimes_{i=1}^{n} \rho_i \| \sigma^{\otimes n} \right) \geq nD_n - \sqrt{nV_n} \Phi^{-1} \left( \varepsilon - \delta - \frac{6T_n}{\sqrt{nV_n^3}} \right) - \log \left( n\Xi(\sigma) \right) - \log \frac{1}{\delta} - F_2(\varepsilon),
\]

where \( F_1, F_2 \) are given in Lemma 21.

Proof (Sketch). We first apply Lemma 21 to replace \( D_n^\varepsilon \) with \( D_n^{\varepsilon + 4\delta} \) (for the upper bounds) and \( D_n^{\varepsilon - \delta} \) (for the lower bound). For this purpose, we note that \( \Xi(\sigma^{\otimes n}) \leq n\Xi(\sigma) \). For the upper bound, this yields

\[
D_n^\varepsilon \left( \bigotimes_{i=1}^{n} \rho_i \| \sigma^{\otimes n} \right) \leq D_n^{\varepsilon + 4\delta} \left( \bigotimes_{i=1}^{n} P^{\rho_i, \sigma} \| \bigotimes_{i=1}^{n} Q^{\rho_i, \sigma} \right) + \log \left( n\Xi(\sigma) \right) + 4 \log \frac{1}{\delta} + F_1(\varepsilon, \delta)
\]

(Note that the information spectrum divergence on the right-hand side of (B6) is evaluated for classical product distributions \( \bigotimes_{i=1}^{n} P^{\rho_i, \sigma} \) and \( \bigotimes_{i=1}^{n} Q^{\rho_i, \sigma} \). Consider the independent random variables

\[
Z_i := \log \frac{P^{\rho_i, \sigma}(A_i, B_i)}{Q^{\rho_i, \sigma}(A_i, B_i)}, \quad (A_i, B_i) \leftarrow P^{\rho_i, \sigma}
\]

for each \( i \in [n] \). Then, the definition of the information spectrum divergence in (B3) yields

\[
D_n^{\varepsilon + 4\delta} \left( \bigotimes_{i=1}^{n} P^{\rho_i, \sigma} \| \bigotimes_{i=1}^{n} Q^{\rho_i, \sigma} \right) = \sup \left\{ R \in \mathbb{R} \mid \Pr \left[ \sum_{i=1}^{n} Z_i \leq R \right] \leq \varepsilon + 4\delta \right\}.
\]

Further, observe that the average mean and variance of \( Z_i \) are respectively given by

\[
\frac{1}{n} \sum_{i=1}^{n} E[Z_i] = \frac{1}{n} \sum_{i=1}^{n} D(P^{\rho_i, \sigma} \| Q^{\rho_i, \sigma}) = \frac{1}{n} \sum_{i=1}^{n} D(\rho_i \| \sigma) = D_n, \quad \text{and}
\]

\[
\frac{1}{n} \sum_{i=1}^{n} \text{Var}[Z_i] = \frac{1}{n} \sum_{i=1}^{n} V(P^{\rho_i, \sigma} \| Q^{\rho_i, \sigma}) = \frac{1}{n} \sum_{i=1}^{n} V(\rho_i \| \sigma) = V_n.
\]

Thus, we apply standard Chebyshev or Berry-Esseen [8, Sec. XVI.5] bounds on the probability in (B7). (See, e.g. [37, Lem. 5], for details.) The proof of the lower bounds proceeds analogously. \qed
3. Uniform Upper Bounds

The following two lemmas give uniform upper bounds on \( V(\rho||\sigma) \) and \( T(\rho||\sigma) \).

**Lemma 23.** Let \( \mathcal{S}_o \subset \mathcal{S} \) and \( \lambda_0 > 0 \). Then, there exists a constant \( V^+(\mathcal{S}_o, \lambda_0) \) such that \( V(\rho||\sigma) \leq V^+ \) for all \( \rho \in \mathcal{S}_o \) and \( \sigma \in \mathcal{S} \) such that \( \lambda_{\min}(\sigma) \geq \lambda_0 \).

**Proof.** First, note that \((\rho, \sigma) \mapsto V(\rho||\sigma)\) is continuous on the compact set \( \mathcal{S}_o \times \{ \sigma \in \mathcal{S} \mid \lambda_{\min}(\sigma) \geq \lambda_0 \} \) since \( \sigma \gg \rho \) everywhere. Thus, we may simply choose

\[ V^+ := \max \{ V(\rho||\sigma) \mid \rho \in \mathcal{S}_o, \ \sigma \in \mathcal{S}, \ \lambda_{\min}(\sigma) \geq \lambda_0 \} \]

**Lemma 24.** Let \( \mathcal{S}_o \subset \mathcal{S} \) and \( \sigma \in \mathcal{S} \) such that \( \lambda_{\min}(\sigma) > 0 \). Then, there exists a constant \( T^+(\mathcal{S}_o, \sigma) \) such that \( T(\rho||\sigma) \leq T^+ \) for all \( \rho \in \mathcal{S}_o \).

**Proof.** We have \( \sigma \gg \rho \) and thus \( Q^{\rho,\sigma} \gg P^{\rho,\sigma} \) for all \( \rho \in \mathcal{S}_o \) since \( \sigma \) is strictly positive. Hence, \( \rho \mapsto T(\rho||\sigma) = T(P^{\rho,\sigma}\|Q^{\rho,\sigma}) \) is continuous and it suffices to define \( T^+ := \max_{\rho \in \mathcal{S}_o} T(\rho||\sigma) \).

For the following, let us define \( V(P) := V(P||\rho(P)) \) and \( T(P) := \sum_{\rho \in \mathcal{S}_o} P(\rho) T(\rho||\rho(P)) \) for all \( P \in \mathcal{P}(\mathcal{S}_o) \) in a discrete set \( \mathcal{S}_o \). These quantities have the following uniform upper bounds:

**Lemma 25.** Let \( \mathcal{S}_o \subset \mathcal{S} \) be discrete. Then, there exists constants \( V^*(\mathcal{S}_o) \) and \( T^*(\mathcal{S}_o) \) such that \( V(P) \leq V^* \) and \( T(P) \leq T^* \) for all \( P \in \mathcal{P}(\mathcal{S}_o) \).

**Proof.** To convince ourselves that the functions \( P \mapsto V(P) \) and \( P \mapsto T(P) \) are continuous, we note that, for all \( P \in \mathcal{P}(\mathcal{S}_o) \) and all \( \rho \in \mathcal{S}_o \) at least one of the following conditions holds 1) \( P(\rho) = 0 \) or 2) \( \rho(P) \gg \rho \). The lemma then follows from the fact that \( \mathcal{P}(\mathcal{S}_o) \) is compact.

4. Proof of Proposition 8

**Proof of Proposition 8.** The first statement relies on the Chebyshev-type inequalities in (B4) in Lemma 22, which for any \( \delta = \frac{1}{\sqrt{n}} \) and for \( n \) sufficiently large such that \( \frac{1}{\sqrt{n}} < \min\{ \varepsilon, \frac{1}{\sqrt{n}} \} \) yield

\[
\left| D_{\delta} \left( \bigotimes_{i=1}^{n} P_i \right) \sigma^{\otimes n} \right| - nD_n \leq \sqrt{\frac{nV_n}{\min\{ 1 - \varepsilon - \frac{1}{\sqrt{n}}, \varepsilon - \frac{1}{\sqrt{n}} \}}} + 3\log n + \mathcal{E}(\sigma)
\]

\[
+ \max \left\{ F_1(\varepsilon, \frac{1}{\sqrt{n}}), F_2(\varepsilon) \right\} \text{.}
\]

Now, we note that \( \mathcal{E}(\sigma) \leq 2\log \frac{1}{\lambda_0} + 1 = O(1) \) and note that

\[
\sqrt{\frac{nV_n}{\min\{ 1 - \varepsilon - \frac{1}{\sqrt{n}}, \varepsilon - \frac{1}{\sqrt{n}} \}}} \leq \sqrt{\frac{nV_n}{\min\{ 1 - \varepsilon - \frac{\delta}{\sqrt{n}}, \varepsilon - \frac{\delta}{\sqrt{n}} \}}} = \sqrt{\frac{nV_n}{\varepsilon^*}} + O(1) \quad \text{and}
\]

\[
\max \left\{ F_1(\varepsilon, \frac{1}{\sqrt{n}}), F_2(\varepsilon) \right\} \leq \max \left\{ F_1(\varepsilon + \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}), F_2(\varepsilon - \frac{1}{\sqrt{n}}) \right\} = O(1) \text{.}
\]

Thus, any choice of \( L_1 > 3 \) will yield the desired result.

Finally, we have \( V_n \leq V^+ \) by Lemma 23 due to the assumption on \( \lambda_{\min}(\sigma) \). We can thus pick the constant

\[
K_1(\varepsilon, \mathcal{S}_o, \lambda_0) > \sqrt{\frac{V^+}{\varepsilon^*}}
\]
uniformly in \( \{\rho_i\}_{i=1}^n \). Finally, for any such choices of \( L_1 \) and \( K_1 \), we find a number \( N_1(\epsilon, S_0, \lambda_0) \) such that the statement holds.

The second statement is based on the Berry Esseen-type inequalities in (B5) in Lemma 22. We prove the upper bound and note that the lower bound follows by an analogous argument. First, we use (B5) and set \( \delta = \frac{1}{\sqrt{n}} \) to establish that

\[
D^n_h \left( \bigotimes_{i=1}^n \rho_i \| \sigma^\otimes n \right) \leq nD_n + \sqrt{nV_n} \Phi^{-1}\left( \epsilon_n + \frac{4}{\sqrt{n}} + \frac{6T_n}{nV_n^2} \right) + 3\log n + \log \Xi(\sigma) + F_1\left( \epsilon_n, \frac{1}{\sqrt{n}} \right)
\]

Now, note that \( V_n \geq \xi \) by assumption of the theorem and \( T_n \leq T^+ (S_0, \sigma) \) by Lemma 24. Since \( \epsilon \mapsto \Phi^{-1}(\epsilon) \) is monotonically increasing, we find

\[
\Phi^{-1}\left( \epsilon_n + \frac{4}{\sqrt{n}} + \frac{6T_n}{nV_n^2} \right) \leq \Phi^{-1}\left( \epsilon + \frac{B}{\sqrt{n}} \right), \quad \text{where} \quad B = 5 + \frac{6T^+ (S_0, \sigma)}{\xi^2}.
\]

Moreover, since \( V_n \leq V^+ \) and \( \epsilon \mapsto \Phi^{-1}(\epsilon) \) is continuously differentiable we find that

\[
\sqrt{nV_n} \Phi^{-1}\left( \epsilon + \frac{B}{\sqrt{n}} \right) \leq \sqrt{nV_n} \Phi^{-1}(\epsilon) + O(1).
\]

by Taylor’s theorem. Collecting the remaining terms as \( 3\log n + O(1) \) and choosing \( L_2 > 3 \) reveals that there exists a constant \( N_2(\epsilon, S_0, \sigma, \xi) \) such that the statement holds.

To confirm the final statement, we need to be a bit more careful because \( \lambda_{\min}(P_{\rho^n}) \) can be arbitrarily close to zero and thus Lemmas 23 and 24 do not apply. However, the proof goes through analogously if we replace them with Lemma 25.

\[\square\]

Appendix C: Proof of Lemma 14

The following construction is likely not optimal in the parameters \( \gamma \) and \( |S^\gamma| \), but it suffices for our purpose and allows us to use previously established results.

**Proof.** First, we employ a construction in [16, Lem. II.4] to establish that, for every \( 0 < \gamma < 1 \), there exists a set of pure states \( \{\psi_i\}_{i \in [K]} \subseteq S_0 \) with cardinality \( K \leq (5/\gamma)^{2d} \) such that the following holds: for every \( \phi \in S_o \), we have \( \min_{i \in [K]} \| \phi - \psi_i \|_1 \leq \gamma \).

Second, consider the set \( \mathcal{P}_m^{>0} \) of \( m \)-types [3] with full support, defined as

\[
\mathcal{P}_m^{>0} := \{ P \in \mathcal{P}([d]) \mid mP(i) \in [m] \text{ for all } i \in [d] \}.
\]

Setting \( m = \lfloor 2d^{\frac{1}{2}} \rfloor \), we will now show that, for every \( P \in \mathcal{P}([d]) \), we have \( \min_{Q \in \mathcal{P}_m^{>0}} \|P - Q\|_1 \leq \gamma \). To see this, we construct a \( Q \in \mathcal{P}_m^{>0} \) for every \( P \) as follows. Start by setting \( Q(i) = \frac{1}{m} \) for all \( i \in [d] \). (Note that \( m > d \) so that the total weight is smaller than one.) Then, pick any index \( i \) for which \( Q(i) < P(i) \) and increase \( Q(i) \) by \( \frac{1}{m} \). Repeat this until \( Q \) is normalized. We observe that

\[
\|P - Q\|_1 = 2\sum_{i:Q(i) > P(i)} Q(i) - P(i) \leq \frac{2d}{m} \leq \gamma
\]

since \( Q(i) - P(i) \) never exceeds \( \frac{1}{m} \) by construction. Note that this choice also ensures that \( \min_i Q(i) \geq \frac{1}{m} \). Furthermore, the number of types is bounded as [3],

\[
|\mathcal{P}_m^{>0}| \leq (m + 1)^{d-1} \leq (2d/\gamma + 2)^{d-1}.
\]

Now, we are ready to define an \( \gamma \)-net for mixed states as follows:

\[
S^\gamma := \{ \tau \in S \mid \tau = \sum_{i=1}^d Q(i)\psi_{\ell(i)}, \ Q \in \mathcal{P}_m^{>0}, \ \ell : [d] \to [K] \}.
\]
We have \(|\mathcal{G}| = K^d \cdot |\mathcal{P}^0| \leq (5/\gamma)^{2d} (2d/\gamma + 2)^{d-1}\). Moreover, let \(\rho \in S(B)\) be an arbitrary state with \(\rho = \sum_i P(i) \phi_i\) its eigenvalue decomposition, where \(\phi_i \in S_o(B)\) are (mutually orthogonal) pure states and \(P \in \mathcal{P}(B)\). Now, choose \(Q \in \mathcal{P}^0\) and \(\ell : [d] \to [K]\) such that

\[
\|P - Q\|_1 \leq \gamma \quad \text{and} \quad \forall i \in [d] : \|\psi_{\ell(i)} - \phi_i\|_1 \leq \epsilon.
\]

For \(\tau = \sum_i Q(i)\psi_{\ell(i)} \in \mathcal{G}\), we then have

\[
\|\rho - \tau\|_1 \leq \sum_i \|P(i)\phi_i - Q(i)\psi_{\ell(i)}\|_1 \leq \sum_i P(i)\|\phi_i - \psi_{\ell(i)}\|_1 + |P(i) - Q(i)| \leq 2\gamma,
\]

where we used the triangle inequality multiple times.

To get the second statement, we employ a continuity result by Audenaert and Eisert [2, Thm. 2], which ensures that \(D(\rho\|\tau) \leq 4\kappa^2 / \beta\), where \(\beta\) is the minimum eigenvalue of \(\tau\), and \(\kappa := 1 \|\rho - \tau\|_1\). By our construction of \(\tau\)—in particular, recall the construction of \(Q \in \mathcal{P}^0\)—we enforce that \(\beta \geq 1/m\). Hence, the above can be further bounded as

\[
D(\rho\|\tau) \leq 4\kappa \cdot \frac{\kappa}{\beta} \leq 4\gamma (2d + 1),
\]

where we used that \(\kappa \leq \gamma\) and \(\kappa/\beta \leq \gamma m = \gamma (2d + 1) \leq 2d + 1\).

Finally, we note that every \(\tau \in \mathcal{G}\) has minimum eigenvalue bounded from below by \(\frac{1}{m} \geq 1/(2d/\gamma + 1) = \gamma/(2d + \gamma)\).

\[\square\]

**Appendix D: Auxiliary Lemmas for Sections V B and V C**

1. **Proof of Lemma 11**

**Proof.** By a simple calculation (or employing the law of total variance), it is easy to verify that

\[
V\left(\bigoplus_{\rho \in S_o} P(\rho) \bigoplus_{\rho \in X_o} P(\rho) \otimes \rho^{(P)}\right) = \sum_{\rho \in S_o} P(\rho) V(\rho\|\rho^{(P)}) + \sum_{\rho \in X_o} P(\rho) \left(D(\rho\|\rho^{(P)}) - \sum_{\rho \in S_o} P(\rho) D(\rho\|\rho^{(P)})\right)^2.
\]

Thus, if we choose \(P \in \Pi(S_o)\) we clearly have \(\rho^{(P)} = \sigma^*(S_o)\) and the second term vanishes due to Property 2 of Theorem 1. \[\square\]

2. **Proof of Lemma 16**

This is a straight-forward generalization of the argument in [29, Lem. 62].

**Proof.** Let \(\Xi'\) be the set of indices for which \(\Delta(\rho_i, \Gamma) > \nu / 2\) holds. Then, we have

\[
\nu < \frac{1}{n} \sum_{i=1}^n \Delta(\rho_i, \Gamma) \leq \frac{1}{n} \sum_{i \in \Xi'} 1 + \frac{1}{n} \sum_{i \not\in \Xi'} \frac{\nu}{2} \leq \frac{|\Xi'|}{n} + \frac{\nu}{2},
\]

from which the condition on the cardinality of \(\Xi'\) follows. \[\square\]
3. Proof of Lemma 18

Proof. First note that all infima can be replaced with minima since the optimization is over compact sets. Denote \( \min_{x \in \Theta} f(x) \) by \( f^* \). Clearly, \( \limsup_{n \to \infty} \min_{x \in \Theta_n} f(x) \leq f^* \) since the inequality holds for every \( n \in \mathbb{N} \) as \( \Theta_n \supseteq \Theta_{\infty} \).

Suppose, for the sake of contradiction that \( \liminf_{n \to \infty} \min_{x \in \Theta_n} f(x) < f^* \). Then, there exists a subsequence indexed by \( \{n_k\}_{k \in \mathbb{N}} \) with the property that \( \min_{x \in \Theta_{n_k}} f(x) < f^* \). For every \( k \in \mathbb{N} \), let \( x_k \in \arg \min_{x \in \Theta_{n_k}} f(x) \) be any minimizer. Since the sets \( \Theta_{n_k} \) are compact, there must exist a converging subsequence indexed by \( \{k_l\}_{l \in \mathbb{N}} \) such that \( \lim_{l \to \infty} x_{k_l} = x^* \). Clearly, \( x^* \) must be in \( \Theta_{\infty} \). However, this leads to a contradiction with \( f(x^*) < f^* = \min_{x \in \Theta_{\infty}} f(x) \). Hence, \( \liminf_{n \to \infty} \min_{x \in \Theta_n} f(x) \geq f^* \).

The limit on the left-hand side of (29) thus exists and equals \( f^* \). \( \square \)

[1] S. Arimoto. On the Converse to the Coding Theorem for Discrete Memoryless Channels. *IEEE Trans. on Inf. Theory*, 19(3):357–359, May 1973. DOI: 10.1109/TIT.1973.1055007.
[2] K. M. R. Audenaert and J. Eisert. Continuity bounds on the quantum relative entropy. *J. Math. Phys.*, 46(10):102104, Oct. 2005. DOI: 10.1063/1.2044667.
[3] I. Csiszár. The Method of Types. *IEEE Trans. on Inf. Theory*, 44(6):2505–2523, Oct. 1998. DOI: 10.1109/18.720546.
[4] N. Datta, M. Mosonyi, M.-H. Hsieh, and F. G. S. L. Brandao. A Smooth Entropy Approach to Quantum Transmission Over Quantum Channels. *IEEE Trans. on Inf. Theory*, 60(3):1562–1572, Mar. 2014. DOI: 10.1109/TIT.2013.2295330.
[5] F. Dupuis, L. Kraemer, P. Faist, J. M. Renes, and R. Renner. Generalized Entropies. In *Proc. of the XVIIth Int. Congress on Math. Phys.*, pages 134–153, Aalborg, Denmark, Nov. 2012. DOI: 10.1142/9789814449243_0008.
[6] F. Dupuis, O. Szehr, and M. Tomamichel. A Decoupling Approach to Classical Data Transmission Over Quantum Channels. *IEEE Trans. on Inf. Theory*, 56(3):1562–1572, Mar. 2014. DOI: 10.1109/TIT.2013.2295330.
[7] H. G. Eggleston. *Convexity*. Cambridge University Press, Cambridge, U.K., 1958.
[8] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, 2nd edition, 1971.
[9] R. G. Gallager. A Simple Derivation of the Coding Theorem and Some Applications. *IEEE Trans. on Inf. Theory*, 11(1):3–18, Jan. 1965. DOI: 10.1109/TIT.1965.1053730.
[10] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, 1968.
[11] T. Han and S. Verdú. Approximation theory of output statistics. *IEEE Trans. on Inf. Theory*, 39(3):752–772, May 1993. DOI: 10.1109/18.256486.
[12] T. S. Han. *Information-Spectrum Methods in Information Theory*. Applications of Mathematics. Springer, 2002.
[13] M. B. Hastings. Superadditivity of Communication Capacity Using Entangled Inputs. *Nat. Phys.*, 5(4):255–257, Mar. 2009. DOI: 10.1038/nphys1224.
[14] M. Hayashi. Information Spectrum Approach to Second-Order Coding Rate in Channel Coding. *IEEE Trans. on Inf. Theory*, 55(11):4947–4966, Nov. 2009. DOI: 10.1109/TIT.2009.2030478.
[15] M. Hayashi and H. Nagaoka. General Formulas for Capacity of Classical-Quantum Channels. *IEEE Trans. on Inf. Theory*, 49(7):1753–1768, July 2003. DOI: 10.1109/TIT.2003.813556.
[16] P. Hayden, D. Leung, P. W. Shor, and A. Winter. Randomizing Quantum States: Constructions and Applications. *Commun. Math. Phys.*, 250(2):1–21, July 2004. DOI: 10.1007/s00220-004-1087-6.
[17] F. Hiai and D. Petz. The Proper Formula for Relative Entropy and its Asymptotics in Quantum Probability. *Commun. Math. Phys.*, 143(1):99–114, Dec. 1991. DOI: 10.1007/BF02100287.
[18] A. Holevo. The Capacity of the Quantum Channel with General Signal States. *IEEE Trans. on Inf. Theory*, 44(1):269–273, Jan. 1998. DOI: 10.1109/18.651037.
[19] A. S. Holevo. Bounds for the Quantity of Information Transmitted by a Quantum Communication Channel. *Probl. Inform. Transm.*, 9(3):177–183, 1973.

[20] K. Li. Second-order asymptotics for quantum hypothesis testing. *Ann. Stat.*, 42(1):171–189, Feb. 2014. DOI: 10.1214/13-AOS1185.

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

[22] M. Müller-Lennert, F. Dupuis, O. Szehr, S. Fehr, and M. Tomamichel. On Quantum Rényi Entropies: A New Generalization and Some Properties. *J. Math. Phys.*, 54(12):122203, June 2013. DOI: 10.1063/1.4838856.

[23] M. Nussbaum and A. Szkola. The Chernoff Lower Bound for Symmetric Quantum Hypothesis Testing. *Ann. Stat.*, 37(2):1040–1057, Apr. 2009. DOI: 10.1214/08-AOS593.

[24] T. Ogawa and H. Nagaoka. Strong Converse to the Quantum Channel Coding Theorem. *IEEE Trans. on Inf. Theory*, 45(7):2486–2489, Nov. 1999. DOI: 10.1109/18.796386.

[25] T. Ogawa and H. Nagaoka. Strong converse and Stein’s lemma in quantum hypothesis testing. *IEEE Trans. on Inf. Theory*, 46(7):2428–2433, Nov. 2000. DOI: 10.1109/18.887855.

[26] M. Ohya, D. Petz, and N. Watanabe. On Capacities of Quantum Channels. *Probability and Mathematical Statistics*, 17(1):179–196, 1997.

[27] K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, New York and London, 1967.

[28] Y. Polyanskiy. *Channel Coding: Non-Asymptotic Fundamental Limits*. PhD thesis, Princeton University, Nov. 2010.

[29] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel Coding Rate in the Finite Blocklength Regime. *IEEE Trans. on Inf. Theory*, 56(5):2307–2359, May 2010. DOI: 10.1109/TIT.2010.2043769.

[30] D. Reeb. Personal Communication, 2013.

[31] J. M. Renes and R. Renner. Noisy Channel Coding via Privacy Amplification and Information Reconciliation. *IEEE Trans. on Inf. Theory*, 57(11):7377–7385, Nov. 2011. DOI: 10.1109/TIT.2011.2162226.

[32] B. Schumacher and M. Westmoreland. Sending Classical Information via Noisy Quantum Channels. *Phys. Rev. A*, 56(1):131–138, July 1997. DOI: 10.1103/PhysRevA.56.131.

[33] B. Schumacher and M. Westmoreland. Optimal signal ensembles. *Phys. Rev. A*, 63(2):022308, Jan. 2001. DOI: 10.1103/PhysRevA.63.022308.

[34] M. Sion. On General Minimax Theorems. *Pacific J. Math.*, 8:171–176, 1958.

[35] V. Strassen. Asymptotische Abschätzungen in Shannons Informationstheorie. In *Trans. Third Prague Conf. Inf. Theory*, pages 689–723, Prague, 1962.

[36] M. Tomamichel and M. Hayashi. A Hierarchy of Information Quantities for Finite Block Length Analysis of Quantum Tasks. *IEEE Trans. on Inf. Theory*, 59(11):7693–7710, Nov. 2013. DOI: 10.1109/TIT.2013.2276628.

[37] M. Tomamichel and V. Y. F. Tan. A Tight Upper Bound for the Third-Order Asymptotics for Most Discrete Memoryless Channels. *IEEE Trans. on Inf. Theory*, 59(11):7041–7051, Nov. 2013. DOI: 10.1109/TIT.2013.2276077.

[38] H. Umegaki. Conditional Expectation in an Operator Algebra. *Kodai Math. Sem. Rep.*, 14:59–85, 1962.

[39] L. Wang and R. Renner. One-Shot Classical-Quantum Capacity and Hypothesis Testing. *Phys. Rev. Lett.*, 108(20), May 2012. DOI: 10.1103/PhysRevLett.108.200501.

[40] M. M. Wilde, A. Winter, and D. Yang. Strong Converse for the Classical Capacity of Entanglement-Breaking and Hadamard Channels. June 2013. arXiv: 1306.1586.

[41] A. Winter. Coding Theorem and Strong Converse for Quantum Channels. *IEEE Trans. on Inf. Theory*, 45(7):2481–2485, 1999. DOI: 10.1109/18.796385.