Moderate deviation analysis for classical communication over quantum channels

Christopher T. Chubb[*1], Vincent Y. F. Tan[†2,3], and Marco Tomamichel[‡1,4]

[1]*Centre for Engineered Quantum Systems, School of Physics, University of Sydney, Sydney, Australia.*
[2]*Department of Electrical and Computer Engineering, National University of Singapore, Singapore.*
[3]*Department of Mathematics, National University of Singapore, Singapore.*
[4]*Centre for Quantum Software and Information, University of Technology Sydney, Sydney, Australia.*

February 15, 2017

Abstract

We analyse families of codes for classical data transmission over quantum channels that have both a vanishing probability of error and a code rate approaching capacity as the code length increases. To characterise the fundamental tradeoff between decoding error, code rate and code length for such codes we introduce a quantum generalisation of the moderate deviation analysis proposed by Altŭg and Wagner as well as Polyanskiy and Verdú. We derive such a tradeoff for classical-quantum (as well as image-additive) channels in terms of the channel capacity and the channel dispersion, giving further evidence that the latter quantity characterises the necessary backoff from capacity when transmitting finite blocks of classical data. To derive these results we also study asymmetric binary quantum hypothesis testing in the moderate deviations regime. Due to the central importance of the latter task, we expect that our techniques will find further applications in the analysis of other quantum information processing tasks.

1 Introduction

The goal of information theory is to find the fundamental limits imposed on information processing and transmission by the laws of physics. One of the early breakthroughs in quantum information theory was the characterisation of the capacity of a classical-quantum (c-q) channel to transmit classical information by Holevo [1,2] and Schumacher-Westmoreland [3]. The *classical capacity* of a quantum channel is defined is the maximal rate (in bits per channel use) at which we can transmit information such that the decoding error vanishes asymptotically as the length of the code increases. However, for many practical applications there are natural restrictions on the code length imposed, for example, by limitations on how much quantum information can be processed coherently. Therefore it is crucial to go beyond the asymptotic treatment and understand the intricate tradeoff between decoding error probability, code rate and code length.

For this purpose, we will study families of codes that *both* have a rate approaching the capacity and an error probability that vanishes asymptotically as the code length n increases. The following tradeoff relation gives a rough illustration of our main result: if the code rate approaches capacity as $\Theta(n^{-t})$ for some $t \in (0, 1/2)$, then the decoding error cannot be smaller than $\exp(-\Theta(n^{1-2t}))$. In fact, we will show that the constants implicit in the Θ notation are determined by a second channel parameter beyond the capacity, called the *channel dispersion*. We will also show that this relation is tight, i.e. there exist families of codes achieving equality asymptotically.

Our work thus complements previous work on the boundary cases corresponding to $t \in \{0, 1/2\}$. The error exponent (or reliability function) of c-q channels (see, e.g., Refs. [4–6]) corresponds to the case $t = 0$ where the rate is bounded away from capacity and the error probability vanishes exponentially in n. This is also called the *large deviations* regime. Moreover, the second-order asymptotics of c-q channels were evaluated by Tomamichel and Tan [7]. They correspond to the case $t = 1/2$ where the rate approaches capacity as $\Theta(n^{-1/2})$ and the error probability is non-vanishing. This is also called the *small deviations* regime.

In the present work, we consider the entire regime in between, which is dubbed the *moderate deviation* regime[1]. The different parameter regimes are illustrated in Figure 1.

---

[*]christopher.chubb@sydney.edu.au
[†]vtan@nus.edu.sg
[‡]marco.tomamichel@uts.edu.au

[1]In the technical analysis, we are considering moderate deviations from the mean of a sum of independent log-likelihood ratios, thus justifying the name emanating from statistics [8, Theorem 3.7.1].
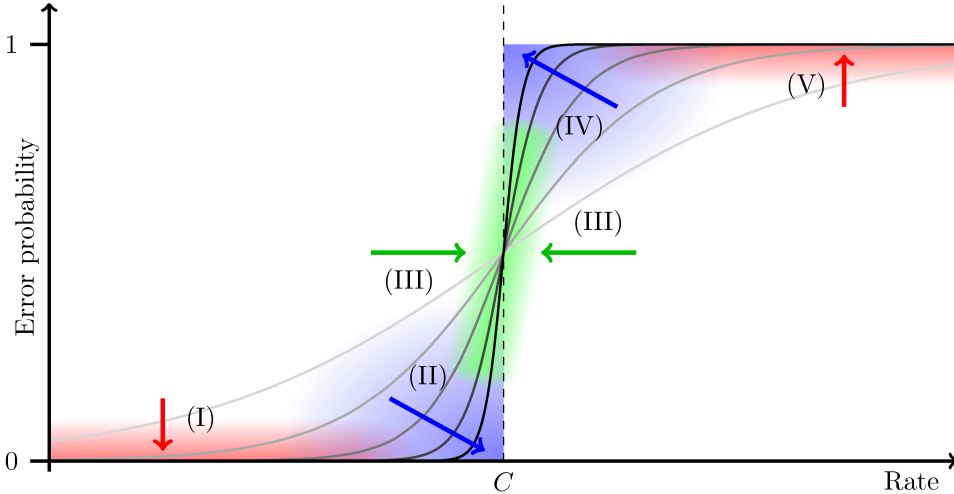
	(I)	(II)	(III)	(IV)	(V)
regime	error exponent	moderate deviation (below capacity)	constant error (second-order)	moderate deviation (above capacity)	strong converse exponent
error prob.	$\exp(-\Theta(n))$	$\exp(-o(n))$ & $\omega(1)$	$\Theta(1)$	$1-\exp(-o(n))$ & $1-\omega(1)$	$1-\exp(-\Theta(n))$
code rate	$C-\Theta(1)$	$C-o(1)$ & $C-\omega(n^{-\frac{1}{2}})$	$C-\Theta(n^{-\frac{1}{2}})$	$C+o(1)$ & $C+\omega(n^{-\frac{1}{2}})$	$C+\Theta(1)$

Figure 1: The figure shows the optimal error probability as a function of the rate, for different block lengths. Darker lines correspond to longer block lengths, and the capacity is denoted by C. The table shows the asymptotics in each region, as the blocklength n goes to infinity. The functions of n implicit in the Θ, o, and ω notation are assumed to be positive-valued.

Main results. Before we present our main results, let us introduce the notion of a *moderate sequence* of real numbers, $\{x_n\}_n$ for $n \in \mathbb{N}$, whose defining properties are that $x_n \to 0$ and $\sqrt{n}\, x_n \to \infty$ as $n \to \infty$.[2] Our two main results concern binary asymmetric quantum hypothesis testing and c-q channel coding.

1. The first result, presented in detail in Section 3, concerns binary quantum hypothesis testing between a pair of of quantum states ρ and σ. We show that for any moderate sequence x_n, there exists a sequence of tests $\{Q_n\}_n$ such that the two kinds of errors satisfy

$$\mathrm{Tr}\,\rho^{\otimes n}(1-Q_n) = e^{-nx_n^2} \quad \text{and} \quad \mathrm{Tr}\,\sigma^{\otimes n}Q_n = \exp\left(-n\Big(D(\rho\|\sigma) - \sqrt{2V(\rho\|\sigma)}\,x_n + o(x_n)\Big)\right), \quad (1)$$

and another sequence of tests $\{Q'_n\}_n$ such that the errors satisfy

$$\mathrm{Tr}\,\rho^{\otimes n}(1-Q'_n) = 1 - e^{-nx_n^2} \quad \text{and} \quad \mathrm{Tr}\,\sigma^{\otimes n}Q'_n = \exp\left(-n\Big(D(\rho\|\sigma) + \sqrt{2V(\rho\|\sigma)}\,x_n + o(x_n)\Big)\right), \quad (2)$$

where $D(\cdot\|\cdot)$ and $V(\cdot\|\cdot)$ denote the relative entropy [9] and relative entropy variance [10,11], respectively. (The reader is referred to the next section for formal definitions of all concepts discussed here.) Most importantly, we show that both of these tradeoffs are in fact optimal.

2. The main result, covered in Section 4, concerns coding over a memoryless classical-quantum channel $\mathcal{W}$. Let us denote by $M^*(\mathcal{W}; n, \varepsilon)$ the maximum $M \in \mathbb{N}$ such that there exists a code transmitting one out of M messages over n uses of the channel $\mathcal{W}$ such that the average probability of error does not exceed ε. For any sequence of tolerated error probabilities $\{\varepsilon_n\}_n$ vanishing sub-exponentially with $\varepsilon_n = e^{-nx_n^2}$, we find that

$$\frac{1}{n}\log M^*(\mathcal{W}; n, \varepsilon_n) = C(\mathcal{W}) - \sqrt{2V_{\min}(\mathcal{W})}\,x_n + o(x_n)\,, \quad (3)$$

$$\frac{1}{n}\log M^*(\mathcal{W}; n, 1-\varepsilon_n) = C(\mathcal{W}) + \sqrt{2V_{\max}(\mathcal{W})}\,x_n + o(x_n)\,, \quad (4)$$

where $C(\cdot)$ denotes the channel capacity and $V_{\min}(\cdot)$ and $V_{\max}(\cdot)$ denote the minimal and maximal channel dispersion as defined in Ref. [7], respectively. This result holds very generally for channels with arbitrary input alphabet and without restriction on the channel dispersion, strengthening also the best known

---

[2]As mentioned above an archetypical moderate sequence is $x_n = \Theta(n^{-t})$ for some $t \in (0, \frac{1}{2})$. The boundary cases are not included — in fact $t = 0$ requires a large deviation analysis whereas $t = \frac{1}{2}$ requires a small deviation analysis.

	asymmetric binary hypothesis testing	channel coding	quantum hypothesis testing	classical-quantum channel coding
large deviation ($<$)	[16]	[17, 18]	[5, 19]	unknown[3]
moderate deviation ($<$)	[20]	[12, 13]	*this work*	*this work*
small deviation	[21]	[21–23]	[10, 11]	[7]
moderate deviation ($>$)	*this work*	*this work*	*this work*	*this work*
large deviation ($>$)	[24, 25]	[26, 27]	[28, 29]	[30]

Table 1: Exposition of related work on finite resource analysis of hypothesis testing and channel coding problems. The rows correspond to different parameter regimes, labelled by the deviation from the critical rate (i.e., the relative entropy for hypothesis testing and the capacity for channel coding problems).

results for classical channels. Moreover, as in Ref. [7], this generality allows us to lift the above result to a statement about coding classical information over image-additive quantum channels and general channels as long as the encoders are restricted to prepare separable states.

Since quantum hypothesis testing underlies many other quantum information processing tasks such as entanglement-assisted classical communication as well as private and quantum communication, we expect that our techniques will have further applications in quantum information theory.

Related work. For classical channels, Altŭg and Wagner [12] first established the best decay rate of the average error probability for a class of discrete memoryless channels (DMCs) when the code rate approaches capacity at a rate slower than $\Theta(n^{-1/2})$. Shortly after the conference version of Ref. [12], Polyanskiy and Verdú [13] relaxed some of the conditions on the class of DMCs and also established the moderate deviations asymptotics for other important classical channels such as the additive white Gaussian noise channel. The other main contributions to the analysis of hypothesis testing, channel coding, quantum hypothesis testing, and c-q channel coding in the different parameter regimes are summarised in Table 1.

From a technical perspective the moderate deviations regime can be approached via a refined large deviations analysis (as was done in Ref. [12]) or via a variation of second-order analysis via the information spectrum method (as was proposed in Ref. [13]). In our work, we mostly follow the latter approach, interspersed with ideas from large deviation theory. In particular, we build on bounds from one-shot information theory by Wang and Renner [14] and use techniques developed for the second-order asymptotics in Ref. [7]. In concurrent work, Cheng and Hsieh [15] provide a moderate deviation analysis for c-q channels via a refined error exponent analysis. Their result holds for c-q channels with finite input alphabets and their techniques are complementary to ours.

2 Preliminaries

2.1 Notation and classical coding over quantum channels

Let $\mathcal{H}$ be a finite-dimensional Hilbert space and denote by $\mathcal{S} := \{\rho \in \mathcal{H} \mid \mathrm{Tr}\,\rho = 1, \rho \geq 0\}$ the quantum states on $\mathcal{H}$. We take $\exp(\cdot)$ and $\log(\cdot)$ to be in an arbitrary but compatible base (such that they are inverses), and denote the natural logarithm by $\ln(\cdot)$. For convenience, we will consider the dimension of this Hilbert space to be a fixed constant, and omit any dependence constants may have on this dimension. For $\rho, \sigma \in \mathcal{S}$ we write $\rho \ll \sigma$ if the support of ρ is contained in the support of σ. For any closed subset $\mathcal{S}_\circ \subseteq \mathcal{S}$, we will denote by $\mathcal{P}(\mathcal{S}_\circ)$ the space of probability distributions supported on $\mathcal{S}_\circ$. We equip $\mathcal{S}$ with the trace metric $\delta_{\mathrm{Tr}}(\rho, \rho') := \frac{1}{2}\|\rho - \rho'\|_1$ and $\mathcal{P}(\mathcal{S})$ with a weak-convergence metric[4] δ_{wc}, such that both are compact metric spaces with

$$f : \mathcal{S} \to \mathbb{R} \text{ continuous} \qquad \Longrightarrow \qquad \mathbb{P} \mapsto \int \mathrm{d}\mathbb{P}(\rho)\, f(\rho) \text{ continuous.} \tag{5}$$

We will use the *cumulative standard normal distribution* function Φ is defined as

$$\Phi(a) := \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, \mathrm{d}x. \tag{6}$$

Following Ref. [7], we consider a general *classical-quantum channel* $\mathcal{W} : \mathcal{X} \to \mathcal{S}$ where $\mathcal{X}$ is any set (without further structure). We define the *image of the channel* as the set $\mathrm{im}\,\mathcal{W} \subset \mathcal{S}$ of all quantum states ρ such that

---

[3]In contrast to classical channels a tight characterisation of the error exponent of c-q channels remains elusive to date even for high rates. See, e.g., Refs. [4–6] for partial progress.

[4]An example of which is the induced Lévy-Prokhorov metric (see, e.g., Section 6 and Theorem 6.4 in Ref. [31]).

$\rho = \mathcal{W}(x)$ for some $x \in \mathcal{X}$. For convenience we assume that our Hilbert space satisfies

$$\mathcal{H} = \bigcup_{\rho \in \operatorname{im} \mathcal{W}} \operatorname{supp}(\rho) \tag{7}$$

such that $\sigma > 0$ is equivalent to $\rho \ll \sigma$ for all $\rho \in \operatorname{im} \mathcal{W}$.

For $M, n \in \mathbb{N}$, an (n, M)-code for a classical-quantum channel $\mathcal{W}$ is comprised of an encoder and a decoder. The *encoder* is a map $e : \{1, 2, \ldots, M\} \to \mathcal{X}^n$ and the *decoder* is a positive operator-valued measure $\{D_m\}_{m=1}^M$ on $\mathcal{H}^{\otimes n}$. Moreover, an (n, M, ε)-code is an (n, M)-code that satisfies

$$\frac{1}{M} \sum_{m=1}^M \operatorname{Tr} \left(\bigotimes_{i=1}^n \mathcal{W}\big(e_i(m)\big) D_m \right) \geq 1 - \varepsilon \,, \tag{8}$$

i.e. the average probability of error does not exceed ε. The finite blocklength achievable region for a channel $\mathcal{W}$ is the set of triples (n, M, ε) for which there exists an (n, M, ε)-code on $\mathcal{W}$. We are particularly interested in the boundary

$$M^*(\mathcal{W}; n, \varepsilon) := \max \left\{ M \in \mathbb{N} : \exists \text{ a } (n, M, \varepsilon)\text{-code for } \mathcal{W} \right\}. \tag{9}$$

Specifically we are going to be concerned with the behaviour of the *maximum rate*, which is defined as $R^*(\mathcal{W}; n, \varepsilon) := \frac{1}{n} \log M^*(\mathcal{W}; n, \varepsilon)$.

2.2 Channel parameters

An important parameter of a channel is the largest rate such that there exists a code of vanishing error probability in the large blocklength limit. This critical rate is known as the *capacity* of a channel $C(\mathcal{W})$, which is defined as

$$C(\mathcal{W}) := \inf_{\varepsilon > 0} \liminf_{n \to \infty} R^*(\mathcal{W}; n, \varepsilon). \tag{10}$$

For classical-quantum channels there exists a *strong converse* bound, which states that the capacity described the asymptotic rate not just for vanishing error probability, but those for non-zero fixed error probabilities as well [32,33]. Together with the original channel coding theorem [1,34], this yields

$$\lim_{n \to \infty} R^*(\mathcal{W}; n, \varepsilon) = C(\mathcal{W}) \quad \text{for all } \varepsilon \in (0, 1). \tag{11}$$

In essence the strong converse tells us that the capacity entirely dictates the asymptotic behaviour of the maximum rate at a fixed error probability. How quickly the rate approaches this asymptotic value for arbitrarily low and high error probabilities are described by the channel *min-dispersion* $V_{\min}(\mathcal{W})$ and *max-dispersion* $V_{\max}(\mathcal{W})$, which are defined respectively as

$$V_{\min}(\mathcal{W}) := \inf_{\varepsilon > 0} \limsup_{n \to \infty} \left(\frac{C(\mathcal{W}) - R^*(\mathcal{W}; n, \varepsilon)}{\Phi^{-1}(\varepsilon)/\sqrt{n}} \right)^2, \tag{12}$$

$$V_{\max}(\mathcal{W}) := \sup_{\varepsilon < 1} \limsup_{n \to \infty} \left(\frac{C(\mathcal{W}) - R^*(\mathcal{W}; n, \varepsilon)}{\Phi^{-1}(\varepsilon)/\sqrt{n}} \right)^2. \tag{13}$$

As with the strong converse, the min and max-dispersions also describe the dispersion at other fixed error probabilities [7]:

$$\lim_{n \to \infty} \left(\frac{C(\mathcal{W}) - R^*(\mathcal{W}; n, \varepsilon)}{\Phi^{-1}(\varepsilon)/\sqrt{n}} \right)^2 = \begin{cases} V_{\min}(\mathcal{W}) & \varepsilon \in (0, 1/2) \\ V_{\max}(\mathcal{W}) & \varepsilon \in (1/2, 1) \end{cases}. \tag{14}$$

2.3 Information quantities

Classically, for two distributions P and Q, the *relative entropy* $D(P\|Q)$ and *relative entropy variance* $V(P\|Q)$ are both defined as the mean and variance of the log-likelihood ratio $\log\big(P/Q\big)$ with respect to the distribution P. In the non-commutative case, for $\rho, \sigma \in \mathcal{S}$ with $\rho \ll \sigma$, these definitions are generalised as [9–11]

$$D(\rho\|\sigma) := \operatorname{Tr} \rho \left(\log \rho - \log \sigma \right), \tag{15}$$

$$V(\rho\|\sigma) := \operatorname{Tr} \rho \big(\log \rho - \log \sigma - D\left(\rho\|\sigma\right) \cdot \operatorname{id} \big)^2. \tag{16}$$

If $\rho \not\ll \sigma$ both quantities are set to $+\infty$.

Following Ref. [7], for a closed set $\mathcal{S}_\circ \in \mathcal{S}$, the *divergence radius*[5] $\chi(\mathcal{S}_\circ)$ is given by

$$\chi(\mathcal{S}_\circ) = \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S}_\circ)} \int \mathrm{d}\mathbb{P}(\rho)\, D\left(\rho \,\middle\|\, \int \mathrm{d}\mathbb{P}(\rho')\, \rho'\right). \tag{17}$$

where $\mathcal{P}(\mathcal{S}_\circ)$ denotes the space of distributions on $\mathcal{S}_\circ$. If we let $\Pi(\mathcal{S}_\circ)$ denote the distributions which achieve the above supremum, we also define the *minimal and maximal peripheral variance*, $v_{\min}(\mathcal{S}_\circ)$ and $v_{\max}(\mathcal{S}_\circ)$, as

$$v_{\min}(\mathcal{S}_\circ) := \inf_{\mathbb{P} \in \Pi(\mathcal{S}_\circ)} \int \mathrm{d}\mathbb{P}(\rho)\, V\left(\rho \,\middle\|\, \int \mathrm{d}\mathbb{P}(\rho')\, \rho'\right), \tag{18}$$

$$v_{\max}(\mathcal{S}_\circ) := \sup_{\mathbb{P} \in \Pi(\mathcal{S}_\circ)} \int \mathrm{d}\mathbb{P}(\rho)\, V\left(\rho \,\middle\|\, \int \mathrm{d}\mathbb{P}(\rho')\, \rho'\right). \tag{19}$$

For the image of a quantum channel, the above three information quantities correspond exactly to the three previously defined channel parameters [7]. Specifically, for $\mathcal{S}_\circ = \overline{\operatorname{im} \mathcal{W}}$, we have

$$C(\mathcal{W}) = \chi(\mathcal{S}_\circ), \qquad V_{\min}(\mathcal{W}) = v_{\min}(\mathcal{S}_\circ), \qquad V_{\max}(\mathcal{W}) = v_{\max}(\mathcal{S}_\circ). \tag{20}$$

2.4 Moderate deviation tail bounds

We now discuss the relevant tail bounds we will require in the moderate deviation regime. Let $\{X_{i,n}\}_{i \leq n}$ be independent zero-mean random variables, and define the average variance as

$$V_n := \frac{1}{n} \sum_{i=1}^{n} \operatorname{Var}[X_{i,n}]. \tag{21}$$

Given certain bounds on the moments and cumulants of these variables, which we will make explicit below, we will see that the probability that the average variable $\frac{1}{n} \sum_{i=1}^{n} X_{i,n}$ deviates from the mean by a moderate sequence $\{t_n\}_n$ decays asymptotically as

$$\ln \Pr\left[\frac{1}{n} \sum_{i=1}^{n} X_{i,n} \geq t_n\right] = -\left(1 + o(1)\right) \frac{n t_n^2}{2 V_n}. \tag{22}$$

Lemma 1 (Moderate deviation lower bound). *If there exist constants $\nu > 0$ and τ such that $\nu \leq V_n$ and*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[|X_{i,n}|^3\right] \leq \tau \tag{23}$$

for all n, then for any $\eta > 0$ there exists a constant $N(\{t_i\}, \nu, \tau, \eta)$ such that, for all $n \geq N$, the probability of a moderate deviation is lower bounded as

$$\ln \Pr\left[\frac{1}{n} \sum_{i=0}^{n} X_{i,n} \geq t_n\right] \geq -(1 + \eta) \frac{n t_n^2}{2 V_n}. \tag{24}$$

Lemma 2 (Moderate deviation upper bound). *If there exists a constant γ such that*

$$\frac{1}{n} \sum_{i=1}^{n} \sup_{s \in [0, 1/2]} \left|\frac{\mathrm{d}^3}{\mathrm{d}s^3} \ln \mathbb{E}\left[e^{s X_{i,n}}\right]\right| \leq \gamma, \tag{25}$$

for all n, then for any $\eta > 0$ there exists a constant $N(\{t_i\}, \gamma, \eta)$ such that, for all $n \geq N$, the probability of a moderate deviation is upper bounded as

$$\ln \Pr\left[\frac{1}{n} \sum_{i=1}^{n} X_{i,n} \geq t_n\right] \leq -\frac{n t_n^2}{2 V_n + \eta}. \tag{26}$$

If V_n has a uniform lower bound, then as $\eta \searrow 0$ the above two bounds sandwich together, giving the two-sided asymptotic scaling of Equation (22). In this case we can see that

$$\sigma\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \sqrt{V_n/n} = \Theta(1/\sqrt{n}) \quad \text{and} \quad \sqrt{\frac{1}{n} \sum_{i=1}^{n} \sigma^2[X_i]} = \sqrt{V_n} = \Theta(1), \tag{27}$$

---

[5]Whilst Equation (17) characterises the divergence radius, we will mostly rely on a more useful form presented in Definition 13.

where $\sigma[\cdot]$ denotes the standard deviation. If we interpret the standard deviation as setting the 'length-scale' on which a distribution decays, then the above two quantities — the deviation of the average, and average[6] of the deviation — set the length-scales of small and large deviation bounds respectively. Using this intuition, we can generalise moderate deviation bounds to give tight two-sided bounds for distributions with arbitrary normalisation, in which V_n is no longer bounded. To do this we will tail bound for deviations which are moderate, *in units of* $\sqrt{V_n}$.

Corollary 3 (Dimensionless moderate deviation bound)**.** *If there exists a γ such that*

$$\frac{1}{nV_n^{3/2}} \sum_{i=1}^{n} \sup_{s \in [0,1/2]} \left| \frac{\mathrm{d}^3}{\mathrm{d}s^3} \ln \mathbb{E}\left[e^{sX_{i,n}} \right] \right| \leq \gamma, \tag{28}$$

for all n, then there exists a constant $N(\{t_i\}, \gamma)$ such that, for all $n \geq N$, we have the two-sided bound

$$-(1+\eta)\frac{nt_n^2}{2} \leq \ln \Pr\left[\frac{1}{n} \sum_{i=1}^{n} X_{i,n} \geq t_n \sqrt{V_n} \right] \leq -(1-\eta)\frac{nt_n^2}{2}. \tag{29}$$

We present proofs of these lemmas in Appendix A.

2.5 Reversing lemma

Intuitively one might expect that moderate deviation bounds can be 'reversed' e.g. that the bound on the probability given the deviation (see Lemmas 1 and 2) of the form

$$\lim_{n \to \infty} \frac{V_n}{nt_n^2} \ln \Pr\left[\frac{1}{n} \sum_{i=1}^{n} X_i \geq t_n \right] = -\frac{1}{2}, \tag{30}$$

is equivalent to a bound on the deviation given the probability

$$\lim_{n \to \infty} \frac{1}{t_n} \inf \left\{ t \in \mathbb{R} \,\middle|\, \frac{V_n}{nt_n^2} \ln \Pr\left[\frac{1}{n} \sum_{i=1}^{n} X_i \geq t \right] \leq -\frac{1}{2} \right\} = 1. \tag{31}$$

We will now see that such an ability to 'reverse' moderate deviation bounds is generic. We do this by considering two quantities A and B defined on the same domain, and considering the infimum value of each quantity for a fixed value of the other.

Lemma 4 (Reversing Lemma)**.** *Let $\{A_i\}_i$ and $\{B_i\}_i$ be sequences of real functions with $\inf_t A_i(t) \leq 0$ and $\inf_t B_i(t) \leq 0$ for all i. If we define $\hat{A}_n(b) := \inf_t \{A_n(t) | B_n(t) \leq b\}$ and $\hat{B}_n(a) := \inf_t \{B_n(t) | A_n(t) \leq a\}$, then*

$$\lim_{n \to \infty} \frac{\hat{A}_n(b_n)}{b_n} = 1, \quad \forall \{b_n\} \text{ moderate} \qquad \Longleftrightarrow \qquad \lim_{n \to \infty} \frac{\hat{B}_n(a_n)}{a_n} = 1, \quad \forall \{a_n\} \text{ moderate}. \tag{32}$$

Proof. See Appendix B. $\qquad\square$

3 Hypothesis testing

Whilst the divergence radius characterises the channel capacity, one-shot channel bounds are characterised by a quantity known as the *ε-hypothesis testing divergence* [14]. As the name suggests, as well as being relevant to one-shot channel coding bounds, the hypothesis testing divergence also has an operational interpretation in the context of hypothesis testing of quantum states. We will start by considering a moderate deviation analysis of this quantity.

3.1 Hypothesis testing divergence

Consider a hypothesis testing problem, in which ρ and σ correspond to the null and alternative hypotheses respectively. A test between these hypotheses will take the form of a POVM $\{Q, I - Q\}$, where $0 \leq Q \leq I$. For a given Q, the type-I and type-II error probabilities are given by

$$\alpha(Q; \rho, \sigma) := \mathrm{Tr}(I - Q)\rho, \qquad \beta(Q; \rho, \sigma) := \mathrm{Tr}\, Q\sigma. \tag{33}$$

---

[6]More specifically the root-mean-square

If we define the smallest possible type-II error given a type-I error at most ε as

$$\beta_\varepsilon(\rho\|\sigma) := \min_{0 \leq Q \leq \mathbb{I}} \{\beta(Q; \rho, \sigma) \,|\, \alpha(Q; \rho, \sigma) \leq \varepsilon\}, \tag{34}$$

then the ε-hypothesis testing divergence is defined as

$$D_{\mathrm{h}}^\varepsilon(\rho\|\sigma) := -\log\frac{\beta_\varepsilon(\rho\|\sigma)}{1 - \varepsilon}. \tag{35}$$

We note that the denominator of $1 - \varepsilon$ follows the normalisation in [35] such that $D_{\mathrm{h}}^\varepsilon(\rho\|\rho) = 0$ for all ρ.

An obvious extension of this hypothesis problem is to the case of n copies of each state, i.e. a hypothesis test between $\rho^{\otimes n}$ and $\sigma^{\otimes n}$, or more generally between two product states $\otimes_{i=1}^n \rho_i$ and $\otimes_{i=1}^n \sigma_i$. A second-order analysis of the ε-hypothesis testing divergence for a non-vanishing ε was given in [10, 11].

Theorem 1 (Moderate deviation of the hypothesis testing divergence). *For any moderate sequence $\{a_n\}_n$ and states $\{\rho_n\}_n$ and $\{\sigma_n\}_n$ such that both $\lambda_{\min}(\sigma_i)$ and $V(\rho_i\|\sigma_i)$ are both uniformly bounded away from zero, the ε_n- and $(1-\varepsilon_n)$-hypothesis testing divergences of non-uniform product states for $\varepsilon_n = e^{-na_n^2}$ scale as*

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \bigotimes_{i=1}^n \sigma_i\right) = D_n - \sqrt{2V_n}\, a_n + o(a_n), \tag{36}$$

$$\frac{1}{n}D_{\mathrm{h}}^{1-\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \bigotimes_{i=1}^n \sigma_i\right) = D_n + \sqrt{2V_n}\, a_n + o(a_n), \tag{37}$$

where $D_n := \frac{1}{n}\sum_{i=1}^n D(\rho_i\|\sigma_i)$ and $V_n := \frac{1}{n}\sum_{i=1}^n V(\rho_i\|\sigma_i)$. More specifically for any ρ and σ such that $\rho \ll \sigma$, the hypothesis testing divergences of uniform product states scale as

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\rho^{\otimes n}\,\middle\|\,\sigma^{\otimes n}\right) = D(\rho\|\sigma) - \sqrt{2V(\rho\|\sigma)}\, a_n + o(a_n), \tag{38}$$

$$\frac{1}{n}D_{\mathrm{h}}^{1-\varepsilon_n}\left(\rho^{\otimes n}\,\middle\|\,\sigma^{\otimes n}\right) = D(\rho\|\sigma) + \sqrt{2V(\rho\|\sigma)}\, a_n + o(a_n). \tag{39}$$

In Section 3.3 we will bound the regularised hypothesis testing divergences towards the relative entropy (the *inward bound*), and in Section 3.4 we will bound them away (the *outward bound*).

Remark 1. *For sequences ε_n bounded away from zero and one the second-order expansion in Refs. [10, 11] yields*

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\rho^{\otimes n}\,\middle\|\,\sigma^{\otimes n}\right) = D(\rho\|\sigma) + \sqrt{\frac{V(\rho\|\sigma)}{n}}\,\Phi^{-1}(\varepsilon_n) + O\left(\frac{\log n}{n}\right), \tag{40}$$

where Φ denotes the cumulative distribution function of the standard normal. As already pointed out in Ref. [13], for small ε_n we have $\Phi^{-1}(\varepsilon_n) \approx \sqrt{-2\ln\varepsilon_n}$. Ignoring all higher order terms, the substitution $\varepsilon_n = e^{-na_n^2}$ into (40) then recovers the expression in (38). In this sense the two results thus agree at the boundary between small and moderate deviations.

Remark 2. *A similarly argument can be sketched at the boundary between moderate and large deviations. The quantum Hoeffding bound [5, 19] states that if $\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}(\rho^{\otimes n}\|\sigma^{\otimes n}) \leq D(\rho\|\sigma) - r$ for some small $r > 0$ then ε_n drops exponentially in n with the exponent given by*

$$\sup_{0 \leq \alpha < 1} \frac{\alpha - 1}{\alpha}\left[D(\rho\|\sigma) - r - D_\alpha(\rho\|\sigma)\right], \tag{41}$$

where $D_\alpha(\rho\|\sigma)$ is the Petz' quantum Rényi relative entropy [36]. For sufficiently small r, the expression in (41) attains its supremum close to $\alpha = 1$ and we can thus approximate $D_\alpha(\rho\|\sigma) \approx D(\rho\|\sigma) + \frac{\alpha-1}{2}V(\rho\|\sigma)$ by its Taylor expansion [37]. Evaluating this approximate expression yields

$$\varepsilon_n = e^{-n\frac{r^2}{2V(\rho\|\sigma)}}. \tag{42}$$

up to leading order in r. Substituting $r = \sqrt{2V(\rho\|\sigma)}a_n$ into (42) then recovers (38). An essentially equivalent argument is also applicable to the strong converse exponent derived in Ref. [28].

3.2 Nussbaum-Szkoła distributions

To allow us to apply a moderate deviation analysis to the quantum hypothesis testing divergence, we leverage the results of Ref. [10] which allow us to reduce the hypothesis testing divergence of quantum states to a quantity known as the information spectrum divergence of certain classical distributions, known as the Nussbaum-Szkoła distributions.

Definition 5 (Nussbaum-Szkoła distributions [38]). *The Nussbaum-Szkoła distributions for a pair of states ρ and σ are given by*

$$P^{\rho,\sigma}(a,b) = r_a \left| \langle \phi_a | \psi_b \rangle \right|^2 \qquad \text{and} \qquad Q^{\rho,\sigma}(a,b) = s_b \left| \langle \phi_a | \psi_b \rangle \right|^2 \tag{43}$$

where the states are eigendecomposed as $\rho = \sum_a r_a |\phi_a\rangle\langle\phi_a|$ and $\sigma = \sum_b s_b |\psi_b\rangle\langle\psi_b|$.

The power of the Nussbaum-Szkoła distributions lies in their ability to reproduce both the divergence and variance of the underlying quantum states

$$D(\rho\|\sigma) = D(P^{\rho,\sigma}\|Q^{\rho,\sigma}), \qquad \text{and} \qquad V(\rho\|\sigma) = V(P^{\rho,\sigma}\|Q^{\rho,\sigma}). \tag{44}$$

As well as capturing these asymptotic quantities, the hypothesis testing relative entropy, which arises one-shot channel coding bounds, can also be captured by the Nussbaum-Szkoła distributions. Specifically this is done via the *information spectrum divergence*, which is defined for two classical distributions P and Q by a tail bound on the log-likelihood ratio as

$$D_s^\varepsilon(P\|Q) := \sup \left\{ R \ \middle| \ \Pr_{X \leftarrow P}\left[\log \frac{P(X)}{Q(X)} \leq R \right] \leq \varepsilon \right\}. \tag{45}$$

Inserting the Nussbaum-Skzoła distributions, we find that the (classical) information spectrum divergence approximates the (quantum) hypothesis testing divergence.

Lemma 6 (Thm. 14, Ref. [10]). *There exists a universal constant K such that for any states ρ and σ with $\lambda_{\min}(\sigma) \geq \lambda$ and $\varepsilon < 1/2$, we find that $D_h^\varepsilon(\rho\|\sigma)$ is bounded as*

$$D_h^\varepsilon(\rho\|\sigma) \leq D_s^{2\varepsilon}(P^{\rho,\sigma}\|Q^{\rho,\sigma}) + \log \frac{1-\varepsilon}{\varepsilon^3(1-2\varepsilon)} + \log K \lceil \ln(1/\lambda) \rceil \tag{46}$$

$$D_h^\varepsilon(\rho\|\sigma) \geq D_s^{\varepsilon/2}(P^{\rho,\sigma}\|Q^{\rho,\sigma}) - \log \frac{1}{\varepsilon(1-\varepsilon)} - \log K \lceil \ln(1/\lambda) \rceil, \tag{47}$$

and $D_h^{1-\varepsilon}(\rho\|\sigma)$ is bounded as

$$D_h^{1-\varepsilon}(\rho\|\sigma) \leq D_s^{1-\varepsilon/2}(P^{\rho,\sigma}\|Q^{\rho,\sigma}) + \log \frac{1-\varepsilon/2}{\varepsilon^4} + \log K \lceil \ln(1/\lambda) \rceil \tag{48}$$

$$D_h^{1-\varepsilon}(\rho\|\sigma) \geq D_s^{1-2\varepsilon}(P^{\rho,\sigma}\|Q^{\rho,\sigma}) - \log \frac{1}{\varepsilon^2} - \log K \lceil \ln(1/\lambda) \rceil. \tag{49}$$

As the information spectrum divergence is defined in terms of a tail bound, we will bound these quantities using the moderate deviation tail bounds of Section 2.4. To do this, we will start by showing that the log-likelihood ratio of Nussbaum-Szkoła distributions is sufficiently well behaved, specifically that its cumulant generating function has bounded derivatives.

Lemma 7 (Bounded cumulants). *For $\lambda > 0$, there exists constants $C_k(\lambda)$ such that the cumulant generating function $h(t) := \ln \mathbb{E}\left[e^{tZ} \right]$ of the log-likelihood ratio $Z := \log P^{\rho,\sigma}/Q^{\rho,\sigma}$ for $\lambda_{\min}(\sigma) \geq \lambda$ is smooth and has uniformly bounded derivatives in a neighbourhood of the origin*

$$\sup_{|t| \leq 1/2} \left| \frac{\partial^k}{\partial t^k} h(t) \right| \leq C_k. \tag{50}$$

We present a proof of this lemma in Appendix C.

3.3 Inward bound

Proposition 8 (Inward bound). *For any constants $\lambda, \eta > 0$, there exists a constant $N(\{a_i\}, \lambda, \eta)$ such that, for $n \geq N$, the hypothesis testing divergence can be bounded for any states $\{\rho_i\}_i$ and $\{\sigma_i\}_i$ with $\lambda_{\min}(\sigma_i) \geq \lambda$ as*

$$\frac{1}{n} D_h^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \middle\| \bigotimes_{i=1}^n \sigma_i \right) \geq D_n - \sqrt{2V_n} a_n - \eta a_n,$$

$$\frac{1}{n} D_h^{1-\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \middle\| \bigotimes_{i=1}^n \sigma_i \right) \leq D_n + \sqrt{2V_n}\, a_n + \eta a_n.$$

where $D_n := \frac{1}{n}\sum_{i=1}^n D(\rho_i\|\sigma_i)$ and $V_n := \frac{1}{n}\sum_{i=1}^n V(\rho_i\|\sigma_i)$.

Proof. Firstly, let Z_i be the log-likelihood ratios

$$Z_i := \log \frac{P^{\rho_i,\sigma_i}(A_i, B_i)}{Q^{\rho_i,\sigma_i}(A_i, B_i)}, \qquad (A_i, B_i) \leftarrow P^{\rho_i,\sigma_i}. \tag{51}$$

In terms of these log-likelihood ratios, the lower and upper bound on the ε_n- and $(1 - \varepsilon_n)$-hypothesis testing divergences respectively from Lemma 6 become

$$D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^{n} \rho_i \middle\| \bigotimes_{i=1}^{n} \sigma_i \right) \geq \sup \left\{ R \;\middle|\; \Pr \left[\sum_{i=1}^{n} Z_i \leq R \right] \leq \varepsilon_n/2 \right\} - \log \frac{1}{\varepsilon_n(1-\varepsilon_n)} - \log Kn\lceil \ln(1/\lambda) \rceil, \tag{52}$$

$$D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^{n} \rho_i \middle\| \bigotimes_{i=1}^{n} \sigma_i \right) \leq \sup \left\{ R \;\middle|\; \Pr \left[\sum_{i=1}^{n} Z_i \leq R \right] \leq 1 - \varepsilon_n/2 \right\} + \log \frac{1 - \varepsilon_n/2}{\varepsilon_n^4} + \log Kn\lceil \ln(1/\lambda) \rceil. \tag{53}$$

Recalling that $\varepsilon_n := e^{-na_n^2}$, we can see that in both cases the error terms scale like $\Theta(na_n^2)$ and $\Theta(\log n)$ respectively, which are both $o(na_n)$. As such, there must exist an $N_1(\{a_i\}, \lambda, \eta)$ such that, for $n \geq N_1$, these error terms are bounded by $\eta na_n/2$ as

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^{n} \rho_i \middle\| \bigotimes_{i=1}^{n} \sigma_i \right) \geq \frac{1}{n} \sup \left\{ R \;\middle|\; \Pr \left[\sum_{i=1}^{n} Z_i \leq R \right] \leq \varepsilon_n/2 \right\} - \eta a_n/2, \tag{54}$$

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^{n} \rho_i \middle\| \bigotimes_{i=1}^{n} \sigma_i \right) \leq \frac{1}{n} \sup \left\{ R \;\middle|\; \Pr \left[\sum_{i=1}^{n} Z_i \leq R \right] \leq 1 - \varepsilon_n/2 \right\} + \eta a_n/2. \tag{55}$$

Next we want to apply the tail bounds of Section 2.4. To this end, we will start by defining zero-mean variables $X_i := Z_i - D(\rho_i\|\sigma_i)$. In terms of these variables, the above bounds take the form

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^{n} \rho_i \middle\| \bigotimes_{i=1}^{n} \sigma_i \right) \geq D_n - \inf \left\{ t \;\middle|\; \Pr \left[\frac{1}{n} \sum_{i=1}^{n} (-X_i) \geq t \right] \leq \varepsilon_n/2 \right\} - \eta a_n/2, \tag{56}$$

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^{n} \rho_i \middle\| \bigotimes_{i=1}^{n} \sigma_i \right) \leq D_n + \inf \left\{ t \;\middle|\; \Pr \left[\frac{1}{n} \sum_{i=1}^{n} (+X_i) \geq t \right] \leq \varepsilon_n/2 \right\} + \eta a_n/2. \tag{57}$$

By Lemma 7 there exists constants $\bar{V}(\lambda)$ and $\gamma(\lambda)$, such that $V_i \leq \bar{V}$ and

$$\sup_{t \in [0,1/2]} \left| \frac{\mathrm{d}^3}{\mathrm{d}s^3} \ln \mathbb{E} \left[e^{s(\pm X_i)} \right] \right| \leq \gamma \tag{58}$$

for all i. If we let $t_n := \left(\sqrt{2V_n} + \eta/2 \right) a_n$, then Lemma 2 gives an $N_2(\{a_i\}, \lambda, \eta)$ such that, for $n \geq N_2$, the tail probability is bounded as

$$\ln \Pr \left[\frac{1}{n} \sum_{i=1}^{n} (\pm X_i) \geq t_n \right] \leq \frac{-nt_n^2}{2V_n + \eta^2/3} \tag{59}$$

$$\leq -\frac{\left(\sqrt{2V_n} + \eta/2 \right)^2}{2V_n + \eta^2/5} na_n^2 \tag{60}$$

$$\leq -\frac{2V_n + \eta^2/4}{2V_n + \eta^2/5} na_n^2 \tag{61}$$

$$= -\left(1 + \frac{\eta^2}{40V_n + 4\eta^2} \right) na_n^2 \tag{62}$$

$$\leq -\left(1 + \frac{\eta^2}{40\bar{V} + 4\eta^2} \right) na_n^2. \tag{63}$$

$$\tag{64}$$

As $\eta^2/(40\bar{V} + 4\eta)$ is a constant and $na_n^2 \to \infty$, there must exist a constant $N_3(\{a_i\}, \lambda, \eta)$ such $n \geq N_3$ implies

$$\ln \Pr \left[\frac{1}{n} \sum_{i=1}^{n} (\pm X_i) \geq t_n \right] \leq -na_n^2 - 1 = \ln(\varepsilon_n/2), \tag{65}$$

and therefore that

$$\inf \left\{ t \;\middle|\; \Pr \left[\frac{1}{n} \sum_{i=1}^{n} (\pm X_i) \geq t \right] \leq \varepsilon_n/2 \right\} \leq t_n. \tag{66}$$

9

Putting everything together, we get that for any $n \geq N(\{a_i\}, \lambda, \eta) := \max\{N_1, N_2, N_3\}$ we have

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \geq D_n - \sqrt{2V_n}\, a_n - \eta a_n,$$

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \leq D_n + \sqrt{2V_n}\, a_n + \eta a_n.$$

as required. $\square$

3.4 Outward bound

Proposition 9 (Outward bound). *For any constants $\lambda, \eta > 0$, there exists a constant $N(\{a_i\}, \lambda, \eta)$ such that, for $n \geq N$, the hypothesis testing divergence can be bounded for any states $\{\rho_i\}_i$ and $\{\sigma_i\}_i$ with $\lambda_{\min}(\sigma_i) \geq \lambda$ as*

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \leq D_n + \eta a_n,$$

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \geq D_n - \eta\, a_n.$$

where $D_n := \frac{1}{n} \sum_{i=1}^n D(\rho_i \| \sigma_i)$ and $V_n := \frac{1}{n} \sum_{i=1}^n V(\rho_i \| \sigma_i)$. Moreover, if there also exists a constant $\nu > 0$ such that $V_i \geq \nu$ for all i, then there exists an $N'(\{a_i\}, \lambda, \nu, \eta)$ such that, for $n \geq N'$, the hypothesis testing divergence is more tightly bounded as

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \leq D_n - \sqrt{2V_n}\, a_n + \eta a_n,$$

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \geq D_n + \sqrt{2V_n}\, a_n - \eta a_n.$$

Proof. Similar to Proposition 8, we will start by taking the upper and lower bounds on the ε_n- and $(1 - \varepsilon_n)$-hypothesis testing divergences respectively from Lemma 6. This gives that there exists an $N_1(\{a_i\}, \lambda, \eta)$ such that, for $n \geq N_1$, we have

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \leq D_n - \inf\left\{ t \, \middle| \, \Pr\left[\frac{1}{n} \sum_{i=1}^n (-X_i) \geq t \right] \leq 2\varepsilon_n \right\} + \eta a_n/2, \tag{67}$$

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \geq D_n + \inf\left\{ t \, \middle| \, \Pr\left[\frac{1}{n} \sum_{i=1}^n (+X_i) \geq t \right] \leq 2\varepsilon_n \right\} - \eta a_n/2. \tag{68}$$

where $X_i := Z_i - D(\rho_i \| \sigma_i)$.

Firstly, applying Chebyshev's inequality two standard deviations below the mean gives us that

$$\Pr\left[\frac{1}{n} \sum_{i=1}^n (\pm X_i) \geq -2\sqrt{V_n/n} \right] \geq 3/4 \geq 2\varepsilon_n, \tag{69}$$

and so we conclude that

$$\inf\left\{ t \, \middle| \, \Pr\left[\frac{1}{n} \sum_{i=1}^n (\pm X_i) \geq t \right] \leq 2\varepsilon_n \right\} \geq -2\sqrt{V_n/n}. \tag{70}$$

By Lemma 7, V_n must be bounded $V_n \leq \bar{V}(\lambda)$, and thus $\sqrt{V_n/n} = \mathcal{O}(1/\sqrt{n}) = o(a_n)$. As such, there must exist an $N_2(\{a_i\}, \lambda, \eta)$ such that $n \geq N_2$ implies $2\sqrt{V_n/n} \leq \eta a_n/2$. Inserting this tail bound, we get that for any $n \geq N(\{a_i\}, \lambda, \eta) := \max\{N_1, N_2\}$ that

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \leq D_n - \eta a_n, \tag{71}$$

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \, \middle\| \, \bigotimes_{i=1}^n \sigma_i \right) \geq D_n + \eta a_n, \tag{72}$$

as required.

If there also exists an $\nu > 0$ such that $V_i \geq \nu$, then we can use a more refined moderate deviation bound. Specifically, Lemma 7 gives us a bound on the absolute third moment of X_i, which allows us to apply Lemma 1. If we let $t_n := (\sqrt{2V_n} - \eta/2)a_n$ and assume $\eta < \sqrt{8\nu}$ such that $\{t_n\}_n$ is moderate, then this gives us that there exists an $N_3(\{a_i\}, \lambda, \nu, \eta)$ such that, for any $n \geq N_3$, the tail probabilities are bounded

$$\ln \Pr\left[\frac{1}{n}\sum_{i=1}^{n}(\pm X_i) \geq t_n\right] \geq -\left(1 + \eta/\sqrt{2\bar{V}}\right)\frac{nt_n^2}{2V_n} \tag{73}$$

$$\geq -\frac{\left(1 + \eta/\sqrt{2\bar{V}}\right)\left(\sqrt{2V_n} - \eta/2\right)^2}{2V_n}na_n^2 \tag{74}$$

$$\geq -\frac{\left(1 + \eta/\sqrt{2V_n}\right)\left(\sqrt{2V_n} - \eta/2\right)^2}{2V_n}na_n^2 \tag{75}$$

$$\geq -\left(1 - \frac{5\eta^2}{8\bar{V}}\right)na_n^2. \tag{76}$$

Once again, the second term in the parenthesis is a non-zero constant, and thus there must exist an $N_4(\{a_i\}, \lambda, \eta)$ such that

$$\log \Pr\left[\frac{1}{n}\sum_{i=1}^{n}(\pm X_i) \geq t_n\right] \geq -na_n^2 + 1 = \ln 2\varepsilon_n, \tag{77}$$

allowing us to conclude $\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(\pm X_i) \geq t_n\right] \geq 2\varepsilon_n$, and therefore

$$\inf\left\{t \,\middle|\, \Pr\left[\frac{1}{n}\sum_{i=1}^{n}X_i \geq t\right] \leq 2\varepsilon_n\right\} \geq t_n. \tag{78}$$

Inserting this into the above bounds, we find that for any $n \geq N'(\{a_i\}, \lambda, \nu, \eta) := \max\{N_1, N_3, N_4\}$, we have the desired final bound

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^{n}\rho_i \,\middle\|\, \bigotimes_{i=1}^{n}\sigma_i\right) \leq D_n - \sqrt{2V_n}\,a_n + \eta a_n,$$

$$\frac{1}{n}D_{\mathrm{h}}^{1-\varepsilon_n}\left(\bigotimes_{i=1}^{n}\rho_i \,\middle\|\, \bigotimes_{i=1}^{n}\sigma_i\right) \geq D_n + \sqrt{2V_n}\,a_n - \eta a_n.$$

$\square$

4 Channel Coding

We are now going to show how the above moderate deviation bounds can be applied to the capacity of a classical-quantum channel.

Theorem 2 (Moderate deviation of c-q channels)**.** *For any moderate sequence $\{a_n\}_n$ and memoryless c-q channel $\mathcal{W}$ with capacity $C(\mathcal{W})$ and min-dispersion $V_{\min}(\mathcal{W})$, being operated at an error probability no larger than $\varepsilon_n := e^{-na_n^2}$, the optimal rate deviates below the capacity as*

$$R^*(\mathcal{W}; n, \varepsilon_n) = C(\mathcal{W}) - \sqrt{2V_{\min}(\mathcal{W})}\,a_n + o(a_n). \tag{79}$$

Conversely, if the channel has max-dispersion $V_{\max}$ and is operated error probability no larger than $1 - \varepsilon_n$, then the optimal rate deviates above the capacity as

$$R^*(\mathcal{W}; n, 1 - \varepsilon_n) = C(\mathcal{W}) + \sqrt{2V_{\max}(\mathcal{W})}\,a_n + o(a_n). \tag{80}$$

If either the min- or max-dispersion is non-zero, an application of Lemma 4 gives an equivalent formulation in terms of the minimal error probability at a given rate.

Corollary 10. *For any moderate sequence $\{s_n\}_n$, the error probability for a code with min-dispersion $V_{\min} > 0$ deviating below capacity by s_n scales as*

$$\lim_{n \to \infty} \frac{1}{ns_n^2} \ln \varepsilon^*(\mathcal{W}; n, C - s_n) = -\frac{1}{2V_{\min}}. \tag{81}$$

Similarly, for a code with max-dispersion $V_{\max} > 0$ deviating above capacity by s_n, the error probability scales

$$\lim_{n \to \infty} \frac{1}{ns_n^2} \ln\big(1 - \varepsilon^*(\mathcal{W}; n, C + s_n)\big) = -\frac{1}{2V_{\max}}. \tag{82}$$

Remark 3. *Recall that our definition of c-q channels does not put any restriction on the input set. In particular, this set may be comprised of quantum states itself such that the c-q channel is just a representation of a quantum channel. Hence, as pointed out in Ref. [7], our results immediately also apply to classical communication over general image-additive channels [39] as well as classical communication over quantum channels with encoders restricted to prepare separable states. We refer the reader to Corollaries 6 and 7 of Ref. [7] for details.*

We will split the proof of Theorem 2 in two, in Section 4.1 we will prove a lower bound on the maximum rate ('achievability'), followed in Section 4.2 by a corresponding the upper bound ('optimality'). For the rest of this section, we will fix the channel $\mathcal{W}$, and omit any dependencies on $\mathcal{W}$ from here on for notational convenience.

4.1 Achievability

For achievability, we will use a lower bound on the ε-one-shot rate that is essentially due to Nagaoka and Hayashi [40] who analyzed the coding problem using the information spectrum method.

Lemma 11 (Theorem 1 of Ref. [14]). *If we have a c-q channel which maps from a finite message space Y as $y \mapsto \rho^{(y)}$, then the maximum rate with error probability at most ε and $1 - \varepsilon$, $R^*(\varepsilon)$ and $R^*(1 - \varepsilon)$ respectively, are lower bounded*

$$R^*(\varepsilon) \geq \sup_{P_Y} D_{\mathrm{h}}^{\varepsilon/2}\left(\pi_{YZ} \| \pi_Y \otimes \pi_Z\right) - \log \frac{8(2 - \varepsilon)}{\varepsilon} \tag{83}$$

$$R^*(1 - \varepsilon) \geq \sup_{P_Y} D_{\mathrm{h}}^{1-2\varepsilon}\left(\pi_{YZ} \| \pi_Y \otimes \pi_Z\right) - \log \frac{8(1 - \varepsilon)}{\varepsilon} \tag{84}$$

where π_{YZ} is the joint state of the input and output, with inputs chosen according to the distribution P_Y

$$\pi_{YZ} := \sum_{y \in Y} P_Y(y) |y\rangle\langle y|_Y \otimes \rho_Z^{(y)}. \tag{85}$$

Proposition 12 (Channel coding: Achievability). *For any moderate sequence $\{a_n\}_n$ and error probability $\varepsilon_n := e^{-na_n^2}$, the rate is at least*

$$R^* (n, \varepsilon_n) \geq C - \sqrt{2V_{\min}}a_n + o(a_n). \tag{86}$$

Similarly, at error probability $1 - \varepsilon_n$, the rate is at least

$$R^* (n, 1 - \varepsilon_n) \geq C + \sqrt{2V_{\max}}a_n + o(a_n). \tag{87}$$

Proof. Let X be our, possibly infinite, message space. By Lemma 3 of Ref. [7], there exists a finite subset $Y \subseteq X$, and a distribution Q_Y thereon, such that $D(\rho\|\sigma) = C$ and $V(\rho\|\sigma) = V_{\min}$ for states

$$\rho := \sum_{y \in Y} Q_Y(y) |y\rangle\langle y| \otimes \rho^{(y)} \qquad \text{and} \qquad \sigma := \sum_{y \in Y} Q_Y(y) |y\rangle\langle y| \otimes \sum_{y' \in Y} Q_Y(y')\rho^{(y')}. \tag{88}$$

Clearly by restricting the message space we can only ever decrease the rate. By applying Lemma 11 to the restriction of the message space to Y, we can lower bound the maximum rate of the full code. Applying this reasoning to n memoryless applications of our channel we find

$$nR^*(n, \varepsilon_n) \geq \sup_{P_{Y^n}} D_{\mathrm{h}}^{\varepsilon_n/2}\left(\pi_{Y^n Z^n} \| \pi_{Y^n} \otimes \pi_{Z^n}\right) - \log \frac{8(2 - \varepsilon_n)}{\varepsilon_n}. \tag{89}$$

Substituting in both the error probability, which is no larger than $\varepsilon_n = e^{-na_n^2}$, and a product distribution $Q_{Y^n}(\vec{y}) := \prod_{i=1}^{n} Q_Y(y_i)$ then we get

$$R^*(n, \varepsilon_n) \geq \frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n/2}\left(\rho^{\otimes n} \| \sigma^{\otimes n}\right) + \mathcal{O}(a_n^2). \tag{90}$$

Applying Proposition 8, we get an overall bound on the rate of

$$R^*(n, \varepsilon_n) \geq C - \sqrt{2V_{\min}}\, a_n + o(a_n). \tag{91}$$

If instead we were to take a distribution Q_Y such that $V(\rho\|\sigma) = V_{\max}$, then the same arguments would allow us to use Proposition 9 to analogously give

$$R^*(n, 1 - \varepsilon_n) \geq C + \sqrt{2V_{\max}}\, a_n + o(a_n). \tag{92}$$

$\square$

4.2 Optimality

Similar to the second-order analysis of Ref. [7], we are going to do this by relating the capacity and one-shot maximum rates to geometric quantities known as the divergence radius and divergence centre.

Definition 13 (Divergence radius and centre). For some set of states $\mathcal{S}_0 \subseteq \mathcal{S}$, the *divergence radius* $\chi(\mathcal{S}_0)$ and *divergence centre* $\sigma^*(\mathcal{S}_0)$ are defined as

$$\chi(\mathcal{S}_0) := \inf_{\sigma \in \mathcal{S}} \sup_{\rho \in \mathcal{S}_0} D(\rho \| \sigma), \qquad \sigma^*(\mathcal{S}_0) := \arg \min_{\sigma \in \mathcal{S}} \sup_{\rho \in \mathcal{S}_0} D(\rho \| \sigma). \tag{93}$$

Similarly the *ε-hypothesis testing divergence radius* $\chi_{\mathrm{h}}^{\varepsilon}(\mathcal{S}_0)$ is defined as

$$\chi_{\mathrm{h}}^{\varepsilon}(\mathcal{S}_0) := \inf_{\sigma \in \mathcal{S}} \sup_{\rho \in \mathcal{S}_0} D_{\mathrm{h}}^{\varepsilon}(\rho \| \sigma). \tag{94}$$

Whilst we have seen that the divergence radius captures the capacity of a channel, the ε-hypothesis testing divergence radius approximates the one-shot capacity.

Lemma 14 (Proposition 5 of [7]). *For $\mathcal{I} := \overline{\mathrm{im}\,\mathcal{W}}$, the maximum rate with error probability at most ε, $R^*(\varepsilon)$, is upper bounded as*

$$R^*(\varepsilon) \leq \chi_{\mathrm{h}}^{2\varepsilon}(I) + \log \frac{2}{1 - 2\varepsilon}. \tag{95}$$

Similarly for an error probability $1 - \varepsilon$, the maximum rate is upper bounded as

$$R^*(1 - \varepsilon) \leq \chi_{\mathrm{h}}^{1 - \varepsilon/2}(I) + \log \frac{2(2 - \varepsilon)}{\varepsilon^2}. \tag{96}$$

If we take $\mathcal{I}_n := \overline{\mathrm{im}\,\mathcal{W}^{\otimes n}}$ to be the closure of the image of n uses of this channel, then we can extend this bound on the one-shot rate to the n-shot rate as

$$nR^*(n, \varepsilon_n) \leq \chi_{\mathrm{h}}^{2\varepsilon_n}(\mathcal{I}_n) + \log \frac{2}{1 - \varepsilon_n}, \tag{97}$$

$$nR^*(n, 1 - \varepsilon_n) \leq \chi_{\mathrm{h}}^{2\varepsilon_n}(\mathcal{I}_n) + \log \frac{2(2 - \varepsilon_n)}{\varepsilon_n^2}. \tag{98}$$

As we are considering memoryless c-q channels, $\mathcal{I}_n$ simply consists of elementwise tensor products of $\mathcal{I}$

$$\mathcal{I}_n = \left\{ \bigotimes_{i=1}^{n} \rho_i \,\middle|\, \rho_i \in \mathcal{I} \right\}. \tag{99}$$

Once again we are going to take a_n to be an arbitrary moderate sequence, and $\varepsilon_n := e^{-na_n^2}$. Expanding this out, this gives bounds on the rate of

$$R^*(n, \varepsilon_n) \leq \inf_{\sigma^n} \sup_{\{\rho_i\} \subseteq \mathcal{I}} \frac{1}{n} D_{\mathrm{h}}^{2\varepsilon_n} \left(\bigotimes_{i=1}^{n} \rho_i \,\middle\|\, \sigma^n \right) + \frac{1}{n} \log \frac{2}{1 - \varepsilon_n}, \tag{100}$$

$$R^*(n, 1 - \varepsilon_n) \leq \inf_{\sigma^n} \sup_{\{\rho_i\} \subseteq \mathcal{I}} \frac{1}{n} D_{\mathrm{h}}^{1 - \varepsilon_n/2} \left(\bigotimes_{i=1}^{n} \rho_i \,\middle\|\, \sigma^n \right) + \frac{1}{n} \log \frac{2(2 - \varepsilon)}{\varepsilon_n^2}. \tag{101}$$

A standard approach now is to pick a state σ^n, such that we can bound the above quantities for arbitrary sequences $\{\rho_i\}$ using the moderate deviation analysis of the hypothesis testing divergence presented in Section 3. To do this we need to consider two cases. The *high cases* are those in which the empirical relative entropy corresponding to $\{\rho_i\}_{i=1}^{n}$ is close to capacity, and the *low cases* are those in which the empirical relative entropy corresponding to $\{\rho_i\}_{i=1}^{n}$ is far from capacity. Specifically, for some constant γ that will be chosen later, the n which correspond to high and low cases are denoted by $H(\gamma)$ and $L(\gamma)$, respectively. They are defined as

$$H(\gamma) := \left\{ n \,\middle|\, \frac{1}{n} \sum_{i=1}^{n} D(\rho_i \| \bar{\rho}_n) \geq C - \gamma \right\} \qquad \text{and} \qquad L(\gamma) := \left\{ n \,\middle|\, \frac{1}{n} \sum_{i=1}^{n} D(\rho_i \| \bar{\rho}_n) < C - \gamma \right\} \tag{102}$$

such that $H(\gamma)$ and $L(\gamma)$ bipartition $\mathbb{N}$ for all γ.

Before employing a moderate deviation bound, we are going to construct a separable state σ^n that will allow us two different moderate deviation analyses for low and high sequences, such that we can obtain the required bounds in both cases. A convenient choice of σ^n would be $\sigma^n = \bar{\rho}_n^{\otimes n}$ where $\bar{\rho}_n := \frac{1}{n} \sum_{i=1}^{n} \rho_i$, but the order of the infimum and supremum require σ^n to be chosen to be independent of the sequence $\{\rho_i\}$. Instead we are going to construct σ^n from a mixture of states that lie in a covering of $\mathcal{S}$, and the divergence centre $\sigma^*(\mathcal{I})$.

The following lemma is based on a construction in Lemma II.4 of Ref. [41].

Lemma 15 (Lemma 18 of Ref. [7]). *For every $\delta \in (0,1)$, there exists a set $\mathcal{C}^\delta \subset \mathcal{S}$ of size*

$$\left|\mathcal{C}^\delta\right| \leq \left(\frac{20(2d+1)}{\delta}\right)^{2d^2} \left(\frac{8d(2d+1)}{\delta}+2\right)^{d-1} \leq \left(\frac{90d}{\delta^2}\right)^{2d^2} \tag{103}$$

such that, for every $\rho \in \mathcal{S}$ there exists a state $\tau \in \mathcal{C}^\delta$ such that

$$D(\rho\|\tau) \leq \delta \qquad and \qquad \lambda_{\min}(\tau) \geq \frac{\delta}{8d(2d+1)+\delta} \geq \frac{\delta}{25d^2}. \tag{104}$$

Given this covering upon states, we now want to take σ^n to be the separable state given by a mixture over such a covering, and the divergence centre

$$\sigma^n(\gamma) := \frac{1}{2}\sigma^*(I)^{\otimes n} + \frac{1}{2\left|\mathcal{C}^{\gamma/4}\right|} \sum_{\tau \in \mathcal{C}^{\gamma/4}} \tau^{\otimes n}. \tag{105}$$

Using the inequality

$$D_{\mathrm{h}}^\varepsilon\left(\rho \,\big\|\, \mu\sigma + (1-\mu)\sigma'\right) \leq D_{\mathrm{h}}^\varepsilon(\rho\|\sigma) - \log\mu \tag{106}$$

we will be able to bound divergences with respect to σ^n by those divergences with respect to either elements of $\mathcal{C}^{\gamma/4}$, or σ^*.

We will start by considering the low case. We will see that this case only accounts for hypothesis testing relative entropies which are below the capacity by a constant amount.

Lemma 16 (Low case). *For any $\gamma > 0$, there exists a constant $N(\{a_i\}, \gamma)$ such that*

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\bigg\|\, \sigma_n(\gamma)\right) \leq C - \gamma/4, \tag{107}$$

$$\frac{1}{n}D_{\mathrm{h}}^{1-\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\bigg\|\, \sigma_n(\gamma)\right) \leq C - \gamma/4, \tag{108}$$

for any $\{\rho_i\}_i \subset I$, $n \in L(\gamma)$ and $n \geq N$.

Proof. We are going to start by considering the ε_n-hypothesis testing divergence. Take τ_n to be the closest element in $\mathcal{C}^{\gamma/4}$ to $\bar{\rho}_n$, such that $D(\bar{\rho}_n\|\tau_n) \leq \gamma/4$. Splitting out the τ_n term from $\sigma_n(\gamma)$, we have

$$D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\bigg\|\, \sigma_n(\gamma)\right) \leq D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\bigg\|\, \tau_n^{\otimes n}\right) + \log 2\left|\mathcal{C}^\gamma\right| \tag{109}$$

$$\leq D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\bigg\|\, \tau_n^{\otimes n}\right) + 2d^2 \log\left(\frac{120d}{\gamma^2}\right). \tag{110}$$

As the final term depending on $\left|\mathcal{C}^{\gamma/4}\right|$ is independent of n, there must exist a constant $N_1(\gamma)$ such that $2d^2 \log(120d/\gamma^2) \leq n\gamma/4$ for any $n \geq N_1$, and thus that

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\bigg\|\, \sigma_n(\gamma)\right) \leq \frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\bigg\|\, \tau_n^{\otimes n}\right) + \gamma/4. \tag{111}$$

Applying Proposition 9 to the ε_n-hypothesis testing relative entropy with respect to τ_n we get that there exists an $N_2(\{a_i\}, \gamma)$ such that

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\bigg\|\, \tau_n^{\otimes}\right) \leq \frac{1}{n}\sum_{i=1}^n D(\rho_i\|\tau_n) + \gamma/4, \tag{112}$$

for any $n \geq N_2$. As for the divergence terms given with respect to τ_n, we can rearrange them in terms of divergences relative to the sequence mean $\bar{\rho}_n$ using the information geometric Pythagorean theorem, yielding

$$\sum_{i=1}^n D(\rho_i\|\tau_n) = \sum_{i=1}^n \mathrm{Tr}\,\rho_i(\log\rho_i - \log\bar{\rho}_n) + \sum_{i=1}^n \mathrm{Tr}\,\rho_i(\log\bar{\rho}_n - \log\tau_n) \tag{113}$$

$$= \sum_{i=1}^n D(\rho_i\|\bar{\rho}_n) + nD(\bar{\rho}_n\|\tau_n) \tag{114}$$

$$\leq \sum_{i=1}^n D(\rho_i\|\bar{\rho}_n) + n\gamma/4. \tag{115}$$

14

If we let $N(\{a_i\}, \gamma) := \max\{N_1, N_2\}$, then pulling the above results together we see that for any $n \geq N$

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma) \right) \leq \frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \tau_n^{\otimes n} \right) + \gamma/4 \tag{116}$$

$$\leq \frac{1}{n} \sum_{i=1}^n D(\rho_i \| \tau_n) + 2\gamma/4 \tag{117}$$

$$\leq \frac{1}{n} \sum_{i=1}^n D(\rho_i \| \bar{\rho}_n) + 3\gamma/4. \tag{118}$$

Finally, since $n \in L(\gamma)$ the average relative entropy is bounded away from capacity, and we arrive at the bound:

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma) \right) \leq C - \gamma/4. \tag{119}$$

As we only relied on Proposition 9 to bound the the regularized hypothesis testing divergence to within a constant of the average relative entropy, we could perform a similar analysis for the $(1 - \varepsilon_n)$-hypothesis testing divergence using Proposition 8 instead, which gives

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma) \right) \leq C - \gamma/4. \tag{120}$$

$\square$

Now that we have dealt with cases far from capacity, we turn our attention to the high cases.

Lemma 17 (High case). *For any $\eta > 0$, there exist constants $\Gamma(\eta)$ and $N(\{a_i\}, \eta)$, such that*

$$\frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\Gamma) \right) \leq C - \sqrt{2V_{\min}}\, a_n + \eta a_n \tag{121}$$

for any $\{\rho_i\}_i \subset I$, $n \in H(\Gamma)$ and $n \geq N$. Similarly, the $(1 - \varepsilon_n)$-hypothesis testing relative entropy is bounded

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\Gamma) \right) \leq C + \sqrt{2V_{\max}}\, a_n + \eta a_n. \tag{122}$$

Proof. Splitting out the σ^* factor within $\sigma_n(\gamma)$ gives

$$D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma) \right) \leq D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma^{*\otimes n} \right) + \log 2, \tag{123}$$

$$D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma) \right) \leq D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma^{*\otimes n} \right) + \log 2. \tag{124}$$

As $\frac{1}{n} \log 2 = o(a_n)$, there exists an $N_1(\{a_i\})$ such that $n \geq N_1$ implies

$$D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma) \right) \leq D_{\mathrm{h}}^{\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma^{*\otimes n} \right) + \eta a_n/3, \tag{125}$$

$$D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma) \right) \leq D_{\mathrm{h}}^{1-\varepsilon_n} \left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma^{*\otimes n} \right) + \eta a_n/3. \tag{126}$$

We now wish to employ a moderate deviation result. We will start by addressing the ε_n-hypothesis testing divergence. For the weaker bound of Proposition 9 we will have no required bounds on $\frac{1}{n} \sum_{i=1}^n V(\rho_i \| \sigma^*)$, but for the stronger bound we will need a uniform lower bound.

If $V_{\min} \leq \eta^2/18$, then the weakened bound of Proposition 9 is sufficient, giving an $N_2(\{a_n\}, \eta)$ such that $n \geq N_2$ implies

$$D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma)\right) \leq \frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma^{*\otimes n}\right) + \eta a_n/3 \tag{127}$$

$$\leq \frac{1}{n}\sum_{i=1}^n D(\rho_i\|\sigma^*) + 2\eta a_n/3 \tag{128}$$

$$\leq \frac{1}{n}\sum_{i=1}^n D(\rho_i\|\sigma^*) - \sqrt{2V_{\min}}\, a_n + \eta a_n \tag{129}$$

$$\leq C - \sqrt{2V_{\min}}\, a_n + \eta a_n. \tag{130}$$

Next we need to consider the case where $V_{\min} > \eta^2/18$. To do this, we will need to establish a lower bound on $\frac{1}{n}\sum_{i=1}^n V(\rho_i\|\sigma^*)$, which places it near $V_{\min}$. The min-dispersion is defined for distributions which exactly achieve capacity; we will now consider an analogous quantity for distributions which are *near* capacity. Specifically

$$V_{\min}(\gamma) := \inf_{P \in \mathcal{P}(\mathcal{I})} \left\{ \int \mathrm{d}P(\rho)\, V\left(\rho\|\sigma^*\right) \,\middle|\, \int \mathrm{d}P(\rho)\, D\left(\rho \,\middle\|\, \int \mathrm{d}P(\rho')\, \rho'\right) \geq C - \gamma \right\}. \tag{131}$$

By definition of the channel dispersion we have that $V_{\min}(0) = V_{\min}$. By Lemma 22 of Ref. [7] we can strengthen this to $\lim_{\gamma \to 0^+} V_{\min}(\gamma) = V_{\min}$, and so for any $\eta > 0$ there must exist a constant $\Gamma(\eta)$ such that

$$\sqrt{2V_{\min}(\Gamma)} \geq \sqrt{2V_{\min}} - \eta/3. \tag{132}$$

As $V_{\min} \geq \eta^2/18$, this implies that $V_{\min}(\Gamma) > 0$.

Next, let P_n be the empirical distribution corresponding to the set $\{\rho_i\}_{i=1}^n$, i.e. $P_n(\rho) := \frac{1}{n}\sum_{i=1}^n \delta(\rho - \rho_i)$. For all $n \in H(\Gamma)$, these distributions are near capacity

$$\int \mathrm{d}P_n(\rho)\, D\left(\rho \,\middle\|\, \int \mathrm{d}P_n(\rho')\, \rho'\right) = \frac{1}{n}\sum_{i=1}^n D(\rho_i\|\bar{\rho}_n) \geq C - \Gamma, \tag{133}$$

and so we can lower bound the average variance with respect to the divergence centre

$$\frac{1}{n}\sum_{i=1}^n V(\rho_i\|\sigma^*) = \int \mathrm{d}P(\rho)\, V(\rho\|\sigma^*) \geq V_{\min}(\Gamma) > 0. \tag{134}$$

Using this lower bound, we can apply the stronger bound from Proposition 9 to give a constant $N_3(\{a_i\}, \eta)$, such that, for every $n \in H(\Gamma)$ and $n \geq N_3$, the hypothesis testing divergence is upper bounded

$$D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma)\right) \leq \frac{1}{n} D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma^{*\otimes n}\right) + \eta a_n/3 \tag{135}$$

$$\leq \frac{1}{n}\sum_{i=1}^n D(\rho_i\|\sigma^*) - \sqrt{\frac{2}{n}\sum_{i=1}^n V(\rho_i\|\sigma^*) a_n} + 2\eta a_n/3 \tag{136}$$

$$\leq \frac{1}{n}\sum_{i=1}^n D(\rho_i\|\sigma^*) - \sqrt{2V_{\min}(\Gamma)}\, a_n + 2\eta a_n/3 \tag{137}$$

$$\leq C - \sqrt{2V_{\min}}\, a_n + \eta a_n. \tag{138}$$

Performing a similar argument for $V_{\max}$, we construct a function

$$V_{\max}(\gamma) := \sup_{P \in \mathcal{P}(\mathcal{I})} \left\{ \int \mathrm{d}P(\rho)\, V\left(\rho\|\sigma^*\right) \,\middle|\, \int \mathrm{d}P(\rho)\, D\left(\rho \,\middle\|\, \int \mathrm{d}P(\rho')\, \rho'\right) \geq C - \gamma \right\}, \tag{139}$$

and define a Γ such that

$$\sqrt{2V_{\max}(\Gamma)} \leq \sqrt{2V_{\max}} + \eta/3. \tag{140}$$

Following through the rest of the argument, and employing Proposition 8, we also get a bound on the $(1 - \varepsilon_n)$-hypothesis testing divergence

$$\frac{1}{n} D_{\mathrm{h}}^{1-\varepsilon_n}\left(\bigotimes_{i=1}^n \rho_i \,\middle\|\, \sigma_n(\gamma)\right) \leq C + \sqrt{2V_{\max}} a_n + \eta a_n. \tag{141}$$

$$\square$$

Proposition 18 (Channel coding: Optimality). *For any moderate sequence $\{a_n\}_n$ and error probability $\varepsilon_n :=$ $e^{-na_n^2}$, the rate is upper bounded as*

$$R^*\left(n, \varepsilon_n\right) \leq C - \sqrt{2V_{\min}}\, a_n + o(a_n). \tag{142}$$

For error probability $(1 - \varepsilon_n)$ the rate is similarly upper bound as

$$R^*\left(n, 1 - \varepsilon_n\right) \leq C + \sqrt{2V_{\max}}\, a_n + o(a_n). \tag{143}$$

Proof. Applying Lemmas 16 and 17, we get that there exist constants $\Gamma(\eta)$ and $N_1(\{a_i\}, \eta)$ such that

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^{n} \rho_i \,\middle\|\, \sigma_n(\Gamma)\right) \leq \begin{cases} C - \Gamma/4 & n \in L(\Gamma) \\ C - \sqrt{2V_{\min}}\, a_n + \eta a_n & n \in H(\Gamma) \end{cases} \tag{144}$$

for any $n \geq N_1$. As Γ is a constant, there must exist some $N_2(\{a_i\}, \eta)$ such that $\Gamma/4 \geq \sqrt{2V_{\min}}a_n$. As such, for any $n \geq \max\{N_1, N_2\}$, high or low, we have

$$\frac{1}{n}D_{\mathrm{h}}^{\varepsilon_n}\left(\bigotimes_{i=1}^{n} \rho_i \,\middle\|\, \sigma_n(\Gamma)\right) \leq C - \sqrt{2V_{\min}}\, a_n + \eta a_n. \tag{145}$$

Pulling this bound back to Equation (100), we have

$$R^*(n, \varepsilon_n) \leq \inf_{\sigma^n \in \mathcal{S}^n} \sup_{\{\rho_i\} \subseteq I} \frac{1}{n}D_{\mathrm{h}}^{2\varepsilon_n}\left(\bigotimes_{i=1}^{n} \rho_i \,\middle\|\, \sigma^n(\Gamma)\right) + \frac{1}{n}\log\frac{2}{1 - \varepsilon_n} \tag{146}$$

$$\leq C - \sqrt{2V_{\min}}a_n + \eta a_n + \frac{1}{n}\log\frac{2}{1 - \varepsilon_n}. \tag{147}$$

Finally, noting that $1/n = o(a_n)$, there must exist a constant $N_3(\{a_i\}, \eta)$ such that $n \geq N_3$ implies

$$\frac{1}{n}\log\frac{2}{1 - \varepsilon_n} \leq \eta a_n. \tag{148}$$

We can therefore conclude that, for $n \geq \max\{N_1, N_2, N_3\}$, we get the overall upper bound

$$R^*(n, \varepsilon_n) \leq C - \sqrt{2V_{\min}}\, a_n + 2\eta a_n. \tag{149}$$

As this is true for arbitrary $\eta > 0$, we can take $\eta \searrow 0$ and conclude

$$R^*(n, \varepsilon_n) \leq C - \sqrt{2V_{\min}}a_n + o(a_n) \tag{150}$$

as required. A similar analysis for the $(1 - \varepsilon_n)$-error regime shows

$$R^*\left(n, 1 - \varepsilon_n\right) \leq C + \sqrt{2V_{\max}}a_n + o(a_n). \tag{151}$$

□

5 Conclusion

The main result of this paper is to give a second order approximation of the non-asymptotic fundamental limit for classical information transmission over a quantum channel in the moderate deviations regime, as in Eqns. 3 and 4:

$$\frac{1}{n}\log M^*(\mathcal{W}; n, \varepsilon_n) = C(\mathcal{W}) - \sqrt{2V_{\min}(\mathcal{W})}\, x_n + o(x_n), \tag{152}$$

$$\frac{1}{n}\log M^*(\mathcal{W}; n, 1 - \varepsilon_n) = C(\mathcal{W}) + \sqrt{2V_{\max}(\mathcal{W})}\, x_n + o(x_n). \tag{153}$$

Along the lines of third and fourth order approximations for classical channel coding in the fixed error regime (see, e.g., Refs. [42–44]), a natural question to ask is whether we can expand this further and resolve the term $o(x_n)$. A preliminary investigation suggests the conjecture that $o(x_n) = O(x_n^2) + O(\log n)$ and that at least some of the implicit constants can be determined precisely. We leave this for future work.

Due to the central importance of binary asymmetric quantum hypothesis testing we expect our techniques to have applications also to other quantum channel coding tasks. In particular, source coding [45,46], entanglement-assisted classical coding [47] as well as quantum [48] and private coding [49] over quantum channels have recently been analyzed in the small deviations regime by relating the problem to quantum hypothesis testing. An extension of these results to moderate deviations using our techniques thus appears feasible.

6 References

[1] A. S. Holevo, "The Capacity of the Quantum Channel with General Signal States," *IEEE Transactions on Information Theory* **44**, 269–273, (1998).

[2] A. S. Holevo, "Bounds for the Quantity of Information Transmitted by a Quantum Communication Channel," *Problems of Information Transmission* **9**, 3, 177–183, (1973).

[3] B. Schumacher and M. Westmoreland, "Sending Classical Information via Noisy Quantum Channels," *Physical Review A* **56**, 131–138, (1997).

[4] A. Holevo, "Reliability function of general classical-quantum channel," *IEEE Transactions on Information Theory* **46**, 2256–2261, arXiv:quant-ph/9907087, (2000).

[5] M. Hayashi, "Error Exponent in Asymmetric Quantum Hypothesis Testing and Its Application to Classical-Quantum Channel coding," *Physical Review A* **76**, 62301, arXiv:quant-ph/0611013, (2006).

[6] M. Dalai, "Lower Bounds on the Probability of Error for Classical and Classical-Quantum Channels," *IEEE Transactions on Information Theory* **59**, 8027–8056, arXiv:1201.5411, (2013).

[7] M. Tomamichel and V. Y. F. Tan, "Second-Order Asymptotics for the Classical Capacity of Image-Additive Quantum Channels," *Communications in Mathematical Physics* **338**, 103–137, arXiv:1308.6503, (2015).

[8] A. Dembo and O. Zeitouni, Large Deviations Techniques and Applications. Stochastic Modelling and Applied Probability, Springer, 2 ed., (1998).

[9] H. Umegaki, "Conditional expectation in an operator algebra. IV. Entropy and information," *Kodai Mathematical Seminar Reports* **14**, 2, 59–85, (1962).

[10] M. Tomamichel and M. Hayashi, "A Hierarchy of Information Quantities for Finite Block Length Analysis of Quantum Tasks," *IEEE Transactions on Information Theory* **59**, 7693–7710, arXiv:1208.1478, (2013).

[11] K. Li, "Second-Order Asymptotics for Quantum Hypothesis Testing," *Annals of Statistics* **42**, 171–189, arXiv:1208.1400, (2014).

[12] Y. Altug and A. B. Wagner, "Moderate Deviations in Channel Coding," *IEEE Transactions on Information Theory* **60**, 4417–4426, arXiv:1208.1924, (2014).

[13] Y. Polyanskiy and S. Verdu, "Channel dispersion and moderate deviations limits for memoryless channels," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1334–1339, IEEE, (2010).

[14] L. Wang and R. Renner, "One-Shot Classical-Quantum Capacity and Hypothesis Testing," *Physical Review Letters* **108**, 200501, arXiv:1007.5456, (2012).

[15] H.-C. Cheng and M.-H. Hsieh, "Moderate Deviation Analysis for Classical-Quantum Channels and Quantum Hypothesis Testing," arXiv:1701.03195, (2016).

[16] W. Hoeffding, "Asymptotically Optimal Tests for Multinomial Distributions," *Annals of Mathematical Statistics* **36**, 2, 369–401, (1965).

[17] R. G. Gallager, Information Theory and Reliable Communication. New York: Wiley, , New York(1968).

[18] I. Csiszár and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge University Press, 2nd ed., (2011).

[19] H. Nagaoka, "The Converse Part of The Theorem for Quantum Hoeffding Bound," arXiv:quant-ph/0611289, (2006).

[20] I. Sason, "Moderate deviations analysis of binary hypothesis testing," in *2012 IEEE International Symposium on Information Theory Proceedings*, (Cambridge, MA), 821–825, IEEE, , Cambridge, MA(2012), arXiv:1111.1995.

[21] V. Strassen, "Asymptotische Abschätzungen in Shannons Informationstheorie," in *Trans. Third Prague Conference on Information Theory*, (Prague), 689–723, (1962).

[22] M. Hayashi, "Information Spectrum Approach to Second-Order Coding Rate in Channel Coding," *IEEE Transactions on Information Theory* **55**, 4947–4966, arXiv:0801.2242, (2009).

[23] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Transactions on Information Theory* **56**, 2307–2359, (2010).

[24] I. Csiszár and G. Longo, "On the Error Exponent for Source Coding and for Testing Simple Statistical Hypotheses," *Studia Scientiarum Mathematicarum Hungarica* **6**, 181–191, (1971).

[25] T. S. Han and K. Kobayashi, "The Strong Converse Theorem for Hypothesis Testing," *IEEE Transactions on Information Theory* **35**, 1, 178–180, (1989).

[26] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Transactions on Information Theory* **19**, 357–359, (1973).

[27] G. Dueck and J. Korner, "Reliability function of a discrete memoryless channel at rates above capacity (Corresp.)," *IEEE Transactions on Information Theory* **25**, 82–85, (1979).

[28] M. Mosonyi and T. Ogawa, "Quantum Hypothesis Testing and the Operational Interpretation of the Quantum Rényi Relative Entropies," *Communications in Mathematical Physics* **334**, 1617–1648, arXiv:1309.3228, (2015).

[29] M. Mosonyi and T. Ogawa, "Two approaches to obtain the strong converse exponent of quantum hypothesis testing for general sequences of quantum states," *IEEE Transactions on Information Theory*, arXiv:1407.3567, (2014).

[30] M. Mosonyi and T. Ogawa, "Strong Converse Exponent for Classical-Quantum Channel Coding," arXiv:1409.3562, (2014).

[31] K. R. Parthasarathy, Probability Measures on Metric Spaces. Academic Press, New York and London, (1967).

[32] A. Winter, "Coding Theorem and Strong Converse for Quantum Channels," *IEEE Transactions on Information Theory* **45**, 2481–2485, arXiv:1409.2536, (2014).

[33] T. Ogawa and H. Nagaoka, "Strong converse to the quantum channel coding theorem," *IEEE Transactions on Information Theory* **45**, 2486–2489, arXiv:quant-ph/9808063, (1999).

[34] B. Schumacher and M. D. Westmoreland, "Optimal signal ensembles," *Physical Review A* **63**, 1–5, arXiv:quant-ph/9912122, (1999).

[35] F. Dupuis, L. Kraemer, P. Faist, J. M. Renes, and R. Renner, "Generalized Entropies," in *Proc. XVIIth International Congress on Mathematical Physics*, (Aalborg, Denmark), 134–153, (2012), arXiv:1211.3141.

[36] D. Petz, "Quasi-entropies for finite quantum systems," *Reports on Mathematical Physics* **23**, 57–65, (1986).

[37] M. S. Lin and M. Tomamichel, "Investigating Properties of a Family of Quantum Renyi Divergences," *Quantum Information Processing* **14**, 1501–1512, arXiv:1408.6897, (2014).

[38] M. Nussbaum and A. Szkoła, "The Chernoff lower bound for symmetric quantum hypothesis testing," *The Annals of Statistics* **37**, 1040–1057, arXiv:quant-ph/0607216, (2009).

[39] M. Fukuda, I. Nechita, and M. M. Wolf, "Quantum Channels With Polytopic Images and Image Additivity," *IEEE Transactions on Information Theory* **61**, 1851–1859, arXiv:1408.2340, (2015).

[40] H. Nagaoka and M. Hayashi, "An Information-Spectrum Approach to Classical and Quantum Hypothesis Testing for Simple Hypotheses," *IEEE Transactions on Information Theory* **53**, 534–549, (2007).

[41] P. Hayden, D. Leung, P. W. Shor, and A. Winter, "Randomizing Quantum States: Constructions and Applications," *Communications in Mathematical Physics* **250**, 371–391, arXiv:quant-ph/03071, (2004).

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

[43] M. Tomamichel and V. Y. F. Tan, "A Tight Upper Bound for the Third-Order Asymptotics for Most Discrete Memoryless Channels," *IEEE Transactions on Information Theory* **59**, 7041–7051, arXiv:1212.3689, (2013).

[44] P. Moulin, "The Log-Volume of Optimal Codes for Memoryless Channels, Asymptotically Within A Few Nats," *IEEE Transactions on Information Theory*, (Published online), arXiv:1311.0181, (2017).

[45] N. Datta and F. Leditzky, "Second-Order Asymptotics for Source Coding, Dense Coding, and Pure-State Entanglement Conversions," *IEEE Transactions on Information Theory* **61**, 582–608, arXiv:1403.2543, (2015).

[46] F. Leditzky and N. Datta, "Second-Order Asymptotics of Visible Mixed Quantum Source Coding via Universal Codes," *IEEE Transactions on Information Theory* **62**, 4347–4355, arXiv:1407.6616, (2016).

[47] N. Datta, M. Tomamichel, and M. M. Wilde, "On the second-order asymptotics for entanglement-assisted communication," *Quantum Information Processing* **15**, 2569–2591, arXiv:1405.1797, (2016).

[48] M. Tomamichel, M. Berta, and J. M. Renes, "Quantum coding with finite resources," *Nature Communications* **7**, 11419, arXiv:1504.04617, (2016).

[49] M. M. Wilde, M. Tomamichel, and M. Berta, "Converse bounds for private communication over quantum channels," *IEEE Transactions on Information Theory*, (Published online), arXiv:1602.08898, (2017).

[50] L. V. Rozovsky, "Estimate from Below for Large-Deviation Probabilities of a Sum of Independent Random Variables with Finite Variances," *Journal of Mathematical Sciences* **109**, 6, 2192–2209, (2002).

[51] S.-h. Lee, V. Y. F. Tan, and A. Khisti, "Streaming Data Transmission in the Moderate Deviations and Central Limit Regimes," *IEEE Transactions on Information Theory* **62**, 6816–6830, arXiv:1512.06298, (2016).

A Moderate deviation tail bounds

A.1 Lower bound

Here we apply the lower bound of Ref. [50], which gives a Berry-Esseen-type inequality with multiplicative error.

Lemma 19 (Theorem B2, Ref. [50])**.** *There exists universal constants κ_1, κ_2 such that, for independent zero-mean variables $X_1, \ldots, X_n$ with*

$$V_n := \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}\,[X_i] \qquad and \qquad T_n := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[|X_i|^3\right], \tag{154}$$

and a t_n bounded

$$\sqrt{\frac{V_n}{n}} \le t_n \le \frac{V_n^2}{T_n}, \tag{155}$$

the probability that the average variable $\frac{1}{n}\sum_{i=1}^{n} X_i$ deviates above the mean by t_n is lower bounded

$$\ln \Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \ge t_n\right] \ge \ln \Phi\left(-\sqrt{\frac{nt_n^2}{V_n}}\right) - \frac{\kappa_1 T_n n t_n^3}{V_n^3} + \ln\left(1 - \frac{\kappa_2 T_n t_n}{V_n^2}\right). \tag{156}$$

Given this Lemma, we can now prove the desired lower bound on the moderately deviating tail.

Proof of Lemma 1. First we note that the bound on the average third absolute moment also imposes a bound on the average variance

$$V_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X^2\right] \tag{157}$$

$$< \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[|X|^3 + 1\right] \tag{158}$$

$$= T_n + 1 \tag{159}$$

$$\le \tau + 1. \tag{160}$$

As $\{t_i\}_i$ is moderate, and the moments are bounded $\nu \leq V_n \leq \tau+1$ and $T_n \leq \tau$, there must exist an $N_1(\{t_i\}, \nu, \tau)$ such that

$$\sqrt{\frac{\tau+1}{n}} \leq t_n \leq \frac{\nu^2}{\tau} \tag{161}$$

for $n \geq N_1$. Applying Lemma 19, we have that for $n \geq N_1$

$$\ln \Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,n} \geq t\right) \geq \ln \Phi\left(-\sqrt{\frac{nt_n^2}{V_n}}\right) - \frac{\kappa_1 \tau}{\nu^3} nt_n^3 + \ln\left(1 - \frac{\kappa_2 \tau}{\nu^{3/2}} t_n\right). \tag{162}$$

As $nt_n^2 \to \infty$ and $V_n \leq \tau+1$, there must exist a constant $N_2(\{t_i\}, \tau)$ such that $n \geq N_2$ implies $nt_n^2/V_n \geq 1$. Using the standard bound $\ln \Phi(-x) \geq -x^2/2 - \ln\sqrt{8\pi}x$ for $x \geq 1$, we find

$$\ln \Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,n} \geq t\right) \geq -\frac{nt_n^2}{2V_n} - \ln\sqrt{8\pi \frac{nt_n^2}{V_n}} - \frac{\kappa_1 \tau}{\nu^3} nt_n^3 + \ln\left(1 - \frac{\kappa_2 \tau}{\nu^{3/2}} t_n\right). \tag{163}$$

As t_n is moderate, we have that the first term $-nt_n^2/2V_n$ dominates as $n \to \infty$ in the above. As such, for any $\eta > 0$, there must exist an $N_3(\{t_n\}, \nu, \tau, \eta)$ such that, for all $n \geq N_3$, the other terms are smaller than this dominant term by a multiplicative factor of $\eta > 0$, such that

$$\ln\sqrt{8\pi \frac{nt_n^2}{V_n}} + \frac{\kappa_1 \tau}{V_n^3} nt_n^3 - \ln\left(1 - \frac{\kappa_2 \tau}{V_n^{3/2}} t_n\right) \leq \eta \frac{nt_n^2}{2V_n}. \tag{164}$$

We conclude that for $n \geq N(\{t_i\}, \nu, \tau, \eta) := \max\{N_1, N_2, N_3\}$

$$\ln \Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_{i,n} \geq t_n\right] \geq -(1+\eta)\frac{nt_n^2}{2V_n}. \tag{165}$$

$\square$

A.2   Upper bound

For the upper bound we are going to use a proof technique similar to that used to prove Cramér's and Gartner-Ellis theorems in the large deviation regime (see, e.g., Ref. [8]), and for Lemma 4 of Ref. [51] in the iid moderate deviation regime. However, our approach differs from that in Ref. [51] because we do not want to assume that the average variance, V_n, is bounded away from zero.

Proof of Lemma 2. Let $h_n(s)$ be the average cumulant generating function $h_n(s) = \frac{1}{n}\sum_{i=1}^{n} \ln \mathbb{E}\left[e^{sX_{i,n}}\right]$, such that

$$h_n(0) = 0, \qquad h_n'(0) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_{i,n}] = 0, \qquad h_n''(0) = \frac{1}{n}\sum_{i=1}^{n} \text{Var}[X_{i,n}] = V_n. \tag{166}$$

For our tail bound we are going to employ a Chernoff bound. Specifically for any $\alpha > 0$, an application of the Markov inequality gives

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_{i,n} \geq t_n\right] = \Pr\left[e^{\alpha t_n \sum_{i=1}^{n} X_{i,n}} \geq e^{\alpha n t_n^2}\right] \leq \frac{\mathbb{E}\left[e^{\alpha t_n \sum_{i=1}^{n} X_{i,n}}\right]}{e^{\alpha n t_n^2}}. \tag{167}$$

Using the independence of $\{X_i\}$, the above bound can be expressed in terms of the average cumulant generating function as

$$\ln \Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_{i,n} \geq t_n\right] \leq -n\left(\alpha t_n^2 - h_n(\alpha t_n)\right). \tag{168}$$

In general our choice of α will depend on n. If we assume for the moment that α is bounded then, as $t_n \to 0$, there exists a constant $N_1(\{t_i\}, \alpha)$ such that $n \geq N_1$ implies $\alpha t_n \leq 1/2$. Applying Taylor's theorem for such n, specifically a second-order expansion with the error in Lagrange form, gives that there exists an $s \in [0, \alpha t_n] \subseteq [0, 1/2]$ such that

$$h_n(s) = h_n(0) + \alpha t_n h_n'(0) + \alpha^2 t_n^2 h_n''(0)/2 + \alpha^3 t_n^3 h_n'''(s)/6 \tag{169}$$
$$\leq \alpha^2 t_n^2 V_n/2 + \alpha^3 t_n^3 \gamma/6. \tag{170}$$

Plugging this Taylor expansion in to our Chernoff bound above gives

$$\frac{1}{nt_n^2} \ln \Pr\left[\frac{1}{n}\sum_{i=1}^n X_{i,n} \geq t_n\right] \leq \left(\alpha^2 V_n/2 - \alpha\right) + \alpha^2 \gamma t_n/6. \tag{171}$$

We now need to choose our value of α. An obvious choice would be $\alpha = 1/V_n$, which gives the tightest possible asymptotic bound. As we have not imposed a lower bound on V_n, this value is not necessarily bounded, and therefore could render the previous Taylor expansion invalid. Instead we will slightly modify this choice such that the Taylor expansion is still valid, whilst only changing the final bound by the introducing of an η. Specifically we will take

$$\alpha^{-1} := \sqrt{V_n + \eta/4}\left(\sqrt{V_n + \eta/4} + \sqrt{\eta/4}\right). \tag{172}$$

As required, this choice of α is bounded independent V_n as $\alpha \leq 2/\eta$, meaning that the previous Taylor expansion was indeed valid, and that $N_1 = N_1(\{t_i\}, \eta)$. Plugging this choice in to Equation (171) gives

$$\ln \Pr\left[\frac{1}{n}\sum_{i=1}^n X_{i,n} \geq t_n\right] \leq -\frac{nt_n^2}{2V_n + \eta/2} + \frac{2\gamma}{3\eta^2}nt_n^3. \tag{173}$$

Similar to Lemma 1, the bound on the third derviative of cumulant function bounds the variances as $\mathrm{Var}[X_i] \leq \gamma + 1$. Given this, there must exist a constant $N_2(\{t_i\}, \gamma, \eta)$ such that $n \geq N_2$ implies

$$-\frac{1}{2V_n + \eta/2} + \frac{2\gamma}{3\eta^3}t_n \leq -\frac{1}{2V_n + \eta}. \tag{174}$$

We conclude therefore that for any $n \geq N(\{t_i\}, \gamma, \eta) := \max\{N_1, N_2\}$ we have the desired tail bound

$$\ln \Pr\left[\frac{1}{n}\sum_{i=1}^n X_i \leq t_n\right] \leq -\frac{nt_n^2}{2V_n + \eta}. \tag{175}$$

$\square$

A.3 Dimensionless bound

The non-dimensional bound follows as a corollary of the two previous bound, where we explicitly use the possible dependence of our random variables $X_{i,n}$ on n.

Proof of Corollary 3. Starting with random variables $\{X_{i,n}\}_{i \leq n}$, define rescaled variables as $\tilde{X}_{i,n} := X_{i,n}/\sqrt{V_n}$ for all $i \leq n$. This scaling has the property that it normalises the average variance

$$\tilde{V}_n := \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[\tilde{X}_{i,n}] = 1. \tag{176}$$

As well as this, we can see the dimensionless assumption on $X_{i,n}$

$$\frac{1}{nV_n^{3/2}}\sum_{i=1}^n \sup_{s \in [0,1/2]} \left|\frac{\mathrm{d}^3}{\mathrm{d}s^3}\ln\mathbb{E}\left[e^{sX_{i,n}}\right]\right| \leq \gamma, \tag{177}$$

is equivalent to the bound on $\tilde{X}_{i,n}$

$$\frac{1}{n}\sum_{i=1}^n \sup_{s \in [0,1/2]} \left|\frac{\mathrm{d}^3}{\mathrm{d}s^3}\ln\mathbb{E}\left[e^{s\tilde{X}_{i,n}}\right]\right| \leq \gamma. \tag{178}$$

Noticing that

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n X_{i,n} \leq t_n\sqrt{V_n}\right] = \Pr\left[\frac{1}{n}\sum_{i=1}^n X_{i,n}/\sqrt{V_n} \leq t_n\right] = \Pr\left[\frac{1}{n}\sum_{i=1}^n \tilde{X}_{i,n} \leq t_n\right], \tag{179}$$

we can simply apply the existing tail bounds of Lemmas 1 and 2 to $\tilde{X}_{i,n}$, giving that, for any $\eta > 0$, there must exist a constant $N(\{t_i\}, \gamma, \eta)$ such that

$$-(1+\eta)\frac{nt_n^2}{2} \leq \ln \Pr\left[\frac{1}{n}\sum_{i=1}^n X_{i,n} \leq t_n\sqrt{V_n}\right] \leq -(1-\eta)\frac{nt_n^2}{2}. \tag{180}$$

$\square$

B   Proof of Lemma 4

Lemma 20. *Let A and B be two real functions both defined on the same domain , with*

$$\hat{A}(b) := \inf_t \{A(t)|B(t) \le b\} \qquad and \qquad \hat{B}(b) := \inf_t \{B(t)|A(t) \le a\}, \tag{181}$$

then for any $a \ge \inf_t A(t)$ and $\delta > 0$

$$\hat{A}(\hat{B}(a) + \delta) \le a \tag{182}$$

$$\hat{A}(\hat{B}(a) - \delta) \ge a. \tag{183}$$

Proof. By the definition of the infimum in $\hat{B}(\cdot)$, there must exist a $t^\star$ such that $A(t^\star) \le a$ and $B(t^\star) \le \hat{B}(a) + \delta$. Hence we can upper bound

$$\hat{A}(\hat{B}(a) + \delta) = \inf_s \left\{ A(s) \middle| B(s) \le \hat{B}(a) + \delta \right\} \le A(t^\star) \le a. \tag{184}$$

Next, suppose $\hat{A}(\hat{B}(a) - \delta) \le a - \varepsilon$ for some $\varepsilon > 0$. By definition of the infimum in $\hat{A}(\cdot)$, there must therefore exist an $s^\star$ such that $B(s^\star) \le \hat{B}(a) - \delta$ and $A(s^\star) \le a$. This in turn allows us to upper bound

$$\hat{B}(a) = \inf_t \{B(t)|A(t) \le a\} \le B(s^\star). \tag{185}$$

We can therefore conclude that $\hat{B}(x) \le \hat{B}(a) - \delta$, proving $\hat{A}(\hat{B}(a) - \delta) > a - \varepsilon$ by contradiction. As this is true for arbitrarily small ε, we therefore conclude

$$\hat{A}(\hat{B}(x) - \delta) \ge a. \tag{186}$$

$\square$

Proof of Lemma 4. By swapping both A_n and B_n we can see that the forward and backwards directions of this proof are equivalent, as such we will only consider the forward direction. First, we assume that

$$\lim_{n \to \infty} \frac{\hat{B}_n(a_n)}{a_n} = 1, \quad \forall a_n \text{ moderate.} \tag{187}$$

Next we split the proof of the limit into upper bounding the limit superior, and lower bounding the limit inferior.

Take any moderate sequence b_n, and let $a_n := b_n/2$ and $b'_n := \hat{B}_n(a_n) + b_n/n$. By Lemma 20 we have that $\hat{A}_n(b'_n) = \hat{A}_n(\hat{B}_n(a_n) + b_n/n) \le a_n$. By assumption we then have that

$$\lim_{n \to \infty} \frac{b'_n}{b_n} = \lim_{n \to \infty} \frac{\hat{B}_n(a_n) + 2a_n/n}{2a_n} = \lim_{n \to \infty} \frac{\hat{B}_n(a_n)}{2a_n} = \frac{1}{2}. \tag{188}$$

As a result we have, for sufficiently large n, that $b'_n \le b_n$. Using this we can bound the limit superior:

$$\limsup_{n \to \infty} \frac{\hat{A}_n(b_n)}{b_n} \le \lim_{n \to \infty} \frac{a_n}{b'_n} \tag{189}$$

$$= \lim_{n \to \infty} \frac{a_n}{\hat{B}_n(a_n)} = 1. \tag{190}$$

Moreover, if we take $a_n := 2b_n$ and $b'_n := \hat{B}_n(a_n) - b_n/n$ then, by an analogous argument, Lemma 20 gives us $\hat{A}_n(b'_n) \ge a_n$, and the assumption gives us $b'_n \ge b_n$ for sufficiently large n. As such we can also bound the limit inferior

$$\liminf_{n \to \infty} \frac{\hat{A}_n(b_n)}{b_n} \ge \lim_{n \to \infty} \frac{a_n}{b'_n} = \lim_{n \to \infty} \frac{a_n}{\hat{B}_n(a_n)} = 1. \tag{191}$$

$\square$

C   Proof of Lemma 7

Proof. Consider the moment generating function $m(t) := \mathbb{E}\left[e^{tZ}\right]$, such that $h(t) = \ln m(t)$. Similar to the relationship between cumulants and central moments, the derivatives of $h(t)$ can be expressed in terms of derivatives of $m(t)$:

$$h = \ln m \tag{192}$$

$$h' = \frac{m'}{m} \tag{193}$$

$$h'' = \frac{m''m - m'm'}{m^2} \tag{194}$$

$$h''' = \frac{m'''mm - 3m''m'm + 2m'm'm'}{m^3} \tag{195}$$

$$\vdots \tag{196}$$

As such, if we bound $m(t)$ away from zero, proving that the derivatives of $m(t)$ are uniformly bounded would imply the same about $h(t)$. Noticing that $\sum_a r_a = 1$ implies $\sum_a r_a^2 \geq 1/d$ and $\lambda \leq 1/d$, we can see that m is bounded away from zero for any $t \leq 1$:

$$m(t) = \sum_{a,b} r_a \left|\langle\phi_a|\psi_b\rangle\right|^2 (r_a/s_b)^t \tag{197}$$

$$\geq \frac{1}{\lambda^t} \sum_{a,b} \left|\langle\phi_a|\psi_b\rangle\right|^2 r_a^{t+1} \tag{198}$$

$$\geq \frac{1}{\lambda^t} \sum_a r_a^{1+t} \tag{199}$$

$$\geq \frac{1}{\lambda^t} \sum_a r_a^2 \tag{200}$$

$$\geq \frac{1}{d\lambda} \geq 1 \tag{201}$$

Next we can use the bound $\sup_{x\in[0,1]} \sqrt{x}\ln^k(1/x) = (2k/e)^k$, to bound the derivatives of the moment generating function for $|t| \leq 1/2$

$$\left|m^{(k)}(t)\right| = \left|\sum_{a,b} r_a \left|\langle\phi_a|\psi_b\rangle\right|^2 (r_a/s_b)^t \ln^k (r_a/s_b)\right| \tag{202}$$

$$\leq \sum_{a,b} r_a \left|\langle\phi_a|\psi_b\rangle\right|^2 (r_a/s_b)^t \left|\ln^k (r_a/s_b)\right| \tag{203}$$

$$\leq \sum_{a,b} r_a \left|\langle\phi_a|\psi_b\rangle\right|^2 (r_a/s_b)^t \left[\ln^k(1/s_b) + \ln^k(1/r_a)\right] \tag{204}$$

$$\leq \sum_{a,b} \left|\langle\phi_a|\psi_b\rangle\right|^2 \sqrt{r_a/s_b} \left[\ln^k(1/s_b) + \ln^k(1/r_a)\right] \tag{205}$$

$$\leq \max_{\substack{\lambda\leq s\leq 1 \\ 0\leq r\leq 1}} \sqrt{r/s} \left[\ln^k(1/s) + \ln^k(1/r)\right] \tag{206}$$

$$\leq \frac{1}{\sqrt{\lambda}} \max_{\substack{\lambda\leq s\leq 1 \\ 0\leq r\leq 1}} \left[\ln^k(1/s) + \sqrt{r}\ln^k(1/r)\right] \tag{207}$$

$$\leq \frac{1}{\sqrt{\lambda}} \left[\ln^k\left(\frac{1}{\lambda}\right) + (2k/e)^k\right] =: C_k. \tag{208}$$

$\square$