A Separation between Divergence and Holevo Information for Ensembles

Rahul Jain *
U. Waterloo

Ashwin Nayak †
U. Waterloo & Perimeter

Yi Su ‡
U. Waterloo

February 10, 2008

Abstract

The notion of *divergence information* of an ensemble of probability distributions was introduced by Jain, Radhakrishnan, and Sen [5, 7] in the context of the "substate theorem". Since then, divergence has been recognized as a more natural measure of information in several situations in quantum and classical communication.

We construct ensembles of probability distributions for which divergence information may be significantly smaller than the more standard Holevo information. As a result, we establish that lower bounds previously shown for Holevo information are weaker than similar ones shown for divergence information.

*School of Computer Science, and Institute for Quantum Computing, University of Waterloo, 200 University Ave. W., Waterloo, ON N2L 3G1, Canada. Email: rjain@cs.uwaterloo.ca. Research supported in part by ARO/NSA USA.

†Department of Combinatorics and Optimization, and Institute for Quantum Computing, University of Waterloo, 200 University Ave. W., Waterloo, ON N2L 3G1, Canada. E-mail: anayak@math.uwaterloo.ca. Research supported in part by NSERC Canada, CIFAR, MITACS, QuantumWorks, and an ERA from the Province of Ontario. A.N. is also Associate Member, Perimeter Institute for Theoretical Physics, Waterloo, Canada. Research at Perimeter Institute for Theoretical Physics is supported in part by the Government of Canada through NSERC and by the Province of Ontario through MRI.

‡Department of Pure Mathematics, University of Waterloo, 200 University Ave. W., Waterloo, ON N2L 3G1, Canada. E-mail: y6su@student.math.uwaterloo.ca. Research supported in part by an NSERC Canada Undergraduate Research Award.

1 Introduction

In this article, we study the relationship between two different measures of information contained in an ensemble of probability distributions. The first measure, *Holevo information*, is a standard notion from information theory, and is equivalent to the notion of *mutual information* between two random variables. Consider jointly distributed random variables XY, with X taking values in a sample space $\mathcal{X}$. Consider the ensemble of distributions $\mathcal{E} = \{(\lambda_i, Y_i) : i \in \mathcal{X}\}$, where $\lambda_i = \Pr(X = i)$, and $Y_i = Y|(X = i)$, obtained by conditioning on values assumed by X. The Holevo information of the ensemble is given by $\chi(\mathcal{E}) = \mathrm{I}(X : Y) = \mathbb{E}_{i \sim X} \mathrm{S}(Y_i \| Y)$, where $\mathrm{S}(\cdot\|\cdot)$ measures the relative entropy of a random variable (equivalently, distribution) with respect to another. This notion may be extended to ensembles of quantum states (see, e.g., the text [11]), and the term 'Holevo information' is derived from the literature in quantum information theory.

The second measure, *divergence information*, was introduced by Jain, Radhakrishnan, and Sen [5, 7]. It arises in the study of relative entropy, and its connection with a "substate property". The *observational divergence* (or simply *divergence*) of two classical distributions P, Q on the same finite sample space is $\max_E P(E) \log_2(P(E)/Q(E))$, where E ranges over all events. We may view this as a (scaled) measure of the factor by which P may exceed Q for an event of interest. The notion of *divergence information* is derived from this as $\mathrm{D}(\mathcal{E}) = \mathbb{E}_{i \sim X} \mathrm{D}(Y_i \| Y)$, in analogy with Holevo information. A quantum generalisation of this measure may also be defined [7].

Relative entropy and Holevo (or mutual) information have been studied extensively in communication theory and beyond (see, e.g, [2]) as they arise in a variety of applications. Since the discovery of the substate theorem [5], divergence is being recognized as a more natural measure of information in a growing number of applications [7, Section 1]. The applications include privacy trade-offs in communicatioin protocols for computing relations [6] and *bit-string commitment* [3], and the communication complexity of *remote state preparation* [4]. In particular, divergence captures, up to a constant factor, the substate property for probability distributions. It thus becomes relevant in every application where the substate theorem is used.

We construct ensembles of probability distributions (equivalently, jointly distributed random variables) for which the Holevo and divergence information are quantitatively different.

Theorem 1.1 *For every positive integer N, and real number $k \geq 1$ such that $N > 2^{36k^2}$, there is an ensemble $\mathcal{E}$ of distributions over a sample space of size N such that $\mathrm{D}(\mathcal{E}) = k$ and $\chi(\mathcal{E}) = \Theta(k \log \log N)$.*

A more precise statement of this theorem (Theorem 3.1) and related results may be found in Section 3.

The ensembles we construct satisfy the property that the ensemble average (i.e., the distribution of the random variable Y in the description above) is uniform. We show that the above separation is essentially the best possible whenever the ensemble average is uniform (Theorem 3.5). The result also applies to ensembles of quantum states, where the ensemble average is the completely mixed state (Theorem 3.6). We leave open the possibility of larger separations for classical or quantum ensembles with non-uniform averages.

The difference between the two measures demonstrated by Theorem 1.1 shows that in certain applications, divergence is quantitatively a more relevant measure of information. In Appendix A, we describe two applications where functionally similar lower bounds have been established in terms of both measures. This article shows that the lower bounds in terms of divergence information are, in fact, stronger.

In prior work on the subject, Jain *et al.* [7, Appendix A] compare relative entropy and divergence for classical as well as quantum states. For pairs of distributions P, Q over a sample space of size N, they show that $\mathrm{D}(P\|Q) \leq \mathrm{S}(P\|Q) + 1$, and $\mathrm{S}(P\|Q) \leq \mathrm{D}(P\|Q) \cdot (N-1)$. This extends to the corresponding measures of information in an ensemble: $\mathrm{D}(\mathcal{E}) \leq \chi(\mathcal{E}) + 1$ and $\chi(\mathcal{E}) \leq \mathrm{D}(\mathcal{E}) \cdot (N-1)$. They show qualitatively similar relations for ensembles of quantum states. In addition, they construct a pair of distributions P, Q such that $\mathrm{S}(P\|Q) = \Omega(\mathrm{D}(P\|Q) \cdot N)$. However, they do not translate their construction to a similar separation for *ensembles* of probability distributions. Our work fills this gap for ensembles (of classical or quantum states) with a uniform average.

2 Preliminaries

Here, we summarise our notation and the information-theoretic concepts we encounter in this work. We refer the reader to the text by Cover and Thomas [2] for a deeper treatment of (classical) information theory. While the bulk of this article pertains to classical information theory, as mentioned in Section 1, it is motivated by studies in (and has implications for) quantum information. We refer the reader to the text [11] for an introduction to quantum information.

For a positive integer N, let $[N]$ represent the set $\{1, \ldots, N\}$. We view probability distributions over $[N]$ as vectors in $\mathbb{R}^N$. The probability assigned by distribution P to a sample point $i \in [N]$ is denoted by p_i (i.e., with the same letter in small case). We denote by $P^{\downarrow}$ the distribution obtained from P by composing it with a permutation π on $[N]$ so that $p_i^{\downarrow} = p_{\pi(i)}$ and $p_1^{\downarrow} \geq p_2^{\downarrow} \geq \cdots \geq p_N^{\downarrow}$. For an event $E \subseteq [N]$, let $P(E) = \sum_{i \in E} p_i$ denote the probability of that event. We denote the uniform distribution over $[N]$ by U_N. The expected value of a function $f : [N] \rightarrow \mathbb{R}$ with respect to the distribution P over $[N]$ is abbreviated as $\mathbb{E}_P f$.

We appeal to the *majorisation* relation for some of our arguments. The relation tells us which of two given distributions is "more random".

Definition 2.1 (Majorisation) *Let P, Q be distributions over $[N]$. We say that P* majorises *Q, denoted as $P \succeq Q$, if*

$$\sum_{j=1}^{i} p_j^{\downarrow} \quad \geq \quad \sum_{j=1}^{i} q_j^{\downarrow},$$

for all $i \in [N]$.

The following is straightforward.

Fact 2.1 *Any probability distribution P on $[N]$ majorises U_N, the uniform distribution over $[N]$.*

Throughout this article, we use 'log' to denote the logarithm with base 2, and 'ln' to denote the logarithm with base e.

Definition 2.2 (Entropy, relative entropy) *Let P, Q be probability distributions on $[N]$. The* entropy *of P is defined as $\mathrm{H}(P) \stackrel{\text{def}}{=} -\sum_{i=1}^{N} p_i \log p_i$. The* relative entropy *between P, Q, denoted*

$\mathrm{S}(P\|Q)$, *is defined as*

$$\mathrm{S}(P\|Q) \quad \stackrel{\mathrm{def}}{=} \quad \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} \; .$$

Note that the relative entropy with respect to the uniform distribution is connected to entropy as $\mathrm{S}(P\|\mathrm{U}_N) = \log N - \mathrm{H}(P)$.

We can formalise the connection between majorisation and randomness through the following fact.

Fact 2.2 *If P, Q are distributions over $[N]$ such that P majorises Q, i.e. $P \succeq Q$, then $\mathrm{H}(P) \leq \mathrm{H}(Q)$.*

The notion of *observational divergence* was defined by Jain, Radhakrishnan, and Sen [5] in the context of the "substate theorem".

Definition 2.3 (Observational divergence) *Let P, Q be probability distributions on $[N]$. Then the* observational divergence *between them, denoted $\mathrm{D}(P\|Q)$, is defined as*

$$\mathrm{D}(P\|Q) \quad \stackrel{\mathrm{def}}{=} \quad \max_{f:[N]\to[0,1]} \, (\mathbb{E}_P f) \, \log \frac{\mathbb{E}_P f}{\mathbb{E}_Q f} \; .$$

Note that we allow the quantity to take the value $+\infty$. Throughout the paper we refer to 'observational divergence' as simply 'divergence'.

Divergence $\mathrm{D}(P\|Q)$ is always non-negative, and it is finite precisely when the support of P is contained in the support of Q [5]. Due to convexity, the divergence between two distributions is attained by the characteristic function of an event.

Lemma 2.3 $\mathrm{D}(P\|Q) \quad = \quad \max_{E\subseteq[N]} \; P(E) \log \frac{P(E)}{Q(E)}$.

Proof: Let $\mathcal{F}$ denote the (convex) set of functions from $[N]$ to $[0,1]$. The extreme points of $\mathcal{F}$ are precisely the characteristic functions of events in $[N]$. For an extreme point, say the characteristic function f_E of the event $E \subseteq [N]$, we have $\mathbb{E}_P f_E = P(E)$.

If the divergence is $+\infty$, then there is an event for which the right hand side also takes the value $+\infty$. So assume that the divergence is finite. In this case, the right hand side also is finite, as the support of P is contained in the support of Q. By restricting $f : [N] \to [0,1]$ to characteristic functions of events, we see that $\mathrm{D}(P\|Q)$ is at least the expression on the right hand side above.

For the inequality in the other direction, we note that the function

$$g(x) \quad = \quad (ax+b) \log \left(\frac{ax+b}{cx+d} \right)$$

defined on $[0,1]$ is convex in x, for any $a, b, c, d \in \mathbb{R}$ such that $ax + b \geq 0$ and $cx + d > 0$ when $x \in [0,1]$. Therefore, the function $g(x)$ attains its maximum at either $x = 0$ or at $x = 1$.

The convexity of $g(x)$ implies that for any $\alpha \in [0,1]$, and functions $f, f' \in \mathcal{F}$, we have

$$\begin{aligned}
& (\mathbb{E}_P(\alpha f + (1-\alpha) f')) \; \log \frac{\mathbb{E}_P(\alpha f + (1-\alpha) f')}{\mathbb{E}_Q(\alpha f + (1-\alpha) f')} \\
= \quad & (\alpha(\mathbb{E}_P f - \mathbb{E}_P f') + \mathbb{E}_P f') \; \log \frac{\alpha(\mathbb{E}_P f - \mathbb{E}_P f') + \mathbb{E}_P f'}{\alpha(\mathbb{E}_Q f - \mathbb{E}_Q f') + \mathbb{E}_Q f'} \\
\leq \quad & \max \left\{ (\mathbb{E}_P f) \log \frac{\mathbb{E}_P f}{\mathbb{E}_Q f}, \quad (\mathbb{E}_P f') \log \frac{\mathbb{E}_P f'}{\mathbb{E}_Q f'} \right\} .
\end{aligned}$$

Thus, the divergence is attained at an extreme point of $\mathcal{F}$. This proves the claim. ■

Henceforth, we only use the equivalent definition of divergence given by this lemma.

The divergence of any distribution with respect to the uniform distribution is bounded.

Lemma 2.4 *For any probability distribution P on $[N]$, we have $0 \leq \mathrm{D}(P\|\mathrm{U}_N) \leq \log N$.*

Proof: Consider the event E which achieves the divergence between P and U_N. W.l.o.g., the event E is non-empty. Therefore $P(E) \geq \mathrm{U}_N(E) \geq 1/N$, and $0 \leq \mathrm{D}(P\|U_N) \leq P(E)\log P(E)N \leq \log N$. ■

We observe that we need only maximise over N events to calculate divergence with respect to the uniform distribution.

Lemma 2.5 *For any probability distribution P on $[N]$ such that $P^{\downarrow} = P$, i.e., $p_1 \geq p_2 \geq \cdots \geq p_N$, we have*

$$\mathrm{D}(P\|\mathrm{U}_N) \quad = \quad \max_{i\in[N]} P([i]) \log \frac{N \cdot P([i])}{i} \ .$$

Proof: By definition of observational divergence, the RHS above is bounded by $\mathrm{D}(P\|\mathrm{U}_N)$. For the inequality in the other direction, we note that the probability $P(E)$ of any event E with size $n_E = |E|$ is bounded by $P([n_E])$, the probability of the first n_E elements in $[N]$. We thus have

$$\begin{aligned} \mathrm{D}(P\|Q) &= \max_{E\subseteq[N]} P(E) \log \frac{N \cdot P(E)}{n_E} \\ &\leq \max_{E\subseteq[N]} P(E) \log \frac{N \cdot P([n_E])}{n_E} \\ &\leq \max_{E\subseteq[N]} P([n_E]) \log \frac{N \cdot P([n_E])}{n_E}, \end{aligned}$$

since P majorises U_N (Fact 2.1) and $P([n_E]) \geq \frac{n_E}{N}$. This is equivalent to the RHS in the statement of the lemma. ■

Definition 2.4 (Ensemble) *An* ensemble *is a sequence of pairs $\{(\lambda_j, Q_j) \ : \ j \in [M]\}$, for some integer M, where $\Lambda = (\lambda_j) \in \mathbb{R}^M$ is a probability distribution on $[M]$ and Q_j are probability distributions over the same sample space.*

Definition 2.5 (Holevo information) *The* Holevo information *of an ensemble $\mathcal{E} = \{(\lambda_j, Q_j) \ : \ j \in [M]\}$, denoted as $\chi(\mathcal{E})$, is defined as*

$$\chi(\mathcal{E}) \quad \stackrel{\mathrm{def}}{=} \quad \sum_{j=1}^{M} \lambda_j \, \mathrm{S}(Q_j\|Q),$$

where $Q = \sum_{j=1}^{M} \lambda_j Q_j$ is the ensemble average.

Definition 2.6 (Divergence information) *The* divergence information *of an ensemble $\mathcal{E} = \{(\lambda_j, Q_j) \ : \ j \in [M]\}$, denoted as $\mathrm{D}(\mathcal{E})$ is defined as*

$$\mathrm{D}(\mathcal{E}) \quad \stackrel{\mathrm{def}}{=} \quad \sum_{j=1}^{M} \lambda_j \, \mathrm{D}(Q_j\|Q),$$

where $Q = \sum_{j=1}^{M} \lambda_j Q_j$ is the ensemble average.

3 Divergence versus relative entropy

In this section, we describe the construction of an ensemble for which there is a large separation between divergence and Holevo information. The ensemble has the property that the ensemble average is uniform. As a by-product of our construction, we also obtain a bound on the maximum possible separation for ensembles with a uniform average.

We begin with the construction of the ensemble. Let $f_{\mathrm{L}}(k,N) = k(\ln\log(kN) - \ln(6k) + 1) - \log(1 + k\ln 2) - 1 - \frac{1}{\ln 2}$ on point in the positive orthant in $\mathbb{R}^2$ with $Nk > 1$.

Theorem 3.1 *For every integer $N > 1$, and every positive real number $\frac{16}{N} \leq k < \log N$, there is an ensemble $\mathcal{E} = \left\{(\frac{1}{N}, Q_i) \,:\, i \in [N]\right\}$ with $\frac{1}{N}\sum_i Q_i = \mathrm{U}_N$, the uniform distribution over $[N]$, with $\mathrm{D}(\mathcal{E}) \leq k$, and*

$$\chi(\mathcal{E}) \quad \geq \quad f_{\mathrm{L}}(k,N).$$

To construct the ensemble described in the theorem above, we first construct a probability distribution P on $[N]$ with observational divergence $\mathrm{D}(P\|\mathrm{U}_N) \leq k$ such that its relative entropy $\mathrm{S}(P\|\mathrm{U}_N)$ is large as compared with k. Let $f_{\mathrm{U}} = k(\ln\log(Nk) - \ln k + 1)$ be defined on points in the positive orthant of $\mathbb{R}^2$ with $kN > 1$.

Theorem 3.2 *For every integer $N > 1$, and every positive real number $\frac{16}{N} \leq k < \log N$, there is a probability distribution P with $\mathrm{D}(P\|\mathrm{U}_N) = k$, and*

$$f_{\mathrm{L}}(k,N) \quad \leq \quad \mathrm{S}(P\|\mathrm{U}_N) \quad \leq \quad f_{\mathrm{U}}(k,N).$$

The construction of the ensemble is now immediate.

Proof of Theorem 3.1: Let $Q_j = P \circ \pi_j$, where π_j is the cyclic permutation of $[N]$ by $j-1$ places. We endow the set of the N cyclic permutations $\{Q_j \,:\, j \in [N]\}$ of P with the uniform distribution. By construction, the ensemble average is U_N. Since both observational divergence and relative entropy with respect to the uniform distribution are invariant under permutations of the sample space, $\mathrm{D}(\mathcal{E}) = \mathrm{D}(P\|\mathrm{U}_N) \leq k$, and $\chi(\mathcal{E}) = \mathrm{S}(P\|\mathrm{U}_N) \geq f_{\mathrm{L}}(k,N)$. ■

We turn to the construction of the distribution P. Our construction is such that $P^{\downarrow} = P$, i.e., $p_1 \geq p_2 \geq \cdots \geq p_N$. Lemma 2.5 tells us that we need only ensure that

$$P([i]) \log \frac{N \cdot P([i])}{i} \quad \leq \quad k, \qquad \forall\, i \in [N], \tag{1}$$

to ensure $\mathrm{D}(P\|Q) \leq k$. Since $\mathrm{S}(P\|\mathrm{U}_N) = \log N - \mathrm{H}(P)$, we wish to minimise the entropy of P subject to the constraints in Eq. (1). This is equivalent to successively maximising $p_1, p_2, \ldots$, and motivates the following definitions.

Define the function $g(y,x) = y\log(Ny/x) - k$ on the positive orthant of $\mathbb{R}^2$. Consider the function $h : \mathbb{R}^+ \to \mathbb{R}^+$ implicitly defined by the equation $g(h(x),x) = 0$.

Lemma 3.3 *The function $h : \mathbb{R}^+ \to \mathbb{R}^+$ is well-defined, strictly increasing, and concave.*

Proof: Fix an $x \in \mathbb{R}^+$, and consider the function $g_x(y) = g(y,x)$. This function is continuous on $\mathbb{R}^+$, tends to $-k < 0$ as $y \to 0^+$, and tends to ∞ as $y \to \infty$. By Intermediate Value Theorem, for some $y > 0$, we have $g_x(y) = 0$. Moreover, $g_x(y) < -k$ for $0 < y \leq x/N$, and is strictly increasing for $y > x/N\mathrm{e}$ (its derivative is $g_x'(y) = \log\frac{\mathrm{e}Ny}{x}$). Therefore there is a unique y such that $g_x(y) = 0$ and $h(x)$ is well-defined.

The function h satisfies the equation $h\log\frac{Nh}{x} = k$, and therefore the identity

$$x \quad = \quad Nh\exp\left(-\tfrac{k\ln 2}{h}\right).$$

Differentiating with respect to h, we see that

$$\begin{aligned}\frac{dx}{dh} &= N\left(1 + \frac{k\ln 2}{h}\right)\exp\left(-\tfrac{k\ln 2}{h}\right), \text{ and}\\ \frac{d^2x}{dh^2} &= \frac{N(k\ln 2)^2}{h^3}\exp\left(-\tfrac{k\ln 2}{h}\right).\end{aligned}$$

So $\frac{dh}{dx} > 0$ for all $x > 0$, and h is a strictly increasing function. Note also that $\frac{d^2x}{dh^2} > 0$ for all $h > 0$, so x is a convex function of h. Since h is an increasing function, convexity of $x(h)$ implies concavity of $h(x)$. ■

Let $v_0 = 0$. For $i \in [N]$, let $v_i = h(i)$, i.e., $v_i\log\frac{Nv_i}{i} = k$. Let $s_i \stackrel{\text{def}}{=} \min\{1, v_i\}$, for $i \in [N]$. Let $p_1 = s_1$, and $p_i = s_i - s_{i-1}$ for all $2 \le i \le N$. Lemma 3.3 guarantees that these numbers are well-defined. We claim that

Lemma 3.4 *The vector $P = (p_i) \in \mathbb{R}^N$ defined above is a probability distribution, and $P^{\downarrow} = P$, i.e., $p_1 \ge p_2 \ge \cdots \ge p_N$.*

Proof: By definition, we have $v_i > 0$ for all $i \in [N]$. Therefore $s_1 = \min\{1, v_1\} > 0$. Since $h(x)$ is an increasing function in x, the sequence (v_i) is also increasing, so (s_i) is non-decreasing. Therefore $p_i = s_i - s_{i-1} \ge 0$ for $i > 1$.

Now $v_N\log v_N = k > 0$. Since $x\log x \le 0$ for $x \in (0, 1]$, we have $v_N > 1$. So $s_N = \min\{1, v_N\} = 1$. Therefore $\sum_{i=1}^N p_i = s_N = 1$. So P is a probability distribution on $[N]$.

Note that $(v_2/2)\log(Nv_2/2) = k/2 < k$, so $v_1 > v_2/2$. So $s_1 \ge s_2/2$, i.e., $p_1 \ge p_2$. For $i \ge 2$, we have $p_i - p_{i+1} = (s_i - s_{i-1}) - (s_{i+1} - s_i) = 2s_i - s_{i-1} - s_{i+1}$. Since $h(x)$ is concave, so is the function $\min\{1, h(x)\}$. Therefore, $s_i \ge (s_{i-1} + s_{i+1})/2$, and the sequence (p_i) is non-decreasing. ■

The vector $S = (s_i) \in \mathbb{R}^N$ thus represents the (cumulative) distribution function corresponding to P.

Proof of Theorem 3.2: We claim that the probability distribution P constructed above satisfies the properties stated in the theorem.

Since $P^{\downarrow} = P$, by Lemma 2.5, we need only verify that $s_i\log(Ns_i/i) \le k$ for $i \in [N]$. If $s_i = v_i$, then the condition is satisfied with equality. (Note that since $k < \log N$, we have $s_1 = v_1 < 1$.) Else, $s_i = 1 < v_i$, so $s_i\log(Ns_i/i) < v_i\log(Nv_i/i) = k$.

We now bound the relative entropy $\mathrm{S}(P\|\mathrm{U}_N)$ from below. Let n be the smallest positive integer such that $v_{n-1} \le 1$ and $v_n > 1$. Note that $n > 1$. We also have $n \le N$, since $v_N > 1$ (as $v_N\log v_N = k > 0$). Therefore, we have $s_i = v_i$ (equivalently, $Ns_i = i2^{k/s_i}$) for $i \in [n-1]$, and $s_n = 1 < v_n$. Thus, for $1 < i < n$,

$$\begin{aligned}
Np_i &= i2^{\frac{k}{s_i}} - (i-1)2^{\frac{k}{s_{i-1}}} \\
&= 2^{\frac{k}{s_i}} + (i-1)(2^{\frac{k}{s_i}} - 2^{\frac{k}{s_{i-1}}}) \\
&= 2^{\frac{k}{s_i}} + (i-1)2^{\frac{k}{s_{i-1}}}(2^{\frac{k}{s_i} - \frac{k}{s_{i-1}}} - 1) \\
&= 2^{\frac{k}{s_i}} + Ns_{i-1}(2^{\frac{k}{s_i} - \frac{k}{s_{i-1}}} - 1) \\
&\geq 2^{\frac{k}{s_i}} + Ns_{i-1}\left(\frac{k}{s_i} - \frac{k}{s_{i-1}}\right)\ln 2 \\
&= 2^{\frac{k}{s_i}} - \frac{Np_i k}{s_i}\ln 2.
\end{aligned}$$

The penultimate line follows from the inequality $2^x \geq 1 + x\ln 2$ for all $x \in \mathbb{R}$. Thus we have

$$Np_i \quad \geq \quad \frac{2^{\frac{k}{s_i}}}{1 + \frac{k}{s_i}\ln 2}. \tag{2}$$

Since $Np_1 = Ns_1 = 2^{\frac{k}{s_1}}$, this also holds for $i = 1$.

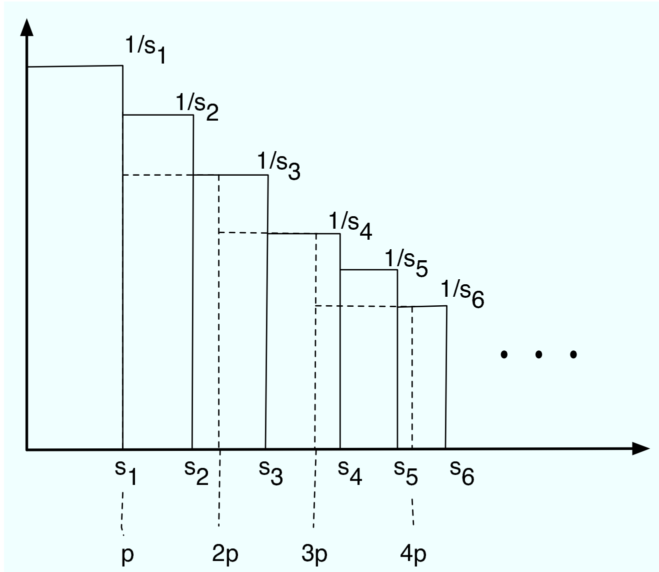
We bound the relative entropy using Eq. (2).

$$\begin{aligned}
\mathrm{S}(P\|\mathrm{U}_N) &= \sum_{i=1}^{N} p_i \log Np_i \quad = \quad \sum_{i=1}^{n} p_i \log Np_i \\
&\geq \sum_{i=1}^{n-1} p_i \log \frac{2^{\frac{k}{s_i}}}{1 + \frac{k}{s_i}\ln 2} + p_n \log Np_n \\
&\geq \sum_{i=1}^{n-1} \frac{p_i k}{s_i} - \sum_{i=1}^{n-1} p_i \log\left(1 + \frac{k\ln 2}{s_i}\right) + p_n \log Np_n. \qquad (3)
\end{aligned}$$

We bound each of the three terms in the RHS of Eq. (3) separately.

We start with $\sum_{i=1}^{n-1} \frac{p_i k}{s_i}$. Let $p = p_1$, and let $m = \left\lfloor \frac{1}{p} \right\rfloor$. For every $j \in [m]$, there is an $i \in [n]$, say $i = i_j$, such that $jp \leq s_{i_j} \leq (j+1)p$. (Otherwise, for some $i > 1$, the probability $p_i = s_i - s_{i-1}$ is strictly larger than p, an impossibility.)

We interpret the sum $\sum_{i=2}^{n-1} \frac{p_i}{s_i} = \sum_{i=2}^{n-1} \frac{s_i - s_{i-1}}{s_i}$ as a Riemann sum approximating the area under the curve $1/x$ between s_1 and s_{n-1} with the area under the solid lines in Figure 3. This area is bounded from below by the area under the dashed lines, which corresponds to the area of rectangles



of uniform width p and height $1/s_{j+1}$ for the jth interval. Thus,

$$\begin{aligned}\sum_{i=1}^{n-1}\frac{p_i k}{s_i} &\geq k + k\sum_{j=1}^{m} p \cdot \frac{1}{s_{i_{j+1}}} \\ &\geq k + k\sum_{j=1}^{m} p \cdot \frac{1}{(j+2)p} \\ &= k + k\sum_{j=1}^{m}\frac{1}{j+2} \\ &\geq k + k\int_{3}^{m+3}\frac{1}{x}dx \\ &= k + k\ln\frac{m+3}{3}. \end{aligned} \tag{4}$$

We lower bound $m = \left\lfloor \frac{1}{p} \right\rfloor$ next. Recall that $g_1(y) = y\log(Ny) - k$ is an increasing function for $y > \frac{1}{\mathrm{e}N}$, and $p = p_1 \geq 1/N$. Consider the value of $g_1(y)$ at the point $q = \frac{2k}{\log kN}$:

$$g_1(q) \quad = \quad \frac{2k}{\log kN}\log\frac{2Nk}{\log kN} - k \quad > \quad 2k\left(1 - \frac{\log\log kN}{\log kN}\right) - k \quad \geq \quad 0,$$

since $kN \geq 16$. As $g_1(q) > g_1(p) > 0$, we have $q > p$. Therefore, $m \geq \frac{1}{p} - 1 \geq \frac{\log kN}{2k} - 1$. Together with Eq. (4), we get

$$\sum_{i=1}^{n-1}\frac{p_i k}{s_i} \quad \geq \quad k(\ln\log kN - \ln 6k + 1). \tag{5}$$

Next, we derive a lower bound for the second term in Eq. (3).

$$
\begin{aligned}
-\sum_{i=1}^{n-1} p_i \log\left(1+\frac{k\ln 2}{s_i}\right) &= -\sum_{i=1}^{n-1} p_i \log(s_i + k\ln 2) + \sum_{i=1}^{n-1} p_i \log s_i \\
&\geq -\log(1+k\ln 2) + \sum_{i=1}^{n-1} p_i \log s_i. \qquad (6)
\end{aligned}
$$

Viewing the second term above as a Riemann sum, we get

$$
\begin{aligned}
\sum_{i=1}^{n-1} p_i \log s_i &\geq \int_0^{s_{n-1}} \log x \; dx \\
&\geq \int_0^1 \log x \; dx \\
&= -\frac{1}{\ln 2}. \qquad (7)
\end{aligned}
$$

Combining Eq. (6) and (7), we get

$$
-\sum_{i=1}^{n-1} p_i \log\left(1+\frac{k\ln 2}{s_i}\right) \geq -\log(1+k\ln 2) - \frac{1}{\ln 2}. \qquad (8)
$$

We bound the third term in Eq. (3) crudely as $p_n \log Np_n \geq -1$. Along with the bounds for the previous two terms, Eq. (5), (8), this shows that

$$
\mathrm{S}(P\|\mathrm{U}_N) \geq f_{\mathrm{L}}(k,N) \stackrel{\text{def}}{=} k(\ln\log kN - \ln 6k + 1) - \log(1+k\ln 2) - 1 - \frac{1}{\ln 2}. \qquad (9)
$$

This proves the lower bound on the relative entropy.

Moving to an upper bound, we have for $i \geq 2$,

$$
\begin{aligned}
Np_i &= i2^{\frac{k}{s_i}} - (i-1)2^{\frac{k}{s_{i-1}}} \\
&= 2^{\frac{k}{s_i}} + (i-1)(2^{\frac{k}{s_i}} - 2^{\frac{k}{s_{i-1}}}) \\
&\leq 2^{\frac{k}{s_i}},
\end{aligned}
$$

since the second term is negative. This also holds for $i=1$, since $p_1 = s_1$ and $s_1 \log Ns_1 = k$. Therefore,

$$
\begin{aligned}
\mathrm{S}(P\|\mathrm{U}_N) &= \sum_{i=1}^{n} p_i \log Np_i \\
&\leq \sum_{i=1}^{n} \frac{kp_i}{s_i} \\
&\leq k + k\int_{s_1}^{1} \frac{1}{s} ds \\
&= k - k\ln s_1 \\
&\leq k + k\ln\left(\frac{\log Nk}{k}\right) \\
&= k(1 - \ln k + \ln(\log Nk)).
\end{aligned}
$$

In the last inequality, we used the lower bound $s_1 \geq k/\log Nk$. ■

The upper and lower bounds on the relative entropy of P with respect to the uniform distribution both behave as $k \log\log Nk$ up to constant factors.

Proof of Theorem 1.1: The dominating term in both of lower bound and upper bound on the relative entropy $\mathrm{S}(P\|\mathrm{U}_N)$, with P as in Theorem 3.2, is $k \ln\log Nk$ when N is large as compared with k. Specifically, when $N > 2^{36k^2}$, we have

$$\frac{1}{2} k \log\log Nk \quad \leq \quad \mathrm{S}(P\|\mathrm{U}_N) \quad \leq \quad 2k \log\log Nk.$$

By hypothesis, $1 \leq k$ and by Lemma 2.4, we have $k \leq \log N$. Thus, $\mathrm{S}(P\|\mathrm{U}_N) = \Theta(\mathrm{D}(P\|\mathrm{U}_N)\log\log N)$. The same holds for the ensembles constructed in Theorem 3.1. ■

The separation we demonstrated above is the best possible for ensembles of distributions that have a uniform average distribution.

Theorem 3.5 *For any positive integer N, and any ensemble $\mathcal{E} = \{(\lambda_j, Q_j) \;:\; j \in [M]\}$ of distributions over $[N]$ such that $\sum_{j=1}^{M} \lambda_j Q_j = \mathrm{U}_N$, we have*

$$\chi(\mathcal{E}) \quad \leq \quad K(2\ln\log N - \ln K + 1) + 16,$$

where $K = \mathrm{D}(\mathcal{E})$.

Proof: Let $\mathrm{D}(Q_j\|\mathrm{U}_N) = k_j$. We show that $\mathrm{S}(Q_j\|\mathrm{U}_N) \leq k_j(2\ln\log N - \ln k_j + 1)$ when $k_j \geq \frac{16}{N}$. When $k_j < \frac{16}{N}$, we have $\mathrm{S}(Q_j\|\mathrm{U}_N) < 16$. Since $k(2\ln\log N - \ln k + 1)$ is a concave function in k, averaging over j with respect to the distribution $\Lambda = (\lambda_j)$ gives the claimed bound.

Fix an j such that $k_j > \frac{16}{N}$. Let $R = Q_j^{\downarrow}$. Note that $\mathrm{D}(R\|\mathrm{U}_N) = k_j$ and $\mathrm{S}(R\|\mathrm{U}_N) = \mathrm{S}(Q_j\|\mathrm{U}_N)$. Consider the distribution P constructed as in Section 3 with $k = k_j$. Using the notation of that section, we have $s_i \log(Ns_i/i) = k_j$ for all $i < n$, and $s_n = 1$. Let $t_i = \sum_{l=1}^{i} r_l$, where $r_l \stackrel{\mathrm{def}}{=} \Pr(R = l)$. By definition, we have $t_i \log(Nt_i/i) \leq k_j = s_i \log(Ns_i/i)$. Since the function $g_i(y) = y\log(Ny/i)$ is strictly increasing for $y \geq i/N\mathrm{e}$, and $t_i \geq i/N$ (Fact 2.1), we have $t_i \leq s_i$ for $i < n$. Since $s_i = 1$ for $i \geq n$, we have $t_i \leq s_i$ for these i as well. In other words, $P \succeq R$. By Fact 2.2, we have $\mathrm{H}(P) \leq \mathrm{H}(R)$. This is equivalent to $\mathrm{S}(R\|\mathrm{U}_N) \leq \mathrm{S}(P\|\mathrm{U}_N)$. By Theorem 3.2, $\mathrm{S}(P\|\mathrm{U}_N) \leq k_j(\ln\log(Nk_j) - \ln k_j + 1)$. Since $k_j \leq \log N$, this is at most $k_j(2\ln\log N - \ln k_j + 1)$. ■

Finally, we observe that this is also the best separation possible for an ensemble of quantum states with a completely mixed ensemble average.

Theorem 3.6 *For any positive integer N, and any ensemble $\mathcal{E} = \{(\lambda_j, \rho_j) \;:\; j \in [M]\}$ of quantum states ρ_j over a Hilbert space of dimension N such that $\sum_{j=1}^{M} \lambda_j \rho_j = \frac{\mathrm{I}}{N}$, the completely mixed state of dimension N, we have*

$$\chi(\mathcal{E}) \quad \leq \quad K(2\ln\log N - \ln K + 1) + 16,$$

where $K = \mathrm{D}(\mathcal{E})$.

Proof: Let Q_j be the probability distribution on $[N]$ corresponding to the eigenvalues of ρ_j. By definition of observational divergence for quantum states, $\mathrm{D}(Q_j\|\mathrm{U}_N) \leq \mathrm{D}(\rho_j\|\frac{\mathrm{I}}{N})$. Further, we have $\mathrm{S}(\rho_j\|\frac{\mathrm{I}}{N}) = \mathrm{S}(Q_j\|\mathrm{U}_N)$. We now apply the same reasoning as in the proof of Theorem 3.5, note that the divergence of the ensemble $\{(\lambda_j, Q_j) \;:\; j \in [M]\}$ is bounded by $\mathrm{D}(\mathcal{E})$, and that the RHS in the statement is a non-decreasing function of K. This gives us the stated bound. (Note that we do not need $\sum_{j=1}^{M} \lambda_j Q_j = \mathrm{U}_N$ to use the reasoning in Theorem 3.5.) ■

References

[1] Harry Buhrman, Matthias Christandl, Patrick Hayden, Hoi-Kwong Lo, and Stephanie Wehner. Security of quantum bit string commitment depends on the information measure. *Physical Review Letters*, 97, 2006. Article no. 250501.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

[3] Rahul Jain. Stronger impossibility results for quantum string commitment. Technical Report arXiv:quant-ph/0506001v4, ArXiv.org Preprint Archive, `http://www.arxiv.org/`, 2005.

[4] Rahul Jain. Communication complexity of remote state preparation with entanglement. *Quantum Information and Computation*, 6(4–5):461–464, July 2006.

[5] Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. Privacy and interaction in quantum communication complexity and a theorem about the relative entropy of quantum states. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 429–438. IEEE Computer Society Press, Los Alamitos, CA, USA, 2002. A more complete version appears as [7].

[6] Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. Prior entanglement, message compression and privacy in quantum communication. In *Proceedings of the 20th Annual IEEE Conference on Computational Complexity*, pages 285–296. IEEE Computer Society Press, Los Alamitos, CA, USA, 2005.

[7] Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. A theorem about relative entropy of quantum states with an application to privacy in quantum communication. Technical Report arXiv:0705.2437v1, ArXiv.org Preprint Archive, `http://www.arxiv.org/`, May 2007.

[8] Adrian Kent. Quantum bit string commitment. *Physical Review Letters*, 90, 2003. Article no. 237901.

[9] Hoi-Kwong Lo and H. F. Chau. Is quantum bit commitment really possible? *Physical Review Letters*, 78:3410–3413, 1997.

[10] Dominic Mayers. Unconditionally secure quantum bit commitment is impossible. *Physical Review Letters*, 78(17):3414–3417, 1997.

[11] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, UK, 2000.

A Implications for quantum protocols

A.1 Quantum string commitment

A *string commitment* scheme is an extension of the well-studied and powerful cryptographic primitive of *bit commitment*. In such schemes, one party, Alice, wishes to commit an entire string $x \in \{0,1\}^n$ to another party, Bob. The protocol is required to be such that Bob not be able to identify the string until it is revealed by Alice. In turn, Alice should not be able to renege on her commitment at the time of revelation. Formally, quantum string commitment protocols are defined as follows [1, 3].

Definition A.1 (Quantum string commitment (QSC)) *Let $P = \{p_x : x \in \{0,1\}^n\}$ be a probability distribution and let B be a measure of information contained in an ensemble of quantum states. A (n, a, b)-B-**QSC** protocol for P is a quantum communication protocol between two parties, Alice and Bob. Alice gets an input $x \in \{0,1\}^n$ chosen according to the distribution P. The starting joint state of the qubits of Alice and Bob is some pure state independent of x. The protocol runs in two phases: the commit phase, followed by the reveal phase. There are no intermediate measurements during the protocol. At the end of the reveal phase, Bob measures his qubits according to a POVM $\{M_y \,:\, y \in \{0,1\}^n\} \cup \{I - \sum_y M_y\}$ to determine the value of the committed string by Alice or to detect cheating. The protocol satisfies the following properties.*

1. **(Correctness)** *Suppose Alice and Bob act honestly. Let ρ_x be the state of Bob's qubits at the end of the reveal phase of the protocol, when Alice gets input x. Then $(\forall x, y)$ $\mathrm{Tr}\ M_y\rho_x = 1$ iff $x = y$, and 0 otherwise.*
2. **(Concealing property)** *Suppose Alice acts honestly, and Bob possibly cheats, i.e., deviates from the protocol in his local operations. Let σ_x be the state of Bob's qubits after the commit phase when Alice gets input x. Then the B information $B(\mathcal{E})$ of the ensemble $\mathcal{E} = \{p_x, \sigma_x\}$ is at most b. In particular, this also holds when both Alice and Bob follow the protocol honestly.*
3. **(Binding property)** *Suppose Bob acts honestly , and Alice possibly cheats. Let $c \in \{0,1\}^n$ be a string in a special cheating register C with Alice that she keeps independent of the rest of the registers till the end of the commit phase. Let τ_c be the state of Bob's qubits at the end of the reveal phase when Alice has c in the cheating register. Let $q_c \stackrel{\mathrm{def}}{=} \mathrm{Tr}\ M_c\tau_c$. Then*

$$\sum_{c \in \{0,1\}^n} p_c q_c \leq 2^{a-n}$$

The idea behind the above definition is as follows. At the end of the reveal phase of an honest run of the protocol Bob identifies x from ρ_x by performing the POVM measurement $\{M_y\}_y \cup \{I - \sum_y M_y\}$. He accepts the committed string to be x iff the observed outcome $y = x$; this happens with probability $\mathrm{Tr}\ M_x\rho_x$. He declares that Alice is cheating if outcome $I - \sum_x M_x$ is observed. Thus, at the end of an honest run of the protocol, with probability 1, Bob accepts the committed string as being exactly Alice's input string. The concealing property ensures that the amount of B information about x that a possibly cheating Bob gets is bounded by b. In *bit*-commitment protocols, the concealing property is quantified in terms of the probability with which Bob can guess Alice's bit. Here we instead use different notions of information contained in the corresponding ensemble. The binding property ensures that when a cheating Alice wishes to postpone committing to a string string until after the commit phase, then she succeeds in forcing an honest Bob to accept her choice with bounded probability (in expectation).

Strong string commitment, in which both parameters a, b above are required to be 0, is impossible for the same reason that of *strong* bit-commitment protocols are impossible [10, 9]. Weaker versions are nonetheless possible, and exhibit a trade-off between the concealing and binding properties. The trade-off between the parameters a and b has been studied by several researchers [8, 1, 3]. Buhrman, Christandl, Hayden, Lo, and Wehner [1] study this trade-off both in the scenario of a single execution of the protocol and also in the asymptotic regime, with an unbounded number of parallel executions of the protocol. In the asymptotic scenario, they show the following result in terms of Holevo information (which is denoted by χ).

Theorem A.1 ([1]) *Let Π be an (n, a_1, b)-χ-**QSC** scheme. Let Π_m represent m independent, parallel executions of Π (so $\Pi_1 = \Pi$). Let a_m represent the binding parameter of Π_m and let $a \stackrel{\text{def}}{=} \lim_{m\to\infty} a_m/m$. Then, $a + b \geq n$.*

Jain [3] shows a similar trade-off result regarding **QSC**s, in terms of the divergence information of an ensemble (denoted by D).

Theorem A.2 ([3]) *For single execution of the protocol of an (n, a, b)-D-**QSC** scheme,*

$$a + b + 8\sqrt{b+1} + 16 \geq n.$$

As mentioned before, for any ensemble $\mathcal{E}$, divergence information is bounded by the Holevo χ-information $\mathrm{D}(\mathcal{E}) \leq \chi(\mathcal{E}) + 1$. This immediately implies:

Theorem A.3 ([3]) *For single execution of the protocol of a (n, a, b)-χ-**QSC** scheme*

$$a + b + 8\sqrt{b+2} + 17 \geq n.$$

As Jain shows, this implies the asymptotic result due to Buhrman *et al.* (Theorem A.1).

The separation that we demonstrate between divergence and Holevo information (Theorem 1.1) shows that for some ensembles over n qubits, $\mathrm{D}(\mathcal{E})$ may be a $\log n$ larger than $\chi(\mathcal{E})$. For such ensembles the binding-concealing trade-off of Theorem A.2 is stronger than that of Theorem A.1.

A.2 Privacy trade-off for two-party protocols for relations

Let us consider two-party protocols between Alice and Bob for computing a relation $f \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. The goal here is to find a $z \in \mathcal{Z}$ such that $(x, y, z) \in f$, when Alice and Bob are given $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, respectively. Jain, Radhakrishnan, and Sen [5] studied to what extent the two parties may solve f while keeping their respective inputs hidden from the other party. They showed the following:

Result A.4 ([6], informal statement) *Let μ be a product distribution on $\mathcal{X} \times \mathcal{Y}$. Let $Q^{\mu, A\to B}_{1/3}(f)$ represent the one-way distributional complexity of f with a single communication from Alice to Bob and distributional error under μ at most $1/3$. Let X and Y represent the random variables corresponding to Alice and Bob's inputs respectively. If there is a quantum communication protocol for f where Bob* leaks *divergence information at most b about his input Y, then Alice leaks divergence information at least $\Omega(Q^{\mu, A\to B}_{1/3}(f)/2^{O(b)})$ about her input X. A similar statement also holds with the roles of Alice and Bob interchanged.*

From the upper bound on the divergence information in terms of Holevo information this immediately implies the following.

Result A.5 ([6], informal statement) *Let μ be a product distribution on $\mathcal{X} \times \mathcal{Y}$. Let $Q^{\mu, A \to B}_{1/3}(f)$ represent the one-way distributional complexity of f with a single communication from Alice to Bob and distributional error under μ at most $1/3$. Let X and Y represent the random variables corresponding to Alice and Bob's inputs respectively. If there is a quantum communication protocol for f where Bob* leaks *Holevo information at most b about his input Y, then Alice leaks Holevo information at least $\Omega(Q^{\mu, A \to B}_{1/3}(f)/2^{O(b)})$ about her input X. A similar statement also holds with the roles of Alice and Bob interchanged.*

It follows from Theorem 1.1 that Result A.4 is much stronger than the second, Result A.5 in case the ensembles arising in the protocol between Alice and Bob has divergence information much smaller than its Holevo information.