The Communication Complexity of Correlation

Prahladh Harsha* Rahul Jain† David McAllester‡ Jaikumar Radhakrishnan§

December 3, 2006

Abstract

We examine the communication required for generating random variables remotely. One party Alice will be given a distribution D, and she has to send a message to Bob, who is then required to generate a value with distribution exactly D. Alice and Bob are allowed to share random bits generated without the knowledge of D. There are two settings based on how the distribution D provided to Alice is chosen.

Average case: D is itself chosen randomly from some set (the set and distribution are known in advance) and we wish to minimize the expected communication in order for Alice to generate a value y, with distribution D. We characterize the communication required in this case based on the mutual information between the the input to Alice and the output Bob is required to generate.

Worst case: D is chosen from a set of distributions $\mathcal{D}$, and we wish to devise a protocol so that the expected communication (the randomness comes from the shared random string and Alice's coin tosses) is small for each $D \in \mathcal{D}$. We characterize the communication required in this case in terms of the channel capacity associated with the set $\mathcal{D}$.

Prior to this work, only the limit (or asymptotic) versions of these results were known, where Alice is given a sequence of distributions $\langle D_1, D_2, \ldots, D_n \rangle$, and Bob is required to generate n values $\langle y_1, y_2, \ldots, y_n$, where y_i has distribution D_i. Here the amortized cost (per D_i) is studied as n tends to infinity. In the case, when the D_i's are iid and some error is allowed Winter [Win] characterized the cost in terms of mutual information. In the case where D_i's are only known to come from some set $\mathcal{D}$ and we require worst case bounds, the Reverse Shannon Theorem of Bennett *et al.* [BSST] characterizes the limiting amortized cost in terms of the channel capacity.

Our results, are for the *one-shot* case, and immediately imply the limit versions shown earlier. We use our one-shot protocol to derive a direct sum result in communication complexity, for which the asymptotic versions do not seem to help. Our result substantially improves the previous such result shown by Jain *et al.* [JRSb].

An essential ingredient in our proofs is an improved rejection sampling procedure that relates the relative entropy between two distributions to the communication complexity of generating one distribution from the other.

*Toyota Technological Institute, Chicago, USA. E-mail: prahladh@tti-c.org.

†Institute for Quantum Computing and Dept. of Computer Science, Univ. of Waterloo, Waterloo CANADA. E-mail: rjain@iqc.ca. Part of the work was done while the author was at the Univ. California, Berkeley.

‡Toyota Technological Institute, Chicago, USA. E-mail: mcallester@tti-c.org.

§Tata Institute of Fundamental Research, Mumbai, INDIA. E-mail: jaikumar@tifr.res.in. Research done while the author was at the Toyota Technological Institute, Chicago.

1 Introduction

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite non-empty sets, and let $(\mathbf{X}, \mathbf{Y})$ be a pair of (correlated) random variables taking values in $\mathcal{X} \times \mathcal{Y}$. Consider the following communication problem between two parties, Alice and Bob. Alice is given a random input $x \in \mathcal{X}$, sampled according to the distribution $\mathbf{X}$. (We use the same symbol to refer to a random variable and its distribution.) Alice needs to transmit a message $\mathbf{M}$ to Bob so that Bob can generate a value $y \in \mathcal{Y}$, that is distributed according to the conditional distribution $\mathbf{Y}|_{\mathbf{X}=x}$ (i.e., the pair (x, y) has joint distribution $(\mathbf{X}, \mathbf{Y})$). How many bits must Alice send Bob in any protocol that accomplishes this? It follows from the data-processing inequality in information theory that this minimum, which we shall call $T(\mathbf{X} : \mathbf{Y})$, is at least the mutual information between, $\mathbf{X}$ and $\mathbf{Y}$, that is,

$$I[\mathbf{X} : \mathbf{Y}] \triangleq H[\mathbf{X}] + H[\mathbf{Y}] - H[\mathbf{X}, \mathbf{Y}],$$

where $H[\mathbf{Z}]$ denotes the Shannon entropy of the random variable $\mathbf{Z}$ (This lower bound is implied by the following sequence of inequalities: $T(\mathbf{X} : \mathbf{Y}) \geq H[\mathbf{M}] \geq I[\mathbf{X} : \mathbf{M}] \geq I[\mathbf{X} : \mathbf{Y}]$, where the last inequality is the data-processing inequality (cf. [CT, Page 32, Theorem 2.8.1]) applied to the Markov chain $\mathbf{X} \to \mathbf{M} \to \mathbf{Y}$.) We can also consider a slightly relaxed version of this problem which allows for some error. More formally, let $T_\lambda(\mathbf{X}, \mathbf{Y})$ denote the minimum expected number of bits Alice needs to send Bob so that the joint distribution generated by the protocol, which we call $(\mathbf{X}, \Pi(\mathbf{X}))$, to be λ-close in total variation distance[1] to the joint distribution $(\mathbf{X}, \mathbf{Y})$.

How good is $I[\mathbf{X} : \mathbf{Y}]$ as a lower bound for $T(\mathbf{X} : \mathbf{Y})$ (or $T_\lambda(\mathbf{X} : \mathbf{Y})$)? Does the complexity of this communication problem provide us a functional interpretation of the information theoretic notion of mutual information?

This problem was first studied by Wyner [Wyn], who considered its asymptotic version (with error), where Alice is given several independently drawn samples $(x_1, \ldots, x_m)$ from the distribution $\mathbf{X}^m$ and Bob needs to generate $(y_1, \ldots, y_m)$ such that the output distribution of $((x_1, y_1), \ldots, (x_m, y_m))$ is λ-close to the distribution $(\mathbf{X}, \mathbf{Y})^m$. Wyner referred to the amortized minimum expected number of bits Alice needs to send Bob as the *common information* $C(\mathbf{X} : \mathbf{Y})$ of the random variables $\mathbf{X}$ and $\mathbf{Y}$, i.e.,

$$C(\mathbf{X} : \mathbf{Y}) \triangleq \liminf_{\lambda \to 0} \left[\lim_{m \to \infty} \frac{T_\lambda(\mathbf{X}^m : \mathbf{Y}^m)}{m} \right]. \tag{1}$$

He then obtained the following remarkable information theoretic characterization of common information.

Theorem 1.1 (Wyner's Theorem [Wyn, Theorem 1.3])

$$C(\mathbf{X} : \mathbf{Y}) = \min_{\mathbf{W}} I[\mathbf{XY} : \mathbf{W}],$$

where the minimum is taken over all random variables $\mathbf{W}$ *such that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are conditionally independent given* $\mathbf{W}$*.*

It can easily be verified that $T(\mathbf{X} : \mathbf{Y}) \geq C(\mathbf{X} : \mathbf{Y}) \geq I[\mathbf{X} : \mathbf{Y}]$ (See Section 6 for a proof of these inequalities). However, both these inequalities can be very loose. To demonstrate how weak these inequalities can be, in Section 6, we give examples of joint distributions $(\mathbf{X}, \mathbf{Y})$ that satisfy $T(\mathbf{X} : \mathbf{Y}) = \omega(C(\mathbf{X} : \mathbf{Y})$ and $C(\mathbf{X} : \mathbf{Y}) = \omega(I[\mathbf{X} : \mathbf{Y}])$. Thus, this seeming natural problem does not offer us the functional characterization for $I[\mathbf{X} : \mathbf{Y}]$ we were initially hoping for.

[1] The total variation distance between two distribution P and Q is defined as $\max_{S \subseteq \mathcal{X}} |P(S) - Q(S)|$, which is also equal to $\frac{1}{2}\|P - Q\|_1$ where $\|\cdot\|_1$ is the ℓ^1-norm

1.1 A characterization of mutual information

Our first result shows that there is such a characterization if Alice and Bob are allowed to share random information, generated independently of Alice's input. In fact, then Alice need send no more than approximately $I(\mathbf{X} : \mathbf{Y})$ bits to Bob. In order to state our result precisely, let us first define the kind of communication protocol Alice and Bob are expected to use.

Definition 1.2 (one-way protocol) *In a one-way protocol, the two parties* Alice *and* Bob *share a random string* $\mathbf{R}$*, and also have private random strings* $\mathbf{R}_A$ *and* $\mathbf{R}_B$ *respectively.* Alice *receives an input* $x \in \mathcal{X}$*. Based on the shared random string* $\mathbf{R}$ *and her own private random string* $\mathbf{R}_A$*, she sends a message* $\mathbf{M}(x, \mathbf{R}, \mathbf{R}_A)$ *to Bob. On receiving the message* $\mathbf{M}$*,* Bob *computes the output* $y = y(\mathbf{M}, \mathbf{R}, \mathbf{R}_B)$*. The protocol is thus specified by the two functions* $\mathbf{M}(x, \mathbf{R}, \mathbf{R}_A)$ *and* $y(\mathbf{M}, \mathbf{R}, \mathbf{R}_B)$ *and the distributions for the random strings* $\mathbf{R}$*,* $\mathbf{R}_A$ *and* $\mathbf{R}_B$*. For such a protocol* Π*, let* $\Pi(x)$ *denote its (random) output when the input given to* Alice *is* x*. Let* $T_\Pi(x)$ *be the expected length of the message transmitted by* Alice *to* Bob*, that is,* $\mathbb{E}[|\mathbf{M}(x, \mathbf{R}, \mathbf{R}_A)|]$*. Note that the private random strings can be considered part of the shared random string if we are not concerned about minimizing the amount of shared randomness.*

Definition 1.3 *Given random variables* $(\mathbf{X}, \mathbf{Y})$*, let*

$$T^R_\lambda(\mathbf{X} : \mathbf{Y}) \stackrel{\Delta}{=} \min_{\Pi} \mathop{\mathbb{E}}_{x \leftarrow \mathbf{X}}[T_\Pi(x)],$$

where Π *ranges over all one-way protocols where* $(\mathbf{X}, \Pi(\mathbf{X}))$ *is* λ*-close in total variation distance to the distribution* $(\mathbf{X}, \mathbf{Y})$*. For the special case when* $\lambda = 0$*, we define* $T^R(\mathbf{X} : \mathbf{Y}) \stackrel{\Delta}{=} T^R_0(\mathbf{X} : \mathbf{Y})$*.*

As in the case of $T(\mathbf{X} : \mathbf{Y})$, it again follows from the data processing inequality that $T^R(\mathbf{X} : \mathbf{Y})$ is bounded below by the mutual information $I[\mathbf{X} : \mathbf{Y}]$. This lower bound is implied by the following sequence of inequalities: $T^R(\mathbf{X} : \mathbf{Y}) \geq H[\mathbf{M}] \geq H[\mathbf{M} \mid \mathbf{R}] \geq I[\mathbf{X} : \mathbf{M} \mid \mathbf{R}] = I[\mathbf{X} : \mathbf{M} \mid \mathbf{R}] + I[\mathbf{X} : \mathbf{R}] = I[\mathbf{X} : \mathbf{MR}] \geq I[\mathbf{X} : \mathbf{Y}]$. We have used the fact that $I[\mathbf{X} : \mathbf{R}] = 0$ since $\mathbf{X}$ and $\mathbf{R}$ are independent. The last inequality is the data-processing inequality (cf. [CT, Page 32, Theorem 2.8.1]) applied to the Markov chain $\mathbf{X} \to (\mathbf{M}, \mathbf{R}) \to \mathbf{Y}$.

Our first result shows that this lower bound is essentially tight, giving a characterization of mutual information (modulo some lower order logarithmic terms[2]).

Result 1 (characterization of mutual information)

$$I[\mathbf{X} : \mathbf{Y}] \leq T^R(\mathbf{X} : \mathbf{Y}) \leq I[\mathbf{X} : \mathbf{Y}] + 2\lg(I[\mathbf{X} : \mathbf{Y}] + 1) + O(1).$$

We have an additive $2 \lg I[\mathbf{X} : \mathbf{Y}]$ term in the upper bound because our proof of the result employs a prefix-free encoding of integers that requires $\ln n + 2 \lg \lg n$ bits to encode the integer n. By using an encoding that requires $\ln n + (1 + \varepsilon) \lg \lg n$ bits, the constant 2 can be improved to $(1 + \varepsilon)$ for any $\varepsilon > 0$.

The above result does not place any bound on the amount of randomness that Alice and Bob need to share. In fact, there exist distributions $(\mathbf{X}, \mathbf{Y})$ for which our proof of Result 1 requires Alice and Bob to share a random string of unbounded length. However, using technique from network flows, we can bound the amount of shared randomness by $O(\lg |\mathcal{X}| + \lg |\mathcal{Y}|)$ if we are allowed to increase the expected communication by an additive factor of $O(\lg \lg(|\mathcal{X}| + \mathcal{Y}|))$ (proof not included in paper for want of space).

[2] All logarithms (denoted by lg) in this paper are with respect to base 2.

1.2 Generating one distribution from another

The main tool in our proof of Theorem 1 is a sampling procedure for generating one distribution from another. This sampling procedure is of independent interest because it relates the *relative entropy* between two distributions and the communication complexity of generating one distribution from the other.

Definition 1.4 (relative entropy) *The* relative entropy *or* Kullback-Leibler divergence *between two probability distributions P and Q on a finite set $\mathcal{X}$ is*

$$S(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \lg \frac{P(x)}{Q(x)}.$$

Observe that $S(P\|Q)$ is finite if and only if the support of distribution P (i.e., the set of points $x \in \mathcal{X}$ such that $P(x) > 0$) is contained in the support of the distribution Q; in that case it is zero iff $P = Q$, but otherwise always positive.

Let P and Q be two distributions such that the relative entropy $S(P\|Q)$ is finite. We consider the problem of generating a sample according to P from a sequence of samples drawn according to Q. Let $\langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots \rangle$ be a sequence of samples, drawn independently, each with distribution Q. The idea is to generate an index $\mathbf{i}^*$ (a random variable depending on the sample) so that the sample $\mathbf{x}_{\mathbf{i}^*}$ has distribution P. For example, if P and Q are identical, then we can set $\mathbf{i}^* = 1$ and be done. It is easy to show that for any such procedure

$$\mathbb{E}[\ell(\mathbf{i}^*)] \geq S(P\|Q) - 1,$$

where $\ell(\mathbf{i}^*)$ is the length of the binary encoding of $\mathbf{i}^*$. We show that there, in fact, exists a procedure that almost achieves this lower bound. More formally, we have

Lemma 1.5 (rejection sampling lemma) *Let P and Q be two distributions such that $S(P\|Q)$ is finite. There exists a sampling procedure $\mathcal{P}$ which on input a sequence $\langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots \rangle$ of independently drawn samples from the distribution Q outputs an index $\mathbf{i}^*$ such that the sample $\mathbf{x}_{\mathbf{i}^*}$ is distributed according to the distribution P and the expected encoding length of the index $\mathbf{i}^*$ is at most*

$$S(P\|Q) + 2\lg(S(P\|Q) + 1) + O(1),$$

where the expectation is taken over the sample sequence and the internal random coins of the procedure $\mathcal{P}$.

As in the case of Result 1, the constant 2 can be improved to any constant $(1 + \varepsilon)$ for any $\varepsilon > 0$.

1.3 Reverse Shannon theorem

In Result 1, we considered the communication cost averaged over $x \in \mathcal{X}$, chosen according to the distribution of $\mathbf{X}$. We now consider the worst-case communication over all $x \in \mathcal{X}$ (but we still average over the random choices of the protocol). The setting is similar to the earlier section. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite non-empty sets as before. Let $\mathcal{P}_\mathcal{Y}$ be the set of all probability distributions on the set $\mathcal{Y}$. A channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is a function $E : \mathcal{X} \to \mathcal{P}_\mathcal{Y}$. The Shannon capacity of such a channel is

$$\mathcal{C}(E) \stackrel{\Delta}{=} \max_{(\mathbf{X},\mathbf{Y})} I[\mathbf{X} : \mathbf{Y}],$$

where $(\mathbf{X}, \mathbf{Y})$ is a pair of random variables taking values in $\mathcal{X} \times \mathcal{Y}$ such that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\Pr[\mathbf{Y} = y \mid \mathbf{X} = x] = E(x)(y)$. A simulator for this channel (using a noiseless communication channel and shared randomness) is a one-way protocol Π such that for all $x \in \mathcal{X}$, Bob's output $\Pi(x)$ has distribution $E(x)$. The goal is to minimize the worst-case communication; let

$$T(E) = \min_{\Pi} \max_{x \in \mathcal{X}} T_{\Pi}(x),$$

where the minimum is taken over all valid simulators Π of E. The following relationship between $T(E)$ and $\mathcal{C}(E)$ is easy to show and well known.

Proposition 1.6 $T(E) \geq \mathcal{C}(E)$.

Using the rejection sampling lemma (Lemma 1.5), we can show that this lower bound is essentially tight (modulo some lower order logarithmic terms). A result of this nature is called the Reverse Shannon Theorem as it gives an (optimal) simulation of noisy channels using noiseless channels and shared randomness.

Result 2 (one-shot reverse Shannon theorem) $T(E) \leq \mathcal{C}(E) + 2\lg(\mathcal{C}(E) + 1) + O(1)$.

We call this result, the "one-shot Reverse Shannon Theorem", since asymptotic versions of this result was previously known (See Section 1.5 for a discussion of these asymptotic results).

1.4 A direct-sum result in communication complexity

To present our next result, we need to recall some standard definitions from two-party communication complexity. We refer the reader to the book by Kushilevitz and Nisan [KN] for an excellent introduction to communication complexity. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be finite non-empty sets, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a function. A two-party protocol for computing f consists of two parties, Alice and Bob, who get inputs $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively, and exchange messages in order to compute $f(x, y) \in \mathcal{Z}$. A protocol is said to be k-round, if the two parties exchange at most k messages.

For a distribution μ on $\mathcal{X} \times \mathcal{Y}$, let the ε-error k-round distributional communication complexity of f under μ (denoted by $D_{\varepsilon}^{\mu,k}(f)$), be the number of bits communicated (for the worst-case input) by the best deterministic k-round protocol for f with average error at most ε under μ. Let $R_{\varepsilon}^{\mathrm{pub},k}(f)$, the public-coins k-round randomized communication complexity of f with worst case error ε, be the number of bits communicated (for the worst-case input) by the best k-round public-coins randomized protocol, that for each input (x, y) computes $f(x, y)$ correctly with probability at least $1 - \varepsilon$. Randomized and distributional complexity are related by the following celebrated result of Yao [Yao].

Theorem 1.7 (Yao's minmax principle [Yao]) $R_{\varepsilon}^{\mathrm{pub},k}(f) = \max_{\mu} D_{\varepsilon}^{\mu,k}(f)$

For function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, the t-fold *direct sum* of f, $f^{\oplus t} : \mathcal{X}^t \times \mathcal{Y}^t \to \mathcal{Z}^t$, is defined by $f^{\oplus t}(\langle x_1, \ldots, x_t \rangle, \langle y_1, \ldots, y_t \rangle) \stackrel{\Delta}{=} \langle f(x_1, y_1), \ldots, f(x_t, y_t) \rangle$. It is natural to ask if the communication complexity of $f^{\oplus t}$ is at least t times that of f. This is commonly known as the direct sum question. The direct sum question is a very basic question in communication complexity and had been studied for a long time. Several results were known for this question in restricted settings for deterministic and randomized protocols [KN]. Recently Chakrabarti, Shi, Wirth and Yao [CSWY] studied this question in the *Simultaneous message passing* (SMP) model in which instead of Alice and Bob

communicating among themselves send a message each to a third party Referee who then outputs a z such that $f(x, y) = z$. They showed that in this model, the *Equality* function EQ satisfies the direct sum property. Their result also holds for any function that satisfies a certain robustness requirement. This result was then extended by Jain, Radhakrishnan and Sen [JRSd] to hold for all functions and relations, not necessarily satisfying the robustness requirement, both in the one-way and the SMP model of communication. In another work Jain *et al.* [JRSb] showed, a weaker direct sum result for bounded-round two-way protocols. Their result was the following (here μ is a product distribution on $\mathcal{X} \times \mathcal{Y}$ and k represents the number of rounds):

$$D_{\varepsilon}^{\mu^t,k}(f^{\oplus t}) \geq t\left(\frac{\delta^2}{2k} \cdot D_{\varepsilon+2\delta}^{\mu,k}(f) - 2\right)$$

Using the rejection sampling lemma (Lemma 1.5), we obtain a considerable strengthening of the above result.

Result 3 (direct sum result) *For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, and a product distribution μ on $\mathcal{X} \times \mathcal{Y}$, we have*

$$D_{\varepsilon}^{\mu^t,k}(f^{\oplus t}) \geq \frac{t}{2}\left(\delta D_{\varepsilon+\delta}^{\mu,k}(f) - O(k)\right).$$

Applying Yao's minmax principle (Theorem 1.7), we have:

$$R_{\varepsilon}^{\mathrm{pub},k}(f^{\oplus t}) \geq \max_{\mu}\left(\frac{t}{2}\left(\delta D_{\varepsilon+\delta}^{\mu,k}(f) - O(k)\right)\right).$$

where the maximum above is taken over all product distributions μ on $\mathcal{X} \times \mathcal{Y}$.

1.5 Related work

Asymptotic versions of our Results 1 and 2 were independently shown by Winter [Win] and Bennett *et al.* [BSST] respectively.

Theorem 1.8 ([Win, Theorem 9 and Remark 10]) *For every pair of distributions $(\mathbf{X}, \mathbf{Y})$ and $\lambda > 0$ and n, there exists a one-way protocol Π_n such that the distribution $(\mathbf{X}^n, \Pi_n(\mathbf{X}^n))$ is λ-close in total variation distance to the joint distribution $(\mathbf{X}^n, \mathbf{Y}^n)$ and furthermore,*

$$\max_{\overline{x} \in \mathcal{X}^n} T_{\Pi_n}(\overline{x}) \leq nI[\mathbf{X} : \mathbf{Y}] + O\left(\frac{1}{\lambda}\right) \cdot \sqrt{n}.$$

Theorem 1.9 (reverse Shannon Theorem [BSST]) *Let E be a discrete memoryless channel with Shannon capacity C and $\varepsilon > 0$. Then, for each block size n there is a deterministic simulation protocol Π_n for E^n which makes use of a noiseless channel and prior random information R shared between sender and receiver. The simulation is exactly faithful in the sense that for all n, and for all xbar $\in \mathcal{X}^n$, the output $\Pi_n(\overline{x})$ has the distribution $E^n(\overline{x})$, and it is asymptotically efficient in the sense that*

$$\lim_{n \to \infty} \max_{\overline{x} \in \mathcal{X}^n} \Pr[C_{\Pi_n}(\overline{x}) > n(C(E) + \varepsilon)] = 0.$$

It is to be noted that the asymptotic result of [Win] is slightly stronger than what is stated above in that Winter's result actually bounds the worst case number of bits communicated while our results (and the above statement) bound the expected number of bits communicated. Despite this, these asymptotic results (and their stronger counterparts) follow immediately from our results by routine applications of the law of large numbers.

One-shot vs. asymptotic results In the light of the above, it might seem natural to ask why would one be interested in one-shot versions of known asymptotic results. Our motivation for the one-shot versions is two-fold.

- The asymptotic equipartition property (cf. [CT, Chapter 3]) for distributions states that for sufficiently large n, n independently drawn samples from a distribution X almost always fall in what are called "typical sets". Typical sets have the property that all elements in it are nearly equiprobable and the size of the typical set is approximately $2^{nH[X]}$. Any property that is proved for typical sets will then be true with high probability for a large sequence of independently drawn samples. Thus, to prove the asymptotic results, it suffices to prove the same for typical sets. Thus, one might contend that these asymptotic results are in fact properties of typical sets and it could be the case that the results are in fact, not true for the one-shot case. Our results show that this is not the case and one need not resort to typical sequences to prove them.
- Second, our results provide tools for certain problems in communication complexity (e.g., our improved direct sum result). For such communication complexity applications, the asymptotic versions do not suffice and we require the one-shot versions.

Bounding shared randomness As mentioned earlier, we can bound the shared randomness in Result 1 by $O(\lg |\mathcal{X}| + \lg |\mathcal{Y}|)$ if we are allowed to increase the expected communication by an additive factor of $O(\lg \lg(|\mathcal{X}| + \mathcal{Y}|))$. This raises the natural question of tradeoffs between shared randomness and expected communication. The asymptotic version of this problem was recently solved by Bennett and Winter (Personal Communication [BW]).

Substate Theorem Jain, Radhakrishnan and Sen [JRSa] prove the following result relating the relative entropy between two distributions P and Q to how well a distribution is contained in another.

Theorem 1.10 (classical substate Theorem, [JRSa]) *Let P and Q be two distributions such that $k = S(P\|Q)$ is finite. For all $\varepsilon > 0$ there exists a distribution P' such that $\|P' - P\|_1 \leq \varepsilon$ and $Q = \alpha P' + (1-\alpha)P''$ where P'' is some other distribution and $\alpha = 2^{-O(k/\varepsilon)}$.*

The rejection sampling lemma (Lemma 1.5) is a strengthening of the above theorem (the above theorem follows from Lemma 1.5 by an application of Markov's inequality). In fact, the classical substate theorem can then be used to prove a weaker version of Result 1 which allows for error. More precisely, one can infer (from Theorem 1.10) that $T^R_\lambda(\mathbf{X} : \mathbf{Y}) \leq O(I[\mathbf{X} : \mathbf{Y}])/\lambda)$. It is to be noted that the fundamental contribution of Jain, Radhakrishnan and Sen [JRSa] is actually a quantum analogue of the above substate theorem. It is open if there exist quantum analogues of our results.

Lower Bounds using message compression Chakrabarti and Regev [CR] prove that any randomized cell probe algorithm that solves the approximate nearest search problem on the Hamming cube $\{0,1\}^d$ using polynomial storage and word size $d^{O(1)}$ requires a worst case query time of $\Omega(\lg \lg d / \lg \lg \lg d)$. An important component in their proof of this lower bound is the message compression technique of Jain, Radhakrishnan and Sen [JRSb]. The rejection sampling lemma (Lemma 1.5) can be used to improve message compression of [JRSb], which in turn simplifies the

lower bound argument of Chakrabarti and Regev. It is likely that there are other similar applications.

Organization

The rest of the paper is organized as follows: We first prove Results 1 and 2 assuming the rejection sampling lemma (Lemma 1.5) in Sections 2 and 3 respectively. We then proceed to prove the rejection sampling lemma in Section 4. The Direct Sum Result (Result 3) is then proved in Section 5. Finally, in Section 6, we give examples of joint distributions $(\mathbf{X}, \mathbf{Y})$ that satisfy $T(\mathbf{X} : \mathbf{Y}) = \omega(C(\mathbf{X} : \mathbf{Y}))$ and $C(\mathbf{X} : \mathbf{Y}) = \omega(I[\mathbf{X} : \mathbf{Y}])$.

2 Proof of Result 1

Result 1 follows easily from the rejection sampling lemma (Lemma 1.5) and the following well-known relationship between relative entropy and mutual information.

Fact 2.1 $I[\mathbf{X} : \mathbf{Y}] = \mathbb{E}_{x \leftarrow \mathbf{X}}[S(\mathbf{Y}|_{\mathbf{X}=x} \| \mathbf{Y})]$.

In other words, the mutual information between any two random variables $\mathbf{X}$ and $\mathbf{Y}$ is the average relative entropy between the conditional distribution $\mathbf{Y}|_{\mathbf{X}=x}$ and the marginal distribution $\mathbf{Y}$.

Proof of Result 1: We may assume that the random string shared by Alice and Bob is a sequence of independently drawn samples $\langle \mathbf{y}_1, \mathbf{y}_2, \dots \rangle$ according to the marginal distribution $\mathbf{Y}$. On input $x \in \mathcal{X}$ drawn according to the distribution $\mathbf{X}$, Alice uses the sampling procedure $\mathcal{P}$ (from Lemma 1.5) to sample the conditional distribution $\mathbf{Y}|_{\mathbf{X}=x}$ from the marginal distribution $\mathbf{Y}$ in order to generate the index $\mathbf{i}^*$. (Note that the conditional and marginal distribution always satisfy $S(\mathbf{Y}|_{\mathbf{X}=x} \| \mathbf{Y}) < \infty$). Alice transmits the index $\mathbf{i}^*$ to Bob, who then outputs the sample $\mathbf{y}_{i^*}$ which has the required distribution. The expected number of bits transmitted in this protocol is at most $E_{x \leftarrow \mathbf{X}}\left[S(\mathbf{Y}|_{\mathbf{X}=x} \| \mathbf{Y}) + 2\lg(S(\mathbf{Y}|_{\mathbf{X}=x} \| \mathbf{Y}) + 1) + O(1)\right]$ which (by Fact 2.1 and Jensen's inequality) is at most $I[\mathbf{X} : \mathbf{Y}] + 2\lg(I[\mathbf{X} : \mathbf{Y}] + 1) + O(1)$. ∎

3 Proof of the one-shot reverse Shannon theorem (Result 2)

Fix the the channel E, and let $(\mathbf{X}, \mathbf{Y})$ be the random variables that realize its channel capacity. Let Q be the marginal distribution of $\mathbf{Y}$.

Claim 3.1 *For all $x \in \mathcal{X}$, $S(E(x) \| Q) \leq \mathcal{C}(E)$.*

The existence of a distribution Q with the above property was also shown by Jain [Jai] using a different argument.

Note that the result follows immediately from this claim by invoking the rejection sampling lemma (Lemma 1.5). The resulting protocol uses samples drawn according to Q as shared randomness and on input $x \in \mathcal{X}$ generates a symbol in $\mathcal{Y}$ whose distribution is $E(x)$. The communication required is bounded by $\lg S(E(x) \| Q) + 2\lg(S(E(X) \| Q) + 1) + O(1)$; by Claim 3.1, this is at most $\lg \mathcal{C}(E) + 2\lg(\mathcal{C}(E) + 1) + O(1)$.

Proof of Claim 3.1: For contradiction assume that for some $x_0 \in \mathcal{X}$, we have $S(E(x_0)\|Q) > \mathcal{C}(E)$. We will show that by assigning greater probability to x_0 than it receives in $\mathbf{X}$, we can obtain a pair of random variables $(\mathbf{X}', \mathbf{Y}')$ whose distribution is compatible with the channel, but whose mutual information is strictly more than $\mathcal{C}(E)$—a contradiction. Formally, for $\alpha \in [0,1]$, consider a new random variable $\mathbf{X}_\alpha$ obtained by picking x_0 with probability α and $\mathbf{X}$ with probability $(1-\alpha)$. Let $\mathbf{Y}_\alpha$ be a random variable correlated with $\mathbf{X}_\alpha$ such that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\Pr[\mathbf{Y}_\alpha = y \mid \mathbf{X}_\alpha = x] = E(x)(y)$; let Q_α be the marginal distribution of $\mathbf{Y}_\alpha$. A direct calculation yields:

$$\begin{aligned} I[\mathbf{X}_\alpha : \mathbf{Y}_\alpha] - I[\mathbf{X} : \mathbf{Y}] &= \alpha(S(E(x)\|Q) - I[\mathbf{X} : \mathbf{Y}]) + (1-\alpha)S(Q\|Q_\alpha) + \alpha \sum_{y \in \mathcal{Y}} E(x)(y) \lg \frac{Q(y)}{Q_\alpha(y)} \\ &= \alpha(S(E(x)\|Q) - I[\mathbf{X} : \mathbf{Y}]) + (1-\alpha)S(Q\|Q_\alpha) + O(\alpha^2). \end{aligned}$$

Since, $S(E(x)\|Q) - I[\mathbf{X} : \mathbf{Y}] > 0$, for some small enough $\alpha > 0$, we have $I[\mathbf{X}_\alpha : \mathbf{Y}_\alpha] > I[\mathbf{X} : \mathbf{Y}] = \mathcal{C}(E)$—a contradiction. ∎

This completes the proof of Result 2. ∎

4 The rejection sampling procedure

Let P and Q be two distributions on the set $\mathcal{X}$ such that the relative entropy $S(P\|Q)$ is finite. Recall that we need to design a rejection sampling procedure that on input, a sequence of samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \}$ independently drawn from the distribution Q, outputs an index $\mathbf{i}^*$ such that $\mathbf{x}_{\mathbf{i}^*}$ is distributed according to P and the expected encoding length of the index $\mathbf{i}^*$ is as small as possible.

A natural approach to this would be the greedy method: at each iteration, we fill the distribution P with the best possible sub-distribution of Q, while maintaining that our sum is always less then the distribution P (note that since we are doing rejection sampling, we can only create sub-distributions of Q at each iteration). This greedy approach will generate the required distribution P, but it is not guaranteed to perform well with respect to expected index length. For instance, suppose there exists a $x^* \in \mathcal{X}$ such that $P(x^*) > Q(x^*)$ and are both very small while for all other $x \in \mathcal{X}$ we have $Q(x) \gg 0$ and $P(x) \leq Q(x)$. Then, with high probability, the first sample is unlikely to be x^* (since $Q(x^*)$ is small), while at the same time the first sample suffices to satisfy the probability requirements of all values $x \in \mathcal{X}$ but x^*. We would then have to wait for at least $1/Q(x^*)$ samples, on average, to see the value x^* before fulfilling its probability requirement. Thus, the average length of the index of the greedy method can be as large as $P(x^*)\lg(1/Q(x^*))$ which can be much larger than the relative entropy $S(P\|Q)$.

We show that a variant of the greedy algorithm overcomes this problem and in fact achieves expected index length roughly $S(P\|Q)$. This variant works in several phases and within each phase, the algorithm proceeds greedily to fill the distribution $P/2$ instead of P. Filling $P/2$ instead of P in each phase guarantees that there is significant probability (at least $1/2$ in this case) of seeing samples with low probability (like x^* in the above example). The factor 2 is arbitrary and we could have as well worked with any other constant bounded away from 1. In the following lemma, we describe the behavior of the algorithm within each phase.

Lemma 4.1 *Let P and Q be two distributions on the set $\mathcal{X}$ such that their relative entropy $S(P\|Q)$ is finite. Then there exists is a procedure $\widetilde{\mathcal{P}}$ with the following properties:*

Random Input: *A random string* $\mathbf{R}$ *which is a sequence* $\langle \mathbf{x}_1, \mathbf{x}_2, \ldots \rangle$*, of independent samples each with distribution* Q*;*

Output: *The procedure either aborts (which happens with probability* $1/2$*) or outputs an index* $\mathbf{j}^*$ *that satisfies* $\Pr[\mathbf{x}_{\mathbf{j}^*} = x] = P(x)/2$ *for all* $x \in \mathcal{X}$*. Here the probability is taken over the random string* $\mathbf{R}$ *and the internal random coins of the procedure.*

Expected length: *The expected length of the index* $\mathbf{j}^*$*, conditioned on the fact that the procedure* $\widetilde{\mathcal{P}}$ *does not abort, is at most*

$$S(P\|Q) + 2\lg(S(P\|Q) + 1) + O(1).$$

We defer the construction of $\widetilde{\mathcal{P}}$ to the latter part of this section and first show how the procedure $\mathcal{P}$ claimed in Lemma 1.5 can be constructed given the procedure $\widetilde{\mathcal{P}}$ specified in Lemma 4.1.

Proof of Lemma 1.5: For notational convenience, we will assume that the sequence of samples for $\mathcal{P}$ is indexed by a pair of indices (i, j) instead of a single index i. $\mathcal{P}$ repeatedly invokes $\widetilde{\mathcal{P}}$ in each phase till $\widetilde{\mathcal{P}}$ does not abort in which case $\mathcal{P}$ outputs both the phase number and the output of $\widetilde{\mathcal{P}}$. More formally, we have:

Rejection Sampling Procedure $\mathcal{P}(P, Q)$

Random Input: $\{\mathbf{x}_{i,j} | i, j \in \mathbb{N}\}$ a sequence of samples independently drawn from the distribution Q.

1. For $i \leftarrow 1$ to ∞ do

   Phase i

   (a) Run procedure $\widetilde{\mathcal{P}}$ on the subsequence of samples $\{\mathbf{x}_{i,j} | j \in \mathbb{N}\}$.

   (b) If $\widetilde{\mathcal{P}}$ does not abort, set $\mathbf{i}^* = (i, j)$ where j is the output of $\widetilde{\mathcal{P}}$ and go to Step 2
2. Output $\mathbf{i}^*$

Clearly, the sample $\mathbf{x}_{\mathbf{i}^*}$ is distributed according to the distribution P. Since $\widetilde{\mathcal{P}}$ aborts with probability exactly $1/2$, $\mathcal{P}$ invokes $\widetilde{\mathcal{P}}$ twice on average. The expected length of the index $\mathbf{i}^*$ is the expected length the index j and the expected length of the phase number i, which is at most a constant. This completes the proof of Lemma 1.5.

It remains to justify Lemma 4.1.

Proof of Lemma 4.1: The idea is as follows. The procedure $\widetilde{\mathcal{P}}$ will examine the samples $\langle \mathbf{x}_j : j \in \mathbb{N} \rangle$ sequentially; after examining $\mathbf{x}_j$ it either accepts (by returning the value j for $\mathbf{j}^*$) or moves on to the next sample $\mathbf{x}_{j+1}$. The key is to to assign acceptance probabilities for each step so that $\mathbf{x}_{\mathbf{j}^*}$ has the right distribution. These acceptance probabilities are given by a function $a_j : \mathcal{X} \to [0, 1]$, with the natural interpretation that when the procedure examines $\mathbf{x}_j$ and finds that its value is z, then it accepts it with probability $a_j(z)$. We now define $\langle a_j : j \in \mathbb{N} \rangle$ precisely, then show how the procedure uses them. It will be convenient to inductively define two other sequences along with a_j:

$\langle s_j : j = 0, 1, \ldots \rangle$, where each $s_j \in [0, 1]$; later we will show that s_j is the probability that the procedure halts before it examines the sample $\mathbf{x}_{j+1}$;

$\langle p_j : j = 0, 1, \ldots \rangle$, where $p_j : \mathcal{X} \to [0, 1]$; it will turn out that $p_j(z)$ is the probability that $\mathbf{x}_{\mathbf{j}^*} = z$ and $\mathbf{j}^* > j$.

The three sequences are defined as follows.

Definition 4.2 (acceptance probabilities) *Let P and Q be distributions on $\mathcal{X}$ such that $S(P\|Q) < \infty$.*

1. *$L = \max_{x \in \mathcal{X}} \left\lceil \frac{P(x)}{Q(x)} \right\rceil$, $s_0 = 0$, and $p_0(x) = P(x)/2$ for all $x \in \mathcal{X}$.*
2. *For $j \in \{1, \ldots, L\}$, let*
   - $a_j(x) = \min\left(1, \frac{p_{j-1}(x)}{(1-s_{j-1}) \cdot Q(x)}\right)$,
   - $p_j(x) = p_{j-1}(x) - (1 - s_{j-1}) \cdot Q(x) \cdot a^j(x)$, *and*
   - $s_j = \frac{1}{2} - \sum_{x \in \mathcal{X}} p_j(x)$

We now turn to the construction of $\widetilde{\mathcal{P}}$. As mentioned before $\widetilde{\mathcal{P}}$ tries to greedily fill the sub-distribution $P/2$ and aborts with probability $1/2$.

PROCEDURE $\widetilde{\mathcal{P}}(P, Q)$

RANDOM INPUT: $\langle \mathbf{x}_j : j \in \mathbb{N} \rangle$ a sequence of samples independently drawn from the distribution Q.

1. Compute L and the sequence $\langle a_j : j \in \mathbb{N} \rangle$ as given in Definition 4.2
2. For $j \leftarrow 1$ to L do

   ITERATION (j)

   (a) Examine sample $\mathbf{x}_j$,

   (b) With probability $a_j(\mathbf{x}_j)$, output j and halt.
3. Abort (this happens if the procedure does not accept in any of the L iterations).

The following claim relates the quantities defined in Definition 4.2 to the halting probabilities of $\widetilde{\mathcal{P}}$.

Claim 4.3 *For every $j \in \{0, \ldots, L\}$,*

(a) the probability that $\widetilde{\mathcal{P}}$ halts within j iterations is exactly s_j, which is at most $1/2$ for all j;

(b) for each $x \in \mathcal{X}$, the probability that $\widetilde{\mathcal{P}}$ halts within j iterations and outputs $\mathbf{j}^*$ such that $\mathbf{x}_{\mathbf{j}^*} = x$ is exactly $P(x)/2 - p_j(x)$.

Proof: We will prove the two parts simultaneously by induction on j. At the very beginning (i.e., end of iteration 0), we have $s_0 = 0$ and $p_0(x) = P(x)/2$. Thus, the claim holds at the end of iteration 0.

Suppose the claim holds at the end of j iterations. Then the probability that $\widetilde{\mathcal{P}}$ halts within $(j+1)$ iterations and outputs $\mathbf{j}^*$ such that $x_{\mathbf{j}^*} = x$ is exactly $P(x)/2 - p_j(x) + (1-s_j) \cdot Q(x) \cdot a_j(x) = P(x)/2 - p_{j+1}(x)$. This shows part (b); part (a) then follows immediately from part (b). ∎

For each $x \in \mathcal{X}$, the $p_j(x)$'s decreases monotonically from $P(x)/2$ to 0 and does not change after it has attained the value 0. The following claim describes the rate at which this sequence falls.

Claim 4.4 *For each $x \in \mathcal{X}$ and $j \in \{1, \ldots, L\}$, either $p_j(x) = 0$ or $p_j(x) \leq p_{j-1}(x) - Q(x)/2$.*

Proof: Fix some j and x. We have that $p_j(x) = p_{j-1}(x) - \min\left((1 - s_{j-1}) \cdot Q(x), p_{j-1}(x)\right)$. If the minimum is $p_{j-1}(x)$, then $p_j(x) = 0$. Otherwise, the decrease is $(1 - s^{j-1}) \cdot Q(x)$ which is at least $Q(x)/2$. ∎

Thus, by the end of $j = \left\lceil \frac{P(x)}{Q(x)} \right\rceil \leq L$ iterations, $p_j(x) = 0$. This implies that $s_L = 1/2$. It then follows from Claims 4.3,and 4.4 that $\widetilde{\mathcal{P}}$ either aborts (which happens with probability 1/2) or outputs an index j that satisfies $\Pr[\mathbf{x}_j = x] = P(x)/2$ for all $x \in \mathcal{X}$.

We now bound the expected encoding length of the index j conditioned on $\widetilde{\mathcal{P}}$ not aborting. Suppose the index j output by $\widetilde{\mathcal{P}}$ satisfies $\mathbf{x}_j = x$ for some $x \in \mathcal{X}$. It follows from Claim 4.4 that $j \leq \left\lceil \frac{P(x)}{Q(x)} \right\rceil$. However this event happens with probability exactly $P(x)$. It then follows that

$$\mathbb{E}[\ell(j)] \quad \leq \quad \sum_{x \in \mathcal{X}} P(x) \cdot l\left(\left\lceil \frac{P(x)}{Q(x)} \right\rceil\right)$$

We now fix some prefix-free encoding of the integers such that the encoding of every integer $n > 2$ requires no more than $\ell(n) = \lg n + 2 \lg \lg n + O(1)$ bits. Let $\mathcal{X}_{(P>2Q)}$ be the set of $x \in \mathcal{X}$ that satisfy $P(x) > 2Q(x)$. It then follows that

$$
\begin{aligned}
\mathbb{E}[\ell(j)] &\leq \sum_{x\in\mathcal{X}_{(P>2Q)}} P(x)\cdot\left(\lg\left\lceil\frac{P(x)}{Q(x)}\right\rceil + 2\lg\lg\left\lceil\frac{P(x)}{Q(x)}\right\rceil + O(1)\right) + \sum_{x\notin\mathcal{X}_{(P>2Q)}} P(x)\cdot O(1)\\
&\leq \sum_{x\in\mathcal{X}_{(P>2Q)}} P(x)\cdot\lg\frac{2P(x)}{Q(x)} + 2\sum_{x\in\mathcal{X}_{(P>2Q)}} P(x)\lg\lg\frac{2P(x)}{Q(x)} + O(1)\\
&= \sum_{x\in\mathcal{X}_{(P>2Q)}} P(x)\cdot\lg\frac{P(x)}{Q(x)} + 2\sum_{x\in\mathcal{X}_{(P>2Q)}} P(x)\lg\lg\frac{P(x)}{Q(x)} + O(1)\\
&\leq \sum_{x\in\mathcal{X}_{(P>2Q)}} P(x)\cdot\lg\frac{P(x)}{Q(x)} + 2\lg\left(1+\sum_{x\in\mathcal{X}_{(P>2Q)}} P(x)\lg\frac{P(x)}{Q(x)}\right) + O(1)\\
&\qquad\qquad\qquad

We immediately have the following relationship between $\mathsf{IC}^{\mu,k}_{\varepsilon}$ and $D^{\mu,k}_{\varepsilon}$.

Proposition 5.2 *Let μ be a product distribution on $\mathcal{X} \times \mathcal{Y}$. Then, $\mathsf{IC}^{\mu,k}_{\varepsilon}(f) \leq D^{\mu,k}_{\varepsilon}(f)$.*

Proof: Let Π be a protocol whose communication is $c \stackrel{\Delta}{=} D^{\mu,k}_{\varepsilon}(f)$. Let $\mathbf{M}$ denote the message transcript of Π. Then we have, $c \geq H(\mathbf{M}) \geq I(\mathbf{XY} : \mathbf{M}) \geq \mathsf{IC}^{\mu,k}_{\varepsilon}(f)$. ∎

A key insight of Chakrabarti *et al.* [CSWY] was that one could show (approximately) a relationship in the opposite direction when the inputs are being drawn from the uniform distribution. They showed this for SMP protocols using a kind of message compression. Their result was then extended using different techniques involving the (classical) substate theorem (Theorem 1.10) by Jain *et al.* [JRSb, JRSd]. Using this they showed that messages could be compressed to the amount of information they are carry about the inputs, under all distributions for one-way and SMP protocols and under product distributions for two-way protocols. These message compression results then lead to corresponding direct sum results. Using the rejection sampling lemma (Lemma 1.5), we can considerably strengthen the result of Jain *et al.* [JRSb] for two-way protocols as follows. The dependence on k, the number of rounds, in their result was much worse as mentioned in the Introduction section.

Lemma 5.3 *Let $\varepsilon, \delta > 0$. Let μ be a distribution (not necessarily product) on the $\mathcal{X} \times \mathcal{Y}$ and $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. Then,*

$$D^{\mu,k}_{\varepsilon+\delta}(f) \leq \frac{1}{\delta}\left[2 \cdot \mathsf{IC}^{\mu,k}_{\varepsilon}(f) + O(k)\right].$$

The second ingredient in our proof of Theorem 3 is the direct sum property of information cost, originally observed by Chakrabarti *et al.* [CSWY] for the uniform distribution.

Lemma 5.4 *Let μ be a product distribution on $\mathcal{X} \times \mathcal{Y}$. Then, $\mathsf{IC}^{\mu^t,k}_{\varepsilon}(f^{\oplus t}) \geq t \cdot \mathsf{IC}^{\mu,k}_{\varepsilon}(f)$.*

Before proving these lemmas, let us show that they immediately imply our theorem.

Proof of Theorem 3: Let μ be a product distribution on $\mathcal{X} \times \mathcal{Y}$. Then we have

$$D^{\mu^t,k}_{\varepsilon}(f^{\oplus t}) \geq \mathsf{IC}^{\mu^t,k}_{\varepsilon}(f^{\oplus t}) \geq t \cdot \mathsf{IC}^{\mu,k}_{\varepsilon}(f) \geq \frac{t}{2}\left(\delta D^{\mu,k}_{\varepsilon+\delta}(f) - O(k)\right),$$

where the first inequality follows from Proposition 5.2, the second from Lemma 5.4 and the last from Lemma 5.3. ∎

Proof of Lemma 5.3: Let μ be a distribution on $\mathcal{X} \times \mathcal{Y}$. Fix a private-coins protocol Π that achieves the optimum information cost $\mathsf{IC}^{\mu,k}_{\varepsilon}(f)$. Let $(\mathbf{X}, \mathbf{Y})$ be the random variables representing the inputs of Alice and Bob distributed according to μ. We will use the following notation: $\mathbf{M} = \mathbf{M}(\mathbf{X}, \mathbf{Y})$ will be the transcript of the protocol; for $i = 1, 2, \ldots, k$, $\mathbf{M}_i$ will denote the i-th message of the transcript $\mathbf{M}$ and $\mathbf{M}_{1,i}$ will denote the first i messages in $\mathbf{M}$. Now, we have from the *chain rule* for mutual information (cf. [CT]).

$$I[\mathbf{XY} : \mathbf{M}] = \sum_{i=1}^{k} I[\mathbf{XY} : \mathbf{M}_i \mid \mathbf{M}_{1,i-1}]. \tag{2}$$

We now construct another protocol Π' as follows. The idea is as follows. For $i = 1, 2, \ldots, k$, the party that sent $\mathbf{M}_i$ in Π, will now instead use Result 1 to generate the message $\mathbf{M}_i$ for the other party by sending about $I[\mathbf{XY} : \mathbf{M}_i \mid \mathbf{M}_{1,i-1}]$ bits on the average. Suppose, we have managed to generate the first $i-1$ messages in Π' with distribution exactly as that of $\mathbf{M}_{1,i-1}$, and the (partial) transcript so far is m. For the rest of this paragraph we condition on $\mathbf{M}_{1,i-1} = m$, and describe how the next message is to be generated. Assume that it is Alice's turn to send the next message. We have two observations concerning the distributions involved. First, the prefix m of the transcript has already been generated and hence both parties can condition on this information. In particular, the conditional distribution $(\mathbf{M}_i \mid \mathbf{M}_{1,i-1} = m)$ is known to both Alice and Bob and (pre-generated) samples from it can be used as shared randomness. Second, since Π is a private-coins protocol, for each $x \in \mathcal{X}$, the conditional random variable $(\mathbf{M}_i(x, \mathbf{Y}) \mid \mathbf{M}_{1,i-1}(x, \mathbf{Y}) = m)$, is independent of $\mathbf{Y}$. Hence on input x, Alice knows the distribution of $(\mathbf{M}_i(x, \mathbf{Y}) \mid \mathbf{M}_{1,i-1}(x, \mathbf{Y}) = m)$.

The second observation in particular implies (using chain rule for information),

$$I[\mathbf{XY} : \mathbf{M}_i \mid \mathbf{M}_{1,i-1} = m] = I[\mathbf{X} : \mathbf{M}_i \mid \mathbf{M}_{1,i-1} = m].$$

Thus, by Theorem 1, Alice can arrange for $(\mathbf{M}_i | \mathbf{M}_{1,i-1} = m)$ to be generated on Bob's side by sending at most

$$2I[\mathbf{X} : \mathbf{M}_i \mid \mathbf{M}_{1,i-1} = m] + O(1)$$

bits on the average; the overall communication in the i-th round is the average of this quantity over all choices m, that is, at most

$$2I[\mathbf{XY} : \mathbf{M}_i \mid \mathbf{M}_{1,i-1}] + O(1).$$

By applying this strategy for all rounds, we note from (2) that we obtain a public-coins k-round protocol Π', with expected communication $2I[\mathbf{XY} : \mathbf{M}] + O(k)$ bits, and error at most ε as in Π. Using Markov's inequality, we conclude that the number of bits sent by the protocol is at least $\frac{1}{\delta}$ times this quantity with probability at most δ. By truncating the long runs and then fixing the private random sequences suitably, we obtain a deterministic protocol Π'' with error at most $\varepsilon + \delta$ and communication at most $\frac{1}{\delta}(2I[\mathbf{XY} : \mathbf{M}] + O(k)) = \frac{1}{\delta}(2 \cdot \mathsf{IC}_{\varepsilon}^{\mu,k}(f) + O(k))$. The lemma now follows from this and definition of $D_{\varepsilon+\delta}^{\mu,k}(f)$. ∎

Proof of Lemma 5.4: Let μ be a product distribution on $\mathcal{X} \times \mathcal{Y}$. Fix a k-round private-coins protocol Π for $f^{\oplus t}$ that achieves $\mathsf{IC}_{\varepsilon}^{\mu^t,k}(f^{\oplus t})$. For this protocol Π the input is chosen according to μ^t. We denote this input by $(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_t, \mathbf{Y}_1\mathbf{Y}_2 \cdots \mathbf{Y}_t)$ and note that the $2t$ random variables involved are mutually independent. Let $\mathbf{M}$ denote the transcript of this protocol when run the input $(\mathbf{X}, \mathbf{Y})$. Now, we have from chain rule for mutual information and independence of the $2t$ random variables as above,

$$\mathsf{IC}_{\varepsilon}^{\mu^k,k}(f) = I[\mathbf{XY} : M] \geq \sum_{i=1}^{t} I[\mathbf{X}_i\mathbf{Y}_i : M].$$

We claim that each term in the sum of the form $I[\mathbf{X}_i\mathbf{Y}_i : M]$ is at least $\mathsf{IC}_{\varepsilon}^{\mu,k}$. Indeed, consider the following protocol Π' for f derived from Π. In Π', on receiving the input $(x, y) \in \mathcal{X} \times \mathcal{Y}$, Alice and Bob simulate Π as follows. They insert x and y as the i-th component of their respective

inputs for Π, and generate the remaining components based on the product distribution μ. They can do so using private coins since μ is a product distribution. This results in a k-round private coins protocol Π' for f with error at most ε under μ, since the error of Π was at most ε under μ^k. Clearly, $\mathsf{IC}^{\mu}(\Pi) = I[\mathbf{X}_i\mathbf{Y}_i : \mathbf{M}]$. ∎

6 Separating $T(\mathbf{X} : \mathbf{Y})$, $C(\mathbf{X} : \mathbf{Y})$ and $I[\mathbf{X}, \mathbf{Y}]$

For any pair of random variables $(\mathbf{X}, \mathbf{Y})$, it easily follows from the definitions that $T(\mathbf{X} : \mathbf{Y}) \geq C(\mathbf{X} : \mathbf{Y})$. Furthermore, by Wyner's theorem (Theorem 1.1)

$$C(\mathbf{X} : \mathbf{Y}) = \min_{\mathbf{W}} I[\mathbf{XY} : \mathbf{W}],$$

where $\mathbf{W}$ is such that $\mathbf{X}$ and $\mathbf{Y}$ are independent when conditioned on $\mathbf{W}$. Note, however, that

$$I[\mathbf{XY} : \mathbf{W}] =\geq I[\mathbf{X} : \mathbf{W}] \geq I[\mathbf{X} : \mathbf{Y}].$$

The last inequality is the data-proceesing inequality applied to the Markov chain $\mathbf{X} \to \mathbf{W} \to \mathbf{Y}$. Thus, we have $T(\mathbf{X} : \mathbf{Y}) \geq C(\mathbf{X} : \mathbf{Y}) \geq I[\mathbf{X} : \mathbf{Y}]$. In this section, we will show that both these inequalities are strict for $(\mathbf{X}, \mathbf{Y})$ defined as follows.

Definition 6.1 *Let $\mathbf{W} = (\mathbf{i}, \mathbf{b})$ be a random variable uniformly distributed over the set $[n] \times \{0, 1\}$. Now, let $\mathbf{X}$ and $\mathbf{Y}$ be random variables taking values in $\{0, 1\}^n$, such that*

(a) $\Pr[\mathbf{X} = z \mid \mathbf{W} = (i, b)], \Pr[\mathbf{Y} = z \mid \mathbf{W} = (i, b)] = \begin{cases} 2^{-(n-1)} & z[i] = b \\ 0 & \text{otherwise} \end{cases}$.

(b) $\mathbf{X}$ and $\mathbf{Y}$ are independent when conditioned on $\mathbf{W}$.

Proposition 6.2 *For $(\mathbf{X}, \mathbf{Y})$ defined as above, we have:*

(a) $I[\mathbf{X} : \mathbf{Y}] = O\left(n^{-\frac{1}{3}}\right)$.

(b) $C(\mathbf{X} : \mathbf{Y}) = 2 - I[X : Y] = 2 - O\left(n^{-\frac{1}{3}}\right)$.

(c) $T(\mathbf{X} : \mathbf{Y}) = \Theta(\lg n)$.

Note that in the above example, though $C(\mathbf{X} : \mathbf{Y})$ and $I[\mathbf{X} : \mathbf{Y}]$ differ by a super-constant multiplicative factor, they only differ by a constant additive factor. We can construct another joint distribution $(\mathbf{X}', \mathbf{Y}')$ by taking m-wise independent copies of the joint distribution $(\mathbf{X}, \mathbf{Y})$ (i.e., $(\mathbf{X}', \mathbf{Y}') = (\mathbf{X}, \mathbf{Y})^m$). We then have $I[\mathbf{X}' : \mathbf{Y}'] = I[\mathbf{X}^m : \mathbf{Y}^m] = mI[\mathbf{X} : \mathbf{Y}] = o(m)$ while $C(\mathbf{X}' : \mathbf{Y}') = C(\mathbf{X}^m : \mathbf{Y}^m) = mC(\mathbf{X} : \mathbf{Y}) = \Theta(m)$[3]. This implies, that $C(\mathbf{X}' : \mathbf{Y}')$ and $I[\mathbf{X}' : \mathbf{Y}']$ differ by a super-constant factor both multiplicatively as well as additively.

[3] $C(\mathbf{X}^m : \mathbf{Y}^m) = \liminf_{\lambda \to 0} \lim_{n \to \infty} (T(\mathbf{X}^{mn} : \mathbf{Y}^{mn})/n) = m \cdot \liminf_{\lambda \to 0} \lim_{n \to \infty} (T(\mathbf{X}^{mn} : \mathbf{Y}^{mn})/mn) = mC(\mathbf{X} : \mathbf{Y})$ where the first and third equalities follow from Eq. (1)

Proof of part (a): Given $\mathbf{X} = x$ for some n-bit string x, the conditional distribution $\mathbf{Y}|_{\mathbf{X}=x}$ is given by

$$\Pr[\mathbf{Y} = y \mid \mathbf{X} = x] = \frac{\mathrm{agr}(x, y)}{n2^{n-1}}$$

where $\mathrm{agr}(x, y)$ is the number of bit positions x and y agree on. We can now compute the conditional entropy $H[\mathbf{Y} \mid \mathbf{X}]$ as follows:

$$\begin{aligned}
H[\mathbf{Y} \mid \mathbf{X}] &= -\sum_{x\in\{0,1\}^n} \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \frac{2k}{n2^n} \lg \frac{2k}{n2^n} \\
&= -\sum_{k=0}^{n} \binom{n}{k} \frac{k}{n2^{n-1}} \lg \frac{k}{n2^{n-1}} \\
&= -\sum_{k=1}^{n} \binom{n-1}{k-1} \frac{1}{2^{n-1}} \lg \frac{k}{n2^{n-1}} \\
&= -\sum_{k=0}^{n-1} \binom{n-1}{k} \frac{1}{2^{n-1}} \left[\lg \frac{k+1}{n} - (n-1)\right] \\
&= (n-1) - \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{1}{2^{n-1}} \lg \frac{k+1}{n} \\
&= n + \lg n - 1 - \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{1}{2^{n-1}} \lg(k+1) \\
&\geq n + \lg n - 1 - \left(1 - 2^{-O(n^{1/3})}\right) \cdot \lg \left[\frac{n}{2}\left(1 + \frac{1}{n^{1/3}}\right)\right] - 2^{-O(n^{1/3})} \cdot \lg n \\
&= n + \lg n - 1 - \left(1 - 2^{-O(n^{1/3})}\right) \cdot \left(\lg n - 1 + \frac{\lg e}{n^{1/3}}\right) - 2^{-O(n^{1/3})} \cdot \lg n \\
&\qquad\qquad [\text{since } \lg(1+\delta) \leq \delta \lg e] \\
&= n - O\left(\frac{1}{n^{1/3}}\right)
\end{aligned}$$

Thus, $I[\mathbf{X} : \mathbf{Y}] = H[\mathbf{Y}] - H[\mathbf{Y} \mid \mathbf{X}] = O(n^{-\frac{1}{3}})$. ∎

Proof of part (b): By Wyner's theorem (Theorem 1.1),

$$\begin{aligned}
C(\mathbf{X} : \mathbf{Y}) &= \min_{\mathbf{W}'} I[\mathbf{XY} : \mathbf{W}'] \\
&= H[\mathbf{XY}] - \max_{\mathbf{W}'} H[\mathbf{XY} \mid \mathbf{W}'] \\
&= H[\mathbf{X}] + H[\mathbf{Y}] - I[\mathbf{X} : \mathbf{Y}] - \max_{\mathbf{W}'} H[\mathbf{XY} \mid \mathbf{W}'] \\
&= 2n - I[\mathbf{X} : \mathbf{Y}] - \max_{\mathbf{W}'} H[\mathbf{XY} \mid \mathbf{W}'].
\end{aligned}$$

where the random variable $\mathbf{W}'$ is such that $I[\mathbf{X} : \mathbf{Y} \mid \mathbf{W}'] = 0$. We already know that $I[\mathbf{X} : \mathbf{Y}] = O\left(n^{-\frac{1}{3}}\right)$. So, part (b) will follow if we show

$$\max_{\mathbf{W}'} H[\mathbf{XY} \mid \mathbf{W}'] = 2n - 2. \tag{3}$$

Let $\mathbf{W}'$ be such that $I[\mathbf{X} : \mathbf{Y} \mid \mathbf{W}'] = 0$. Consider any w in the support of $\mathbf{W}'$. Let X_w be the set of $x \in \{0,1\}^n$ such that $\Pr[\mathbf{X} = x \mid \mathbf{W}' = w] > 0$. Similarly, define Y_w. We must have that $|X_w| + |Y_w| \leq 2^n$, since otherwise there exist an x such that $\Pr[\mathbf{X} = x \wedge \mathbf{Y} = \bar{x}] > 0$ where $\bar{x}$ is the n-bit string obtained by complementing each bit of x. This implies that $|X_w \times Y_w| \leq 2^{2n}/4$. Thus,

$$\max_{\mathbf{W}'} H[\mathbf{XY} \mid \mathbf{W}'] \leq 2n - 2.$$

Now, if we $\mathbf{W}'$ is the random variable $\mathbf{W}$ used in Definition 6.1, we have $H[\mathbf{XY} \mid \mathbf{W}] = 2(n-1)$. Hence,

$$\max_{\mathbf{W}'} H[\mathbf{XY} \mid \mathbf{W}'] \geq 2n - 2.$$

This justifies (3) and completes the proof of part (b). ∎

To prove part (c), we will use a theorem of Harper [Har], which states that Hamming balls in the hypercube have the smallest boundary. The following version, due to Frankl and Füredi (see Bollobás [Bol, Theorem 3, page 127]), will be the most convenient for us. First, we need some notation.

Notation. For $x, y \in \{0,1\}^n$, let $d(x,y)$ be the Hamming distance between x and y, that is, the number of positions where x and y differ. For non-empty subsets $\mathcal{A}, \mathcal{B} \subseteq \{0,1\}^n$, let

$$d(\mathcal{A}, \mathcal{B}) \stackrel{\Delta}{=} \min\{d(a,b) : a \in \mathcal{A} \text{ and } b \in \mathcal{B}\}.$$

We say that a subset $S \subseteq \{0,1\}^n$ is a Hamming ball centered at $x \in \{0,1\}^n$ if for all $y, y' \in \{0,1\}^n$, if $y \in S$ and $d(x,y') < d(x,y)$, then $y' \in S$. Let

$$\mathsf{Ball}(x,d) = \{y \in \{0,1\}^n : d(x,y) \leq d\}.$$

Theorem 6.3 *Let $\mathcal{A}$ and $\mathcal{B}$ be non-empty subsets of $\{0,1\}^n$. Then, we can find Hamming balls $\mathcal{A}_0$ and $\mathcal{B}_0$ centered at 0^n and 1^n respectively, such that $|\mathcal{A}_0| = |\mathcal{A}|$, $|\mathcal{B}_0| = |\mathcal{B}|$, and $d(\mathcal{A}_0, \mathcal{B}_0) \geq d(\mathcal{A}, \mathcal{B})$.*

Corollary 6.4 *If $\mathcal{A}$ and $\mathcal{B}$ are non-empty sets of strings such that $d(\mathcal{A}, \mathcal{B}) \geq d \geq 2$, then*

$$\min\{|\mathcal{A}|, |\mathcal{B}|\} \leq \exp\left(-\frac{(d-2)^2}{4}\right) 2^n.$$

Proof: By Theorem 6.3, we may assume that $\mathcal{A}$ and $\mathcal{B}$ are balls centered at 0^n and 1^n. Suppose $|\mathcal{A}| \leq |\mathcal{B}|$, and let r be a non-negative integer such that

$$\mathsf{Ball}(0^n, r) \subseteq \mathcal{A} \subseteq \mathsf{Ball}(0^n, r+1).$$

Then, $2r + d \leq n$, that is, $r + 1 \leq (n - d + 2)/2$. It then follows using the Chernoff bound (see, e.g., Alon and Spencer[AS, Theorem A.1.1, page 263]) that

$$|\mathcal{A}| \leq |\mathsf{Ball}(0^n, r+1)| \leq \exp\left(-\frac{(d-2)^2}{2n}\right).$$

∎

Proof of part (c): It is easy to see that $T(\mathbf{X} : \mathbf{Y}) \leq \lceil \lg n \rceil + 1$: on receiving $x \in \{0,1\}^n$, Alice sends Bob an index $\mathbf{i}$ uniformly distributed in $[n]$ and the bit $x[\mathbf{i}]$; on receiving (i, b), Bob generates a random string $\mathbf{y} \in \{0,1\}$ such that $\mathbf{y}[i] = b$, with each of the 2^{n-1} possibilities being equally likely.

It remains to show that $T(\mathbf{X} : \mathbf{Y}) = \Omega(\lg n)$. It follows from the definition of $T(\mathbf{X} : \mathbf{Y})$ that $T(\mathbf{X} : \mathbf{Y}) \geq \min_{\mathbf{W}'} H[\mathbf{W}']$, where the minimum is over all random variables $\mathbf{W}'$ such that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{W}'$. Thus, it is enough to show that any such $\mathbf{W}'$ has entropy $\Omega(\lg n)$. Let $\mathbf{W}'$ be one such random variable. We show below that for all w

$$\Pr[\mathbf{W}' = w] = O\left(\sqrt{\frac{\lg n}{n}}\right).$$

That is, we show that the min-entropy of $\mathbf{W}'$ is $\Omega(\lg n)$; it follows that the entropy of $\mathbf{W}'$ is $\Omega(\lg n)$

Fix w such that $\alpha \triangleq \Pr[\mathbf{W}' = w] > 0$. Let

$$\begin{aligned} X_w &= \left\{x \in \{0,1\}^n : \Pr[\mathbf{X} = x \mid \mathbf{W}' = w] > 2^{-(n+1)}\right\}; \\ Y_w &= \left\{y \in \{0,1\}^n : \Pr[\mathbf{X} = y \mid \mathbf{W}' = w] > 2^{-(n+1)}\right\}. \end{aligned}$$

Then, for all $x \in X_w$ and $y \in Y_w$, we have

$$\alpha 2^{-2(n+1)} < \Pr[(\mathbf{X}, \mathbf{Y}) = (x, y) \wedge \mathbf{W}' = w] \leq \Pr[(\mathbf{X}, \mathbf{Y}) = (x, y)] = \frac{\mathrm{agr}(x, y)}{n 2^{2n-1}},$$

that is, $\mathrm{agr}(x, y) > \alpha n/8$. Furthermore, since for all x, $\Pr[\mathbf{X} = x \mid \mathbf{W}' = w] \leq 2^{-n}/\alpha$ and $\sum_{x \in \{0,1\}^n} \Pr[\mathbf{X} = x \mid \mathbf{W}' = w] = 1$, we have $|X_w| \geq \alpha 2^{n-1}$. Similarly $|Y_w| \geq \alpha 2^{n-1}$. We thus obtain two sets $X_w, Y_w \subseteq \{0,1\}^n$, each with at least $\alpha 2^{n-1}$ elements, such that every $x \in X_w$ and $y \in Y_w$ satisfies $\mathrm{agr}(x, y) > \alpha n/8$. Our goal is to show that this implies that α is small.

Let Y'_w be the set of strings whose complements belong to Y_w. Since $\mathrm{agr}(x, y) > \alpha n/8$ for all $x \in X_w$ and $y \in Y_w$, the Hamming distance between X_w and Y'_w is more than $\alpha n/8$. By Corollary 6.4, we conclude that

$$\alpha 2^{n-1} \leq \exp\left(-\frac{(\alpha n - 16)^2}{128}\right) 2^n,$$

which implies that $\alpha \leq 15\sqrt{\frac{\ln n}{n}}$, for all large enough n. ∎

References

[AS] Noga Alon and Joel H Spencer, *The Probabilistic Method*, Wiley-Interscience, second ed., 2000.

[BJKS] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar, *An information statistics approach to data stream and communication complexity*, Journal of Computer and System Sciences, 68 (2004), pp. 702–732. doi:10.1016/j.jcss.2003.11.006.

[BSST] Charles H. Bennett, Peter W. Shor, John A. Smolin, and Ashish V. Thapliyal, *Entanglement-assisted capacity of a quantum channel and the reverse shannon theorem*, IEEE Transactions on Information Theory, 48 (2002), pp. 2637–2655. doi:10.1109/TIT.2002.802612.

[BW] Charles H. Bennett, and Andreas Winter, Personal Communication.

[Bol] Béla Bollobás, *Combinatorics: Set Systems, Hypergraphs, Families of Vectors and Combinatorial Probability*, Cambridge University Press, 1986.

[CR] Amit Chakrabarti and Oded Regev, *An optimal randomised cell probe lower bound for approximate nearest neighbour searching*, in Proc. 45th IEEE Symp. on Foundations of Comp. Science, 2004, pp. 473–482. doi:10.1109/FOCS.2004.12.

[CSWY] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Chi-Chih Yao, *Informational complexity and the direct sum problem for simultaneous message complexity*, in Proc. 42nd IEEE Symp. on Foundations of Comp. Science, 2001, pp. 270–278. doi:10.1109/SFCS.2001.959901.

[CT] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

[Har] Lawrence H. Harper, *Optimal numberings and isoperimetric problems on graphs*, J. Combinatorial Theory, 1 (1966), pp. 385–394.

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

[JRSa] 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 Proc. 43rd IEEE Symp. on Foundations of Comp. Science, 2002, pp. 429–438. doi:10.1109/SFCS.2002.1181967.

[JRSb] ——, *A direct sum theorem in communication complexity via message compression*, in Proc. 30th International Colloquium of Automata, Languages and Programming, Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger, eds., vol. 2719 of Lecture Notes in Computer Science, 2003, Springer-Verlag, pp. 300–315.

[JRSc] ——, *A lower bound for the bounded round quantum communication complexity of set disjointness.*, in Proc. 44th IEEE Symp. on Foundations of Comp. Science, 2003, pp. 220–229. doi:10.1109/SFCS.2003.1238196.

[JRSd] ——, *Prior entanglement, message compression and privacy in quantum communication*, in Proc. 20th IEEE Conference on Computational Complexity, 2005, pp. 285–296. doi:10.1109/CCC.2005.24.

[KN] Eyal Kushilevitz and Noam Nisan, *Communication Complexity*, Cambridge University Press, 1997.

[Win] Andreas Winter, *Compression of sources of probability distributions and density operators*, e-print: quant-ph/0208131, 2002. Available from: `http://arxiv.org/abs/quant-ph/0208131`.

[Wyn] Aaron D Wyner, *The common information of two dependent random variables*, IEEE Transactions on Information Theory, 21 (1975), pp. 163–179.

[Yao] Andrew Chi-Chih Yao, *Probabilistic computations: Toward a unified measure of complexity (extended abstract)*, in Proc. 18th IEEE Symp. on Foundations of Comp. Science, 1977, pp. 222–227.

A Bounding the Negative terms

Claim A.1 *Let P and Q be two distributions on the set $\mathcal{X}$. For any set $\mathcal{X}' \subseteq \mathcal{X}$, we have*

$$\sum_{x \in \mathcal{X}'} P(x) \lg \frac{P(x)}{Q(x)} \geq -\frac{\lg e}{e}.$$

Proof: We require the following facts.

- *log-sum inequality:* For non-negative integers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$\sum a_i \lg \frac{a_i}{b_i} \geq \left(\sum a_i\right) \lg \frac{\sum a_i}{\sum b_i}.$$

- The function $x \lg x \geq -(\lg e)/e$ for all $x > 0$

$$\begin{aligned}
\sum_{x \in \mathcal{X}'} P(x) \lg \frac{P(x)}{Q(x)} &= \sum_{x \in \mathcal{X}'} P(x) \lg \frac{P(x)}{Q(x)} + \sum_{x \notin \mathcal{X}'} Q(x) \lg \frac{Q(x)}{Q(x)} \\
&\geq \left(\sum_{x \in \mathcal{X}'} P(x) + \sum_{x \notin \mathcal{X}'} P(x)\right) \lg \left(\frac{\sum_{x \in \mathcal{X}'} P(x) + \sum_{x \notin \mathcal{X}'} P(x)}{\sum_{x \in \mathcal{X}} Q(x)}\right) \\
&= \left(\frac{\sum_{x \in \mathcal{X}'} P(x) + \sum_{x \notin \mathcal{X}'} P(x)}{\sum_{x \in \mathcal{X}} Q(x)}\right) \lg \left(\frac{\sum_{x \in \mathcal{X}'} P(x) + \sum_{x \notin \mathcal{X}'} P(x)}{\sum_{x \in \mathcal{X}} Q(x)}\right) \\
&\geq -\frac{\lg e}{e}
\end{aligned}$$

∎