The Communication Complexity of Correlation

Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan

***Abstract*—Let $\mathcal{X}$ and $\mathcal{Y}$ be finite non-empty sets and (X,Y) a pair of random variables taking values in $\mathcal{X}\times\mathcal{Y}$. We consider communication protocols between two parties, ALICE and BOB, for generating X and Y. ALICE is provided an $x\in\mathcal{X}$ generated according to the distribution of X, and is required to send a message to BOB in order to enable him to generate $y\in\mathcal{Y}$, whose distribution is the same as that of $Y|_{X=x}$. Both parties have access to a shared random string generated in advance. Let $T[X:Y]$ be the minimum (over all protocols) of the expected number of bits ALICE needs to transmit to achieve this. We show that**

$$I[X:Y]\le T[X:Y]\le I[X:Y]+2\log_2(I[X:Y]+1)+O(1).$$

We also consider the worst-case communication required for this problem, where we seek to minimize the average number of bits ALICE must transmit for the worst-case $x\in\mathcal{X}$. We show that the communication required in this case is related to the capacity $\mathcal{C}(E)$ of the channel E, derived from (X,Y), that maps $x\in\mathcal{X}$ to the distribution of $Y|_{X=x}$. We show that the required communication $T(E)$ satisfies

$$\mathcal{C}(E)\le T(E)\le\mathcal{C}(E)+2\log_2(\mathcal{C}(E)+1)+O(1).$$

Using the first result, we derive a direct sum theorem in communication complexity that substantially improves the previous such result shown by Jain, Radhakrishnan and Sen [*In Proc. 30th International Colloquium of Automata, Languages and Programming (ICALP)*, ser. LNCS, vol. 2719. 2003, pp. 300–315].

These results are obtained by employing a rejection sampling procedure that relates the relative entropy between two distributions to the communication complexity of generating one distribution from the other.

***Index Terms*—mutual information, relative entropy, rejection sampling, communication complexity, direct-sum**

I. INTRODUCTION

LET $\mathcal{X}$ and $\mathcal{Y}$ be finite non-empty sets, and let (X,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 X. (We use the same symbol to refer to a random variable and its distribution.) ALICE needs to transmit a message M to BOB so that BOB can generate a value $y\in\mathcal{Y}$, that is distributed according to the conditional distribution $Y|_{X=x}$ (i.e., the pair (x,y) has joint distribution (X,Y)). How many bits must ALICE send BOB in order to accomplish this? It follows from the data processing inequality in information theory that the minimum expected number of bits of communication, which we shall call $T[X:Y]$, is at least the mutual information $I[X:Y]$ between X and Y, that is,

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

where $H[Z]$ denotes the Shannon entropy of the random variable Z. In this paper, we are interested in deriving an upper bound in terms of $I[X:Y]$ on the expected length of the communication, which can be viewed as a functional characterization of the quantity $I[X:Y]$.

One can also consider a version of this problem that allows error. Formally, let $T_\lambda[X:Y]$ denote the minimum expected number of bits ALICE needs to send BOB in a protocol, such that the joint distribution generated by the protocol, which we call $(X,\Pi(X))$, is within λ of (X,Y) in total variational distance. (The total variational distance between distributions P and Q defined over a set $\mathcal{Z}$ is $\frac{1}{2}\sum_{z\in\mathcal{Z}}|P(z)-Q(z)|$.)

This problem was first studied by Wyner [Wyn75], who considered its asymptotic version (with error), where ALICE is given several independently drawn samples $(x_1,\ldots,x_m)$ from the distribution 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 $(X,Y)^m$. Wyner referred to the amortized minimum expected number of bits ALICE needs to send BOB as the *common information* $C[X:Y]$ of the random variables X and Y, i.e.,

$$C[X:Y]\triangleq\liminf_{\lambda\to0}\left[\lim_{m\to\infty}\frac{T_\lambda[X^m:Y^m]}{m}\right]. \qquad \text{(I.1)}$$

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

Theorem I.1 (Wyner's theorem [Wyn75, Theorem 1.3])**.**

$$C[X:Y]=\min_W I[XY:W],$$

where the minimum is taken over all random variables W such that X and Y are conditionally independent given W (in other words, $X\to W\to Y$ forms a Markov chain).

It can easily be verified (see Section VI) that $T[X:Y]\ge C[X:Y]\ge I[X:Y]$. However, as we show in Section VI, both these inequalities can be very loose; in particular, $T[X:Y]$ cannot be bounded above by any linear function in $I[X:Y]$. Thus, this natural approach does not yield a good functional characterization for $I[X:Y]$ we hoped for.

A preliminary version of this paper appeared in *Proc. 22nd IEEE Conference on Computational Complexity*, 2007 [HJMR07].

Toyota Technological Institute, Chicago, USA. Email: prahladh@tti-c.org.

Centre for Quantum Technologies and Department of Computer Science, National University of Singapore. Email: rahul@comp.nus.edu.sg. Part of the work was done while the author was at the Univ. of California, Berkeley, and at the Univ. of Waterloo.

Toyota Technological Institute, Chicago, USA. Email: mcallester@tti-c.org.

Tata Institute of Fundamental Research, Mumbai, INDIA. Email: jaikumar@tifr.res.in. Part of the work was done while the author was at the Toyota Technological Institute, Chicago.

A. Protocols with shared randomness

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 (shared randomness has recently been found useful in a similar information theoretic setting [BSST02]). In fact, then ALICE need send no more than approximately $I[X:Y]$ 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 I.2 (One-way protocol)**.** *In a one-way protocol, the two parties* ALICE *and* BOB *share a random string R, and also have private random strings R_A and R_B respectively.* ALICE *receives an input $x \in \mathcal{X}$. Based on the shared random string R and her own private random string R_A, she sends a message $M(x, R, R_A)$ to* BOB*. On receiving the message M,* BOB *computes the output $y = y(M, R, R_B)$. The protocol is thus specified by the two functions $M(x, R, R_A)$ and $y(M, R, R_B)$ and the distributions for the random strings R, R_A and 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, $T_\Pi(x) = \mathbb{E}[|M(x, R, R_A)|]$. 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.*

One can also consider protocols with multiple rounds of communication. However, if our goal is only to minimize communication, then one can assume without loss of generality that the protocol is one-way. This is because we can include the random strings R_A and R_B as part of the shared random string R, enabling ALICE to determine BOB's responses to her messages on her own. She can then concatenate all her messages and send them in one round.

Definition I.3. *Given random variables (X, Y), let*

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

where Π ranges over all one-way protocols where $(X, \Pi(X))$ is within λ of (X, Y) in total variation distance. For the special case when $\lambda = 0$, we write $T^R[X:Y]$ instead of $T^R_0[X:Y]$.

Our first result shows that $T^R[X:Y]$ and $I[X:Y]$ are closely related.

Result 1 (Characterization of mutual information)**.**

$$I[X:Y] \le T^R[X:Y] \le I[X:Y] + 2\lg(I[X:Y]+1) + O(1).$$

This result provides a functional characterization of $I[X:Y]$ in terms of the communication needed to generate Y from X in the presence of shared randomness. We have the $2\lg(I[X:Y]+1)$ term in the upper bound because our proof of the result employs a prefix-free encoding of integers that requires $\lg n + 2\lg\lg(n+1) + O(1)$ bits to encode the positive integer n. By using an encoding that requires $\lg n + (1+\varepsilon)\lg\lg(n+1) + O(1)$ 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 (X, Y) for which our proof of Result 1 requires ALICE and BOB to share a random string of unbounded length. However, by stating the question in terms of flows in a suitably defined network, we can bound the amount of shared randomness by $O(\lg\lg|\mathcal{X}| + \lg|\mathcal{Y}|)$ provided we allow the expected communication to increase by $O(\lg\lg(|\mathcal{Y}|))$ (see Section VII).

B. Generating one distribution from another

The main tool in our proof of Result 1 is a sampling procedure that relates the *relative entropy* between two distributions and the communication complexity of generating one distribution from the other.

Definition I.4 (Relative entropy)**.** *The* relative entropy *or* Kullback-Leibler divergence *between two probability distributions P and Q on a finite set $\mathcal{X}$ is*

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

Note 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; also, it is zero iff $P = Q$, but is 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 $\langle x_1, x_2, \ldots, x_i, \ldots\rangle$ be a sequence of samples, drawn independently, each with distribution Q. The idea is to generate an index i^* (a random variable depending on the sample) so that the sample x_{i^*} has distribution P. For example, if P and Q are identical, then we can set $i^* = 1$ and be done. It is easy to show (see Proposition IV.3) that for any such procedure

$$\mathbb{E}[\ell(i^*)] \ge S(P\|Q),$$

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

Lemma I.5 (Rejection sampling lemma)**.** *Let P and Q be two distributions such that $S(P\|Q)$ is finite. There exists a sampling procedure* REJ-SAMPLER *which on input a sequence $\langle x_1, x_2, \ldots, x_i, \ldots\rangle$ of independently drawn samples from the distribution Q outputs (with probability 1) an index i^* such that the sample x_{i^*} is distributed according to the distribution P and the expected encoding length of the index 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 REJ-SAMPLER*. The constant 2 can be reduced to $1+\varepsilon$ for any $\varepsilon > 0$.*

C. Reverse Shannon theorem

In Result 1, we considered the communication cost averaged over $x \in \mathcal{X}$, chosen according to the distribution of 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). 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 $E : \mathcal{X} \to \mathcal{P}_{\mathcal{Y}}$ that associates with each $x \in \mathcal{X}$ a probability distribution $E_x \in \mathcal{P}_{\mathcal{Y}}$. The Shannon capacity of such a channel is

$$\mathcal{C}(E) \triangleq \max_{(X,Y)} I[X : Y],$$

where (X, Y) is a pair of random variables taking values in $\mathcal{X} \times \mathcal{Y}$ such that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\Pr[Y = y \mid X = x] = E_x(y)$. 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

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

where the minimum is taken over all valid simulators Π of E. Our next result shows that $T(E)$ and $\mathcal{C}(E)$ are closely related.

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

As in the case of Result 1, the constant 2 can be reduced to $1+\varepsilon$ for any $\varepsilon > 0$. Such a result is called the Reverse Shannon Theorem as it gives an (optimal) simulation of noisy channels using noiseless channels and shared randomness. We use the prefix *one-shot* to distinguish this result from the previously known asymptotic versions. See Section I-E for a discussion of these results.

D. A direct-sum result in communication complexity

Result 1 has an interesting consequence in communication complexity. To state this result, we need to recall some definitions from two-party communication complexity (see Kushilevitz and Nisan [KN97] for an excellent introduction to the area). 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 $D^{\mu,k}_{\varepsilon}(f)$), 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 $R^{\mathrm{pub},k}_{\varepsilon}(f)$, the public-coin 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-coin 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 [Yao77].

Theorem I.6 (Yao's minmax principle [Yao77])**.** $R^{\mathrm{pub},k}_{\varepsilon}(f) = \max_{\mu} D^{\mu,k}_{\varepsilon}(f)$

For function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ and a positive integer t, let $f^{(t)} : \mathcal{X}^t \times \mathcal{Y}^t \to \mathcal{Z}^t$ be defined by

$$f^{(t)}(\langle x_1, \ldots, x_t\rangle, \langle y_1, \ldots, y_t\rangle) \triangleq \langle f(x_1, y_1), \ldots, f(x_t, y_t)\rangle.$$

It is natural to ask if the communication complexity of $f^{(t)}$ 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 are known for this question in restricted settings for deterministic and randomized protocols [KN97]. Recently Chakrabarti, Shi, Wirth and Yao [CSWY01] studied this question in the *simultaneous message passing* (SMP) model in which ALICE and BOB, instead of communicating with each other, 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. Jain, Radhakrishnan and Sen [JRS05] showed that the claim holds for all functions and relations, not necessarily those satisfying the robustness condition, both in the one-way and the SMP model of communication. In another work Jain, Radhakrishnan and Sen [JRS03a] showed a weaker direct sum result for bounded-round two-way protocols under product distributions over the inputs. Their result was the following (here μ is a product distribution on $\mathcal{X} \times \mathcal{Y}$ and k represents the number of rounds):

$$\forall \delta > 0, \quad D^{\mu^t,k}_{\varepsilon}(f^{(t)}) \geq t\left(\frac{\delta^2}{2k} \cdot D^{\mu,k}_{\varepsilon+2\delta}(f) - 2\right)$$

We show that Result 1 implies the following stronger claim.

Result 3 (Direct sum for communication complexity)**.** *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*

$$\forall \delta > 0, \quad D^{\mu^t,k}_{\varepsilon}(f^{(t)}) \geq \frac{t}{2}\left(\delta D^{\mu,k}_{\varepsilon+\delta}(f) - O(k)\right).$$

that $(x, y, z) \in F$.

By applying Yao's minimax principle (Theorem I.6), we obtain

$$\forall \delta > 0, \quad R^{\mathrm{pub},k}_{\varepsilon}(f^{(t)}) \geq \max_{\mu}\left(\frac{t}{2}\left(\delta D^{\mu,k}_{\varepsilon+\delta}(f) - O(k)\right)\right).$$

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

Remark. Such a direct sum result holds even if the two parties are given a relation $F \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, and on input (x, y) are required to produce a $z \in \mathcal{Z}$ such that $(x, y, z) \in F$. Result 3 requires the distribution μ to be a product distribution. If this requirement could be removed, we would be able infer a direct sum result for randomized communication complexity, namely

$$R^{\mathrm{pub},k}_{\varepsilon}(f^{(t)}) \geq \frac{t}{2}\left(\delta R^{\mathrm{pub},k}_{\varepsilon+\delta}(f) - O(k)\right). \tag{I.2}$$

In some cases, however, our result implies this claim: if for some function f, the distribution μ that achieves the maximum in Theorem I.6 when applied to $R^{\mathrm{pub},k}_{\varepsilon+\delta}(f)$ is a product distribution, then (I.2) holds.

E. Related work

The following asymptotic versions of our Results 1 and 2 were shown (independently of our work) by Winter [Win02] and Bennett *et al.* [BSST02] respectively.

Theorem I.7 ([Win02, Theorem 9 and Remark 10])**.** *For every pair of distributions (X, Y) and $\lambda > 0$ and n, there exists a one-way protocol Π_n such that the distribution $(X^n, \Pi_n(X^n))$ is λ-close in total variation distance to the joint distribution (X^n, Y^n) and furthermore,*

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

Theorem I.8 (Reverse Shannon theorem [BSST02])**.** *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 $\overline{x} \in \mathcal{X}^n$, the output $\Pi_n(\overline{x})$ has the distribution $E^n(\overline{x})$, and it is asymptotically efficient in the sense that*

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

Note that the asymptotic result of Winter [Win02] is slightly stronger than what is stated above in Theorem I.7 in that it 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 one should 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. [CT91, 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 sets 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 seem to 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 $O(\lg\lg(|\mathcal{X}|+\mathcal{Y}|))$ (see Section VII). This raises the natural question of tradeoffs between shared randomness and expected communication. The asymptotic version of this problem was recently solved by Cuff [Cuf08], and Bennett and Winter (Personal Communication [BW06]); they determined the precise asymptotic tradeoffs between communication and shared randomness.

Substate Theorem: Jain, Radhakrishnan and Sen [JRS09] prove the following result relating the relative entropy between two distributions P and Q to how well a distribution is contained in another.

Theorem I.9 (Classical substate theorem, [JRS09])**.** *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 I.5) is a strengthening of the above theorem (the above theorem follows from Lemma I.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 and Result 2 which allows for error. More precisely, one can infer (from Theorem I.9) that $T^R_\lambda[X:Y] \leq O(I[X:Y]/\lambda)$ and $T^R_\lambda(E) \leq O(\mathcal{C}(E)/\lambda)$. Note Jain, Radhakrishnan and Sen [JRS09] actually showed a quantum analogue of the above substate theorem. It is open if quantum analogues of our results hold.

Lower Bounds using message compression: Chakrabarti and Regev [CR04] 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 [JRS03a]. The rejection sampling lemma (Lemma I.5) can be used to improve message compression of [JRS03a], 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. Assuming the rejection sampling lemma (Lemma I.5), we first prove Results 1 and 2 in Sections II and III respectively. We then proceed to prove the rejection sampling lemma in Section IV. The Direct Sum Result (Result 3) is then proved in Section V. In Section VI, we give examples of joint distributions (X, Y) that satisfy $T[X:Y] = \omega(C[X:Y])$ and $C[X:Y] = \omega(I[X:Y])$. Finally, in Section VII, we show how to reduce

the shared randomness at the expense of a small additive cost in the expected communication.

II. Proof of Result 1

The inequality $T^R[X:Y] \geq I[X:Y]$ follows from the data processing inequality. Let M denote the message generated by Alice in the optimal protocol. Then, we have $T^R[X:Y] \geq H[M] \geq H[M \mid R] \geq I[X:M \mid R] = I[X:M \mid R] + I[X:R] = I[X:MR] \geq I[X:Y]$, where we use the fact that X and R are independent to conclude that $I[X:R] = 0$, the chain rule for mutual information and the data processing inequality (applied to the Markov chain $X \to (M,R) \to Y$) to conclude that $I[X:MR] \geq I[X:Y]$.

To obtain the second inequality we combine the rejection sampling lemma (Lemma I.5) and the following well-known relationship between relative entropy and mutual information.

Fact II.1. $I[X:Y] = \mathbb{E}_{x \leftarrow X}[S(Y|_{X=x} \| Y)]$.

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

We assume that the random string shared by Alice and Bob is a sequence of independently drawn samples $\langle y_1, y_2, \ldots \rangle$ according to the marginal distribution Y. On input $x \in \mathcal{X}$ drawn according to the distribution X, Alice uses the sampling procedure Rej-Sampler (from Lemma I.5) to sample the conditional distribution $Y|_{X=x}$ from the marginal distribution Y in order to generate the index i^*. (Note that $S(Y|_{X=x} \| Y) < \infty$). Alice transmits the index i^* to Bob, who then outputs the sample y_{i^*} which has the required distribution. The expected number of bits transmitted in this protocol is at most $E_{x \leftarrow X}\left[S(Y|_{X=x} \| Y) + 2\lg(S(Y|_{X=x} \| Y) + 1) + O(1)\right]$ which (by Fact II.1 and Jensen's inequality) is at most $I[X:Y] + 2\lg(I[X:Y]+1) + O(1)$. □

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

Fix the channel E, and let (X,Y) be the pair of random variables that realize its channel capacity.

Consider the first inequality. Let Π be any protocol simulating the channel E. We wish to show that $T_\Pi(x) \geq \mathcal{C}(E)$ for some $x \in X$. Let $M(x)$ be the message generated by Alice on input x. Then, we have (using reasoning similar to the one used for the proof of the first inequality in Result 1) that $C(E) = I[X:\Pi(X)] \leq H[M(X)] \leq \mathbb{E}_{x \leftarrow X}[T_\Pi(x)]$. Thus, there exists an $x \in X$ such that $T_\Pi(x) \geq \mathcal{C}(E)$.

To show the second inequality, we will combine the rejection sampling lemma (Lemma I.5) and the following fact.

Claim III.1 (see Gallager [Gal67, Theorem 4.5.1,p. 91])**.** *Let Q be the marginal distribution of Y. Then, 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 [Jai06] using a different argument.)*

The 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 $S(E_x \| Q) + 2\lg(S(E_x \| Q) + 1) + O(1)$; by Claim III.1, this is at most $\lg \mathcal{C}(E) + 2\lg(\mathcal{C}(E)+1) + O(1)$. This completes the proof of Result 2. □

IV. 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 $\langle x_1, x_2, \ldots \rangle$ independently drawn according to the distribution Q, outputs an index i^* such that x_{i^*} is distributed according to P, and the expected encoding length of the index i^* is as small as possible.

The procedure Rej-Sampler we formally state below examines the samples $\langle x_i : i \in \mathbb{N} \rangle$ sequentially; after examining x_i it either accepts it (by returning the value i for i^*) or moves on to the next sample x_{i+1}. For $x \in \mathcal{X}$ and $i \geq 1$, let $\alpha_i(x)$ denote the probability that the procedure outputs i and $x_i = x$. We wish to ensure that for all $x \in \mathcal{X}$,

$$P(x) = \sum_{i=1}^{\infty} \alpha_i(x).$$

Let $p_i(x) = \sum_{j=1}^{i} \alpha_i(x)$; thus, $p_i(x)$ is the probability that the procedure halts with $i^* \leq i$ and $x_{i^*} = x$. Let $p_i^* = \sum_{x \in \mathcal{X}} p_i(x)$; hence, p_i^* is the probability that the procedure halts within i iterations. These quantities will be determined once $\alpha_i(x)$ is defined. We define $\alpha_i(x)$ (and hence $p_i(x)$ and p_i^*) inductively. For $x \in \mathcal{X}$, let $p_0(x) = 0$. For $i = 1, 2, \ldots$ and for $x \in \mathcal{X}$, let

$$\begin{aligned} \alpha_i(x) &= \min\{P(x) - p_{i-1}(x), (1 - p_{i-1}^*)Q(x)\}; \\ p_i(x) &= p_{i-1}(x) + \alpha_i(x). \end{aligned}$$

The definition of $\alpha_i(x)$ can be understood as follows. The first term $P(x) - p_{i-1}(x)$ ensures that $p_i(x)$ never exceeds $P(x)$. The second term $(1 - p_{i-1}^*)Q(x)$ has the following interpretation. The probability that the procedure enters the i-th iteration and gets to examine x_i is precisely $1 - p_{i-1}^*$. Since, $\Pr[x_i = x] = Q(x)$, the probability that the procedure outputs i after examining $x_i = x$ can be at most $(1 - p_{i-1}^*)Q(x)$. Our definition of $\alpha_i(x)$ corresponds to the *greedy* strategy, that accepts the i-th sample, with as much probability as possible under the constraint that $p_i(x) \leq P(x)$ for all $x \in \mathcal{X}$. The following procedure implements this idea formally.

Rej-Sampler(P, Q)

Random Input: $\langle x_i : i \in \mathbb{N} \rangle$ a sequence of samples independently drawn from the distribution Q.

A. Initialization

a) For each $x \in \mathcal{X}$, set $p_0(x) \leftarrow 0$.

b) Set $p_0^* \leftarrow 0$.

B. For $i \leftarrow 1$ to ∞ do

Iteration (i)

a) Using the definitions above, for each $x \in \mathcal{X}$, compute $\alpha_i(x)$ and $p_i(x)$, and compute p_i^*. Let $\beta_i(x_i) = \frac{\alpha_i(x)}{(1-p_{i-1}^*)Q(x_i)}$.

Note that if we arrive at the i-th iteration,

then $p^*_{i-1} < 1$*; also* $Q(x_i) > 0$*. So,* $\beta_i \in [0,1]$ *is well defined.*

b) Examine sample x_i.

c) With probability $\beta_i(x_i)$ output i and halt.

Note that the probability that this procedure outputs i and $x_i = x$ is precisely $\beta_i(x)(1-p^*_{i-1})Q(x) = \alpha_i(x)$. We have two claims, which we prove below.

Claim IV.1. *For all* x, $\langle p_i(x) : i = 1, 2, \ldots\rangle$ *converges to* $P(x)$.

Claim IV.2. *Let* $\ell : \mathbb{N} \to \{0,1\}^*$ *be a prefix-free encoding of positive integers such that* $|\ell(n)| = \lg n + 2\lg\lg(n+1) + O(1)$*. Then,*

$$\mathbb{E}[|\ell(i^*)|] = S(P\|Q) + 2\lg(S(P\|Q)+1) + O(1).$$

Remark. Li and Vitanyi [LV08] present a sequence of prefix-free binary encodings E_i ($i \geq 0$) of natural numbers. In this sequence, $E_2(n)$ has length at most $\lg n + 2\lg\lg(n+1) + O(1)$, and $E_3(n)$ has length at most $\lg n + \lg\lg(n+1) + O(\lg\lg(1+\lg(n+1)))$. The assumptions in the above claim are satisfied if we take ℓ to be E_2. If we take ℓ to be $E_3(n)$ instead, then we can reduce the constant 2 in the Claim IV.2 to $1+\varepsilon$, for any $\varepsilon > 0$.

Proof of Claim IV.1: Fix an $x \in \mathcal{X}$. We will show that for $i = 1, 2, \ldots,$

$$\alpha_i(x) \geq (P(x) - p_{i-1}(x))Q(x), \tag{IV.1}$$

and then use induction to show that for $i = 0, 1, 2, \ldots,$

$$P(x) - p_i(x) \leq P(x)(1-Q(x))^i. \tag{IV.2}$$

The inequality (IV.2) holds for $i = 0$ because $p_0(x) = 0$, and for the induction step we have

$$\begin{aligned} P(x) - p_i(x) &= P(x) - p_{i-1}(x) - \alpha_i(x) \\ &\leq (P(x) - p_{i-1}(x))(1-Q(x)) \\ &\leq P(x)(1-Q(x))^i. \end{aligned}$$

Since $S(P\|Q) < \infty$, we have that if $P(x) > 0$, then $Q(x) > 0$. Since, $P(x) \geq p_i(x)$, we conclude that $p_i(x)$ converges to $P(x)$. It remains to establish (IV.1). First, observe that

$$\begin{aligned} 1 - p^*_{i-1} &= \sum_{y\in\mathcal{X}} P(y) - \sum_{y\in\mathcal{X}} p_{i-1}(y) \\ &= \sum_{y\in\mathcal{X}} (P(y) - p_{i-1}(y)) \\ &\geq P(x) - p_{i-1}(x). \end{aligned}$$

Now, (IV.1) follows immediately from the definition, $\alpha_i(x) = \min\{P(x) - p_{i-1}(x), (1-p^*_{i-1})Q(x)\}$. □

Proof of Claim IV.2: We show below that

$$\mathbb{E}[\lg i^*] \leq S(P\|Q) + O(1). \tag{IV.3}$$

Then, we have

$$\begin{aligned} \mathbb{E}\left[|\ell(i^*)|\right] &= \mathbb{E}\left[\lg i^* + 2\lg\lg(i^*+1) + O(1)\right] \\ &= \mathbb{E}\left[\lg i^*\right] + 2\,\mathbb{E}\left[\lg\lg(i^*+1)\right] + O(1) \\ &\leq \mathbb{E}\left[\lg i\right] + 2\lg\left(\mathbb{E}\left[\lg(i^*+1)\right]\right) + O(1) \\ &\quad \text{[Jensen's inequality: } \lg(\cdot) \text{ is concave]} \\ &\leq \mathbb{E}\left[\lg i^*\right] + 2\lg\left(\mathbb{E}\left[\lg i^*\right]+1\right) + O(1) \\ &\leq S(P\|Q) + O(1) + 2\lg\left(S(P\|Q) + O(1)\right) \\ &\quad + O(1) \qquad \text{[by (IV.3)]} \\ &= S(P\|Q) + 2\lg\left(S(P\|Q)+1\right) + O(1). \end{aligned}$$

It remains to show (IV.3). We have

$$\mathbb{E}\left[\lg i\right] = \sum_{x\in\mathcal{X}} \sum_{i=1}^{\infty} \alpha_i(x)\cdot\lg i. \tag{IV.4}$$

The term corresponding to $i = 1$ is 0. To bound the other terms we need to obtain a suitable bound on $\alpha_i(x)$. Let $i \geq 2$ and suppose $\alpha_i(x) > 0$. Then for $j = 1, 2, \ldots, i$, $p_{j-1}(x) < P(x)$, implying that for $j = 1, 2, \ldots, i-1$, $\alpha_j(x) = (1 - p^*_{j-1})Q(x) \geq (1-p^*_{i-1})Q(x)$. Thus,

$$P(x) > p_{i-1}(x) = \sum_{j=1}^{i-1} \alpha_j(x) \geq (i-1)(1-p^*_{i-1})Q(x),$$

which implies that

$$i \leq \frac{1}{1-p^*_{i-1}} \cdot \frac{P(x)}{Q(x)} + 1. \tag{IV.5}$$

We have shown this inequality assuming $i \geq 2$ and $\alpha_i(x) > 0$, but it clearly holds when $i = 1$ and $\alpha_1(x) > 0$. Returning to (IV.4) with this, we obtain

$$\begin{aligned} \mathbb{E}\left[\lg i\right] &= \sum_{x\in\mathcal{X}} \sum_{i=1}^{\infty} \alpha_i(x)\cdot\lg i \\ &\leq \sum_{x\in\mathcal{X}} \sum_{i=1}^{\infty} \alpha_i(x)\cdot\lg\left(\frac{1}{1-p^*_{i-1}}\cdot\frac{P(x)}{Q(x)} + 1\right) \\ &\quad \text{[by (IV.5)]} \\ &\leq \sum_{x\in\mathcal{X}} \sum_{i=1}^{\infty} \alpha_i(x)\cdot\lg\left(\frac{1}{1-p^*_{i-1}}\left(\frac{P(x)}{Q(x)}+1\right)\right) \\ &= \sum_{i=1}^{\infty} (p^*_i - p^*_{i-1})\cdot\lg\left(\frac{1}{1-p^*_{i-1}}\right) \\ &\quad + \sum_{x\in\mathcal{X}} P(x)\cdot\lg\left(\frac{P(x)}{Q(x)}+1\right) \\ &\leq \int_0^1 \lg\frac{1}{1-p}\mathrm{d}p + \sum_{x\in\mathcal{X}} P(x)\cdot\lg\left(\frac{P(x)}{Q(x)}+1\right) \\ &= \lg e + \sum_{x\in\mathcal{X}} P(x)\lg\left(\frac{P(x)}{Q(x)}\right) + \\ &\quad \sum_{x\in\mathcal{X}} P(x)\lg\left(1+\frac{Q(x)}{P(x)}\right) \\ &\leq \lg e + S(P\|Q) + \sum_{x\in\mathcal{X}} P(x)\cdot\lg\exp\left(\frac{Q(x)}{P(x)}\right) \\ &\quad \text{[since } 1+t \leq \exp(t)\text{]} \\ &= S(P\|Q) + 2\lg e. \end{aligned}$$

□

The above proof shows that there exists a rejection sampling procedure such that $\mathbb{E}[|\ell(i^*)|] \leq S(P\|Q) + 2\lg(S(P\|Q) + 1) + O(1)$, for a suitable encoding ℓ. We now observe that any such procedure satisfies $\mathbb{E}[|\ell(i^*)|] \geq S(P\|Q)$.

Proposition IV.3. *Fix a prefix-free binary encoding ℓ of positive natural numbers. Any rejection sampling procedure as defined in Section I-B satisfies $\mathbb{E}[|\ell(i^*)|] \geq S(P\|Q)$.*

Proof: For all x and i, define α_i as follows:

$$\alpha_i \triangleq \Pr[i^* = i \wedge x_{i^*} = x].$$

Clearly, $\alpha_i \leq \Pr[x_i = x] = Q(x)$. We thus have

$$\begin{aligned}\mathbb{E}[|\ell(i^*)| \mid x_{i^*} = x] &\geq H[i^* \mid x_{i^*} = x] \\ &= \sum_i \frac{\alpha_i}{P(x)} \lg \frac{P(x)}{\alpha_i} \\ &\geq \lg \frac{P(x)}{Q(x)}.\end{aligned}$$

Thus, $\mathbb{E}[|\ell(i^*)|] = \sum_x P(x)\,\mathbb{E}[|\ell(i^*)| \mid x_{i^*} = x] \geq S(P\|Q)$. □

V. Proof of Direct Sum Result (Result 3)

Below we present our result in the two-party model for computing functions $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. However, the result also holds for protocols computing relations $R \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ in which Alice and Bob given $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively, need to output a $z \in \mathcal{Z}$ such that $(x, y, z) \in R$.

Our proof uses the notion of information cost defined by Chakrabarti *et al.* [CSWY01], and refined in several subsequent works [BJKS04], [JRS03a], [JRS03b], [JRS05].

Definition V.1 (Information cost)**.** *Let Π be a private-coin protocol taking inputs from the set $\mathcal{X} \times \mathcal{Y}$, and let μ be a distribution on the input set $\mathcal{X} \times \mathcal{Y}$. Then, the information cost of Π under μ is*

$$\mathsf{IC}^{\mu}(\Pi) = I[XY : M],$$

where (X, Y) represent the input to the two parties (chosen according to the distribution μ) and M is the transcript of the messages exchanged by the protocol on this input. For a function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, let

$$\mathsf{IC}^{\mu,k}_{\varepsilon}(f) = \min_{\Pi} \mathsf{IC}^{\mu}(\Pi),$$

where Π ranges over all k-round private-coin protocols for f with error at most ε under μ.

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

Proposition V.2. $\mathsf{IC}^{\mu,k}_{\varepsilon}(f) \leq D^{\mu,k}_{\varepsilon}(f)$.

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

A key insight of Chakrabarti *et al.* [CSWY01] was that one could (approximately) show a relationship in the opposite direction when the inputs are being drawn from the uniform distribution. Their result, which was stated for simultaneous message passing (SMP) protocols, was later extended by Jain *et al.* [JRS03a], [JRS05] using the substate theorem (Theorem I.9). The main idea in Jain *et al.* was that messages could be compressed to the amount of information they carried 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 led to corresponding direct sum results. Using Result 1, we can now considerably strengthen the result of Jain *et al.* [JRS03a] for two-way protocols.

Lemma V.3 (Message compression)**.** *Let $\varepsilon, \delta > 0$. Let μ be a distribution (not necessarily product) on $\mathcal{X} \times \mathcal{Y}$ 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 Result 3 is the direct sum property of information cost, originally observed by Chakrabarti *et al.* [CSWY01] for the uniform distribution.

Lemma V.4 (Direct sum for information cost)**.** *Let μ be a product distribution on $\mathcal{X} \times \mathcal{Y}$. Then, $\mathsf{IC}^{\mu^t,k}_{\varepsilon}(f^{(t)}) \geq t \cdot \mathsf{IC}^{\mu,k}_{\varepsilon}(f)$.*

This is the only place in the proof where we require μ to be a product distribution. Before proving these lemmas, let us show that they immediately imply our theorem.

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

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

where the first inequality follows from Proposition V.2, the second from Lemma V.4 and the last from Lemma V.3. □

Proof of Lemma V.3: Let μ be a distribution on $\mathcal{X} \times \mathcal{Y}$. Fix a private-coin protocol Π that achieves the optimum information cost $\mathsf{IC}^{\mu,k}_{\varepsilon}(f)$. Let (X, Y) be the random variables representing the inputs of Alice and Bob distributed according to μ. We will use the following notation: $M = M(X, Y)$ will be the transcript of the protocol; for $i = 1, 2, \ldots, k$, M_i will denote the i-th message of the transcript M and $M_{1,i}$ will denote the first i messages in M. Now, we have from the *chain rule* for mutual information (cf. [CT91]).

$$I[XY : M] = \sum_{i=1}^{k} I[XY : M_i \mid M_{1,i-1}]. \tag{V.1}$$

We now construct another protocol Π' as follows. For $i = 1, 2, \ldots, k$, the party that sent M_i in Π will now instead use Result 1 to generate the message M_i for the other party by sending about $I[XY : M_i \mid M_{1,i-1}]$ bits on the average. Suppose, we manage to generate the first $i - 1$ messages in Π' with distribution exactly as that of $M_{1,i-1}$, and the (partial) transcript so far is m. For the rest of this paragraph we condition on $M_{1,i-1} = m$, and describe how

the next message is 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 $(M_i \mid M_{1,i-1} = m)$ 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-coin protocol, for each $x \in \mathcal{X}$, the conditional random variable $(M_i \mid M_{1,i-1} = m \wedge X = x)$, is independent of $(Y \mid M_{1,i-1} = m \wedge X = x)$. Hence on input x, ALICE knows the distribution of $(M_i(x,Y) \mid M_{1,i-1}(x,Y) = m)$.

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

$$I[XY : M_i \mid M_{1,i-1} = m] = I[X : M_i \mid M_{1,i-1} = m].$$

Thus, by Result 1, ALICE can arrange for $(M_i | M_{1,i-1} = m)$ to be generated on BOB's side by sending at most

$$2I[X : M_i \mid M_{1,i-1} = m] + O(1)$$

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

$$2I[XY : M_i \mid M_{1,i-1}] + O(1).$$

By applying this strategy for all rounds, we note from (V.1) that we obtain a public-coin k-round protocol Π', with expected communication $2I[XY : M] + O(k)$ 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 $\frac{1}{\delta}(2I[XY : M] + O(k)) = \frac{1}{\delta}(2 \cdot \mathsf{IC}^{\mu,k}_{\varepsilon}(f) + O(k))$. The lemma now follows from this and the definition of $D^{\mu,k}_{\varepsilon+\delta}(f)$. □

Proof of Lemma V.4: Let μ be a product distribution on $\mathcal{X} \times \mathcal{Y}$. Fix a k-round private-coin protocol Π for $f^{(t)}$ that achieves $\mathsf{IC}^{\mu^t,k}_{\varepsilon}(f^{(t)})$. For this protocol Π the input is chosen according to μ^t. We denote this input by $(X,Y) = (X_1X_2\cdots X_t, Y_1Y_2\cdots Y_t)$ and note that the $2t$ random variables involved are mutually independent. Let M denote the transcript of this protocol when run on the input (X,Y). Now, we have from chain rule for mutual information and independence of the $2t$ random variables as above,

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

We claim that each term in the sum of the form $I[X_iY_i : M]$ is at least $\mathsf{IC}^{\mu,k}_{\varepsilon}$. 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-coin protocol Π' for f with error at most ε under μ, since the error of Π was at most ε under μ^k. Clearly, $\mathsf{IC}^{\mu}(\Pi) = I[X_iY_i : M]$. □

VI. SEPARATING $T[X : Y]$, $C[X : Y]$ AND $I[X : Y]$

For any pair of random variables (X,Y), it easily follows from the definitions that $T[X : Y] \geq C[X : Y]$. Furthermore, by Wyner's theorem (Theorem I.1)

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

where W is such that X and Y are independent when conditioned on W. Note, however, that

$$I[XY : W] \geq I[X : W] \geq I[X : Y].$$

The first inequality comes from the monotonicity of mutual information which in turn follows from the chain rule for mutual information. The second inequality is the data processing inequality applied to the Markov chain $X \to W \to Y$. Thus, we have $T[X : Y] \geq C[X : Y] \geq I[X : Y]$. In this section, we will show that both these inequalities are strict for (X,Y) defined as follows.

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

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

(b) *X and Y are independent when conditioned on W.*

It is easy to see that X and Y are uniformly distributed n-bit strings (but not independent). Hence, $H[X] = H[Y] = n$.

Proposition VI.2. *For (X,Y) defined as above, we have:*

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

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

(c) $T[X : Y] = \Theta(\lg n)$.

Note that in the above example, though $C[X : Y]$ and $I[X : Y]$ differ by at most 2. However, we can construct another joint distribution (X', Y') by taking m independent copies of the joint distribution (X,Y) (i.e., $(X',Y') = (X,Y)^m$) for some $m = m(n) = \omega(1)$. It then follows from the chain rule for mutual information that $I[X' : Y'] = I[X^m : Y^m] = mI[X : Y] = o(m)$. Furthermore,

$$\begin{aligned} C[X^m : Y^m] &= \liminf_{\lambda \to 0} \lim_{k \to \infty} (T_\lambda[X^{mk} : Y^{mk}]/k) \\ &= m \cdot \liminf_{\lambda \to 0} \lim_{k \to \infty} (T_\lambda[X^{mk} : Y^{mk}]/mk) \\ &= mC[X : Y] = \Theta(m), \end{aligned}$$

where the first and third equalities follow from (I.1). Hence, $C[X' : Y'] = C[X^m : Y^m] = mC[X : Y] = \Theta(m)$. This implies that it is not possible to bound $C[X : Y]$ from above by a linear function in $I[X : Y]$.

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

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

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

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

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

Proof of part (b): By Wyner's theorem (Theorem I.1),

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

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

$$\max_{W'} H[XY\mid W']=2n-2. \tag{VI.1}$$

Let W' be such that $I[X:Y\mid W']=0$. Consider any w in the support of W'. Let X_w be the set of $x\in\{0,1\}^n$ such that $\Pr[X=x\mid W'=w]>0$. Similarly, define Y_w. We must have that $|X_w|+|Y_w|\leq 2^n$, since otherwise there will exist an x such that $\Pr[X=x\wedge Y=\bar{x}]>0$ 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,

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

Now, if we let W' be the random variable W used in Definition VI.1, we have $H[XY\mid W]=2(n-1)$. Hence,

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

This justifies (VI.1) and completes the proof of part (b). □

To prove part (c), we will use a theorem of Harper [Har66], which states that Hamming balls in the hypercube have the smallest boundary. The following version, due to Frankl and Füredi (see Bollobás [Bol86, 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

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

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

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

Theorem VI.3 ([Bol86, Theorem 3, page 127])**.** *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 VI.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*

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

Proof: By Theorem VI.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 [AS00, Theorem A.1.1, page 263]) that

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

□

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

It remains to show that $T[X:Y]=\Omega(\lg n)$. It follows from our definition that $T[X:Y]\geq\min_{W'}H[W']$, where the minimum is over all random variables W' such that X and Y are conditionally independent given W'. Fix such a random variable W'. We show below that for all w,

$$\alpha_w=\Pr[W'=w]=O\left(\frac{\sqrt{\ln n}}{n}\right). \tag{VI.2}$$

That is, we show that the min-entropy of W' is $\Omega(\lg n)$; it follows that the entropy of W' is $\Omega(\lg n)$ as required. It remains to establish (VI.2). Fix w such that $\alpha_w>0$, and let

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

Our proof of (VI.2) is based on two observations.

Claim VI.5. (i) *For all $x \in X_w$ and $y \in Y_w$, we have* $\mathrm{agr}(x,y) > \alpha_w n/8$.
(ii) $|X_w|, |Y_w| \geq \alpha_w 2^{n-1}$

We will justify these claims below. Let us first derive the desired upper bound on α_w from this using Corollary VI.4. Let Y'_w be the set of strings whose bitwise complements belong to Y_w. Since $\mathrm{agr}(x,y) > \alpha_w n/8$ for all $x \in X_w$ and $y \in Y_w$, the Hamming distance between X_w and Y'_w is greater than $\alpha_w n/8$. By Corollary VI.4, we conclude that

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

which implies that $\alpha_w = O\left(\frac{\sqrt{\ln n}}{n}\right)$. It remains to prove Claim VI.5.

First consider Claim VI.5 (i). For all $x \in X_w$ and $y \in Y_w$, we have

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

that is, $\mathrm{agr}(x,y) > \alpha_w n/8$.

Next, consider Claim VI.5 (ii). Note that

$$\sum_{x \in \{0,1\}^n} \Pr[X = x \mid W' = w] = 1.$$

The contribution to the left hand side from $x \notin X_w$ is at most $2^{-(n+1)} \times 2^n \leq \frac{1}{2}$. Thus, the total contribution from $x \in X_w$ is at least $\frac{1}{2}$. Any one $x \in X_w$ contributes $\Pr[X = x \mid W' = w] \leq \Pr[X = x]/\Pr[W' = w] \leq 2^{-n}/\alpha_w$. It follows that $|X_w| \geq \alpha_w 2^{n-1}$. Similarly, we conclude that $|Y_w| \geq \alpha_w 2^{n-1}$. □

VII. Reducing the shared randomness

In the preceding sections, we did not formally bound the amount of shared randomness used by the protocol. We now address this shortcoming, and show how one can reduce the number of shared random bits used substantially, while increasing the communication only slightly. Our main result is the following.

Theorem VII.1. *For all pairs of random variables (X,Y), there is a one-way protocol Π for generating (X,Y) (so that $(X,\Pi(X))$ has the same distribution as (X,Y)) such that*

1) *the expected communication from* Alice *to* Bob *is at most $I[X:Y] + O(\lg(I[X:Y]+1)) + O(\lg\lg|\mathsf{support}(Y)|)$;*
2) *the number of bits of shared randomness read by either party is $O(\lg\lg|\mathsf{support}(X)| + \lg|\mathsf{support}(Y)|)$.*

To justify this theorem, we will present a protocol that satisfies the requirements. We will derive our protocol from a probabilistic argument, which we state using the language of graphs.

Definition VII.2 (Protocol graphs)**.** *A protocol graph $G(M,N)$ is a labeled directed acyclic graph with a source s, a sink t, and two layers in between: V with M vertices and W with N. The source s is connected by an edge to each of the M vertices in the layer V. Each of the N vertices in the third layer W is connected to the sink t. The remaining edges go from V to W. We use E to denote the set of these edges. There is a labeling $\ell : E \to \{0,1\}^*$, such that for each $v \in V$, the labeling ℓ when restricted to the edges incident on v is a prefix-free encoding for those edges.*

In our argument, Alice and Bob will work based on a graph $G(M,N)$, viewing $[M]$ as the set of shared random strings, and $[N]$ as the set over which Bob's output must be distributed. The edges of the graphs and the labels on them will determine how Bob interprets Alice's message. This is made precise in the following definition.

Definition VII.3 (Protocols based on graphs)**.** *Let $G(M,N)$ be a protocol graph and let P be a distribution on $[N]$. In a protocol for P based on G,* Alice *sends a message to* Bob *so as to enable him to generate a string $w \in [N]$ whose distribution is P. Such a protocol operates as follows.*

- **Shared randomness:** Alice *and* Bob *share a random string R picked with uniform distribution from $[M]$, so R has $\lceil \lg M \rceil$ bits;*
- **Message:** Alice *computes a message $m \in \{0,1\}^*$ based on the random string R, her input and her private coins;*
- **Output:** *On receiving the message $m \in \{0,1\}^*$,* Bob *outputs $w \in [N]$, where (R,w) is the unique edge of G incident on R with label m.*

The cost of a protocol is the expected number of bits Alice *transmits.*

Lemma VII.4 (Main lemma)**.** *For all distributions Q on $[N]$ and $M \geq N$, there is a distribution $G(M,N)$ on protocol graphs, such that for each distribution P on $[N]$, with probability greater than $1 - 2^{-M}$, there is a protocol for P based on G with cost $S(P\|Q) + O(\lg(S(P\|Q)+1)) + O(\lg\lg N)$.*

Before proving this lemma, let us first see how this immediately implies Theorem VII.1.

Proof of Theorem VII.1: We apply Lemma VII.4 with Q as the distribution of Y, P as P_x, the distribution of Y conditioned on $X = x$, and $M = \max\{\lceil \lg|\mathsf{support}(X)| \rceil, |\mathsf{support}(Y)|\}$. Since there are at most 2^M choices for $x \in \mathsf{support}(X)$, we conclude from the union bound that there is an instance $\hat{G}$ of the protocol graph $G(M,N)$, such that for each $x \in \mathsf{support}(X)$, there is a protocol Π_x for P_x based on $\hat{G}$, using $O(\lg M)$ bits of shared randomness and $S(P_x\|Q) + O(\lg(S(P_x\|Q)+1)) + O(\lg\lg N)$ bits of communication. Since Bob's actions are determined completely by the protocol graph, he acts in the same way in all these protocols. The protocol for (X,Y) is now straightforward: on input x, Alice sends a message assuming she is executing Π_x, and Bob interprets this message as before using the graph $\hat{G}$, and is guaranteed to output a string $y \in \mathsf{support}(Y)$ with distribution P_x. We thus have a protocol for (X,Y). Furthermore, it follows from Fact II.1 (see Section II) that the cost of this protocol is $I[X:Y] + O(\lg I[X:Y]) + O(\lg\lg N)$. The number of bits of shared randomness is $\lceil \lg M \rceil = O(\lg\lg|\mathsf{support}(X)| + \lg|\mathsf{support}(Y)|)$. □

A. Proof of Lemma VII.4

We view a protocol with low communication as a low-cost flow in a suitably constructed capacitated protocol graph. Then, we will construct a random graph that admits such a low-cost flow with high probability.

Definition VII.5 (Capacities, flows)**.** *Let $G(M,N)$ be a protocol graph and P a distribution on $[N]$. Then, G^P is the capacitated version of G where the edges of the form (s,v) have capacity $\frac{1}{M}$, edges of the form (w,t) have capacity $P(w)$, and all other edges have infinite capacity. An α-flow in G^P is a flow $f : E(G^P) \to [0,1]$ with value (that is, the total flow out of s) at least α. The cost of a flow f is $\sum_{e\in E} f(e)|\ell(e)|$, where $\ell : E \to \{0,1\}^*$ is the labelling of the edges specified in the protocol graph G.*

Proposition VII.6. *Let $G(M,N)$ be a protocol graph and P be a distribution on $[N]$. If there is a 1-flow in G^P with cost C, then there is a protocol for P with shared randomness $O(\lg M)$ and cost C.*

Proof: Fix a flow f in G for P. We will show how ALICE picks the label to transmit in order to enable BOB to generate a string in $[N]$ with the required distribution. If the shared random string is R, ALICE picks the edge $e = (R,w)$ leaving R with probability $Mf(e)$ and transmits its label $\ell(e)$. BOB's actions are now determined, and it is easy to verify that the string he produces has the required distribution. □

Now, Lemma VII.4 follows immediately by combining the above proposition with the following lemma, which is the main technical observation of this section.

Lemma VII.7. *For all distributions Q on $[N]$ and $M \geq N$, there is a random variable $G(M,N)$ taking values in the set of protocol graphs such that for each distribution P on $[N]$, with probability at least $1-2^{-M}$, there is a 1-flow in G^P with cost $S(P\|Q) + O(\lg S(P\|Q)) + O(\lg\lg N)$.*

Proof: To define the random graph $G(M,N)$, we need to state how the edges that go from $V = [M]$ to $W = [N]$ are chosen and labelled. In our random graph, this set, $E(V,W)$, will be the union of two sets E_0 and E_1; the labels of the edges in E_0 begin with a 0 and the labels of the edges in E_1 begin with a 1.

- **The edges in E_0:** $E_0 = [M] \times [N]$, with the edge (i,j) labeled by $0 \cdot [j]$, where $[j]$ denotes the binary encoding of j using $\lceil \lg N \rceil$ bits;
- **The edges in E_1:** The labels of the edges in this set start with a 1. These edges are generated randomly as follows. For each $i \in [M]$ and each $k \in \mathbb{Z}^+$, we have one edge with label $1\cdot\tau(k)$, where $\tau : \mathbb{Z}^+ \to \{0,1\}^*$ is a prefix-free encoding of $\mathbb{Z}^+$ with $|\tau(k)| \leq \lg k + O(\lg(1+\lg k))$. The other end point of this edge is chosen randomly from the set $[N]$ with distribution Q (independently for different i and k).

We wish to show that for all distributions P on $[N]$, with high probability, there is a 1-flow in G^P. We will do this in two steps. First, we will show using the max-flow min-cut theorem that with high probability there is a low-cost $\left(1-\frac{1}{\lg N}\right)$-flow in G for P using the edges in E_1. To turn this into a proper flow, we will send some additional flow along the edges in E_0. Since the total value of the flow along edges in E_0 is $\left(\frac{1}{\lg N}\right)$, this does not significantly increase the cost.

Consider the subgraph G_1 of $G^P(M,N)$ obtained by retaining only those edges (i,j) in E_1 whose labels lie in the set $\{1\cdot\tau(1), 1\cdot\tau(2), \ldots, 1\cdot\tau(L(j))\}$, where $L(j) \triangleq \lceil 3(P(j)/Q(j))\lg^2 N\rceil$. To show that G_1 has a $\left(1-\frac{1}{\lg N}\right)$-flow with high probability, we show that (with high probability) it has no cut of size $1-\frac{1}{\lg N}$, that is the removal of no set of edges of total capacity less than $1-\frac{1}{\lg N}$ can disconnect t from s. Since the edges going between $[M]$ and $[N]$ have infinite capacity, each edge in any such cut is incident on either s or t. Fix a set of edges $\mathcal{C}$ of total capacity less than $1-\frac{1}{\lg N}$. Let $S = \{i \in V : (s,i) \notin \mathcal{C}\}$ and $T = \{j \in W : (j,t) \notin \mathcal{C}\}$. We will show that with high probability $\mathcal{C}$ is not an s-t cut in G_1. Note that $|S| > \frac{M}{\lg N}$ and $\sum_{i\in T} P(i) > \frac{1}{\lg N}$. For $\mathcal{C}$ to be a cut, there should be no edge in G_1 connecting S and T. Fix a vertex $j \in T$, and consider the event $\mathcal{E}_j \equiv$ *there is no edge from S to j in G_1*. Since the edges out of S are chosen independently, $\Pr[\mathcal{E}_j] = (1-Q(j))^{L(j)|S|}$; furthermore the events $(\mathcal{E}_j : j \in T)$ are negatively correlated; in particular, for any set $T' \subseteq [N]$, $\Pr\left[\mathcal{E}_j \mid \bigwedge_{i\in T'\{j\}} \mathcal{E}_i\right] \leq \Pr[\mathcal{E}_j]$. Thus,

$$\begin{aligned}
\Pr[\mathcal{C} \text{ is a cut in } G_1] &\leq \Pr\left[\bigwedge_{j\in T} \mathcal{E}_j\right] \\
&\leq \prod_{j\in T} (1-Q(j))^{L(j)|S|} \\
&\leq \prod_{j\in T} \exp(-L(j)Q(j)|S|) \\
&\qquad\qquad \text{(because } 1-x \leq e^{-x}\text{)} \\
&< \exp\left(-3(\lg^2 N)|S| \sum_{j\in T} P(j)\right) \\
&\leq \exp(-3M).
\end{aligned}$$

Since there are at most 2^M choices for S and at most 2^N choices for T, there are at most 2^{M+N} choices for $\mathcal{C}$. Thus, we have

$$\begin{aligned}
\Pr[G_1^P \text{ has a small cut}] &< 2^{M+N}\exp(-3M) \\
&\leq 2^{-M} \quad (\text{since } M \geq N).
\end{aligned}$$

By the max-flow min-cut theorem, with probability at least $1-2^{-M}$, G_1 has a flow with value at least $1-\frac{1}{\lg N}$ and cost at most

$$\begin{aligned}
\sum_{j\in T} P(j)|\tau(L(j))| &= \sum_{j\in[N]} P(j)\,[\lg L(j) \\
&\qquad + O(\log(\log L(j)+1))] \\
&\leq S(P\|Q) + O(\lg(S(P\|Q)+1)) \\
&\qquad + O(\lg\lg N),
\end{aligned}$$

where the last inequality follows by recalling that $L(j) = \lceil 3(P(j)/Q(j))\lg^2 N\rceil$ and that $\lg(\cdot)$ is a concave function. We convert this $(1-\frac{1}{\lg N})$-flow into a proper flow by using

the edges in E_0 to supply the remaining $\frac{1}{\lg N}$ units. Since the edges in E_0 have labels of length at most $1 + \lceil \lg N \rceil$ and the total flow through these edges is at most $\frac{1}{\lg N}$, the resulting increase in cost is $O(1)$. □

Acknowledgments

Prahladh Harsha, David McAllester and Jaikumar Radhakrishnan participated in the discussions of the Winter 2006 machine learning reading group at TTI-Chicago that led to this work. We thank Lance Fortnow for his help in the proof of Proposition VI.2. We thank Andreas Winter for informing us of his work with Charles Bennett on the asymptotic versions of our results, and Oded Regev for comments and suggestions that greatly improved the presentation in this paper. We thank the referees for their helpful comments.

References

[AS00] Noga Alon and Joel H Spencer. *The Probabilistic Method*. Wiley-Interscience, 2nd edition, 2000. doi:10.1002/0471722154.

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. *An information statistics approach to data stream and communication complexity*. J. Computer and System Sciences, 68(4):702–732, June 2004. (Preliminary Version in *43rd FOCS*, 2002). doi:10.1016/j.jcss.2003.11.006.

[Bol86] Béla Bollobás. *Combinatorics: Set Systems, Hypergraphs, Families of Vectors and Combinatorial Probability*. Cambridge University Press, 1986. doi:10.2277/0521337038.

[BSST02] 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(10):2637–2655, October 2002. (Preliminary Version in *Proc. Quantum Information: Theory, Experiment and Perspectives* Gdansk, Poland, 10 - 18 July 2001). arXiv:quant-ph/0106052, doi:10.1109/TIT.2002.802612.

[BW06] Charles H. Bennett and Andreas Winter, 2006. (Personal Communication).

[CR04] 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 (FOCS)*, pages 473–482. IEEE, 2004. doi:10.1109/FOCS.2004.12.

[CSWY01] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Chi-Chih Yao. *Informational complexity and the diect sum problem for simultaneous message complexity*. In *Proc. 42nd IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 270–278. IEEE, 2001. doi:10.1109/SFCS.2001.959901.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991. doi:10.1002/0471200611.

[Cuf08] Paul Cuff. *Communication requirements for generating correlated random variables*. In *Proc. of IEEE International Symposium on Information Theory (ISIT)*, pages 1393–1397. July 2008. arXiv:0805.0065, doi:10.1109/ISIT.2008.4595216.

[Gal67] Robert G. Gallager. *Information Theory and Reliable Communication*. Wiley Publishers, 1967.

[Har66] Lawrence H. Harper. *Optimal numberings and isoperimetric problems on graphs*. J. Combinatorial Theory, 1(3):385–394, 1966. doi:10.1016/S0021-9800(66)80059-5.

[HJMR07] Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. *The communication complexity of correlation*. In *Proc. 22nd IEEE Conference on Computational Complexity*, pages 10–23. IEEE, 2007. doi:10.1109/CCC.2007.32.

[Jai06] Rahul Jain. *Communication complexity of remote state preparation with entanglement*. Quantum Information and Computation, 6(4–5):461–464, July 2006. arXiv:quant-ph/0504008.

[JRS03a] Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. *A direct sum theorem in communication complexity via message compression*. In Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger, eds., *Proc. 30th International Colloquium of Automata, Languages and Programming (ICALP)*, volume 2719 of *LNCS*, pages 300–315. Springer, 2003. arXiv:cs/0304020, doi:10.1007/3-540-45061-0_26.

[JRS03b] ———. *A lower bound for the bounded round quantum communication complexity of set disjointness*. In *Proc. 44th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 220–229. IEEE, 2003. doi:10.1109/SFCS.2003.1238196.

[JRS05] ———. *Prior entanglement, message compression and privacy in quantum communication*. In *Proc. 20th IEEE Conference on Computational Complexity*, pages 285–296. IEEE, 2005. doi:10.1109/CCC.2005.24.

[JRS09] ———. *A property of quantum relative entropy with an application to privacy in quantum communication*. J. ACM, 56(6):1–32, 2009. (Preliminary version in *43rd FOCS*, 2002). doi:10.1145/1568318.1568323.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1997. doi:10.2277/052102983X.

[LV08] Ming Li and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 3rd edition, 2008. doi:10.1007/978-0-387-49820-1.

[Win02] Andreas Winter. *Compression of sources of probability distributions and density operators*, 2002. arXiv:quant-ph/0208131.

[Wyn75] Aaron D Wyner. *The common information of two dependent random variables*. IEEE Transactions on Information Theory, 21(2):163–179, March 1975.

[Yao77] Andrew Chi-Chih Yao. *Probabilistic computations: Toward a unified measure of complexity (extended abstract)*. In *Proc. 18th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 222–227. IEEE, 1977. doi:10.1109/SFCS.1977.24.