Communication Complexity of Estimating Correlations

Uri Hadar[*]
Tel Aviv University, Israel
urihadar@mail.tau.ac.il

Jingbo Liu
MIT, Cambridge, MA
jingbo@mit.edu

Yury Polyanskiy
MIT, Cambridge, MA
yp@mit.edu

Ofer Shayevitz
Tel Aviv University, Israel
ofersha@eng.tau.ac.il

ABSTRACT

We characterize the communication complexity of the following distributed estimation problem. Alice and Bob observe infinitely many iid copies of ρ-correlated unit-variance (Gaussian or ±1 binary) random variables, with unknown $\rho \in [-1, 1]$. By interactively exchanging k bits, Bob wants to produce an estimate $\hat{\rho}$ of ρ. We show that the best possible performance (optimized over interaction protocol Π and estimator $\hat{\rho}$) satisfies $\inf_{\Pi\hat{\rho}} \sup_\rho \mathbb{E}[|\rho - \hat{\rho}|^2] = k^{-1}(\frac{1}{2\ln 2} + o(1))$. Curiously, the number of samples in our achievability scheme is exponential in k; by contrast, a naive scheme exchanging k samples achieves the same $\Omega(1/k)$ rate but with a suboptimal prefactor. Our protocol achieving optimal performance is one-way (non-interactive). We also prove the $\Omega(1/k)$ bound even when ρ is restricted to any small open sub-interval of $[-1, 1]$ (i.e. a local minimax lower bound). Our proof techniques rely on symmetric strong data-processing inequalities and various tensorization techniques from information-theoretic interactive common-randomness extraction. Our results also imply an $\Omega(n)$ lower bound on the information complexity of the Gap-Hamming problem, for which we show a direct information-theoretic proof.

CCS CONCEPTS

• **Theory of computation → Communication complexity**; Interactive computation; • **Mathematics of computing** → *Information theory*; Probabilistic inference problems.

KEYWORDS

Distributed estimation, Communication complexity, Correlation estimation

ACM Reference Format:
Uri Hadar, Jingbo Liu, Yury Polyanskiy, and Ofer Shayevitz. 2019. Communication Complexity of Estimating Correlations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on the Theory of Computing (STOC '19), June 23–26, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3313276.3316332

1 INTRODUCTION

The problem of distributed statistical inference under communication constraints has gained much recent interest in the theoretical computer science, statistics, machine learning, and information theory communities. The prototypical setup involves two or more remote parties, each observing local samples drawn from of a partially known joint statistical model. The parties are interested in estimating some well-defined statistical property of the model from their data, and to that end, can exchange messages under some prescribed communication model. The communication complexity associated with this estimation problem concerns the minimal number of bits that need to be exchanged in order to achieve a certain level of estimation accuracy. Whereas the sample-complexity of various estimation problems in the centralized case is well studied (see e.g. [LC06],[VT04]), the fundamental limits of estimation in a distributed setup are far less understood, due to the inherent difficulty imposed by the restrictions on the communication protocol.

In this paper, we study the following distributed estimation problem. Alice and Bob observe infinitely many iid copies of ρ-correlated unit variance random variables, that are either binary symmetric or Gaussian, and where the correlation $\rho \in [-1, 1]$ is unknown. By interactively exchanging k bits on a shared blackboard, Bob wants to produce an estimate $\hat{\rho}$ that is guaranteed to be ϵ-close to ρ (in the sense that $\mathbb{E}[(\hat{\rho} - \rho)^2] \leq \epsilon^2$) regardless of the true underlying value of the correlation. We show that the communication complexity of this task, i.e., the minimal number of bits k that need to be exchanged between Alice and Bob to that end, is $\frac{1+o(1)}{2\epsilon^2 \ln 2}$ in both the binary and Gaussian settings, and one-way schemes are optimal. We also prove a local version of the bound, showing that the communication complexity is still $\Theta_\rho(1/\epsilon^2)$ even if the real correlation is within an interval of vanishing size near ρ.

Let us put our work in context of other results in the literature. The classical problem of communication complexity, originally introduced in a seminal paper by Yao for two parties [Yao79], has been extensively studied in various forms and variations, see e.g. [KN96] and references therein. In its simplest (two-party) form, Alice and Bob wish to compute some given function of their local inputs, either exactly for any input or with high probability over some distribution on the inputs, while interactively exchanging the least possible number of bits. While in this paper we also care about the communication complexity of the task at hand, our setting differs from the classical setup in important ways. First, rather than computing a specific function of *finite* input sequences with a small error probability under a known distribution (or in the worst case), we assume an unknown parametric distribution on *infinite* input

[*] Author order is alphabetical.

STOC '19, June 23–26, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6705-9/19/06…$15.00
https://doi.org/10.1145/3313276.3316332

sequences, and wish to approximate the underlying parameter to within a given precision. In a sense, rather than to compute a function, our task is to interactively extract the most valuable bits from the infinite inputs towards our goal. Notwithstanding the above, an appealing way to cast our problem is to require interactively approximating the function

$$f(\mathbf{X},\mathbf{Y}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \qquad (1)$$

for infinite iid (binary or Gaussian) strings $\mathbf{X}, \mathbf{Y}$, to within precision ϵ, which we show requires $\Theta(1/\epsilon^2)$ bits of interactive communication.

In another related aspect, many traditional communication complexity lower bounds are proved via information-theoretic arguments, most notably by bounding the *information complexity* of good protocols over a suitable choice of distribution over the inputs, see e.g. the classical proof for the disjointness problem [BYJKS04]. Our proofs have a similar information-theoretic flavor; in fact, our key technical contribution is connecting a so-called *symmetric strong data-processing inequality (SDPI)*, previously considered in [LCV17] in the context of interactive secret key generation, to interactive hypothesis testing and estimation problems. Loosely speaking, the symmetric SDPI gives the following bound:

$$\textbf{mutual information } \textbf{\textit{interchanged}} \textbf{ between Alice and Bob}$$
$$\leq$$
$$\rho^2 \times \textbf{mutual information } \textbf{\textit{injected}} \textbf{ by Alice and Bob}$$

This is formalized in our Lemmas 9.1, 9.3 and 9.4, where the upper and lower expressions above correspond to R in eq. (87) and S in eq. (88), respectively. In fact, as a side application of this interactive SDPI, we show an $\Omega(n)$ lower bound on information complexity of the Gap-Hamming problem [IW03],[CR12], which has so far resisted an information-theoretic attack; see Remark 3 for details.

There has also been much contemporary interest in distributed estimation with communication constraints under a different context, where a finite number of iid samples from a distribution belonging to some parametric family are observed by multiple remotely located parties, which in turn can communicate with a data center (either one-way or interactively) in order to obtain an estimate of the underlying parameters, under a communication budget constraint, see e.g. [ZDJW13], [BGM$^+$16], [HÖW18]. These works are markedly different from ours: the samples observed by the parties are taken from the *same* distribution, and the main regime of interest is typically where the dimension of the problem is relatively high compared to the number of local samples (so that each party alone is not too useful), but is low relative to the total number of samples observed by all parties (so that centralized performance is good). The goal is then to communicate efficiently in order to approach the centralized performance. This stands in contrast to our case, where each party observes an *unlimited* number of samples drawn from a *different* component of a bivariate, single-parameter distribution, and the difficulty hence lies in the fact that the quantity of interest (correlation) is truly distributed; none of the parties can estimate it alone, and both parties together can estimate it to arbitrary precision in a centralized setup. Hence, the only bottleneck is imposed by communications.

Another line of works closely related to ours has appeared in the information theory literature, limited almost exclusively to one-way protocols. The problem of distributed parameter estimation under communication constraints has been originally introduced in [ZB88], where the authors provided a rate-distortion-type upper bound on the quadratic error in distributively estimating a scalar parameter using one-way communication (possibly to a third party) under a rate constraint in communication-bits per sample, for a limited set of estimation problems. They have studied our Gaussian setup in particular, and the upper bounds we briefly discuss herein can be deduced (albeit non-constructively) from their work (in [HS18] it is shown how to constructively attain the same performance, and also generalize to the vector parameter case). There has been much followup work on this type of problems, especially in the discrete samples case, see [HA98] and references therein. A local and global minimax setup similar to our own (but again for one-way protocols) has been explored in [AB90]. The local minimax bound we obtain (for one-way protocols) was essentially claimed in that paper, but a subtle yet crucial flaw in their proof of the Fisher information tensorization has been pointed out later in [HA98].

Finally, it is worth noting the closely related problem of distributed hypothesis testing for independence under communication constraints. In [AC86], the authors provide an exact asymptotic characterization of the optimal tradeoff between the rate (in bits per sample) of one-way protocols and the false-alarm error exponent attained under a vanishing mis-detect probability. This result has recently been extended to the interactive setup with a finite number of rounds [XK13].

In fact, some of our lower bounds are also based on a reduction to testing independence with finite communication complexity. For a special case of one-way protocols, this problem was recently analyzed in [ST18]. There is also an inherent connection between the problem of testing independence and generating common randomness from correlated sources, cf. [TW15], as well as between the problem of testing independence and hypercontractivity [Pol12]. For common randomness, two recent (and independent) works [GR16, LCV17] showed that to (almost) agree on L (almost) pure bits the minimal two-way communication required is $(1-\rho^2)L$. There are several differences between the results and techniques in these two works. The work [GR16] followed upon earlier [CGMS17] and considers exponentially small probability of error. Their main tool is hypercontractivity (Concurrently, hypercontractivity bounds for one-way protocols in similar common randomness generation models were also obtained independently in [LCV15][LCCV16]). The lower bound in [GR16] was partial, in the sense that the common randomness generated by Alice was required to be a function of her input and not of the transcript. Thus [GR16] [LCV15][LCCV16] all concern settings where one-way protocols are optimal. In contrast, the work [LCV17] followed on a classical work on interactive compression [Kas85] and showed an unrestricted lower bound. In that setting, one-way communication is not optimal for general sources (although it was numerically verified and proved in the limiting cases that one-way protocols are optimal for binary symmetric sources). The main tool in [LCV17] in the small communication regime was the "symmetric strong data-processing inequality". Here we adapt the latter to our problem.

Organization. In Section 2 we formally present the problem and state our main results. Section 3 contains necessary mathematical background. Section 4 proves that the achievability in the Gaussian case implies the achievability in the binary symmetric case (so that we only need to the achievability for Gaussian and converse for binary). Section 5 proves the upper bounds. Section 6 proves the lower bounds in the special case of one-way protocols (as a warm-up), and Section 7 proves the lower bound in the full interactive case, both for the global risks. Section 8 discusses how to extend to the local version by using common randomness. Section 9 gives the technical proof for the symmetric strong data processing inequality in the binary and Gaussian cases.

2 MAIN RESULTS

We define the problem formally as follows. Alice and Bob observe $\mathbf{X}$ and $\mathbf{Y}$ respectively, where $(\mathbf{X}, \mathbf{Y}) \sim P_{XY}^{\otimes n}$. The distribution P_{XY} belongs to one of the two families, parameterized by a single parameter $\rho \in [-1, 1]$:

(1) Binary case: $X, Y \in \{\pm 1\}$ are unbiased and $\mathbb{P}[X = Y] = \frac{1+\rho}{2}$.
(2) Gaussian case: X, Y are unit-variance ρ-correlated Gaussian.

The communication between Alice and Bob proceeds in rounds: First, Alice writes $W_1 = f_1(\mathbf{X})$ on the board. Bob then writes $W_2 = f_2(\mathbf{Y}, W_1)$ and so on where in the r-th round Alice writes W_r if r is odd, and Bob writes W_r if r is even, where in both cases $W_r = f_r(\mathbf{X}, W_1, \dots, W_{r-1})$. We note that, in principle, we allow each function f_r to also depend on a private randomness (i.e. f_r can be a stochastic map of its arguments). We also note that our impossibility results apply to a slightly more general model where there is also a common randomness in the form of a uniform W_0 on $[0, 1]$ pre-written on the board, but we do not need this for our algorithms.

Let $\Pi = (W_1, W_2, \dots)$ be the contents of the board after all of (possibly infinitely many) rounds. We say that the protocol is *k-bit* if the entropy $H(\Pi) \le k$ for any $\rho \in [-1, 1]$. Note that the protocol is completely characterized by the conditional distribution $P_{\Pi|\mathbf{XY}}$.

At the end of communication, Bob produces an estimate $\hat{\rho}(\Pi, \mathbf{Y})$ for the correlation ρ of the underlying distribution. We are interested in characterizing the tradeoff between the communication size k and the worst-case (over ρ) squared-error, which we call *quadratic risk*, in the regime where the number of samples n is arbitrarily large but k is fixed. Explicitly, the quadratic risk of the protocol Π and the estimator $\hat{\rho}$ is given by

$$R_\rho(\Pi, \hat{\rho}) \triangleq \mathbb{E}_\rho \left(\hat{\rho}(\Pi, \mathbf{Y}) - \rho\right)^2, \tag{2}$$

where $\mathbb{E}_\rho$ is the expectation under the correlation value ρ. Similarly, we write $P_{\mathbf{XY}\Pi}^{\rho}$ for the joint distribution corresponding to a fixed value of ρ. The (global) *minimax risk* is defined as

$$R^* \triangleq \inf_{n, \Pi, \hat{\rho}} \sup_{-1 \le \rho \le 1} R_\rho(\Pi, \hat{\rho}), \tag{3}$$

whereas the *local minimax risk* is

$$R^*_{\rho,\delta} \triangleq \inf_{n, \Pi, \hat{\rho}} \sup_{|\rho' - \rho| \le \delta} R_{\rho'}(\Pi, \hat{\rho}). \tag{4}$$

The infima in both the definitions above are taken over all k-bit protocols Π and estimators $\hat{\rho}$, as well as the number of samples n. We will also discuss *one-way protocols*, i.e. where $\Pi = W_1$ consists of a single message from Alice to Bob. We denote the global and local minimax risk in the one-way case by R^{*1} and $R^{*1}_{\rho,\delta}$ respectively.

Our main results are the following.

Theorem 2.1 (Upper bounds). *In* both *the Gaussian* and *the binary symmetric cases with infinitely many samples,*

$$R^*_{\rho,\delta} \le \frac{1}{k}\left(\frac{\left(1-\rho^2\right)^2}{2\ln 2} + o(1)\right), \tag{5}$$

as long as $\delta = o(1)$ (here and after, $o(1)$ means a vanishing sequence indexed by k), and

$$R^* \le \frac{1}{k}\left(\frac{1}{2\ln 2} + o(1)\right). \tag{6}$$

In fact, one-way protocols achieve these upper bounds.

Remark 1. Previously, [HS18] showed that there exists a one-way protocol and an unbiased estimator achieving $R_\rho(\Pi, \hat{\rho}) \le \frac{1}{k}\left(\frac{1-\rho^2}{2\ln 2} + o(1)\right)$ for any ρ. The protocol (in the Gaussian case) sends the index $\operatorname{argmax}_{1 \le i \le 2^k} X_i$ using k bits and employs the super concentration property of the max. Here, the local risk bound (5) is tighter because we can send the index more efficiently using the side information Y^{2^k} and the knowledge of ρ within $o(1)$ error. Such a scheme has the drawback that it is specially designed for a small interval of ρ (as in the definition of the local risk), and hence the performance may be poor outside that small interval. However, we remark that one can achieve the risk $\frac{1}{k}\left(\frac{(1-\rho^2)^2}{2\ln 2} + o(1)\right)$ at any ρ by a *two-way* protocol. Indeed, Alice can use the first round to send $\omega(1) \cap o(k)$ bits to Bob so that Bob can estimate ρ up to $o(1)$ error. Then Bob can employ the one-way local protocol in (5) for the ρ estimated from the first round.

Theorem 2.2 (lower bounds). *In both the Gaussian and binary symmetric cases with infinitely many samples,*

$$R^*_{\rho,\delta} \ge \frac{(1-|\rho|)^2}{2k\ln 2}(1 + o(1)). \tag{7}$$

In particular, since the global risk dominates the local risk at any ρ, we have

$$R^* \ge \frac{1}{k}\left(\frac{1}{2\ln 2} + o(1)\right). \tag{8}$$

Note in particular that theorems Theorem 2.1 and 2.2 have identified the exact prefactor in the global risk.

Remark 2 (Unbiased estimation)**.** We note that the proof of Theorem 2.2 also implies that for any *unbiased* estimator $\hat{\rho}$ of ρ in the binary case it holds that

$$\operatorname{Var} \hat{\rho} \ge \frac{(1-|\rho|)^2}{2k\ln 2}. \tag{9}$$

We further note that an unbiased estimator with $\operatorname{Var} \hat{\rho} = (1 - \rho^2 + o(1))/(2k\ln 2)$ was introduced in [HS18] (and discussed in Section 5 below), establishing the tightness of (9) at $\rho = 0$.

The bound (9) follows from the Cramér-Rao inequality (see e.g. [VT04]) along with the bound we obtain for the Fisher information given in (72). The associated regularity conditions are discussed in Remark 4.

Remark 3 (Gap-Hamming Problem)**.** In the Gap-Hamming problem [IW03], Alice and Bob are given binary length-n vectors ($\mathbf{X}$ and $\mathbf{Y}$) respectively, with the promise that $\#\{i : X_i \neq Y_i\}$ is either $\leq n/2 - \sqrt{n}$ or $\geq n/2 + \sqrt{n}$. They communicate in (possibly infinitely many) rounds to distinguish these two hypotheses. It was shown in [CR12] (later with simplified proofs in [Vid12], [She12]) that the communication complexity of any protocol that solves Gap-Hamming with small error probability is $\Omega(n)$ (an upper bound of $O(n)$ is trivial). However, whereas many interesting functions in communication complexity have information-theoretic lower bounds, Gap-Hamming has so far resisted an information-theoretic proof, with the exception of the single-round case for which a proof based on SDPI is known [Gar18]. It is nevertheless already known that the information complexity of Gap-Hamming is linear, i.e., that $I(\Pi; \mathbf{X}, \mathbf{Y}) = \Omega(n)$ for any Π that solves it, under the uniform distribution on $(\mathbf{X}, \mathbf{Y})$. This is however observed only indirectly, since the smooth rectangle bound used in the original proof is known to be "below" the information complexity, i.e., any lower bound proved using the smooth rectangle bound also yields a lower bound on information complexity. It is therefore of interest to note that our result in particular directly implies a $\Omega(n)$ lower bound on the information complexity of Gap-Hamming (and hence also on its communication complexity).

To see this, we note that the main step in proving our main result is the inequality

$$D(P^{\rho}_{\mathbf{X}\Pi} \| P^{0}_{\mathbf{X}\Pi}) \leq \rho^2 I(\Pi; \mathbf{X}, \mathbf{Y})\,. \tag{10}$$

which is implied by Theorems 7.1 and 7.2, in Section 7. We note that it implies the $\Omega(n)$ lower-bound on distributional communication *and information* complexity of the Gap-Hamming problem, see [CR12] for references and the original proof of $\Omega(n)$. Indeed, let $U \sim \mathrm{Ber}(1/2)$ and given U let $\mathbf{X}, \mathbf{Y}$ have correlation $\rho = (-1)^U \rho_0$, where $\rho_0 = \frac{100}{\sqrt{n}}$. Take Π to be a protocol used for solving the Gap-Hamming problem (which decides whether $\#\{i : X_i \neq Y_i\}$ is $\leq n/2 - \sqrt{n}$ or $\geq n/2 + \sqrt{n}$ with small error probability). Its decision should equal U with high probability, and hence there exists a decision rule based on Π reconstructing U with high probability of success. Thus $I(U; \Pi) = \Omega(1)$, and we further have

$$I(U; \Pi) \leq I(U; \Pi, \mathbf{X}) \leq \frac{1}{2} D(P^{+\rho_0}_{\mathbf{X}\Pi} \| P^{0}_{\mathbf{X}\Pi}) + \frac{1}{2} D(P^{-\rho_0}_{\mathbf{X}\Pi} \| P^{0}_{\mathbf{X}\Pi}), \tag{11}$$

where the last inequality follows from a property of the mutual information ((14) below). Finally, from (10) we get the statement that $H(\Pi) \geq \Omega(\rho_0^{-2}) = \Omega(n)$.

As pointed out by the anonymous reviewer, a more general version of the Gap-Hamming problem concerns the decision between $\#\{i\colon X_i \neq Y_i\} \leq n/2 - g$ and $n/2 + g$ for some $\sqrt{n} \leq g \leq n/2$, and it was shown in [CR12, Proposition 4.4] that the communication complexity is $\Omega(n^2/g^2)$. This result can also be recovered by the above argument. And notably, this result also implies the $R^* = \Omega(1/k)$ lower bound.

3 PRELIMINARIES

3.1 Notation

Lower- and upper-case letters indicate deterministic and random variables respectively, with boldface used to indicate n-dimensional vectors. For any positive integer r, the set $\{1, 2, \ldots, r\}$ is denoted by $[r]$. Let P and Q be two probability distributions over the same probability space. The *KL divergence* between P and Q is

$$D(P \| Q) = \int \log \left(\frac{dP}{dQ} \right) dP \tag{12}$$

with the convention that $D(P\|Q) = \infty$ if P is not absolutely continuous w.r.t. Q. Logarithms are taken to the base 2 throughout, unless otherwise stated. With this definition, the mutual information between two jointly distributed r.v.s $(X, Y) \sim P_{XY}$ can be defined

$$I(X; Y) = D(P_{XY} \| P_X \times P_Y), \tag{13}$$

and it satisfies the "radius" property:

$$I(X; Y) = \inf_{Q_Y} D(P_{Y|X} \| Q_Y | P_X), \tag{14}$$

where the conditioning means taking the expectation of the conditional KL divergence w.r.t. P_X. Given a triplet of jointly distributed r.v.s (X, Y, Z), the conditional mutual information between X and Y given Z is

$$I(X; Y|Z) = D(P_{XY|Z} \| P_{X|Z} \times P_{Y|Z} | P_Z) \tag{15}$$

We will say that r.v.s A, B, C form a Markov chain $A - B - C$, if A is independent of C given B.

3.2 Symmetric strong data-processing inequalities

Given P_{XY}, the standard data processing inequality states that $I(U; Y) \leq I(U; X)$ for any U satisfying $U - X - Y$. Recall that a *strong data processing inequality* (see e.g. [PW17]) is satisfied if there exists $s \in [0, 1)$ depending on P_{XY} such that $I(U; Y) \leq sI(U; X)$ for any U satisfying $U - X - Y$.

The connection between the strong data processing and communication complexity problems is natural, and U can be thought of as the message from Alice to Bob, $I(U; X)$ the communication complexity, and $I(U; Y)$ the information for the estimator. However, the best constant s in the strong data processing inequality is not symmetric (i.e. $s(P_{XY}) = s(P_{YX})$ is not true for general P_{XY}), whereas the performance in an interactive communication problems is by definition symmetric w.r.t. the two parties. An inequality of the following form, termed "symmetric strong data processing inequality" in [LCV17], plays a central role in interactive communication problems:

$$\begin{aligned} &I(U_1; Y) + I(U_2; X|U_1) + I(U_3; Y|U^2) + \ldots \\ &\leq s_\infty [I(U_1; X) + I(U_2; Y|U_1) + I(U_3; X|U^2) + \ldots] \end{aligned} \tag{16}$$

where $U_1, U_2, \ldots$ must satisfy

$$U_r - (X, U^{r-1}) - Y, \quad r \in \{1, 2, \ldots\} \setminus 2\mathbb{Z}, \tag{17}$$

$$U_r - (Y, U^{r-1}) - X, \quad r \in \{1, 2, \ldots\} \cap 2\mathbb{Z}, \tag{18}$$

and where s_∞ depends only on P_{XY}. Clearly $s_\infty(P_{XY}) = s_\infty(P_{YX})$ and $s_\infty \geq s$. A succinct characterization of s_∞ in terms of the "marginally convex envelope" was reported in [LCV17]. Using the Markov assumptions (17)-(18) we can also rewrite (16) as

$$I(X; Y) - I(X; Y|U) \leq s_\infty I(U; X, Y). \tag{19}$$

When X, Y are iid binary symmetric vectors with correlation ρ^2 per coordinate, it was shown in [LCV17] that $s_\infty = \rho^2$, equal to the strong data processing constant. In this paper, we extend the result to Gaussian vectors with correlation ρ per coordinate (Theorem 7.2).

In order to upper bound s_∞ in the binary case, [LCV17] observed that $s_\infty(Q_{XY})$ is upper bounded by the supremum of $s(P_{XY})$ over P_{XY} a "marginally titled" version of Q_{XY}. Indeed, note that the Markov structure in the strong data processing inequality implies that $P_{XY|U=u}(x,y) = f(x)P_{XY}(x,y)$ for some function f. In the case of symmetric strong data processing inequality, the Markov conditions (17) and (18) imply that $P_{XY|U^r=u^r}(x,y) = f(x)P_{XY}(x,y)g(y)$, which naturally lead one to considering the following result:

Lemma 3.1 ([LCV17, Theorem 6]). *Let Q_{XY} be the distribution of a binary symmetric random variables with correlation $\rho \in [-1,1]$, i.e. $Q_{XY}(x,y) = \frac{1}{4}(1+(-1)^{1\{x\neq y\}}\rho)$ for $x,y \in \{0,1\}$. Let $(X,Y) \sim P_{XY}$ have an arbitrary distribution of the form*

$$P_{XY}(x,y) = f(x)g(y)Q_{XY}(x,y)\,.$$

Then for any $U - X - Y - V$ we have

$$I(U;Y) \le \rho^2 I(U;X) \tag{20}$$

$$I(X;V) \le \rho^2 I(Y;V)\,. \tag{21}$$

Lemma 3.1 was proved in [LCV17] by exploring the connection to the maximal correlation coefficient. In Section 9.1 we give another proof using properties of the strong data processing inequalities [PW17].

3.3 Fisher information and Cramèr-Rao inequalities

We recall some standard results from parameter estimation theory. Let θ be a real-valued parameter taking an unknown value in some interval $[a,b]$. We observe some random variable (or vector) X with distribution $P(x|\theta)$ parameterized by θ.

Assume that $P(\cdot|\theta)$ is absolutely continuous with respect to a reference measure μ, for each $\theta \in [a,b]$, and $\frac{dP(\cdot|\theta)}{d\mu}(x)$ is differentiable with respect to $\theta \in (a,b)$ for μ-almost all x. Then the *Fisher information* of θ w.r.t. X, denoted as $\mathrm{I}_F(X;\theta)$, is the variance of the derivative of the log-likelihood w.r.t. θ,

$$\mathrm{I}_F(X;\theta) \triangleq \int \left(\frac{\partial}{\partial\theta}\ln\frac{dP(\cdot|\theta)}{d\mu}(x)\right)^2 dP(x|\theta). \tag{22}$$

We now record some useful facts concerning the Fisher information. First, we recall that the Fisher information encodes the curvature of the KL divergence w.r.t. translation: Let

$$g(\theta,\epsilon) \triangleq D\left(P(x|\theta)\|P(x|\theta+\epsilon)\right) \tag{23}$$

for any $\theta, \theta+\varepsilon \in (a,b)$. The following property is well-known:

Lemma 3.2. *Under suitable regularity conditions, $\frac{\partial}{\partial\varepsilon}g(\theta,\varepsilon)|_{\epsilon=0} = 0$, and*

$$\mathrm{I}_F(X;\theta) = \ln 2\cdot \left.\frac{\partial^2 g(\theta,\varepsilon)}{\partial\varepsilon^2}\right|_{\varepsilon=0}, \tag{24}$$

which implies that

$$\mathrm{I}_F(X;\theta) = 2\ln 2\cdot \lim_{\varepsilon\to 0}\frac{g(\theta,\varepsilon)}{\varepsilon^2}. \tag{25}$$

Remark 4. The "regularity conditions" in Lemma 3.2 (and Lemma 3.3 below) are to ensure that one can apply the dominated convergence theorem to exchange certain integrals and differentiations in the calculus. See for example [Kul, Section 2.6] for details. In particular, these conditions are fulfilled if $\sup_{x,\theta}\frac{dP(\cdot|\theta)}{d\mu}(x) < \infty$, $\inf_{x,\theta}\frac{dP(\cdot|\theta)}{d\mu}(x) > 0$, and $\sup_{x,\theta}\frac{\partial^m}{\partial\theta^m}\left[\frac{dP(\cdot|\theta)}{d\mu}(x)\right] < \infty$ for $m = 1,2,3$. In the interactive estimation problem, these conditions are always satisfied for sources $(\mathbf{X},\mathbf{Y})$ on finite alphabets (even if the message alphabets are not finite). Indeed, suppose that $\mathbf{X},\mathbf{Y}$ are binary vectors, and that Alice performs an estimation. Let the reference measure $\mu = P^0(\Pi,\mathbf{X})$ be the distribution under $\rho = 0$. We have that

$$\frac{dP^\rho}{d\mu}(\Pi,\mathbf{x}) = \frac{\sum_{\mathbf{y}}P(\Pi|\mathbf{x},\mathbf{y})P^\rho(\mathbf{x},\mathbf{y})}{\sum_{\mathbf{y}}P(\Pi|\mathbf{x},\mathbf{y})P^0(\mathbf{x},\mathbf{y})} \le \sup_{\mathbf{x},\mathbf{y}}\frac{P^\rho(\mathbf{x},\mathbf{y})}{P^0(\mathbf{x},\mathbf{y})} \tag{26}$$

is bounded by a value independent of Π. Similarly,

$$\frac{dP^\rho}{d\mu}(\Pi,\mathbf{x}) \ge \inf_{\mathbf{x},\mathbf{y}}\frac{P^\rho(\mathbf{x},\mathbf{y})}{P^0(\mathbf{x},\mathbf{y})}, \tag{27}$$

$$\frac{\partial^m}{\partial\rho^m}\left[\frac{dP^\rho}{d\mu}(\Pi,\mathbf{x})\right] \le \sup_{\mathbf{x},\mathbf{y}}\frac{\frac{\partial^m}{\partial\rho^m}P^\rho(\mathbf{x},\mathbf{y})}{P^0(\mathbf{x},\mathbf{y})}. \tag{28}$$

The Fisher information can be used to lower bound the *expected* quadratic risk of estimating θ from X under a prior distribution on θ.

Lemma 3.3 (Bayesian Cramér-Rao inequality, see e.g. [VT04]). *Let λ be an absolutely continuous density on a closed interval $\mathcal{J} \subseteq [a,b]$, and assume λ vanishes at both endpoints of $\mathcal{J}$. If $P(x|\theta)$ satisfies suitable regularity conditions and $\mathrm{I}_F(X;\theta) < \infty$ for almost all θ,*

$$\mathbb{E}_{\theta\sim\lambda}\mathbb{E}_\theta(\hat{\theta}(X)-\theta)^2 \ge \frac{1}{\mathrm{I}^\lambda + \mathbb{E}_{\theta\sim\lambda}\mathrm{I}_F(X;\theta)} \tag{29}$$

for any estimator $\hat{\theta}$, where $\mathrm{I}^\lambda = \int_{\mathcal{J}}\frac{\lambda'^2}{\lambda}d\theta$.

A common choice of prior (see e.g. [Tsy09]) is

$$\lambda = \frac{2}{|\mathcal{J}|}\lambda_0\left(\frac{\theta-\theta_0}{|\mathcal{J}|/2}\right) \tag{30}$$

where θ_0 is the center of the interval $\mathcal{J}$, and $\lambda_0(x) = \cos^2(\pi x/2)$ for $-1 \le x \le 1$ and 0 otherwise. This prior satisfies $\mathrm{I}^\lambda = (2\pi/|\mathcal{J}|)^2$.

4 REDUCTION OF BINARY TO GAUSSIAN

In this section we show that an achievability scheme for iid Gaussian vector can be converted to a scheme for binary vector by a preprocessing step and applying the central limit theorem (CLT). We remark that a similar argument was used in [LCV17] in the context of common randomness generation.

Lemma 4.1. *Suppose that $(\Pi,\hat{\rho})$ is a scheme for iid sequence of Gaussian pairs at some length n, and the message alphabet size $|\Pi| < \infty$. Then there exists a scheme $(\Pi^T,\hat{\rho}^T)$ for iid sequence of binary symmetric pairs of length T, for each $T = 1,2,\dots,$ such that*

$$\lim_{T\to\infty}H(\Pi^T) = H(\Pi), \quad \forall\rho\in[-1,1], \tag{31}$$

$$\lim_{T\to\infty}R_\rho(\Pi^T,\hat{\rho}^T) \le R_\rho(\Pi,\hat{\rho}), \quad \forall\rho\in[-1,1], \tag{32}$$

where ρ denotes the correlation of the Gaussian or binary pair.

Proof. Let $(A_l, B_l)_{l=1}^t$ be an iid sequence of binary symmetric random variables with correlation ρ, and put

$$X^{(t)} := \frac{A_1 + \cdots + A_t}{\sqrt{t}} + a_t N, \tag{33}$$

$$Y^{(t)} := \frac{B_1 + \cdots + B_t}{\sqrt{t}} + a_t N', \tag{34}$$

where N and N' are standard Gaussian random variables, and N, N', (X^t, Y^t) are independent. By the central limit theorem, we can choose some $a_t = o(1)$ such that the distribution of $(X^{(t)}, Y^{(t)})$ converges to the Gaussian distribution P_{XY} in total variation (Proposition 4.2 below). Now let $T = nt$ and suppose that $(A_l, B_l)_{l=1}^T$ is an iid sequence of binary symmetric pairs. The above argument shows that Alice and Bob can process locally to obtain iid sequence of length n, which convergences to the iid sequence of Gaussian pairs of correlation ρ in the total variation distance. After this preprocessing step, Alice and Bob can apply the given scheme $(\Pi, \hat{\rho})$. Then (31) follows since entropy is continuous w.r.t. the total variation on finite alphabets, and (32) follows since we can assume without loss of generality that $\hat{\rho}$ is bounded. Note that we have constructed $(\Pi^T, \hat{\rho}^T)$ only for T equal to a multiple of n; however this restriction is obviously inconsequential. □

Proposition 4.2. *There exist $a_t = o(1)$ such that $X^{(t)}$ and $Y^{(t)}$ defined in (33) and (34) converges to the Gaussian distribution P_{XY} in total variation.*

Proof. By the convexity of the relative entropy, we can upper bound the KL divergence by the Wasserstein 2 distance:

$$D(X^{(t)}, Y^{(t)} \| X + a_t N, Y + a_t N')$$
$$\leq \frac{1}{2a_t^2} W_2^2 \left(\left[\frac{A_1 + \cdots + A_t}{\sqrt{t}}, \frac{B_1 + \cdots + B_t}{\sqrt{t}} \right], [X, Y] \right) \tag{35}$$

However, $\frac{A_1+\cdots+A_t}{\sqrt{t}}$ and $\frac{B_1+\cdots+B_t}{\sqrt{t}}$ converge to P_{XY} under Wasserstein 2 distance, since this is equivalent to convergence in distribution in the current context where a uniformly integrable condition is satisfied (see e.g. [Vil03, Theorem 7.12]) [1]. Thus there exists $a_t = o(1)$ such that (35) vanishes. By Pinsker's inequality, this implies that $(X^{(t)}, Y^{(t)})$ converges to the Gaussian distribution $(X + a_t N, Y + a_t N')$ in total variation. However, as long as $a_t = o(1)$ we have that $(X + a_t N, Y + a_t N')$ converges to (X, Y). The conclusion then follows by the triangle inequality of the total variation. □

5 PROOF OF THE UPPER BOUNDS (THEOREM 2.1)

Before the proof, let us observe the suboptimality of a naive scheme. Consider the binary case for example (the Gaussian case is similar). Suppose that Alice just sends her first k samples $X_1, \ldots, X_k$. This would let Bob, by computing the empirical average $\hat{\rho}_{\text{emp}} = \frac{1}{k} \sum_j X_j Y_j$, achieve a risk of

$$\mathbb{E}_\rho[|\rho - \hat{\rho}_{\text{emp}}|^2] = \frac{1-\rho^2}{k}. \tag{36}$$

Clearly (36) is not sufficient for the upper bounds in Theorem 2.1. To improve it, we now recall the "max of Gaussian scheme" in [HS18]. By a central limit theorem argument we can show that binary estimation is easier than the Gaussian counterpart (see Lemma 4.1). Hence we only need to prove the achievability for the Gaussian case. Alice observes the first 2^k Gaussian samples, and transmits to Bob, using exactly k bits, the *index* W of the maximal one, i.e.

$$W = \operatorname*{argmax}_{i \in [2^k]} X_i. \tag{37}$$

Upon receiving the index W, Bob finds his corresponding sample Y_W and estimates the correlation using

$$\hat{\rho}_{\max} = \frac{Y_W}{\mathbb{E} X_W}. \tag{38}$$

Recall the following result [HS18], for which we reproduce the short proof and then explain how the local upper bound will follow with a modification of the proof.

Theorem 5.1 ([HS18]). *The estimator $\hat{\rho}_{\max}$ is unbiased with*

$$R_\rho(W, \hat{\rho}_{\max}) = \frac{1}{k} \left(\frac{1-\rho^2}{2\ln 2} + o(1) \right). \tag{39}$$

Proof. It is easy to check that $\hat{\rho}_{\max}$ is unbiased. In order to compute its variance, we need to compute the mean and variance of X_W, which is the maximum of 2^k iid standard normal r.v.s. From extreme value theory (see e.g. [DN04]) applied to the normal distribution, we obtain

$$\mathbb{E} X_W = \sqrt{2\ln(2^k)}(1 + o(1)) \tag{40}$$

$$\mathbb{E} X_W^2 = 2\ln(2^k)(1 + o(1)) \tag{41}$$

$$\operatorname{Var} X_W = O\left(\frac{1}{\ln(2^k)} \right). \tag{42}$$

Therefore, for $Z \sim \mathcal{N}(0,1)$ we have that

$$\operatorname{Var} \hat{\rho}_{\max} = \frac{1}{(\mathbb{E} X_W)^2} \operatorname{Var}(\rho X_W + \sqrt{1-\rho^2} Z) \tag{43}$$

$$= \frac{1}{(\mathbb{E} X_W)^2} (\rho^2 \operatorname{Var} X_W + 1 - \rho^2) \tag{44}$$

$$= \frac{1}{2k\ln 2} (1 - \rho^2 + o(1)). \tag{45}$$

□

Taking $\rho = 0$ in (39) establishes the global upper bound (6). However, achieving the local risk upper bound in (5) is trickier, since a direct application of (39) is loose by a factor of $(1-\rho^2)$. The trick is to send the index W more efficiently using the side information. More precisely, Alice looks for the maximum sample out of 2^k samples as before. Bob sifts his corresponding samples, marking only those where $Y_k > \rho \cdot \sqrt{k \cdot 2\ln 2} \cdot (1 - o(1))$. Note that here ρ is as in the definition of the local risk (4), and the true correlation is within $o(1)$ error to ρ. It is easy to check that with sufficiently high probability (sufficiently here (and below) meaning that the complimentary probability has a negligible effect on the ultimate variance), there are $2^{k(1-\rho^2)(1+o(1))}$ such marked samples that also include the one corresponding to Alice's maximum. Also, by symmetry these marked samples are uniformly distributed among the total 2^k samples. Hence, Alice can describe the $k \cdot (1-\rho^2) \cdot (1 + o(1))$ most

[1] Alternatively, see [MT74] for a direct proof of the central limit theorem under the Wasserstein metric.

significant bits (say) of the index of her maximal sample, which will reveal this index to Bob with sufficiently high probability. This yields a $(1-\rho^2)$ factor saving in communication, and the claim follows.

Remark 5. We note that the above risk can also be achieved directly in Hamming space (without appealing to the CLT). Alice sets some parameter $\tilde{\rho} \in [-1,1]$ to be optimized later, and partitions her data to m blocks of size n. She then finds the first block whose sum is exactly $n\tilde{\rho}$ (recall the samples are in $\{-1,1\}$), which exists with sufficiently high probability for $m = 2^{n(\frac{1}{2}-h(\frac{1-\tilde{\rho}}{2})+o(1))}$ (otherwise, she picks the first block). Bob sifts his corresponding blocks, marking only those with sum $n\rho\tilde{\rho}(1+o(1))$. Alice encodes the index of her chosen block using $\log m = n(\frac{1}{2}-h(\frac{1-\tilde{\rho}}{2})+o(1))$ bits, and sends only the $n(h(\frac{1-\rho\tilde{\rho}}{2})-h(\frac{1-\tilde{\rho}}{2}))$ most significant bits, so that Bob can resolve the index with sufficiently high probability. Bob then finds the sum of his corresponding block, and divides it by $n\tilde{\rho}$ to obtain his estimator for ρ. It is straightforward to check that this procedure results in a variance of $\frac{1}{k}\cdot\left((1-\rho^2)(h(\frac{1-\rho\tilde{\rho}}{2})-h(\frac{1-\tilde{\rho}}{2}))/\tilde{\rho}^2+o(1)\right)$, where $h(q) = -q\log_2 q - (1-q)\log_2(1-q)$ is the binary entropy function. Optimizing over $\tilde{\rho}$ yields that $\tilde{\rho} \longrightarrow 0$ is (not surprisingly) optimal, and the obtained variance is the same as the one achieved by the modified Gaussian maximum estimator above.

6 PROOF OF THE LOWER BOUNDS IN THEOREM 2.2 (ONE-WAY CASE)

In this section we prove the lower bounds on the global and local one-way risks, R^{*1} and $R^{*1}_{\rho,\delta}$, in the binary case. The Gaussian case will then follow from the central limit theorem argument in Lemma 4.1. Of course, the one-way lower bound is a special case of the interactive case in Section 7; we separate the discussion simply because the one-way case is conceptually easily and the proof does not need the symmetric strong data processing inequality (modulus certain technical issues pertaining the continuity of Fisher information which we will discuss).

We note that in the one-way setting, the following Markov chain holds:

$$\Pi - \mathbf{X} - \mathbf{Y}. \tag{46}$$

Note that regardless of ρ the marginal distribution of $\mathbf{X}$ (and thus of Π) is the same. Let $P^{\rho}_{\Pi\mathbf{Y}}$ denote the joint distribution of $(\Pi,\mathbf{Y})$ when the correlation is equal to ρ. Note that under $\rho = 0$ we have that Π and $\mathbf{Y}$ are independent. Thus, via (13) we obtain

$$D(P^{\rho}_{\Pi\mathbf{Y}} \| P^{0}_{\Pi\mathbf{Y}}) = I(\Pi;\mathbf{Y})\,. \tag{47}$$

Furthermore, from (90) we get

$$I(\Pi;\mathbf{Y}) \le \rho^2 I(\Pi;\mathbf{X}) \le \rho^2 H(\Pi) \le \rho^2 k. \tag{48}$$

Thus using the connection between the KL divergence and the Fisher information in Lemma 3.2, we obtain

$$\mathrm{I}_F(\Pi,\mathbf{Y};\rho=0) \le k2\ln 2\,. \tag{49}$$

Now, suppose that we can show a continuity result for the Fisher information at $\rho = 0$, in the sense of

$$\limsup_{\rho\to 0}\sup_{\Pi}\mathrm{I}_F(\Pi,\mathbf{Y};\rho) \le k2\ln 2 \tag{50}$$

then a standard application of the Bayesian Cramér-Rao bound would imply the global risk. Indeed, applying Lemma 3.3 with (e.g.) the prior specified in (30) over $\mathcal{J} = [\rho-\delta,\rho+\delta]$, we obtain

$$R^{*1} \ge \frac{1}{k}\left(\frac{1}{2\ln 2} - o(1)\right), \tag{51}$$

for $\delta \in o(1)\cap\omega(1/\sqrt{k})$, establishing (8) for the special case of one-way protocols.

While the continuity claim in (50) is intuitive enough, to rigorously show it we need to resort to a device to be discussed in Section 8, which will allow us to reduce the problem of testing against an arbitrary ρ to testing against independence. Specifically, in Corollary 8.1 we will show that using common randomness this can be generalized to yield

$$D(P^{\rho_1}_{\Pi\mathbf{Y}} \| P^{\rho_0}_{\Pi\mathbf{Y}}) \le \left(\frac{\rho_1-\rho_0}{1-|\rho_0|}\right)^2 k \tag{52}$$

for any $\rho_1 \in [-1,1]$ and $\rho_0 \in [\frac{\rho_1-1}{2},\frac{\rho_1+1}{2}]$. Again applying Lemma 3.2, we obtain

$$\mathrm{I}_F(\Pi,\mathbf{Y};\rho) \le \frac{2k\ln 2}{(1-|\rho|)^2} \tag{53}$$

for any $\rho \in (-1,1)$. This justifies the continuity claim (50). Moreover, applying the Bayesian Cramér-Rao (Lemma 3.3) with (e.g.) the prior specified in (30) over $\mathcal{J} = [\rho-\delta,\rho+\delta]$, we obtain

$$R^{*1}_{\rho,\delta} \ge \frac{1}{k}\left(\frac{(1-|\rho|)^2}{2\ln 2} - o(1)\right) \tag{54}$$

which is the desired local risk lower bound for the special case of one-way protocols.

7 PROOF OF LOWER BOUNDS IN THEOREM 2.2 (INTERACTIVE CASE)

For the interactive case, our approach is again to upper bound the KL divergence between the distributions of the r.v.s available (to either Alice or Bob) under $\rho \ne 0$, and under $\rho = 0$. This is accomplished by Theorem 7.1 and Theorem 7.2 below, which can be viewed as generalizations of (47) and (48).

Theorem 7.1. *Consider an arbitrary interactive protocol $P_{\Pi|\mathbf{XY}}$ and let $P_{\mathbf{XY}\Pi}$ be the induced joint distribution. Let $\bar{P}_{\mathbf{XY}\Pi} = P_{\mathbf{X}} \times P_{\mathbf{Y}} \times P_{\Pi|\mathbf{XY}}$ be the joint distribution induced by the same protocol, but when the $\mathbf{X}$ and $\mathbf{Y}$ are taken to be independent (but with same marginals). Then*

$$\max\{D(P_{\Pi\mathbf{X}}\|\bar{P}_{\Pi\mathbf{X}}),\, D(P_{\Pi\mathbf{Y}}\|\bar{P}_{\Pi\mathbf{Y}})\} \le I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}|\Pi), \tag{55}$$

where information quantities are computed with respect to $P_{\mathbf{XY}\Pi}$. Moreover, the bound (55) continues to hold also when the protocol Π contains an arbitrary common randomness (i.e. public coin) W_0 independent of $(\mathbf{X},\mathbf{Y})$.

By saying that the protocol contains common randomness we mean that $\Pi = (W_0, W_1, \ldots, W_r)$ where W_0 is the common randomness and $W_1, \ldots, W_r$ are the exchanged messages. The extension to the case of common randomness will be useful in Section 8 where we reduce the problem of testing against an arbitrary ρ to testing against independence.

Proof of Theorem 7.1. First, since

$$D(P_{\Pi \mathbf{X} W_0} \| \bar{P}_{\Pi \mathbf{X} W_0}) = D(P_{\Pi \mathbf{X}|W_0} \| \bar{P}_{\Pi \mathbf{X}|W_0} | P_{W_0}), \tag{56}$$

it suffices to prove the same upper bound for

$$D(P_{\Pi \mathbf{X}|W_0=w_0} \| \bar{P}_{\Pi \mathbf{X}|W_0=w_0}). \tag{57}$$

In other words, it suffices to prove the theorem for the case where the common randomness W_0 is empty. Under this assumption, note that the RHS of (55) is equal to

$$\begin{aligned} & I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}|\Pi) \\ &= I(\mathbf{X};\Pi) + I(\mathbf{Y};\Pi) - I(\mathbf{X},\mathbf{Y};\Pi) && (58) \\ &= \mathbb{E}\left[\log \frac{P_{\mathbf{X}\Pi}(\mathbf{X},\Pi) P_{\mathbf{Y}\Pi}(\mathbf{Y},\Pi) P_{\mathbf{X}\mathbf{Y}}(\mathbf{X},\mathbf{Y}) P_{\Pi}(\Pi)}{P_{\mathbf{X}}(\mathbf{X}) P_{\Pi}(\Pi) P_{\mathbf{Y}}(\mathbf{Y}) P_{\Pi}(\Pi) P_{\mathbf{X}\mathbf{Y}\Pi}(\mathbf{X},\mathbf{Y},\Pi)}\right] && (59) \\ &= \mathbb{E}\left[\log \frac{P_{\mathbf{X}\Pi}(\mathbf{X},\Pi) P_{\mathbf{Y}|\Pi}(\mathbf{Y}|\Pi)}{\bar{P}_{\mathbf{X}\mathbf{Y}\Pi}(\mathbf{X},\mathbf{Y},\Pi)}\right] && (60) \\ &= \mathbb{E}\left[\log \frac{P_{\mathbf{X}\Pi}(\mathbf{X},\Pi)}{\bar{P}_{\mathbf{X}\Pi}(\mathbf{X},\Pi)} + \log \frac{P_{\mathbf{Y}|\Pi}(\mathbf{Y}|\Pi)}{\bar{P}_{\mathbf{Y}|\mathbf{X}\Pi}(\mathbf{Y}|\mathbf{X},\Pi)}\right] && (61) \\ &= D(P_{\mathbf{X}\Pi} \| \bar{P}_{\mathbf{X}\Pi}) + \mathbb{E}\left[\log \frac{P_{\mathbf{Y}|\Pi}(\mathbf{Y}|\Pi)}{\bar{P}_{\mathbf{Y}|\mathbf{X}\Pi}(\mathbf{Y}|\mathbf{X},\Pi)}\right] && (62) \\ &\geq D(P_{\mathbf{X}\Pi} \| \bar{P}_{\mathbf{X}\Pi})\,. && (63) \end{aligned}$$

where all expectations are taken with respect to $P_{\mathbf{X}\mathbf{Y}\Pi}$ and the last step is by non-negativity of divergence $D(P_{\mathbf{Y}|\Pi=\pi} \| \bar{P}_{\mathbf{Y}|\Pi=\pi,\mathbf{X}=\mathbf{x}})$ for all $\pi, \mathbf{x}$, which in turn uses the Markov chain $\mathbf{X} - \Pi - \mathbf{Y}$ under $\bar{P}_{\mathbf{X}\Pi\mathbf{Y}}$. In all, (63) proves part of (55). To prove the same bound on $D(P_{\Pi\mathbf{Y}} \| \bar{P}_{\Pi\mathbf{Y}})$ we can argue by symmetry (it may seem that symmetry is broken by the fact that $\mathbf{X}$ sends W_1 first, but this is not true: W_1 can be empty), or just perform a straightforward modification of step (61). □

Remark 6. It can be seen that for a one-way protocol we have equality in (55). This explains why our impossibility bound can be essentially achieved by a one-way protocol (e.g., see Theorem 5.1), and suggests that this is the only possibility.

Remark 7. After completion of this work, we found out that in a slightly different form Theorem 7.1 has previously appeared in [XK13, Equation (4)]. Our proof is slightly simpler.

Theorem 7.2. *Let Π be any interactive protocol, possibly containing a common randomness W_0, in either the Gaussian or the binary symmetric case. Then*

$$I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}|\Pi) \leq \rho^2 I(\Pi;\mathbf{X},\mathbf{Y}). \tag{64}$$

The proof of Theorem 7.2 is given in Section 9.

Remark 8. The following notions of external and internal information costs were introduced in [CSWY01] and [BBCR10] respectively:

$$\mathsf{IC}^{\text{ext}}{}_P(\Pi) := I(\Pi;\mathbf{X},\mathbf{Y}); \tag{65}$$

$$\mathsf{IC}_P(\Pi) := \mathsf{IC}^{\text{ext}}{}_P(\Pi) - [I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}|\Pi)]. \tag{66}$$

Using the Markov chain conditions of the messages

$$W_i - (\mathbf{X}, W^{i-1}) - \mathbf{Y}, \quad i \in [r] \setminus 2\mathbb{Z}, \tag{67}$$

$$W_i - (\mathbf{Y}, W^{i-1}) - \mathbf{X}, \quad i \in [r] \cap 2\mathbb{Z} \tag{68}$$

we will be able to write the external and internal information as sums of information gains in each round of communication:

$$I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}|\Pi) = \sum_{i \in [r]\setminus 2\mathbb{Z}} I(W_i;\mathbf{Y}|W^{i-1}) + \sum_{i \in [r]\cap 2\mathbb{Z}} I(W_i;\mathbf{X}|W^{i-1}), \tag{69}$$

$$I(\Pi;\mathbf{X},\mathbf{Y}) = \sum_{i \in [r]\setminus 2\mathbb{Z}} I(W_i;\mathbf{X}|W^{i-1}) + \sum_{i \in [r]\cap 2\mathbb{Z}} I(W_i;\mathbf{Y}|W^{i-1}), \tag{70}$$

which are useful later in some proofs.

8 REDUCTION OF TESTING AGAINST ARBITRARY ρ TO TESTING AGAINST INDEPENDENCE

The results in Theorem 7.1 and Theorem 7.2 only (directly) applies to testing against independence, and hence are insufficient for handling the local risks at an arbitrary ρ. Fortunately, for binary and Gaussian vectors, there is a simple device of translating the correlations by leveraging the common randomness, so that the general problem is reduced to the case of testing independence solved in Theorem 7.1 and Theorem 7.2. More precisely, we obtained the following result:

Corollary 8.1. *Let $P^{\rho_0}_{\mathbf{X}\mathbf{Y}}$ (resp. $P^{\rho_1}_{\mathbf{X}\mathbf{Y}}$) be the joint distribution for Gaussian or binary symmetric vector sources under correlation ρ_0 (resp. ρ_1). Let $P_{\Pi|\mathbf{X}\mathbf{Y}}$ be an arbitrary protocol. Then for any $\rho_1 \in [-1,1]$ and $\rho_0 \in [\frac{\rho_1-1}{2}, \frac{\rho_1+1}{2}]$,*

$$\max\{D(P^{\rho_1}_{\Pi\mathbf{X}} \| P^{\rho_0}_{\Pi\mathbf{X}}),\ D(P^{\rho_1}_{\Pi\mathbf{Y}} \| P^{\rho_0}_{\Pi\mathbf{Y}})\} \leq \left(\frac{\rho_1 - \rho_0}{1 - |\rho_0|}\right)^2 k. \tag{71}$$

In particular, this bounds the Fisher information in the case of finite-length binary vectors as

$$\max\{\mathrm{I}_F(\Pi,\mathbf{X};\rho), \mathrm{I}_F(\Pi,\mathbf{Y};\rho)\} \leq \frac{2k\ln 2}{(1-|\rho|)^2}. \tag{72}$$

Proof. From Theorems 7.1 and 7.2 we have

$$\max\{D(P^{\rho}_{\Pi\mathbf{X}} \| P^{0}_{\Pi\mathbf{X}}),\ D(P^{\rho}_{\Pi\mathbf{Y}} \| P^{0}_{\Pi\mathbf{Y}})\} \leq \rho^2 I(\Pi;\mathbf{X},\mathbf{Y}) \leq \rho^2 k. \tag{73}$$

The proof uses a device of shifting the correlation by introducing common randomness. Suppose that $\mathbf{X}$ and $\mathbf{Y}$ are iid binary or Gaussian vectors of length n, where the correlation between X_i and Y_i is 0 under $P^{(0)}$ and $\rho := \frac{\rho_1-\rho_0}{1-|\rho_0|}$ under $P^{(1)}$, for each $i \in [n]$. We define the common randomness W_0 independent of $\mathbf{X}$ and $\mathbf{Y}$ as follows:

- Gaussian case: Let $W_0 = \mathbf{Z}$, where $Z_i \sim \mathcal{N}(0,1)$ are iid, and define
$$X'_i = \alpha Z_i + \sqrt{1-\alpha^2} X_i \tag{74}$$
$$Y'_i = s\alpha Z_i + \sqrt{1-\alpha^2} Y_i \tag{75}$$
for some $\alpha \in [-1,1]$ and $s \in \{-1,1\}$.
- Binary case: Let $W_0 = (\mathbf{B}, \mathbf{Z})$ where $\mathbf{B}$ is independent of $\mathbf{Z}$, $B_i \sim \mathrm{Ber}(\alpha)$ over $\{0,1\}$ are iid and $Z_i \sim \mathrm{Ber}(\frac{1}{2})$ over $\{-1,1\}$ are iid. Put
$$X'_i = B_i Z_i + (1-B_i) X_i \tag{76}$$
$$Y'_i = sB_i Z_i + (1-B_i) Y_i \tag{77}$$
for some $\alpha \in [0,1]$ and $s \in \{-1,1\}$.

In both cases, it can be verified that by appropriately choosing s and α, the correlation between X'_i and Y'_i equals ρ_0 under $P^{(0)}$ and ρ_1 under $P^{(1)}$. Now, consider any protocol $\Pi = (W_0, W^r)$ for the source $\mathbf{X}, \mathbf{Y}$ which includes the common randomness W_0. We have

$$D(P^{(1)}_{W_0W^r\mathbf{Y}'}\|P^{(0)}_{W_0W^r\mathbf{Y}'}) \le D(P^{(1)}_{W_0W^r\mathbf{Y}}\|P^{(0)}_{W_0W^r\mathbf{Y}}) \tag{78}$$

$$\le \rho^2 I(W_0, W^r; \mathbf{X}, \mathbf{Y}) \tag{79}$$

$$\le \rho^2 I(W^r; \mathbf{X}, \mathbf{Y}|W_0) \tag{80}$$

$$\le \rho^2 k \tag{81}$$

where (78) follows since $P^{(1)}_{\mathbf{Y}'|W_0W^r\mathbf{Y}} = P^{(0)}_{\mathbf{Y}'|W_0W^r\mathbf{Y}} = P_{\mathbf{Y}'|W_0\mathbf{Y}}$ (note $\mathbf{Y}'$ is a (deterministic) function of $(W_0, \mathbf{Y})$), and (79) follows from Theorem 7.1 and Theorem 7.2. Observe that $D(P^{(1)}_{W_0W^r\mathbf{Y}'}\|P^{(0)}_{W_0W^r\mathbf{Y}'})$ is exactly $D(P^{\rho_1}_{\Pi\mathbf{Y}}\|P^{\rho_0}_{\Pi\mathbf{Y}})$ which we wanted to upper bound. Repeating the same steps for $D(P^{\rho_1}_{\Pi\mathbf{X}}\|P^{\rho_0}_{\Pi\mathbf{X}})$ establishes (71) for both the Gaussian and binary cases.

The bound on the Fisher information in the binary case (72) follows from Lemma 3.2. □

Remark 9. While we expect that the same bound in (72) continues to hold in the Gaussian case, the regularization condition required in the transition from the KL divergence bound to the Fisher information bound appears difficult to justify in the Gaussian case (see Lemma 3.2 and the ensuing remark).

9 PROOF OF THE SYMMETRIC STRONG DATA PROCESSING INEQUALITY

This section proves Theorem 7.2, which states that the symmetric strong data processing inequality constant is bounded by ρ^2 in the case of binary symmetric or Gaussian vectors. We first outline the proof, and then supplement the key lemmas used.

Proof of Theorem 7.2. First, note that we only need to prove the case where the common randomness W_0 is empty. Indeed, since Π includes W_0 and since W_0 is independent of $(\mathbf{X}, \mathbf{Y})$, we have $I(\mathbf{X}; \mathbf{Y}|\Pi) = I(\mathbf{X}; \mathbf{Y}|W_0, \Pi)$ and $I(\Pi; \mathbf{X}, \mathbf{Y}) = I(\Pi; \mathbf{X}, \mathbf{Y}|W_0)$, hence (64) will follow if we establish

$$I(\mathbf{X}; \mathbf{Y}) - I(\mathbf{X}; \mathbf{Y}|\Pi, W_0 = w_0) \le \rho^2 I(\Pi; \mathbf{X}, \mathbf{Y}|W_0 = w_0). \tag{82}$$

for each w_0. Using the Markov chains satisfied by the messages we have

$$I(\mathbf{X}; \mathbf{Y}) - I(\mathbf{X}; \mathbf{Y}|W^r) = \sum_{i\in[r]\cap 2\mathbb{Z}} I(W_i; \mathbf{X}|W^{i-1}) + \sum_{i\in[r]\setminus 2\mathbb{Z}} I(W_i; \mathbf{Y}|W^{i-1}) \tag{83}$$

$$I(W^r; \mathbf{X}, \mathbf{Y}) = \sum_{i\in[r]\cap 2\mathbb{Z}} I(W_i; \mathbf{Y}|W^{i-1}) + \sum_{i\in[r]\setminus 2\mathbb{Z}} I(W_i; \mathbf{X}|W^{i-1}). \tag{84}$$

Then the result for the binary and Gaussian cases follow respectively from Lemma 9.1 and Lemma 9.4, as well as the tensorization property Lemma 9.3, stated and proved below. □

9.1 Binary case

Our goal is to prove Lemma 9.1, which follows from Lemma 3.1 stated earlier and proved in this section.

Lemma 9.1 ([LCV17]). *Let $X, Y \in \{1, -1\}$ be equiprobably distributed with correlation ρ. Consider any random variables U^r, $r \in \mathbb{Z}$, satisfying*

$$U_i - (X, U^{i-1}) - Y, \quad i \in [r] \setminus 2\mathbb{Z}, \tag{85}$$

$$U_i - (Y, U^{i-1}) - X, \quad i \in [r] \cap 2\mathbb{Z}. \tag{86}$$

Define

$$R(P_{U^rXY}) := \sum_{i\in[r]\cap 2\mathbb{Z}} I(U_i; X|U^{i-1}) + \sum_{i\in[r]\setminus 2\mathbb{Z}} I(U_i; Y|U^{i-1}); \tag{87}$$

$$S(P_{U^rXY}) := \sum_{i\in[r]\cap 2\mathbb{Z}} I(U_i; Y|U^{i-1}) + \sum_{i\in[r]\setminus 2\mathbb{Z}} I(U_i; X|U^{i-1}). \tag{88}$$

Then $R(P_{U^rXY}) \le \rho^2 S(P_{U^rXY})$.

Proof. It suffices to show that the ratio of the i-th term on the right side of (87) to the i-th term on the right side of (88) is upper-bounded by ρ^2 for any i. Consider without loss of generality any $i \in 2\mathbb{Z}$. Note that by inducting on i and using the Markov chain conditions satisfied by U^r, we observe that $P_{YX|U^{i-1}=u^{i-1}}$ has the property that

$$\frac{\mathrm{d}P_{YX|U^{i-1}=u^{i-1}}}{\mathrm{d}P_{XY}} = f(x)g(y), \quad \forall x, y \tag{89}$$

for some functions f and g. Then using Lemma 3.1 we conclude that for each u^i we have $\frac{I(U_i;X|U^{i-1}=u^{i-1})}{I(U_i;Y|U^{i-1}=u^{i-1})} \le \rho^2$. □

The following result is used in the proof of Lemma 3.1. We state it in the general vector case, though we only need the scalar (X, Y) case.

Lemma 9.2. *Let X, Y be binary $\mathbb{P}[X = 1] = 1 - \mathbb{P}[X = 0] = p$ and let $\mathbb{P}[Y \neq X|X] = \frac{1-\rho}{2}$, $\rho \in [-1, 1]$. Consider $(\mathbf{X}, \mathbf{Y})$ to be n iid copies of (X, Y). Then for any random variables U, V such that $U - \mathbf{X} - \mathbf{Y} - V$ we have*

$$I(U; \mathbf{Y}) \le \rho^2 I(U; \mathbf{X}) \tag{90}$$

$$I(\mathbf{X}; V) \le \rho^2 I(\mathbf{Y}; V). \tag{91}$$

Proof. We first recall that a result known as tensorization (due to [AG76] in this context) allows to only check $n = 1$ case. For $n = 1$, the first part (90) is the standard inequality dating back to [AG76], see [PW17] for a survey. To show inequality (91), we apply Theorem 21 in [PW17], which establishes the following. Let A be a binary input and $B \sim P$ when $A = 0$ and $B \sim Q$ when $A = 1$, where $P = (P(v), v = 0, 1, \ldots)$ and $Q = (Q(v), v = 0, 1, \ldots)$ are two arbitrary distributions. Then for any $U - A - B$ we have

$$I(U; B) \le I(U; A)\left(1 - \left(\sum_v \sqrt{P(v)Q(v)}\right)^2\right). \tag{92}$$

(The bound is tight, cf. [PW17, Remark 8], whenever B is binary.) Applying this result to $A = Y$ and $B = X$ and denoting $q = p\rho + \frac{1-\rho}{2}$ we get

$$\sum_v \sqrt{P(v)Q(v)} = \frac{\sqrt{1-\rho^2}}{2\sqrt{q(1-q)}} \ge \sqrt{1-\rho^2}\,.$$

□

Proof of Lemma 3.1. Due to symmetry, it suffices to prove only the first inequality. Computing $P_{Y|X}$ and applying (92) we need to prove

$$\sum_{y \in \{0,1\}} \sqrt{Q_{Y|X}(y|0)Q_{Y|X}(y|1)} \frac{g(y)}{\sqrt{g_0 g_1}} \geq \sqrt{1-\rho^2}\,,$$

where $g_x = \sum_{y'} g(y')Q_{Y|X}(y'|x), x \in \{0,1\}$. Note that for all y, $\sqrt{Q_{Y|X}(y|0)Q_{Y|X}(y|1)} = \sqrt{\frac{1-\rho^2}{4}}$. By rescaling g so that $\sum_y g(y) = 1$ we get that $g_0 + g_1 = 1$ and hence $\sqrt{g_0 g_1} \leq \frac{1}{2}$, as required. □

9.2 Tensorization

The bound in Lemma 3.1 does not tensorize. That is, if Q_{XY} in the lemma is replaced by $Q_{XY}^{\otimes n}$, then $\sup_{P_{UXY}} \frac{I(U;Y^n)}{I(U;X^n)}$ can be strictly larger than ρ^2. Thus the cases of binary symmetric and Gaussian vectors cannot be proved via Lemma 3.1 as in the case of a pair of binary variables. This is a subtle issue that makes the proof of Theorem 7.2 somewhat nontrivial. Luckily, the symmetric strong data procesing constant tensorizes:

Lemma 9.3. *Let $(\mathbf{X},\mathbf{Y}) := (X_j, Y_j)_{j=1}^n \sim \otimes_{j=1}^n P_{X_j Y_j}$ for any given $P_{X_j Y_j}$, $j = 1, \ldots n$. Consider any random variables W^r, $r \in \mathbb{Z}$, satisfying*

$$W_i - (\mathbf{X}, W^{i-1}) - \mathbf{Y}, \quad i \in [r] \setminus 2\mathbb{Z}, \tag{93}$$
$$W_i - (\mathbf{Y}, W^{i-1}) - \mathbf{X}, \quad i \in [r] \cap 2\mathbb{Z}. \tag{94}$$

Then

$$\frac{R(P_{\mathbf{XY}W^r})}{S(P_{\mathbf{XY}W^r})} \leq \max_{1 \leq i \leq n} \sup_{P_{U^r|X_j Y_j}} \frac{R(P_{X_j Y_j U^r})}{S(P_{X_j Y_j U^r})} \tag{95}$$

where $P_{U^r|X_j Y_j}$ is such that

$$U_i - (X_j, U^{i-1}) - Y_j, \quad i \in [r] \setminus 2\mathbb{Z}, \tag{96}$$
$$U_i - (Y_j, U^{i-1}) - X_j, \quad i \in [r] \cap 2\mathbb{Z}. \tag{97}$$

Proof. Note that by induction it suffices to consider $n = 2$. Define

$$U_i := (W_i, Y_2), \quad i = 1, 2, \ldots, r; \tag{98}$$
$$\bar{U}_i := (W_i, X_1), \quad i = 1, 2, \ldots, r. \tag{99}$$

Then note that the Markov chains

$$U_i - (U^{i-1}, X_1) - Y_1, \quad i \in [r] \setminus 2\mathbb{Z}, \tag{100}$$
$$\bar{U}_i - (\bar{U}^{i-1}, X_2) - Y_2, \quad i \in [r] \setminus 2\mathbb{Z}, \tag{101}$$
$$U_i - (U^{i-1}, Y_1) - X_1, \quad i \in [r] \cap 2\mathbb{Z}, \tag{102}$$
$$\bar{U}_i - (\bar{U}^{i-1}, Y_2) - X_2, \quad i \in [r] \cap 2\mathbb{Z}, \tag{103}$$

are satisfied. Moreover,

$$\begin{aligned}
&R(P_{W^r X^2 Y^2}) \\
&= \sum_{i \in [r] \setminus 2\mathbb{Z}} [I(W_i; Y_2 | W^{i-1}) + I(W_i; Y_1 | W^{i-1}, Y_2)] \\
&\quad + \sum_{i \in [r] \cap 2\mathbb{Z}} [I(W_i; X_2 | W^{i-1}, X_1) + I(W_i; X_1 | W^{i-1})] \qquad (104) \\
&= \sum_{i \in [r] \setminus 2\mathbb{Z}} [I(W_i, X_1; Y_2 | W^{i-1}, X_1) - \Delta_i + I(W_i, Y_2; Y_1 | W^{i-1}, Y_2)] \\
&\quad + \sum_{i \in [r] \cap 2\mathbb{Z}} [I(W_i, X_1; X_2 | W^{i-1}, X_1) + I(W_i, Y_2; X_1 | W^{i-1}, Y_2) - \Delta_i] \qquad (105) \\
&= R(P_{U^r X_1 Y_1}) + R(P_{\bar{U}^r X_2 Y_2}) - \sum_{i=1}^r \Delta_i \qquad (106) \\
&= R(P_{U^r X_1 Y_1}) + R(P_{\bar{U}^r X_2 Y_2}) - I(X_1; Y_2 | W^r) \qquad (107)
\end{aligned}$$

where we have defined $\Delta_i := I(X_1; Y_2 | W^i) - I(X_1; Y_2 | W^i)$, and in (106) we have used the independence $Y_1 \perp Y_2$ for the $i = 1$ base case. Next, with similar algebra we obtain

$$\begin{aligned}
&S(P_{W^r X^2 Y^2}) \\
&= \sum_{i \in [r] \setminus 2\mathbb{Z}} [I(W_i; X_1 | W^{i-1}) + I(W_i; X_2 | W^{i-1}, X_1)] \\
&\quad + \sum_{i \in [r] \cap 2\mathbb{Z}} [I(W_i; Y_1 | W^{i-1}, Y_2) + I(W_i; Y_2 | W^{i-1})] \qquad (108) \\
&= \sum_{i \in [r] \setminus 2\mathbb{Z}} [I(W_i, Y_2; X_1 | W^{i-1}, Y_2) - \Delta_i + I(W_i, X_1; X_2 | W^{i-1}, X_1)] \\
&\quad + \sum_{i \in [r] \cap 2\mathbb{Z}} [I(W_i, Y_2; Y_1 | W^{i-1}, Y_2) + I(W_i, X_1; Y_2 | W^{i-1}, X_1) - \Delta_i] \qquad (109) \\
&= S(P_{U^r X_1 Y_1}) + S(P_{\bar{U}^r X_2 Y_2}) - \sum_{i=1}^r \Delta_i \qquad (110) \\
&= S(P_{U^r X_1 Y_1}) + S(P_{\bar{U}^r X_2 Y_2}) - I(X_1; Y_2 | W^r). \qquad (111)
\end{aligned}$$

Then the claim (95) follows.

□

Remark 10. Above, we followed the original method of proof proposed in a classical paper of Kaspi [Kas85], which essentially builds on Csiszár-sum identity in multiuser information theory. This method has been used recently in testing for independence [XK13] and common randomness extraction [LCV17]. A similar method was applied in [BBCR10] to a problem of (approximately) reconstructing a function of two correlated iid strings.

9.3 Gaussian case

To obtain the same lower bound in the Gaussian case, we can use the result for binary symmetric sequence and apply a central limit theorem argument.

Lemma 9.4. *Let X and Y be jointly Gaussian with correlation ρ. Consider any random variables U^r, $r \in \mathbb{Z}$, $|U^r| < \infty$, satisfying*

$$U_i - (X, U^{i-1}) - Y, \quad i \in [r] \setminus 2\mathbb{Z}, \tag{112}$$

$$U_i - (Y, U^{i-1}) - X, \quad i \in [r] \cap 2\mathbb{Z}. \tag{113}$$

Then $R(P_{U^r XY}) \le \rho^2 S(P_{U^r XY})$.

Proof. We claim the following continuity result: If $P^{(t)}_{XY}$ converges to P_{XY} in the total variation distance, then $R(P^{(t)}_{U^r XY})$ and $S(P^{(t)}_{U^r XY})$ converge to $R(P_{U^r XY})$ and $S(P_{U^r XY})$ respectively, where $P^{(t)}_{U^r XY} := P^{(t)}_{U^r|XY} P_{XY}$. The following will establish the lemma. Let $(A_l, B_l)_{l=1}^t$ be an iid sequence of binary symmetric random variables with correlation ρ, and put $X^{(t)} := \frac{A_1 + \dots + A_t}{\sqrt{t}} + a_t N$ and $Y^{(t)} := \frac{B_1 + \dots + B_t}{\sqrt{t}} + a_t N'$, where N and N' are standard Gaussian random variables, and N, N', (X^t, Y^t) are independent. By the central limit theorem, we can choose some $a_t = o(1)$ such that the distribution of $(X^{(t)}, Y^{(t)})$ converges to the Gaussian distribution P_{XY} in total variation (Proposition 4.2). Now suppose that the claim is not true, then there exists $P_{U^r|XY}$ satisfying the required Markov chains and $|\mathcal{U}^r| < \infty$ such that $R(P_{U^r XY}) > \rho^2 S(P_{U^r XY})$. The continuity claim implies that $R(P^{(t)}_{U^r XY}) > \rho^2 S(P^{(t)}_{U^r XY})$ for some t. However, using the data processing inequality of mutual information it is easy to see that $R(P_{U^r A^t B^t}) > R(P^{(t)}_{U^r XY})$ and that $S(P_{U^r A^t B^t}) < S(P^{(t)}_{U^r XY})$. Thus $R(P_{U^r A^t B^t}) > \rho^2 S(P_{U^r A^t B^t})$, which is in contradition with Lemma 9.1 and Lemma 9.3.

It remains to prove the continuity claim. Note that for each u^r, $(x, y) \mapsto P_{U^r|XY}(u^r|x, y)$ is a measurable function taking values in $[0, 1]$. Thus the convergence in total variation implies that $\lim_{t\to\infty} P^{(t)}_{U^r}(u^r) = P_{U^r}(u^r)$ and hence

$$\lim_{t\to\infty} H_{P^{(t)}}(U^r) = H_P(U^r), \tag{114}$$

where the subscripts of H denote the distributions with respect to which the entropies are computed. Moreover,

$$(x, y) \mapsto P_{U^r|XY}(u^r|x, y) \ln P_{U^r|XY}(u^r|x, y)$$

is also a bounded measurable function, so

$$\begin{aligned} &\lim_{t\to\infty} \mathbb{E}\left[P_{U^r|XY}(u^r|X^{(t)}, Y^{(t)}) \ln P_{U^r|XY}(u^r|X^{(t)}, Y^{(t)})\right] \\ &= \mathbb{E}\left[P_{U^r|XY}(u^r|X, Y) \ln P_{U^r|XY}(u^r|X, Y)\right], \end{aligned} \tag{115}$$

and summing over u^r shows that

$$\lim_{t\to\infty} H_{P^{(t)}}(U^r|X, Y) = H_P(U^r|X, Y). \tag{116}$$

Note that (114) and (116) imply the convergence of $R(P^{(t)}_{U^r XY}) := I_{P^{(t)}}(U^r; X, Y)$. Now,

$$S(P^{(t)}_{U^r XY}) = I_{P^{(t)}}(U^r; X) + I_{P^{(t)}}(U^r; Y) - I_{P^{(t)}}(U^r; X, Y), \tag{117}$$

and hence it remains to show that

$$\lim_{t\to\infty} H_{P^{(t)}}(U^r|X) = H_P(U^r|X), \tag{118}$$

$$\lim_{t\to\infty} H_{P^{(t)}}(U^r|Y) = H_P(U^r|Y). \tag{119}$$

By symmetry we only need to prove (118). Let us construct a coupling of $P_{U^r XY}$ and $P^{(t)}_{U^r XY}$ as follows. First construct a coupling such that $(X^{(t)}, Y^{(t)}) = (X, Y)$ with probability $\delta_t := \frac{1}{2}|P^{(t)}_{XY} - P_{XY}|$. Let E be the indicator of the event $(X^{(t)}, Y^{(t)}) \neq (X, Y)$. When $E = 0$, generate $U^{r(t)} = U^r$ according to $P_{U^r|XY}(\cdot|X, Y)$. When $E = 1$, generate $U^{r(t)}$ according to $P_{U^r|XY}(\cdot|X^{(r)}, Y^{(r)})$ and U^r according to $P_{U^r|XY}(\cdot|X, Y)$ independently. Then note that under either $P_{U^r XY}$ or $P^{(t)}_{U^r XY}$,

$$|H(U^r|X) - H(U^r E|X)| \le H(E). \tag{120}$$

Moreover,

$$\begin{aligned} &H(U^r, E|X) \\ &= H(U^r|X, E) \qquad (121) \\ &= \mathbb{P}[E = 1]H(U^r|X, E = 1) + \mathbb{P}[E = 0]H(U^r|X, E = 0), \qquad (122) \end{aligned}$$

hence

$$|H(U^r, E|X) - (1 - \delta_t)H(U^r|X, E = 0)| \le \delta_t \log|\mathcal{U}^r|. \tag{123}$$

However, for any $\mathcal{A} \in \mathcal{X}$ and u^r,

$$\frac{\mathbb{P}[U^r = u^r, X \in \mathcal{A}, E = 0]}{\mathbb{P}[X \in \mathcal{A}, E = 0]} = \frac{\mathbb{P}[U^{r,(t)} = u^r, X^{(t)} \in \mathcal{A}, E = 0]}{\mathbb{P}[X^{(t)} \in \mathcal{A}, E = 0]},$$

implying that $P^{(t)}_{U^r|X=x,E=0}(u^r) = P_{U^r|X=x,E=0}(u^r)$ for each x and u^r, and hence $H_{P^{(t)}}(U^r|X, E = 0) = H_P(U^r|X, E = 0)$. Thus (120) and (123) imply that

$$\begin{aligned} &|H_{P^{(t)}}(U^r|X) - H_P(U^r|X)| \\ &\le 2\delta_t \log|\mathcal{U}^r| + 2\left[\delta_t \log\frac{1}{\delta_t} + (1 - \delta_t)\log\frac{1}{1 - \delta_t}\right] \qquad (124) \end{aligned}$$

and (118) follows since $\delta_t \to 0$. □

Remark 11. In [LCV17], we first computed the symmetric SDPI for the binary symmetric distribution and then proved the converse for secret key generation for the binary source. Then we proved a converse for the Gaussian source using the following reduction argument: Using a large block of binary symmetric random variables we can simulate a joint distribution converging to the Gaussian distribution in total variation. Thus the error probability of the operational problem cannot be too different under the simulated distribution and the true Gaussian distribution. In the present paper, we used a different argument to prove something stronger: the symmetric SPDI constant is equal to ρ^2 for the Gaussian distribution; this of course implies the converse for the operational problem.

10 ACKNOWLEDGMENTS

We thank anonymous reviewer for the idea of reducing the general case to $(\rho, 0)$, cf. Corollary 8.1. We note that this is precisely the method introduced by Rahman-Wagner [RW12], though they only used it in a one-way setting. We thank Yihong Wu, Mark Braverman, Rotem Oshman, Himanshu Tyagi and Sahasranand KR for useful discussions. This material is based upon work supported by the National Science Foundation CAREER award under grant agreement CCF-12-53205, the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370, the European Research Council, under grant agreement 639573, the Israeli Science Foundation, under grant agreement 1367/14, the MIT IDSS Wiener Fellowship and the Yitzhak and Chaya Weinstein Research Institute for Signal Processing.

REFERENCES

[AB90] Rudolf Ahlswede and Marat V. Burnashev. On minimax estimation in the presence of side information about remote data. *The Annals of Statistics*, pages 141–171, 1990.

[AC86] Rudolf Ahlswede and Imre Csiszár. Hypothesis testing with communication constraints. *IEEE transactions on information theory*, 32(4), 1986.

[AG76] Rudolf Ahlswede and Peter Gács. Spreading of sets in product spaces and hypercontraction of the Markov operator. *Ann. Probab.*, pages 925–939, 1976.

[BBCR10] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *2010 ACM International Symposium on Theory of Computing*, pages 67–76, 2010.

[BGM$^+$16] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020. ACM, 2016.

[BYJKS04] Ziv Bar-Yossef, Thathachar S Jayram, Ravi Kumar, and D Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[CGMS17] Clément L. Canonne, Venkatesan Guruswami, Raghu Meka, and Madhu Sudan. Communication with imperfectly shared randomness. *IEEE Transactions on Information Theory*, 63(10):6799–6818, 2017.

[CR12] Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of gap-Hamming-distance. *SIAM Journal on Computing*, 41(5):1299–1317, 2012.

[CSWY01] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Yao. Information complexity and the direct sum problem for simultaneous message complexity. In *42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 270–278, 2001.

[DN04] Herbert A. David and Haikady N. Nagaraja. *Order Statistics*. Wiley Series in Probability and Statistics. Wiley, 2004.

[Gar18] Ankit Garg. Private communications. 2018.

[GR16] Venkatesan Guruswami and Jaikumar Radhakrishnan. Tight bounds for communication-assisted agreement distillation. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 50. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[HA98] Te Sun Han and Shun-ichi Amari. Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory*, 44(6):2300–2324, 1998.

[HÖW18] Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. *arXiv preprint arXiv:1802.08417*, 2018.

[HS18] Uri Hadar and Ofer Shayevitz. Distributed estimation of Gaussian correlations. *arXiv preprint arXiv:1805.12472*, 2018.

[IW03] Piotr Indyk and David Woodruff. Tight lower bounds for the distinct elements problem. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 283–288. IEEE, 2003.

[Kas85] Amiram Kaspi. Two-way source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, 31(6):735–740, 1985.

[KN96] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1996.

[Kul] S. Kullback. *Information Theory and Statistics*. New York: Dover, 1968 (originally published in 1959 by John Wiley).

[LC06] Erich L. Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[LCCV16] Jingbo Liu, Thomas A. Courtade, Paul Cuff, and S. Verdú. Smoothing Brascamp-Lieb inequalities and strong converses for cr generation. In *Proc. 2016 IEEE International Symposium on Information Theory*, Barcelona, Spain, July 10–15, 2016.

[LCV15] Jingbo Liu, Paul Cuff, and Sergio Verdú. Secret key generation with one communicator and a one-shot converse via hypercontractivity. In *Proc. 2015 IEEE International Symposium on Information Theory*, Hong-Kong, China, June 15–19, 2015.

[LCV17] Jingbo Liu, Paul Cuff, and Sergio Verdú. Secret key generation with limited interaction. *IEEE Transactions on Information Theory*, 63(11):7358–7381, 2017.

[MT74] Hiroshi Murata and Hiroshi Tanaka. An inequality for certain functional of multidimensional probability distributions. *Hiroshima Mathematical Journal*, 4(1):75–81, 1974.

[Pol12] Yury Polyanskiy. Hypothesis testing via a comparator. In *Proc. 2012 IEEE International Symposium on Information Theory*, pages 2206–2210. IEEE, 2012.

[PW17] Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and Bayesian networks. In *Convexity and Concentration*, pages 211–249. Springer, 2017.

[RW12] Md Saifur Rahman and Aaron B Wagner. On the optimality of binning for distributed hypothesis testing. *IEEE Transactions on Information Theory*, 58(10):6282–6303, 2012.

[She12] Alexander A Sherstov. The communication complexity of gap hamming distance. *Theory of Computing*, 8(1):197–208, 2012.

[ST18] KR Sahasranand and Himanshu Tyagi. Extra samples can reduce the communication for independence testing. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 2316–2320. IEEE, 2018.

[Tsy09] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer-Verlag New York, 2009.

[TW15] Himanshu Tyagi and Shun Watanabe. Converses for secret key agreement and secure computing. *IEEE Transactions on Information Theory*, 61(9):4809–4827, 2015.

[Vid12] Thomas Vidick. A concentration inequality for the overlap of a vector on a large set. *Chicago Journal of Theoretical Computer Science*, 1:1–12, 2012.

[Vil03] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[VT04] Harry L. Van Trees. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[XK13] Yu Xiang and Young-Han Kim. Interactive hypothesis testing against independence. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 2840–2844. Citeseer, 2013.

[Yao79] Andrew Chi-Chih Yao. Some complexity questions related to distributive computing(preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, STOC '79, pages 209–213, New York, NY, USA, 1979. ACM.

[ZB88] Zhen Zhang and Toby Berger. Estimation via compressed information. *IEEE transactions on Information theory*, 34(2):198–211, 1988.

[ZDJW13] Yuchen Zhang, John Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.