Detecting Correlations with Little Memory and Communication

Yuval Dagan YUVAL.DAGAN@WEIZMANN.AC.IL
Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the problem of identifying correlations in multivariate data, under information constraints: Either on the amount of memory that can be used by the algorithm, or the amount of communication when the data is distributed across several machines. We prove a tight trade-off between the memory/communication complexity and the sample complexity, implying (for example) that to detect pairwise correlations with optimal sample complexity, the number of required memory/communication bits is at least quadratic in the dimension. Our results substantially improve those of Shamir (2014), which studied a similar question in a much more restricted setting. To the best of our knowledge, these are the first provable sample/memory/communication trade-offs for a practical estimation problem, using standard distributions, and in the natural regime where the memory/communication budget is larger than the size of a single data point. To derive our theorems, we prove a new information-theoretic result, which may be relevant for studying other information-constrained learning problems.

1. Introduction

Information constraints play a key role in statistical learning and estimation problems. One always-present constraint is the sample size: We attempt to infer something about an underlying distribution, given only a finite amount of data sampled from that distribution. Indeed, the sample complexity for tackling various statistical problems is a central area in learning theory and statistics. However, in many situations, we are faced with additional information-based constraints, besides the sample complexity. For example, in practice the amount of *memory* used by the learning algorithm might be limited. In other cases, we might wish to solve a distributed version of the learning problem, where the data is randomly partitioned across several machines. Since communication between machines is invariably slow and expensive compared to internal processing, we might wish to solve the problem using a bounded amount of *communication*.

In recent years, an emerging body of literature has attempted to formally study the effect of such memory and communication constraints in learning problems. In many cases, it turns out that one can still solve a given problem with less memory or communication, but at the cost of a larger sample complexity. Thus, a fascinating question is whether such trade-offs are unavoidable, and what is the optimal trade-off.

In this paper, we study memory, communication, and sample complexity trade-offs for detecting correlations in multivariate data, one of the simplest and most common statistical estimation problems. In this problem, we are given a sequence of i.i.d. samples $\mathbf{x}_1, \mathbf{x}_2, \ldots$ from some zero-mean distribution over $\mathbb{R}^d$, and our goal is to detect correlated coordinates. For simplicity, let us focus for now on the case of pairwise correlations, and assume that for some pair of coordinates $(i,j) \in \{1,\ldots,d\}^2$, $\mathbb{E}\left[x_i x_j\right] = \rho > 0$, whereas for any other pair of coordinates (i',j'), $\mathbb{E}\left[x_{i'} x_{j'}\right] = 0$.

In the absence of memory or communication constraints, and given a sample $\mathbf{x}_1,\ldots,\mathbf{x}_t$, a simple approach is to compute the empirical average $\frac{1}{t}\sum_{l=1}^{t} x_{l,i}x_{l,j}$ for every possible coordinate pair, use concentration of measure to bound the difference between this empirical average and the true expectation with high probability, and thus determine which of these subsets is indeed correlated. For example, Hoeffding's inequality and a union bound implies that if all coordinates are bounded in $[-1,+1]$ almost surely, and $t = \Omega\left(\log(d)/\rho^2\right)$, then with arbitrarily high constant probability over the sample $\mathbf{x}_1,\ldots,\mathbf{x}_t$, there will be a unique coordinate pair (i,j) for which $\left|\frac{1}{t}\sum_{l=1}^{t} x_{l,i}x_{l,j}\right| \geq \frac{\rho}{2}$, and this pair corresponds to the correlated coordinates.

Although this approach is quite reasonable in terms of the sample complexity, it requires us to compute and maintain $\binom{d}{2} \approx d^2$ averages, which can be problematic in memory/communication-constrained variants of the problem: For an algorithm which streams over the data, we need at least $\Omega(d^2)$ memory to keep track of all the possible correlations. Similarly, when the data is distributed across several machines, we need at least $\Omega(d^2)$ bits of communication, to compute the empirical averages of every pair of coordinates.

What can we do if our memory or communication budget is less than $\Omega(d^2)$? Considering the case of streaming, memory-bounded algorithms first, a trivial solution is not to estimate all $\binom{d}{2}$ averages at once, but rather a smaller group of averages at a time. For example, if we only have enough memory to estimate one average, we can start with estimating the empirical correlation of coordinates 1 and 2, until we are sufficiently confident whether they are correlated or not, then move to coordinates 1 and 3, and so on. However, if we stream over the data, the price we pay is a larger sample size: In general, if we have s bits of memory, the approach above requires a sample size of $t = \tilde{O}\left(d^2/\left(\rho^2 s\right)\right)$ to detect the correlation with any constant probability (where the $\tilde{O}$ notation hides factors logarithmic in d, ρ[1]). In other words, the approach we just described satisfies $ts = \tilde{O}\left(d^2/\rho^2\right)$. More generally, if the algorithm is allowed to perform ℓ passes over the same data $\mathbf{x}_1,\ldots,\mathbf{x}_t$, then with the same approach, we can detect the correlation assuming

$$ts\ell = \tilde{O}\left(d^2/\rho^2\right) \ .$$

A natural question is whether this naive approach is improvable. Can we have an algorithm where the product of the memory size s and the total number of data points processed $t\ell$ is smaller, perhaps less than quadratic in the dimension d?

An analogous situation occurs in the context of distributed algorithms with communication constraints: If each machine has n i.i.d. data points, and can send s bits of communication, we can split the m machines to $\tilde{O}(d^2/s)$ groups, and have the machines in each group broadcast the empirical average of a different subset of $\tilde{\Theta}(s)$ coordinate pairs. Aggregating these averages and outputting the pair with the highest empirical correlation, we will succeed with any constant probability as long as $\frac{m}{d^2/s}\cdot n \geq \tilde{\Omega}\left(\frac{1}{\rho^2}\right)$ (namely, as long as we can compute the empirical average of at least $\tilde{\Omega}(1/\rho^2)$ data points, for each and every coordinate pair).This implies that the protocol will succeed, with the total number ms of bits communicated at most

$$\tilde{O}\left(d^2/(n\rho^2)\right).$$

[1] For example, we need $O(\log(1/\rho))$ bits of precision to determine if an average is above ρ, and $O(\log(d))$ bits to index coordinates. We note that sometimes such logarithmic factors can be reduced with various tricks (e.g., Luo (2005)), but these are not the focus of our paper.

Note that the non-trivial regime here is $n\rho^2 \ll 1$ (otherwise, any single machine can detect the correlation based on its own data, without any communication). In this regime, we see that the protocol above requires communication complexity quadratic in the dimension d. Again, it is natural to ask whether this simple approach can be improved, and whether the quadratic dependence on d is avoidable.

Perhaps surprisingly, we show in this paper that these approaches are in fact optimal (up to logarithmic factors), and establish a tight trade-off between sample complexity and memory/communication complexity for detecting correlations. Moreover, we show this for simple, natural data distributions; under minimal algorithmic assumptions; and for both pairwise and higher-order correlations (see below for a discussion of related results). In a nutshell, our contributions are the following:

- We prove that if the correlation ρ is sufficiently small (polynomially in d), then for any algorithm with s bits of memory, which performs at most ℓ passes over a sample of size t, we must have $ts\ell = \tilde{\Omega}(d^2/\rho^2)$ for it to detect the correlated coordinates. Also, in a distributed setting, a communication of $\tilde{\Omega}\left(d^2/\left(n\rho^2\right)\right)$ bits is necessary in general. This matches the upper bounds described above up to logarithmic factors. We prove these results for two families of natural distributions: over binary vectors in $\{-1,+1\}^d$, and for Gaussian distributions over $\mathbb{R}^d$.

- For binary vectors, we actually provide a more general result, which applies also to higher-order correlations. Specifically, we assume that there is some unique set I of indices such that $\mathbb{E}\prod_{i\in I} X_i = \rho$, and I comes from some known family of k possible subsets (the previous bullet refers to the special case where $|I| = 2$, and the family of $k = \binom{d}{2}$ coordinate pairs). Assuming ρ is polynomially small in k, we show that in the memory-constrained setting, $ts\ell = \tilde{\Omega}(k/\rho^2)$, and in the communication-constrained setting, $\tilde{\Omega}\left(k/\left(n\rho^2\right)\right)$ bits are required. This directly generalize the results from the previous bullet, and establishes that one cannot in general improve over the naive approach of estimating the correlation separately for every candidate set I.

- To obtain our theorems, we develop a general information-theoretic result, which may be of independent interest and can be roughly stated as follows: Assume that $\mu_0, \mu_1, \ldots, \mu_k$ are distributions over the same sample space, which are close to each other in the sense that for any $1 \le i \le k$ and any event E, $|\mu_i(E)/\mu_0(E) - 1| \le \rho$. Additionally, assume that $\mu_1, \ldots, \mu_k$ are pairwise uncorrelated, in the sense that for any $i \neq j$, $\int \frac{d\mu_i}{d\mu_0}\frac{d\mu_j}{d\mu_0} d\mu_0 = \int \frac{d\mu_i}{d\mu_0} d\mu_0 \int \frac{d\mu_j}{d\mu_0} d\mu_0 = 1$. Then any algorithm for identifying the distribution μ_i given a sample requires either $ts\ell = \tilde{\Omega}(k/\rho^2)$ in the memory-constrained setting, or $\tilde{\Omega}(k/(n\rho^2))$ bits of communication in a communication-constrained setting. This can be seen as generalizing the main technical result of Braverman et al. (2016) (Theorem 4.4), which proved a related lower bound in the context of communication constraints, assuming that the k distributions are defined over a product space. Here, we essentially replace independence assumptions by a weaker pairwise uncorrelation assumption, which is crucial for proving our results.

Related Work

The question of proving lower bounds on learning under memory and communication constraints has been receiving increasing attention recently, and related questions have long been studied in theoretical computer science and other fields. Thus, it is important to emphasize the combination of assumptions that place our setting apart from most other works:

- *The task is a statistical learning problem, based on i.i.d. examples from some underlying distribution*: For example, there is a large literature on memory lower bounds for streaming algorithms (see for instance Alon et al. (1996); Bar-Yossef et al. (2002); Muthukrishnan (2005) and references therein). However, these mostly focus on problems which are not standard learning problems, and/or that the data stream is adversarially generated rather than stochastically generated (which makes proving lower bounds easier). Similarly, there are many results on communication complexity (see for instance Kushilevitz and Nisan (1997)), but most of them refer to non-learning problems, or where the data is adversarially generated and distributed across machines (rather than randomly, which again makes lower bounds easier to prove).

- *Memory/communication budget is larger than the size of a single data point*: This is arguably the most common regime in practice. There are several works which studied the more constrained setting, where the memory or communication budget is smaller than the size of a single data point (but still larger than the required output), for problems such as sparse mean estimation, sparse regression, detecting low-rank subspaces, and multi-armed bandits (Shamir, 2014; Steinhardt and Duchi, 2015; Crouch et al., 2016; Braverman et al., 2016). Also, there has been a line of works on hypothesis testing and statistical estimation with finite memory, in a regime where the memory is insufficient to precisely express the required output (see Hellman and Cover (1970); Leighton and Rivest (1986); Ertin and Potter (2003); Kontorovich (2012) and references therein).

- *Results are for a standard, natural estimation problem, and where multiple communication rounds / passes over the data are allowed*: A breakthrough line of recent works (Raz, 2016, 2017; Moshkovitz and Moshkovitz, 2017a,b; Kol et al., 2017; Garg et al., 2017; Beame et al., 2017) showed that for binary classification problems, which satisfy certain combinatorial or algebraic conditions, any one-pass streaming algorithm would require either quadratic memory (in the dimension), or exponential sample size. So far, these conditions were shown to hold for learning parities and variants thereof (all strongly involving Boolean computations over $\mathbb{Z}_2$). Although such problems are very important in learning theory, they are arguably synthetic in nature and not commonly encountered in practice. In this paper, we focus on detecting correlations, which is a standard and common estimation problem. Moreover, whereas the results above apply to memory-constrained, one-pass algorithms, our results apply to both memory and communication constraints, and where multiple passes / communication rounds are allowed (building on techniques developed in Braverman et al. (2016)). On the flip side, the gaps we show in the required sample size (with and without information constraints) are polynomial in the dimension, whereas the results above imply exponential gaps. We discuss the differences and the similarities in more depth in Sec. C.

Perhaps the work closest to ours is Shamir (2014), which also studied the problem of detecting correlations with memory/communication constraints, and showed trade-offs between the memory/communication complexity and the sample complexity. For example, in the context of memory constraints, that paper showed that there exists a distribution over d-dimensional vectors, with a particular correlation value ρ (depending on d), such that detecting the correlation is statistically feasible given $O(d^2\log^2(d))$ examples, but any one-pass algorithm with only $s \ll d^2/\log^2(d)$ bits of memory requires a strictly larger sample size of at $\Omega(d^4/s)$ examples. However, that result is weaker than ours in several respects: First, it applies to a much more restrictive family of algorithms (where only one round of communication is allowed in the communication-constrained setting, and only one pass over the data in the memory-constrained setting). Second, it only applies to a certain carefully-tailored and unnatural family of data distributions, and does not imply communication/memory/sample trade-offs in the context, say, of vectors with bounded or Gaussian entries. Third, the result only holds for a particular choice of the correlation parameter ρ (depending on the other problem parameters), rather than holding for any small enough correlation. Fourth, the result is specific to pairwise correlations, whereas we prove more general results, applying to higher-order correlations and (potentially) to other information-constrained learning problems. Moreover, proving these results require fundamentally new ideas, which we develop in this paper.

Finally, for pairwise correlations, the problem we study is closely related to the *light-bulb problem*, proposed by Leslie Valiant at the very first COLT conference (Valiant, 1988). That problem is equivalent to identifying a pairwise correlation in data drawn from the d-dimensional Boolean cube. However, while we ask whether $o(d^2)$ memory/communication is possible, Valiant asked whether $o(d^2)$ *runtime* is possible. For the light-bulb problem, the best algorithm we are aware of (Valiant, 2015) requires a runtime of only $O(d^{1.62})$. However, a close inspection of the results indicates that this only applies when the correlation parameter ρ is close to being an absolute constant (a regime which also makes the communication/memory-constrained setting easier – see Remark 11). Although communication/memory complexity and computational complexity are not the same, our results suggest that no algorithm for the light-bulb problem can run in time $o(d^2)$ (as a function of d), if the correlation to be detected is small enough.

Our paper is structured as follows: In Sec. 2, we introduce notation and necessary definitions. In Sec. 3, we present our main results, and in Sec. 4, we sketch our main proof ideas and techniques. Full proofs are provided in Appendix A, and some additional results are provided in Appendix B.

2. Preliminaries

For any integer $k \geq 1$, the notation $[k]$ denotes the set $\{1, \ldots, k\}$. We use the standard $O(), \Omega()$ big-O notation to hide constants, and $\tilde{O}(), \tilde{\Omega}()$ to hide constants as well as polylogarithmic factors. For any distribution μ and any integer $n \geq 1$, define by μ^n the distribution over n i.i.d samples from μ.

2.1. Communication protocols and memory-limited algorithms

In the context of communication-constrained algorithms, we consider a multi-party setting where there are $m \geq 1$ parties/machines, and each party receives an input visible only to her (i.e. a sample of data points). The parties communicate using broadcast messages with the goal of calculating some function over all of the inputs. A *protocol* defines the communication between the parties: which party is to speak next and which message she should send as a function of her input, the message history and some randomness. The *communication complexity* of a protocol is the maximal number of bits sent in this protocol, where the maximum is over all possible inputs and over the randomness of the protocol[2]. The *transcript* of a protocol contains all the messages sent.

[2] It is well-known that worst-case and average-case communication complexity are equivalent up to constants, so our lower bounds also apply to the communication complexity in expectation over the inputs and the randomness of the protocol. To see this, note that if there is a protocol π with expected communication complexity b, succeeding with probability $9/10$, then by Markov's inequality, a protocol π' which simulates π and stops after $10b$ bits of communication still succeeds with probability $8/10$, and has maximal communication complexity $10b$.

Definition 1 *Let $m, n \geq 1$ be integers, let $k \geq 2$ be an integer and let $\mu_1, \ldots, \mu_k$ be distributions on the same sample space. An (m, n)-*protocol identifying $\mu \in \{\mu_1, \ldots, \mu_k\}$ with error ε *is an m-party communication protocol where each party receives as an input an independent set of n i.i.d. samples from the same distribution μ_i. Additionally, for any $i \in [k]$, the protocol outputs the index i of the distribution μ_i which generated the data, with probability at least $1 - \varepsilon$.*

We emphasize that the protocols we consider are not restricted in terms of the number of messages sent or the number of communication rounds: We are only interested in the overall communication complexity, namely the total number of bits sent between machines.

In the memory-constrained setting, we consider an algorithm which is allowed to perform ℓ passes over t data points sampled i.i.d. from some distribution, with a memory limitation of s bits:

Definition 2 *Let $t, s, \ell \geq 1$ be integers and let $\mu_1, \ldots, \mu_k$ be distributions on the same sample space. A (t, s, ℓ)-*algorithm identifying $\mu \in \{\mu_1, \ldots, \mu_k\}$ with error ε *is an algorithm receiving t i.i.d. samples from μ_i for some $1 \leq i \leq k$. This algorithm goes over all samples sequentially in ℓ passes, using at most s bits of memory (formally, letting $x_1, x_2, \ldots, x_{t\ell}$ be ℓ copies of the data set in order, we assume the algorithm can be written recursively as $u_{i+1} = f_i(x_i, u_i)$, where $u_i \in \{0,1\}^s$ for all i denotes the memory of the algorithm after handling example x_i, f_i is an arbitrary function, and the output is a function of $u_{t\ell+1}$). For any $i \in [k]$, the algorithm outputs the index i of the distribution μ_i generating the data, with probability at least $1 - \varepsilon$.*

2.2. Centered families of distributions

For our results, we will consider families of distributions which are all close to one another, in the following sense:

Definition 3 *Let $0 < \rho < 1$ be a number, let $k \geq 2$ be an integer and let $\mu_1, \ldots, \mu_k$ be distributions on the same sample space Ω and the same set of events $\mathcal{F}$. We say that $\{\mu_1, \ldots, \mu_k\}$ is a ρ-*centered family of distributions *(or $\mathrm{CD}(\rho)$ for brevity), if there exists a distribution μ_0 on the same sample space and the same set of events such that for any event $E \in \mathcal{F}$ and any $i \in [k]$,*

$$(1 - \rho)\mu_0(E) \leq \mu_i(E) \leq (1 + \rho)\mu_0(E).$$

We say that $\{\mu_1, \ldots, \mu_k\}$ is centered around μ_0.

3. Main results

Our results are based on two general theorems, which establish the difficulty of distinguishing generic distributions under communication and memory constraints respectively. These theorems are presented in Subsection 3.1. We then apply them to the problem of detecting correlations, for distributions over binary vectors (Subsection 3.2) and for Gaussian distributions (Subsection 3.3).

3.1. A General Theorem

Let $\{\mu_1, \ldots, \mu_k\}$ be a $\mathrm{CD}(\rho)$ family of probability distributions centered around μ_0 (namely, $|\mu_i(E)/\mu_0(E) - 1| \leq \rho$ for any $i \in [k]$ and any event E). The following theorem establishes that under a certain technical condition (Eq. (1)), any (m, n) protocol would require a lot of communication to identify the distribution from which the input data is sampled:

Theorem 4 *There exist positive numerical constants C, C' such that the following holds. Let $\{\mu_1, \ldots, \mu_k\}$ be a $\mathrm{CD}(\rho)$ family of distributions centered around μ_0, let $m, n \geq 1$ be integers such that $\rho \leq (n \ln k)^{-1/2}/C'$. If*

$$\sum_{S \subseteq [k]\colon |S| \geq 2} n^{-|S|/2} \rho^{-|S|} \left| \mathbb{E}_{A \sim \mu_0} \prod_{i \in S} \left(\frac{\mu_i(A)}{\mu_0(A)} - 1 \right) \right| \leq \frac{1}{n} \,, \tag{1}$$

then any (m, n)-protocol identifying $\{\mu_1, \ldots, \mu_k\}$ with error $1/3$ has a communication complexity of at least

$$\frac{k}{C\rho^2 n \log(k/(n\rho^2))}.$$

In particular, Eq. (1) holds if there exists an integer $\ell \geq 2$ such that all the terms in Eq. (1) corresponding to $|S| \leq \ell$ are zero, and $n \geq k^{2(\ell+1)/(\ell-1)}$.

The proof appears in Subsection A.2, whereas Lemma 20 is its main ingredient. To explain the intuition, let B_i (for $i \in [k]$) be the random variable $\frac{\mu_i(A)}{\mu_0(A)}$, where A is sampled from μ_0, and note that its expectation is always 1. Eq. (1) corresponds to requiring B_i to be approximately uncorrelated when n is large enough, namely

$$\sum_{S \subseteq [k]\colon |S| \geq 2} n^{-|S|/2} \rho^{-|S|} \left| \mathbb{E} \prod_{i \in S} (B_i - \mathbb{E}[B_i]) \right| \leq \frac{1}{n} \,.$$

The last part of the theorem simply states that this indeed holds, if the B_i random variables are uncorrelated up to order ℓ, and n is large enough. In particular, for large n, pairwise uncorrelation ($\ell = 2$) is sufficient. The theorem implies that if the distributions are "uncorrelated" in this sense, then the task of identifying $\mu \in \{\mu_1, \ldots, \mu_k\}$ requires a communication complexity of $\tilde{\Omega}(k/(n\rho^2))$. Crucially, the required communication scales linearly with the number of distributions k, and is no better than what we would need for solving k completely independent problems, each involving distinguishing only two such distributions.

We now turn from communication complexity to memory complexity. The following theorem establishes a lower bound on the product of the sample size, memory, and number of data passes for any memory-constrained algorithm which identifies $\mu_1, \ldots, \mu_k$:

Theorem 5 *There exist positive numerical constants $C^{(2)}, C^{(3)}$ such that the following holds. Let $\{\mu_1, \ldots, \mu_k\}$ be a $\mathrm{CD}(\rho)$ family centered around μ_0, and let $t, s, \ell \geq 1$ be integers. Assume that there exists $n \leq C^{(2)}/(\rho^2 \log k)$ such that the conditions of Theorem 4 hold, with respect to k, n and ρ. Then any (t, s, ℓ)-algorithm identifying $\mu_1, \ldots, \mu_k$ with $1/3$ error satisfies*

$$ts\ell \geq \frac{k}{C^{(3)} \rho^2 \log k}.$$

The proof of the theorem is a simple reduction to the communication complexity lower bound of Thm. 4: Given a (t, s, ℓ) algorithm, and any m, n such that $mn \geq t$, one can create an (m, n) protocol which simulates the algorithm in a distributed setting as follows: Fixing some arbitrary order over the parties, each party in turn simulates the (t, s, ℓ) algorithm over its data. Once the party exhausts her data, the state of this algorithm (consisting of at most s bits) is transmitted to the next party, which continues to simulate the algorithm, and so on. Once t data points have been processed in this manner, the current party transmits the algorithm's state back to the first party, which starts simulating the next pass of the (t, s, ℓ) algorithm. This continues until ℓ such passes are done. Then, the output of the protocol is set as the output of the simulated (t, s, ℓ) algorithm. The overall communication complexity is at most $ts\ell/n$, so by Thm. 4 (assuming its conditions are fulfilled), we must have

$$\frac{ts\ell}{n} \geq \frac{k}{C\rho^2 n \log(k/(n\rho^2))}. \tag{2}$$

In particular, picking $m = k$ and $n = C^{(2)}/(\rho^2 \log k)$ for any constant $C^{(2)} \leq C'^{-2}$ concludes the proof.

We finish this subsection with two additional remarks:

Remark 6 (Identification vs. binary decision) *In the results of this paper, we focus on the problem of identifying an underlying distribution, under the promise that it belongs to a certain family of distributions $\mu_1, \ldots, \mu_k$ (e.g., which pair of coordinates are correlated). An arguably easier task is to decide whether the underlying distribution is either some fixed μ_0 or one of $\mu_1, \ldots, \mu_k$ (e.g., whether there exists a correlated pair of coordinates or not). However, our lower bounds apply to that task as well, with an almost identical proof.*

Remark 7 (Data access) *Our memory-based bounds assume that the algorithm performs one or more passes over the data. An even weaker assumption might be that the algorithm can access the data in an arbitrary order (i.e. has random access). However, proving a super-linear (in dimension) memory lower bound in this setting would imply a super-linear lower bound on the runtime of any random-access Turing machine, and unfortunately, this is related to difficult questions in computational complexity (see Raz (2016, Section 1.2) for a related discussion).*

3.2. Binary Vectors

Having establishes our main technical results, we now turn to derive concrete bounds in the context of detecting correlations. In this subsection, we begin with the case of distributions over binary vectors, where the goal is to detect some unique (pairwise or higher-order) correlation. Concretely, fix some $0 < \rho < 1$, and define the sample space as $\Omega = \{-1, 1\}^d$ for some $d \geq 2$. Let $\mathcal{I}$ be the set of all nonempty subsets of $\{1, \ldots, d\}$. For any $I \in \mathcal{I}$, let $\mu_{I,\rho}$ be the distribution over Ω defined by

$$\mu_{I,\rho}((x_1, \ldots, x_d)) = 2^{-d}(1 + \rho \prod_{i \in I} x_i).$$

Namely, $\mu_{I,\rho}$ samples with probability $\frac{1}{2}(1+\rho)$ an element uniformly from all elements with an even number of -1 values in the coordinates corresponding to I and with probability $\frac{1}{2}(1-\rho)$ it samples an element with an odd number of -1 values in I. Note that $\mu_{I,\rho}$ encodes a unique correlated subset of indices in the following manner (the proof appears in Subsection A.3.2):

Lemma 8 *For any set $I' \in \mathcal{I}$, $I' \neq \emptyset$, it holds that $\mathbb{E}_{X \sim \mu_{I,\rho}} \prod_{i \in I'} X_i = \begin{cases} \rho & I' = I \\ 0 & I' \neq I \end{cases}$.*

For any subset $\mathcal{U} \subseteq \mathcal{I}$ and $0 < \rho < 1$, let $\mathcal{P}_{\mathcal{U},\rho} = \{\mu_{I,\rho} \colon I \in \mathcal{U}\}$. We apply Theorems 4 and 5 on the problem of identifying an underlying distribution μ, promised to belong to the family $\mathcal{P}_{\mathcal{U},\rho}$, to get communication and memory lower bounds (the proof appears in Subsection A.3.1).

Theorem 9 *Fix some $\mathcal{U} \subseteq \mathcal{I}$ which satisfies $|\mathcal{U}| \geq 2$. Fix integers $m, n \geq 1$ such that $n \geq |\mathcal{U}|^6$, and a positive $\rho \leq n^{-1/2}\ln^{-1/2}|\mathcal{U}|/C$, where C is a numerical constant. Then any (m, n) protocol identifying $\mu \in \mathcal{P}_{\mathcal{U},\rho}$ with $1/3$ error has a communication complexity of at least*

$$\frac{|\mathcal{U}|}{C\rho^2 n \log(|\mathcal{U}|^2/(n\rho^2))}.$$

For example, the case of detecting pairwise correlations corresponds to choosing $\mathcal{U} = \{I \in \mathcal{I}\colon |I| = 2\}$. Since $|\mathcal{U}| = \binom{d}{2} = \Omega(d^2)$, this gives us a lower bound of $\tilde{\Omega}\left(\frac{d^2}{\rho^2 n} - m\right)$, or $\tilde{\Omega}\left(\frac{d^2}{\rho^2 n}\right)$. This is optimal up to logarithmic factors, as shown by the upper bound discussed in the introduction. More generally, for order-r correlations (for some constant $r \geq 2$), we simply pick $\mathcal{U} = \{I \in \mathcal{I}\colon |I| = r\}$, and since $|\mathcal{U}| = \binom{d}{r} = \Omega(d^r)$ in this case, the theorem implies a communication complexity of $\tilde{\Omega}\left(\frac{d^r}{\rho^2 n}\right)$. Again, this is tight up to logarithmic factors, using a straightforward generalization of the protocol for the pairwise case.

Next, we state the analogue of Thm. 9 for the memory-constrained setting (derived from Thm. 9 by the same communication-to-memory reduction discussed earlier):

Theorem 10 *There exist numerical constants $C, C' > 0$ such that the following holds. For any $\mathcal{U} \subseteq \mathcal{I}$ such that $|\mathcal{U}| \geq C'$, any ρ such that $0 \leq \rho \leq |\mathcal{U}|^{-3}\ln^{-1/2}|\mathcal{U}|C^{-1}$, and any integers $t, s, \ell \geq 1$, it holds that any (t, s, ℓ)-algorithm identifying $\mu \in \mathcal{P}_{\mathcal{U},\rho}$ with $1/3$ error satisfies*

$$ts\ell \geq \frac{|\mathcal{U}|}{C\ln|\mathcal{U}|\rho^2}.$$

As a special case, the theorem implies that for detecting pairwise correlations, $ts\ell = \Omega(d^2/\rho^2)$, and for order-$r$ correlations, $ts\ell = \Omega(d^r/\rho^2)$. For example, assuming the number of passes ℓ is constant, it implies that we cannot successfully detect the correlation, unless either the memory is large (on order d^r), or the number of samples used is much larger than what is required without memory constraints (i.e. $\Omega(\log(d)/\rho^2)$) for any constant r).

Remark 11 (Constraints on problem parameters) *Theorem 10 requires the correlation ρ to be sufficiently small compared to $|\mathcal{U}|$. Such an assumption is necessary to get a strong lower bound: To see this, consider the case of detecting a pairwise correlation in binary vectors with memory constraints. If we can store $\tilde{O}(d/\rho^2)$ bits in memory, then we can simply collect and store $\tilde{O}(1/\rho^2)$ data points, and the empirical correlations in this data will reveal the true correlated coordinates with high probability. Thus, to prove an $\tilde{\Omega}(d^2)$ memory lower bound (as we do here), the correlation ρ* must *be smaller than $\tilde{O}(d^{-1/2})$. Similarly, in a communication constrained setting, note that a communication budget of $\tilde{O}(d/\rho^2)$ bits enables the players to exchange $\tilde{O}(1/\rho^2)$ data points and find the correlation. Hence, in order to prove a communication lower bound of $\tilde{\Omega}\left(d^2/\left(n\rho^2\right)\right)$, one has to assume that $n = \tilde{\Omega}(d)$. That being said, in the theorems above we require a stronger bounds on ρ and n than what these arguments imply. In Appendix B, we show that these requirements can be weakened to some extent, for the case of $\mathcal{U} = \{I \in \mathcal{I}\colon |I| = r\}$, $r \geq 2$. Precisely characterizing the parameter regimes where non-trivial lower bounds are possible is left to future work.*

3.3. Gaussian Distribution

Having discussed distributions supported on binary vectors, we now turn to prove similar results for another cannonical family of distributions, namely Gaussian distributions on $\mathbb{R}^d$. In what follows, we focus on pairwise correlations (since a multivariate Gaussian distribution is uniquely determined by its mean and covariance matrix, there is no sense in discussing higher-order correlations as in the binary case).

Define $\mathcal{I}_2 = \{S \subseteq [d] \colon |S| = 2\}$. Fix some $d \geq 3$ and $0 < \sigma < 1$. For any set $I \in \mathcal{I}_2$, let $\eta_{I,\sigma}$ denote the zero-mean Gaussian distribution on $\mathbb{R}^d$, with covariance matrix $\Sigma_{I,\sigma}$ defined as follows:

$$\Sigma_{I,\sigma}(i,j) = \begin{cases} 1 & i = j \\ \sigma & I = \{i,j\} \\ 0 & \text{otherwise.} \end{cases}$$

In words, each individual coordinate has a variance of 1, and each pair of distinct coordinates are uncorrelated, except for the pair (i,j) with a correlation σ. Let $\mathcal{G}_\sigma = \{\eta_{I,\sigma} \colon I \in \mathcal{I}_2\}$ be the set of all $\binom{d}{2}$ distributions defined this way. The following theorems are analogues of Theorems 9 and 10 for the case of pairwise correlations (the proof appears in Subsection A.4):

Theorem 12 *Fix some $n, m \geq 1$ and $0 < \sigma < 1$, such that $n \geq Cd^6$ for some numerical constant $C > 0$ and $\sigma \leq n^{-1/2} \ln^{-1/2} d \ln^{-1}(dnm/\sigma)/C$. Any (m,n)-protocol identifying $\eta \in \mathcal{G}_\sigma$ with $1/6$ error has a communication complexity of at least*

$$\frac{d^2}{C\sigma^2 \ln^2(nmd/\sigma) \ln(d/(n\sigma^2))n} .$$

Theorem 13 *There exist numerical constants $C, C' > 0$ such that the following holds. If $d \geq C'$, then for any σ such that $0 < \sigma \leq \left(Cd^3 \ln^{1/2} d \ln(d/\sigma)\right)^{-1}$ and any integers $t, s, \ell \geq 1$, it holds that any (t,s,ℓ)-algorithm identifying $\mu \in \mathcal{G}_\sigma$ with $1/6$ error satisfies*

$$ts\ell \geq \frac{d^2}{C\sigma^2 \ln^3 d \ln^2(1/\sigma)}.$$

Whereas for binary vectors, our results are a direct corollary of Theorem 4, the proofs in the Gaussian case are more involved, because no family of distinct Gaussian distributions satisfy the $\mathrm{CD}(\rho)$ property from Definition 3. Instead, we need to work with truncated Gaussian distributions (which do satisfy this property), with some determinant calculations required to verify the conditions of Theorem 4. We then reduce the resulting bound on truncated distributions to non-truncated ones, to get Thm. 12. Thm. 13 is derived from Thm. 12 by the same communication-to-memory reduction discussed earlier.

4. Proof Ideas

In this section, we sketch the main ideas in the proof of Thm. 4, on which all our other results are based, and ignoring various technical issues. For simplicity, we discuss it in terms of the simpler problem of deciding whether the underlying distribution is μ_0 or one of $\mu_1, \ldots, \mu_k$ (as described in Remark 6). In particular, a successfull protocol for this problem should allow us to distinguish between μ_0 and μ_i, for all i, without knowing i beforehand. The crux of our proof lies in showing that these k tasks are "essentially" independent, in the sense that any protocol which solves all of them requires a communication of $\tilde{\Omega}(k)$ times the required communication for solving a single task. Formally, let Π be the transcript (the aggregation of all messages sent) of the protocol; let $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(m)}\right)$ denote the m (i.i.d.) sample sets given to the m parties in the protocol, where $X^{(j)}$ is the input of party j; and let $P_{\Pi|\mathbf{X}\sim\mu_b^{mn}}$ denote the distribution of the transcript Π conditioned on the inputs being distributed μ_i^{mn}, for $i \in [k]$. It is easy to show that any protocol which successfully distinguishes between μ_0 and μ_i must satisfy $d_{TV}\left(P_{\Pi|\mathbf{X}\sim\mu_0^{mn}}, P_{\Pi|\mathbf{X}\sim\mu_i^{mn}}\right) = \Omega(1)$, where d_{TV} is total variation distance[3]. In particular, this implies that

$$\sum_{i=1}^{k} d_{TV}\left(P_{\Pi|\mathbf{X}\sim\mu_0^{mn}}, P_{\Pi|\mathbf{X}\sim\mu_i^{mn}}\right)^2 = \Omega(k)\ . \tag{3}$$

The proof proceeds by showing that the communication complexity (times an $\tilde{O}(n\rho^2)$ factor) upper bounds the left-hand side above, namely the sum of total variations over all k individual tasks. This implies that the communication complexity is $\tilde{\Omega}(k/(n\rho^2))$ as required.

Intuitively, this assertion is true if the tasks are independent, so that information about one task does not convey information on another task. A concrete example (studied in Shamir (2014); Steinhardt and Duchi (2015); Braverman et al. (2016)) is sparse mean estimation, where the goal is to distinguish a zero-mean product distribution on $\mathbb{R}^k$, from similar product distributions where a few of the coordinates are slightly biased. Here, we can think of μ_i as the distribution where coordinate i is slightly biased. Since this is a product distribution, statistics about one coordinate reveals no information about the statistics of other coordinates, so any single party has to send some information on all coordinates in order for a protocol to succeed – hence the communication complexity must scale linearly with k. This idea lies at the heart of the papers mentioned above, and works well when the communication/budget is smaller than the dimension.

Unfortunately, this idea cannot be used as-is for showing lower bounds larger than the dimension. For example, in the context of pairwise correlations on d-dimensional data, an $\Omega(d^2)$ lower bound would require constructing a distribution over inputs $\mathbf{x}$, so that if we consider the $d \times d$ matrix $\mathbf{x}\mathbf{x}'$, at least $\Omega(d^2)$ of its entries has a joint product distribution. But this is impossible, since this matrix is always of rank 1, so no subset of more than $O(d)$ entries can be mutually independent. Shamir (2014), which also studied correlations, circumvented this difficulty with an ad-hoc construction involving extremely sparse vectors, but as discussed in the introduction, the end result has several deficiencies.

Our main technical contribution is to show how one can circumvent this hurdle, by relaxing the independence assumption to the milder technical assumption stated in Thm. 4, which only involves approximate uncorrelation and does apply to our problem.

The proof proceeds by fixing a party j, and constructing a Markov chain $\Pi \to X^{(j)} \to Y \to Z$, where Π is the transcript of the protocol; $X^{(j)}$ is the data of party j, and $Z = (Z_1, \ldots, Z_k), Y = (Y_1, \ldots, Y_k)$ are carefully-constructed binary random vectors, defined as follows:

- The transcript Π is distributed as if the inputs of all players are drawn from μ_0, and the input $X^{(j)}$ of player j is distributed μ_0^n.

[3] The total variation distance between two distributions with densities p and q is $\frac{1}{2}\int |p(x) - q(x)|\, dx$.

- The probability for each Y_i to equal 1 is a certain function of $\mu_i^n(X^{(j)})/\mu_0^n(X^{(j)})$.
- Z_i equals Y_i after flipping it with probability $\frac{1}{2} - \tilde{\Theta}\left(\rho\sqrt{n}\right)$. Additionally, $X^{(j)}$ is distributed roughly μ_i^n conditioned on $Z_i = 1$ (recall that $X^{(j)} \sim \mu_0^n$ unconditionally). Such a construction is possible from Bayes rule and the fact that $\mu_0^n(X^{(j)})$ and $\mu_i^n(X^{(j)})$ are close up to a multiplicative factor of $1 \pm \tilde{O}\left(\sqrt{n}\rho\right)$ for most values of $X^{(j)}$.

These random variables are constructed so that the following properties are satisfied:

- $Y_1, \ldots, Y_k$ are approximately independent, in the sense that $\sum_{i=1}^k I(\Pi; Y_i) \leq \tilde{O}(1) \cdot I(\Pi; Y)$, where $I()$ denotes mutual information. Intuitively, this is due to a central limit phenomenon: If we consider the k random variables $\frac{1}{\sqrt{n}} \log \frac{\mu_i^n(X^{(j)})}{\mu_0^n(X^{(j)})}$ for $i \in k$, they have an asymptotically Gaussian distribution as $n \to \infty$. Moreover, Eq. (1) in the theorem statement ensures that they are almost pairwise uncorrelated, but for Gaussian random variables, uncorrelation is equivalent to independence. Since each Y_i is a function of the corresponding random variable, it follows that $Y_1, \ldots, Y_k$ are approximately independent for large enough n.
- Since Z_i equals Y_i after a nearly-unbiased random coin flip, and these are both binary random variables, one can show the strong data processing inequality[4] $I(\Pi; Z_i) \leq O(n\rho^2) \cdot I(\Pi; Y_i)$. Combined with the previous item and the fact that $I(\Pi; Y) \leq I\left(\Pi; X^{(j)}\right)$ by the data processing inequality, we get that
$$\sum_{i=1}^k I(\Pi; Z_i) \ \leq\ \tilde{O}(n\rho^2) \cdot I(\Pi; X^{(j)}) \ .$$
- The construction of the Markov chain as defined above implies that the distribution of the transcript Π, conditioned on $Z_i = 1$, is close to the distribution of Π conditioned on the input of party j being drawn from μ_i^n. In particular, $\mathrm{h}^2(P_\Pi, P_{\Pi|X^{(j)}\sim\mu_i^n}) \approx \mathrm{h}^2(P_\Pi, P_{\Pi|Z_i=1})$, where h^2 denotes the squared Hellinger distance[5]. Recall that unconditionally, Π is distributed as if all input comes from μ_0. By existing results (Bar-Yossef et al. (2004, Lemma 6.2), Braverman et al. (2016, Lemma 2) and Jayram (2009)), we have that $\mathrm{h}^2(P_\Pi, P_{\Pi|\mathbf{X}\sim\mu_i^{nm}}) \ \leq\ O(1) \sum_{j=1}^m \mathrm{h}^2(P_\Pi, P_{\Pi|X^{(j)}\sim\mu_i^n})$ as well as $\mathrm{h}^2(P_\Pi, P_{\Pi|Z_i=1}) \leq O(1) I(\Pi; Z_i)$. Together with the previous item, we get that
$$\sum_{i=1}^k \mathrm{h}^2(P_\Pi, P_{\Pi|\mathbf{X}\sim\mu_i^{nm}}) \ \leq\ \tilde{O}(n\rho^2) \sum_{j=1}^m I(\Pi, X^{(j)}). \tag{4}$$

Since $X^{(1)}, \ldots, X^{(m)}$ are independent, the right-hand side of Eq. (4) is at most $\tilde{O}(n\rho^2) I(\Pi, \mathbf{X})$, which is at most $\tilde{O}(n\rho^2)$ times the communication complexity of the protocol. Also, the left-hand side of Eq. (4) can be shown to be at least $\sum_{i=1}^k d_{TV}(P_\Pi, P_{\Pi|\mathbf{X}\sim\mu_i^{nm}})^2/2$, which by Eq. (3), is at least $\Omega(k)$. Combining everything, we get that the communication complexity is at least $\tilde{\Omega}\left(k/\left(n\rho^2\right)\right)$ as required.

[4] The data processing inequality states that for any Markov chain $U \to V \to W$, $I(U; W) \leq I(U; V)$. In some cases, one can show a strong (strict) inequality, as the inequality used here.

[5] The squared Hellinger distance between random variables with densities p and q is $\frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$.

References

Rudolf Ahlswede and Peter Gács. Spreading of sets in product spaces and hypercontraction of the markov operator. *The annals of probability*, pages 925–939, 1976.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *STOC*, 1996.

Z. Bar-Yossef, T. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. In *FOCS*, 2002.

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.

Paul Beame, Shayan Oveis Gharan, and Xin Yang. Time-space tradeoffs for learning from small test spaces: Learning low degree polynomial functions. *arXiv preprint arXiv:1708.02640*, 2017.

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.

Harald Cramér. *Mathematical methods of statistics (PMS-9)*, volume 9. Princeton university press, 2016.

Michael Crouch, Andrew McGregor, Gregory Valiant, and David P Woodruff. Stochastic streams: Sample complexity vs. space complexity. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 57. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

E. Ertin and L. Potter. Sequential detection with limited memory. In *Statistical Signal Processing, 2003 IEEE Workshop on*, pages 585–588, 2003.

Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014.

Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. *arXiv preprint arXiv:1708.02639*, 2017.

M. Hellman and T. Cover. Learning with finite memory. *Annals of Mathematical Statistics*, pages 765–782, 1970.

TS Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 562–573. Springer, 2009.

Gillat Kol, Ran Raz, and Avishay Tal. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1067–1080. ACM, 2017.

L. Kontorovich. Statistical estimation with bounded memory. *Statistics and Computing*, 22(5): 1155–1164, 2012.

Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, 1997.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

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

F. Leighton and R. Rivest. Estimating a probability using finite memory. *Information Theory, IEEE Transactions on*, 32(6):733–742, 1986.

Zhi-Quan Luo. Universal decentralized estimation in a bandwidth constrained sensor network. *IEEE Transactions on information theory*, 51(6):2210–2219, 2005.

Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Conference on Learning Theory*, pages 1516–1566, 2017a.

Michal Moshkovitz and Dana Moshkovitz. Mixing implies strong lower bounds for space bounded learning. *Electronic Colloquium on Computational Complexity (ECCC)*, 24:116, 2017b.

S. Muthukrishnan. *Data streams: Algorithms and applications*. Now Publishers Inc, 2005.

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

C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pages 235–247. Springer, 1992.

Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 266–275. IEEE, 2016.

Ran Raz. A time-space lower bound for a large class of learning problems. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 24, page 20, 2017.

Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*, pages 163–171, 2014.

Jacob Steinhardt and John Duchi. Minimax rates for memory-bounded sparse linear regression. In *Conference on Learning Theory*, pages 1564–1587, 2015.

Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory*, pages 1490–1516, 2016.

Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *Journal of the ACM (JACM)*, 62(2):13, 2015.

LG Valiant. Functionality in neural nets. In *Proceedings of the first annual workshop on Computational learning theory*, pages 28–39. Morgan Kaufmann Publishers Inc., 1988.

Yuchen Zhang, Martin J Wainwright, and John C Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

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.

Appendix A. Proofs

A.1. Proof Preliminaries

For any integers $0 \leq a \leq b$, define $\binom{b}{\leq a} = \sum_{i=0}^{a} \binom{b}{i}$.

A.1.1. Probability distances

We will use two distance functions between probability distributions.

Definition 14 *The* total variation distance *between two probability measures with densities p, q on the same sample space Ω is defined as*

$$d_{TV}(p,q) = \frac{1}{2}\int_{x\in\Omega} |p(x) - q(x)|dx = \sup_F |P(F) - Q(F)|$$

where the supremum is over all events.

Definition 15 *The* squared Hellinger distance *between two probability measures with densities p and q on the same sample space Ω is defined as*

$$\mathrm{h}^2(p,q) = \frac{1}{2}\int_{x\in\Omega} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx.$$

The Hellinger distance *is defined as* $\mathrm{h}(p,q) = \sqrt{\mathrm{h}^2(p,q)}$.

These distances are defined for discrete random variables in a similar manner. Both the total variation distance and the (non squared) Hellinger distance are f-divergences[6] which satisfy the triangle inequality. Additionally, these distances are polynomially equivalent:

Proposition 16 *For any distributions p and q,*

$$\mathrm{h}^2(p,q) \leq d_{TV}(p,q) \leq \sqrt{2}\mathrm{h}(p,q). \tag{5}$$

[6] An f-divergence: a function of two distributions p and q which can be written as $\int f(dp/dq)dq$ for a convex function f with $f(1) = 0$

A.1.2. DATA PROCESSING INEQUALITY

We will frequently use the notation P_X to denote the distribution of a random variable X, where $P_X(x)$ denotes $\Pr[X = x]$. Additionally, define a *channel* $P_{Y|X}$ as a random function which gets as an input a member x of some sample space and outputs a random Y according to the distribution $P_{Y|X=x}$. We can compose a channel $P_{Y|X}$ over a distribution P_X to get a new distribution, $P_{Y|X} \circ P_X$, the distribution over the output of the channel given that its input is distributed P_X. Similarly, we can compose channels together.

Definition 17 *A* Markov chain $X_0 \to X_1 \to \cdots \to X_n$ *consists of a distribution* P_{X_0} *and channels* $P_{X_1|X_0}, \dots, P_{X_n|X_{n-1}}$. *It induces a joint distribution* $P_{X_0 \cdots X_n} = P_{X_0(X_1|X_0)\cdots(X_n|X_{n-1})}$ *accordingly.*

The entropy of a random variable X is denoted $H(X)$ and the mutual information between the random variables X and Y is defined as

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

The data processing inequality states that an information cannot increase while being transfered across a channel. It has two formulations: one in terms of mutual information and one in terms of f-divergence.

Proposition 18 (Data processing inequality) *The following hold:*

1. *For any Markov chain* $W \to X \to Y$, $I(W;Y) \le I(W;X)$.
2. *Let* P_{X_1} *and* P_{X_2} *be distributions on the same sample space* Ω *and let* $P_{Y|X}$ *be a channel getting its input from* Ω. *Then for any* f*-divergence* d_f,

$$d_f(P_{Y|X} \circ P_{X_1}, P_{Y|X} \circ P_{X_2}) \le d_f(P_{X_1}, P_{X_2}).$$

A.2. Proof of Theorem 4

Assume a sample space $\Omega \subseteq \{-1,1\}^k$, and assume a distribution μ_0 over Ω which satisfies that for any $i \in \{1, \dots .k\}$, the probability for an element $x = (x_1, \dots, x_k)$ to satisfy $x_i = 1$ equals $1/2$. Given $0 < \rho < 1$, one can define the distributions $\mu_1, \dots, \mu_k$, where $\mu_i(x) = (1 + \rho x_i)\mu_0(x)$ for all x. We say that $\{\mu_1, \dots, \mu_k\}$ is a *binary centered familiy of distributions* (or $\mathrm{BCD}(\rho)$ for brevity), *centered around* μ_0.

We prove Theorem 4 for all $\mathrm{CD}(\rho)$ families of distributions, however, it is sufficient to prove this theorem under a weaker condition on the distributions, namely, that the distributions are $\mathrm{BCD}(\rho)$: if Theorem 4 is correct for all $\mathrm{BCD}(\rho)$ distributions then it is correct for all $\mathrm{CD}(\rho)$ distributions. This can be shown using a reduction: for every $\mathrm{CD}(\rho)$ family $\{\eta_1, \dots, \eta_k\}$, there is a $\mathrm{BCD}(\rho)$ family $\{\mu_1, \dots, \mu_k\}$ and a transformation $P_{\eta|\mu}$ transforming each μ_i to η_i, namely, $\eta_i = P_{\eta|\mu} \circ \mu_i$ for all $1 \le i \le k$. This transformation does not change the high order correlations: for all $S \subseteq [k]$,

$$\mathbb{E}_{X\sim\mu_0} \prod_{i\in S} \left(\frac{\mu_i(X)}{\mu_0(X)} - 1 \right) = \mathbb{E}_{Y\sim\eta_0} \prod_{i\in S} \left(\frac{\eta_i(Y)}{\eta_0(Y)} - 1 \right), \tag{6}$$

hence the condition Eq. (1) applies for $\{\mu_1, \dots, \mu_k\}$ if and only if it applies for $\{\eta_1, \dots, \eta_k\}$. Given an input to the μ-problem the parties can privately transform it to an η-input and simulate an η-protocol.

Lemma 19 *Let $\{\eta_1, \dots, \eta_k\}$ be a $\mathrm{CD}(\rho)$ family. There exists a $\mathrm{BCD}(\rho)$ family $\{\mu_1, \dots, \mu_k\}$ and a channel $P_{\eta|\mu}\colon \Omega_\mu \to \Omega_\eta$ such that for all $1 \le i \le k$, $\eta_i = P_{\eta|\mu} \circ \mu_i$, where Ω_μ and Ω_η are the sample spaces of the μ-family and the η-family respectively. Additionally, Eq. (6) holds for all $S \subseteq [k]$.*

Lemma 19 is prooved in Subsection A.2.3. Assume for the rest of the proof that $\{\mu_1, \dots, \mu_k\}$ is a $\mathrm{BCD}(\rho)$ family of distributions. We present two main lemmas. In the first lemma, we assume a setting that there is just one party which gets some input $X \in \Omega^n$ and outputs Π. We bound the distance between the distribution of Π conditioned on X being distributed μ_0^n or μ_i^n. The distance is bounded in terms of the amount of information that Π reveals on X. This lemma contains the main technical contribution of this paper.

Lemma 20 *Let $\{\mu_1, \dots, \mu_k\}$ be a $\mathrm{BCD}(\rho)$ family on a sample space $\Omega \subseteq \{-1, 1\}^k$ centered around μ_0. Let $P_{\Pi|X}$ be some channel getting an input $X \in \Omega^n$. Under the assumptions of Theorem 4 on n, ρ and k,*

$$\sum_{i=1}^{k} \mathrm{h}^2(P_{\Pi|X\sim\mu_0^n}, P_{\Pi|X\sim\mu_i^n}) \le Cn\rho^2 \log(k^2/(n\rho^2))(I_{X\sim\mu_0^n}(\Pi; X) + 1),$$

for some numerical constant $C > 0$.

Lemma 20 is proved in Subsection A.2.1. The next lemma utilizes results of Jayram (2009) and Braverman et al. (2016) to show that Lemma 20 implies Thm. 4. The tools developed in the prior work derive bounds on settings where there is just a single party who sends some output to settings with multiple communicating parties.

Lemma 21 *Let $\mu_0, \dots, \mu_k$ be probability distributions on the sample space Ω such that for every channel $P_{\Pi|X}$ with input in Ω^n:*

$$\sum_{i=1}^{k} \mathrm{h}^2(P_{\Pi|X\sim\mu_0^n}, P_{\Pi|X\sim\mu_i^n}) \le \beta(I_{X\sim\mu_0^n}(\Pi; X) + 1),$$

for some $\beta > 0$. Then any $1/3$-error (m, n) protocol identifying $\mu \in \{\mu_1, \dots, \mu_k\}$ has a communication complexity of at least Ck/β for some numerical constant $C > 0$.

Lemma 21 is proved in Subsection A.2.2. Combining Lemma 20 and Lemma 21 gives us the lower bound on the communication complexity in Theorem 4 . To conclude the proof, it remains to prove that the condition stated at the end of the theorem is indeed sufficient for Eq. (1) to hold. This is shown in the following lemma:

Lemma 22 *For any integer $\ell \ge 2$, if $n \ge k^{2(\ell+1)/(\ell-1)}$ then the sum of all terms in Eq. (1) corresponding to $|S| > \ell$ is at most $1/(2n)$.*

Proof Under the assumptions of the lemma,

$$\sum_{S\subseteq[k]\colon |S|\ge\ell+1} n^{-|S|/2}\rho^{-|S|} \left| \mathbb{E}_{A\sim\mu_0} \prod_{i\in S} (\mu_i(A)/\mu_0(A) - 1) \right|$$

$$\le \sum_{S\subseteq[k]\colon |S|\ge\ell+1} n^{-|S|/2} = \sum_{r=\ell+1}^{k} \binom{k}{r} n^{-r/2} \le \sum_{r=\ell+1}^{k} \frac{k^r}{r!} n^{-r/2} \le \frac{1}{n} \sum_{r=\ell+1}^{k} \frac{1}{r!} \le \frac{1}{2n}, \qquad (7)$$

where the LHS of Eq. (7) follows from the definition of a $\mathrm{CD}(\rho)$ family: it always holds that $|\mu_i(A)/\mu_0(A) - 1| \le \rho$. ■

Subsection A.2.1 contains the proof of Lemma 20, Subsection A.2.2 contains the proof of Lemma 21 and Subsection A.2.3 contains the proof of Lemma 19.

A.2.1. Proof of Lemma 20

For the majority of our calculations we will assume that some high probability event holds. In what follows we give intuitive explanation about this event and why it holds and then more preceise definition and proof. Recall that the input $x \in \Omega^n$ contains n samples from $\Omega = \{-1,1\}^k$ and define $x_{j,i}$ to be bit i of sample j, for $1 \le i \le k$ and $1 \le j \le n$. Note that $x_{j,i} \in \{-1,1\}$, hence, for any i, any distribution μ over Ω and any $t > 0$, Hoeffding's bound implies that

$$\Pr_{x\sim\mu^n}\left[\left|\sum_{j=1}^n x_{j,i} - \mathbb{E}_{x\sim\mu^n}\left[\sum_{j=1}^n x_{j,i}\right]\right| > \sqrt{n}t\right] \le 2e^{-t^2/2}. \tag{8}$$

In particular, taking $t = \sqrt{2\ln(k^2)}$ and performing a union bound over $i = 1,\dots,n$, one obtaines that with probability at least $1 - 2/k$, Eq. (8) holds for all $i = 1,\dots,k$. We would like to replace μ by $\mu_{i'}$ for $i' = 0,1,\dots,k$. Note that for any $i' = 0,1,\dots,k$, it holds that

$$\left|\mathbb{E}_{x\sim\mu_{i'}^n}\left[\sum_{j=1}^n x_{j,i}\right]\right| = \left|n\mathbb{E}_{y\sim\mu_{i'}}[y_i]\right| \le n\rho \le \sqrt{n},$$

by definition of a $\mathrm{BCD}(\sigma)$ family of distributions and by the requrement $\rho \le \sqrt{n}$. This implies that for any $i' = 0,1,\dots,k$,

$$\Pr_{x\sim\mu_{i'}^n}\left[x \in \mathcal{T}''\right] \ge 1 - 2/k, \tag{9}$$

where $\mathcal{T}''$ is the set of all $x \in \Omega^n$ which satisfies that for all $i \in \{1,\dots,k\}$, $\left|\sum_{j=1}^n x_{j,i}\right| \le \sqrt{n}\sqrt{2\ln(k^2)} + \sqrt{n}$.

If $x \in \mathcal{T}''$ then for any i', $\mu_{i'}^n(x)$ close to $\mu_0^n(x)$. Indeed, recall that for any $y \in \Omega$, $\mu_i(y)/\mu_0(y) = 1 + y_i\rho$. Hence,

$$\mu_i^n(x)/\mu_0^n(x) = \prod_{j=1}^n(1+\rho x_{j,i}) \approx 1 + \sum_{j=1}^n \rho x_{j,i} = 1 \pm O(\rho\sqrt{n\log k}),$$

and recall that $\rho = O(1/\sqrt{n\log k})$. To conclude, there exists some set $\mathcal{T}''$ such that for any $i' \in \{0,1,\dots,k\}$, $\mu_{i'}(x \in \mathcal{T}'') \ge 1 - 2/k$ and for any $x \in \mathcal{T}''$ and any $i \in \{1,\dots,k\}$, $\mu_i^n(x)/\mu_0^n(x) = 1 \pm O(\sqrt{n\log k}\rho)$.

Next, we formalize the above intuition and prove a result which holds with a slightly higher probability. First, we can assume that the constant C' in the statement of Theorem 4 is sufficiently large such that

$$\rho \le \frac{1}{2\sqrt{n}\left(2\sqrt{2\ln(8k^2)}+3\right)}. \tag{10}$$

Denote by $\mathcal{T}$ the set of all samples $x \in \Omega^n$ such that for all $1 \leq i \leq n$,

$$\left| \frac{\mu_i^n(x)}{\mu_0^n(x)} - 1 \right| \leq \alpha,$$

where α is a positive number which satisfies the equation

$$\alpha = \left(2\sqrt{2\ln(2k^2/\alpha^2)} + 3\right) \rho\sqrt{n}. \tag{11}$$

In the next two lemmas, we will show that $\alpha = \tilde{\Theta}(\sqrt{n}\rho)$ and that additionally, for all $0 \leq i' \leq k$, $\mu_{i'}^n(\mathcal{T}^c) \leq \alpha^2/k$. Hence, we change the above claim by replacing $1 - 2/k$ with $1 - \alpha^2/k$.

Lemma 23 *There is a unique positive number α which satisfies this equation and*

$$3\sqrt{n}\rho \leq \alpha \leq \min\left\{1/2, \rho\sqrt{n}\left(2\sqrt{2\ln(2k^2/(9n\rho^2))} + 3\right)\right\}.$$

Proof Such an α exists: if $\alpha \to 0$ then the RHS of Eq. (11) goes to ∞. If $\alpha = 1/2$ then, from Eq. (10), the RHS of Eq. (11) equals

$$\left(2\sqrt{2\ln(2k^2/\alpha^2)} + 3\right) \rho\sqrt{n} = \left(2\sqrt{2\ln(8k^2)} + 3\right) \rho\sqrt{n} \leq 1/2.$$

Hence, by the intermediate value theorem, there exists a value of $0 < \alpha \leq 1/2$ which satisfies the equation. Since the RHS is monotonically decreasing in α whenever $\alpha > 0$, there is just one solution for $\alpha > 0$.

For the last inequalities, it holds from definition that $\alpha \geq 3\rho\sqrt{n}$ and substituting α with $3\rho\sqrt{n}$ in its definition implies that

$$\alpha = \left(2\sqrt{2\ln(2k^2/\alpha^2)} + 3\right) \rho\sqrt{n} \leq \rho\sqrt{n}\left(2\sqrt{2\ln(2k^2/(9n\rho^2)} + 3\right).$$

■

Lemma 24 *For all $0 \leq i' \leq k$, $\mu_{i'}^n(\mathcal{T}^c) \leq \alpha^2/k$.*

Proof Define $p = \frac{\alpha^2}{k}$ and $a = \sqrt{2\ln(2k/p)} + 1$ and let $\mathcal{T}'$ be the set of all $x \in \Omega^n$ such that for all $1 \leq i \leq k$,

$$\left|\sum_{j=1}^n x_{j,i}\right| \leq \sqrt{n}a.$$

The proof is divided into two main claims:

1. For all $0 \leq i' \leq k$, $\mu_{i'}^n\left((\mathcal{T}')^c\right) \leq p$.

2. It holds that $\mathcal{T}' \subseteq \mathcal{T}$.

This two claims suffice to conclude the proof. We start by proving the first claim. Note that from definition of a $\mathrm{BCD}(\rho)$ family of distributions, for any $j \in [n]$ and $i \in [k]$, $\Pr_{A\sim\mu_0^n}[A_{j,i} = 1] = \Pr_{A\sim\mu_0^n}[A_{j,i} = -1] = 1/2$. Next, note that for any $i \in [k]$, if $A \sim \mu_{i'}^n$, then by the definition of a $\mathrm{BCD}(\rho)$ family, $\Pr_{A\sim\mu_{i'}^n}[A_{j,i} = 1] \le (1+\rho)\Pr_{A\sim\mu_0^n}[A_{j,i} = 1] \le \frac{1}{2}(1+\rho)$, hence $\mathbb{E}_{A\sim\mu_{i'}^n}[A_{j,i}] \le \rho$, and similarly, $\mathbb{E}_{A\sim\mu_{i'}^n}[A_{j,i}] \ge -\rho$. We conclude that for any $0 \le i' \le k$, $j \in [n]$ and $i \in [k]$, $\left|\mathbb{E}_{A\sim\mu_{i'}^n}[A_{j,i}]\right| \le \rho$. Hoeffding's inequality states that if $A_1, \dots, A_n$ are independent random variables getting values in $[-1, 1]$, then for any $\beta > 0$,

$$\Pr\left[\left|\sum_{j=1}^n A_j - \mathbb{E}\sum_{j=1}^n A_j\right| > \sqrt{n}\beta\right] \le 2e^{-\beta^2/2}.$$

Fix $0 \le i' \le k$ and $i \in [k]$, and let A be a random variable distributed $\mu_{i'}^n$. Then,

$$\begin{aligned}\Pr\left[\left|\sum_{j=1}^n A_{j,i}\right| > \sqrt{n}a\right] &\le \Pr\left[\left|\sum_{j=1}^n A_{j,i} - \mathbb{E}\sum_{j=1}^n A_{j,i}\right| + \left|\mathbb{E}\sum_{j=1}^n A_{j,i}\right| > \sqrt{n}a\right] \\ &\le \Pr\left[\left|\sum_{j=1}^n A_{j,i} - \mathbb{E}\sum_{j=1}^n A_{j,i}\right| > \sqrt{n}\sqrt{2\ln(2k/p)}\right] \qquad (12) \\ &= \frac{p}{k}, \qquad (13)\end{aligned}$$

where Eq. (12) follows from $\rho \le n^{-1/2}$ (see Eq. (10)) which implies that $\left|\mathbb{E}\sum_{j=1}^n A_{j,i}\right| \le n\rho \le \sqrt{n}$; and Eq. (13) follows from Hoeffding's inequality. A union bound over $i \in [k]$ implies that

$$\mu_{i'}^n\left((\mathcal{T}')^c\right) \le \sum_{i=1}^k \Pr\left[\left|\sum_{j=1}^n A_{j,i}\right| > \sqrt{n}a\right] \le p.$$

Next, we show that $\mathcal{T}' \subseteq \mathcal{T}$. Fix $x \in \mathcal{T}'$ and $i \in [k]$, and we will show that $|\mu_i^n(x)/\mu_0^n(x) - 1| \le \alpha$ to conclude the proof. Note that

$$\rho\sqrt{n}a \le \left(\sqrt{2\ln(2k^2/\alpha^2)} + 1\right)\rho\sqrt{n} \le \alpha/2 \le 1/4, \qquad (14)$$

where the second inequality follows from the definition of α and the third inequality follows from the bound $\alpha \le 1/2$ which Let $\ell = \left|\sum_{j=1}^n x_{j,i}\right|$ and $b \in \{-1, 1\}$ be the sign of $\sum_{j=1}^n x_{j,i}$ ($b = 1$ if the sum equals zero). By definition of $\mathcal{T}'$, $\ell \le \sqrt{n}a$. There are $\frac{n+\ell}{2}$ values of j for which $x_{j,i} = b$ and $\frac{n-\ell}{2}$ values for which $x_{j,i} = -b$. It holds that

$$\begin{aligned}\frac{\mu_i^n(x)}{\mu_0^n(x)} &= (1+b\rho)^{(n+\ell)/2}(1-b\rho)^{(n-\ell)/2} \qquad (15) \\ &= (1-\rho^2)^{(n-\ell)/2}(1+b\rho)^{\ell} \\ &\le (1+\rho)^{\ell} \le (1+\rho)^{\sqrt{na}} \le e^{\rho\sqrt{na}} \le 1 + 2\rho\sqrt{n}a,\end{aligned}$$

where Eq. (15) follows from the fact that by definition of a BCD(ρ) family, for any $x \in \Omega$, $\mu_i(x)/\mu_0(x) = 1 + \rho x_i$; one before the last inequality follows from $1 + s \le e^s$ for all $s \in \mathbb{R}$; and the last inequality follows from $e^s \le 1 + 2s$ for all $0 \le s \le 1$ and the from Eq. (14). Bounding from below,

$$\begin{aligned}\frac{\mu_i^n(x)}{\mu_0^n(x)} &= (1-\rho^2)^{(n-\ell)/2}(1+b\rho)^\ell \ \ge\ (1-\rho^2)^{n/2}(1-\rho)^{\sqrt{n}a} \\ &\ge 1 - \rho^2 n/2 - \rho\sqrt{n}a \ \ge\ 1 - \rho\sqrt{n}(a+1/2), \end{aligned} \tag{16}$$

where the last inequality follows from $\rho \le n^{-1/2}$ which follows from Eq. (10). In conclusion, Eq. (14) and Eq. (16) imply that

$$\left| \frac{\mu_i^n(x)}{\mu_0^n(x)} - 1 \right| \le 2a\rho\sqrt{n} \le \alpha, \tag{17}$$

where the last inequality follows from Eq. (14). This confirms that $\mathcal{T}' \subseteq \mathcal{T}$ as required. ■

To give intuition for the next part of the proof, assume the false assumption that $|\mu_i^n(x)/\mu_0^n(x) - 1| \le \alpha$ for all $x \in \Omega^n$ (instead of only when $x \in \mathcal{T}$). Define a Markov chain $X \to Y \to Z$ as follows: first, X is drawn from μ_0^n. Then, given X, $Y = (Y_1, \dots, Y_k) \in \{-1,1\}^k$ is drawn such that

$$\Pr\left[Y_i = 1 \mid X\right] = \frac{1}{2} + \frac{\mu_i^n(X)/\mu_0^n(X) - 1}{4\alpha}$$

and each bit of Y is distributed independently conditioned on X. Note that due to the assumption $|\mu_i^n(X)/\mu_0^n(x) - 1| \le \alpha$ it holds that $0 \le \Pr[Y_i = 1 \mid X] \le 1$ as required. Next, we define $Z = (Z_1, \dots, Z_k) \in \{-1,1\}^n$ as follows: conditioned on Y, each bit Z_i equals Y_i with probability $\frac{1}{2}(1+2\alpha)$ and otherwise $Z_i = -Y_i$; additionally, the bits of Z are independent conditioned on Y. A simple calculation shows that for any X, $\Pr[Z_i = 1,\ X] = \mu_i^n(X)/2$. Summing over X, one obtains that $\Pr[Z_i = 1] = 1/2$. Using Bayes' rule, $\Pr[X \mid Z_i = 1] = \mu_i^n(X)$. To sum up, one obtains the following properties:

- The random variable X is distributed μ_0^n. Conditioned on $Z_i = 1$, X is distributed μ_i^n.
- The random variable Z is uniform.
- The random variable Z_i is a noisy version of Y_i.

Due to the fact that $|\mu_i^n(x)/\mu_0^n(x) - 1| \le \alpha$ only for $X \in \mathcal{T}$, one cannot define the channel $Y \mid X$ as defined above, or otherwise it will not hold that $0 \le \Pr[Y_i = 1 \mid X] \le 1$. Hence, we change the definition of $P_{Y|X}$. Define the function $\psi \colon \mathbb{R} \to \mathbb{R}$ by

$$\psi(s) = \begin{cases} -1 & s \le -1 \\ s & -1 \le s \le 1 \\ 1 & s \ge 1 \end{cases}. \tag{18}$$

The function ψ should be viewed as "the identity except for some exceptional cases", where the exceptional cases correspond to $X \notin \mathcal{T}$, as will be clear next. Define the channel $P_{Y|X}$ as follows:

given X, each coordinate of Y is set independently to -1 or 1, where for any coordinate i [7],

$$P_{Y|X}(Y_i \mid X) = \frac{1}{2} + \frac{1}{4} Y_i \psi\left(\frac{\mu_i^n(X)/\mu_0^n(X) - 1}{\alpha}\right). \tag{19}$$

Note that for $X \in \mathcal{T}$, the function ψ behaves as the identity and we obtain the previous definition of $P_{Y|X}$. The following lemma characterizes the joint distribution P_{XYZ}, which satisfies an approximate version of the desired properties listed above.

Lemma 25 *The following holds for the distribution P_{XYZ}:*

1. $P_{XY}(X,Y) = 2^{-k}\mu_0^n(X)\prod_{i=1}^k \left(1 + \frac{1}{2}Y_i\psi\left(\frac{\mu_i^n(X)/\mu_0^n(X)-1}{\alpha}\right)\right)$.
2. $P_{XZ}(X,Z) = 2^{-k}\mu_0^n(X)\prod_{i=1}^k \left(1 + \alpha Z_i\psi\left(\frac{\mu_i^n(X)/\mu_0^n(X)-1}{\alpha}\right)\right)$.
3. *For all* $x \in \mathcal{T}$, $P_{XZ_i}(X,1) = \mu_i^n(x)/2$.
4. *For all* $1 \le i \le k$:
$$\left|P_{Z_i}(1) - \frac{1}{2}\right| \le \max_{0\le i\le k} \mu_i^n(\mathcal{T}^c) \le \frac{\alpha^2}{k}.$$
5. *For all* $1 \le i \le k$:
$$\left|P_{Y_i}(1) - \frac{1}{2}\right| = \left|P_{Z_i}(1) - \frac{1}{2}\right|/(2\alpha) \le \frac{\alpha}{2k}.$$

Before proving this lemma we will prove an auxiliary lemma.

Lemma 26 *Let $A \to B$ be a Markov chain, where $A, B \in \{-1,1\}$ are binary random variables. Assume that $P_A(1) = (1+a)/2$ and assume that $P_{B|A}$ is a channel that flips its input with probability $(1-b)/2$ for some $a, b \in [-1,1]$. Then $P_B(B) = (1+Bab)/2$.*

Proof The proof is by calculation:

$$P_B(1) = P_{AB}(1,1) + P_{AB}(-1,1) = (1+a)(1+b)/4 + (1-a)(1-b)/4 = (1+ab)/2.$$

Additionally, $P_B(-1) = 1 - P_B(1) = (1-ab)/2$. ■

Proof [Proof of Lemma 25]

We will prove the lemma items one by one. The first item follows from definition of $X \to Y$. For proving the second item, fix some $x \in \Omega^n$, and note that conditioned on $X = x$, each Y_i is binary as defined in Eq. (19). It holds that $P_{Z_i|Y_i}$ is a channel that flips its input Y_i with probability $(1-2\alpha)/2$, therefore applying Lemma 26 with $P_A = P_{Y_i|X=x}$, $P_{B|A} = P_{Z_i|Y_i}$, $a = \frac{1}{2}\psi\left(\frac{\mu_i^n(X)/\mu_0^n(X)-1}{\alpha}\right)$ and $b = 2\alpha$, we get that

$$P_{Z_i|X}(Z_i \mid x) = \frac{1}{2}\left(1 + \alpha Z_i \psi\left(\frac{\mu_i^n(x)/\mu_0^n(x) - 1}{\alpha}\right)\right). \tag{20}$$

[7] We assume μ_0 has full support, otherwise we can remove from Ω all elements x with $\mu_0(x) = 0$: by definition of a BCD(ρ) family, for all $1 \le i \le k$ it also holds that $\mu_i(x) = 0$.

Note that the bits of Z are independent conditioned on X: bits of Y_i are independent conditioned on X and each Z_i depends only on Y_i. Hence,

$$P_{XZ}(X,Z) = P_X(X)P_{Z|X}(Z \mid X) = P_X(X)\prod_{i=1}^{k} P_{Z_i|X}(Z_i \mid X), \tag{21}$$

and the second item follows from Eq. (20), Eq. (21) and the fact that $P_X(X) = \mu_0^n(X)$ by definition.

The third item is proved as follows:

$$\begin{aligned} P_{XZ_i}(x,1) &= P_X(x)P_{Z_i|X}(1|x) \\ &= \mu_0^n(x)\frac{1}{2}\left(1+\alpha\psi\left(\frac{\mu_i^n(x)/\mu_0^n(x)-1}{\alpha}\right)\right) \qquad (22)\\ &= \mu_0^n(x)\frac{1}{2}\left(1+\alpha\left(\frac{\mu_i^n(x)/\mu_0^n(x)-1}{\alpha}\right)\right) \qquad (23)\\ &= \frac{1}{2}\mu_i^n(x), \end{aligned}$$

where Eq. (22) follows from the fact that $X \sim \mu_0^n$ by definition of X and from Eq. (20); and Eq. (23) follows from the fact that whenever $X \in \mathcal{T}$, $|\mu_i^n(X)/\mu_0^n(X) - 1| \le \alpha$ by definition of $\mathcal{T}$, hence by definition of ψ,

$$\psi\left(\frac{\mu_i^n(X)/\mu_0^n(X)-1}{\alpha}\right) = \frac{\mu_i^n(X)/\mu_0^n(X)-1}{\alpha}.$$

To prove the fourth item,

$$P_{Z_i}(1) \ge \sum_{x\in\mathcal{T}} P_{Z_iX}(1,x) = \frac{1}{2}\mu_i^n(\mathcal{T}) = \frac{1}{2} - \frac{1}{2}\mu_i^n(\mathcal{T}^c), \tag{24}$$

where the first equation follows from the third item. Additionally,

$$\begin{aligned} P_{Z_i}(1) &= \sum_{x\in\mathcal{T}} P_{Z_iX}(1,x) + \sum_{x\notin\mathcal{T}} P_{Z_iX}(1,x) \\ &\le \sum_{x\in\mathcal{T}} \frac{1}{2}\mu_i^n(x) + \sum_{x\notin\mathcal{T}} P_X(x) \qquad (25)\\ &= \frac{1}{2}\mu_i(\mathcal{T}) + \sum_{x\notin\mathcal{T}} \mu_0^n(x) \qquad (26)\\ &\le \frac{1}{2} + \mu_0^n(\mathcal{T}^c), \qquad (27) \end{aligned}$$

where Eq. (25) follows from the third item and Eq. (26) follows from the fact that $X \sim \mu_0^n$ by definition. Eq. (24) and Eq. (27) imply that $|P_{Z_i}(1) - 1/2| \le \max_{0\le i\le k} \mu_i^n(\mathcal{T}^c)$, and the fourth item follow from Lemma 24.

The fifth item follows from Lemma 26 and the fact that $P_{Z_i|Y_i}$ flips its input with probability $(1-2\alpha)/2$: substitute $A = Y_i$, $B = Z_i$, $P_{Y_i}(1) = \frac{1}{2}(1+a)$ and $b = 2\alpha$. The lemma implies that $P_{Z_i}(1) = \frac{1}{2}(1+ab)$. Hence,

$$P_{Y_i}(1) - \frac{1}{2} = \frac{a}{2} = \frac{ab}{2}\frac{1}{b} = \left(P_{Z_i} - \frac{1}{2}\right)\frac{1}{2\alpha}.$$

■

Next, we claim that the coordinates of Y are almost independent. An intuitive explanation was given in Section 4, using the central limit theorem. However, due to the slow convergence guarantees of the central limit theorem, we did not find how to apply it without requiring ρ to be exponentially small in k. Hence, we have an ad-hoc proof. It defines two auxiliary random variables, X' and Y'. The variable X' is uniform on $\{-1,1\}^k$, and in particular, its coordinates are independent. The random variable Y' is constructed from X' the same way that Y is constructed from X. Due to the fact that Y'_i depends only on X'_i, the coordinates of Y' are also independent. We compare the distribution of Y with the distribution of Y' and show that if the high-order correlations between the coordinates of X are low, then the distribution of Y is similar to the distribution of Y'. Assumption Eq. (1) of Theorem 4 assures that these higher order correlations are low. These claims are stated formally in the next lemma, where we prove that the entropy of Y is almost the entropy of a random variable uniform over $\{-1,1\}^k$.

Lemma 27 *There exists some absolute constant C such that*

$$H(Y) \geq k - C.$$

Proof We will show that for all $y \in \{-1,1\}^k$, $P_Y(y) \leq \frac{C'}{2^k}$ for some numerical constant $C' > 0$. This will imply that

$$H(Y) = \sum_{y\in\{-1,1\}^k} P_Y(y) \log \frac{1}{P_Y(y)} \geq \sum_{y\in\{-1,1\}^k} P_Y(y) \log \frac{2^k}{C'} = \log \frac{2^k}{C'} = k - \log C' \tag{28}$$

and complete the proof.

First, we give an equivalent definition to the channel $P_{Y|X}$ (note the original definition is in Eq. (19)):

$$P_{Y_i|X}(y_i|x) = \frac{1}{2}\left(1 + \frac{y_i}{2}\psi\left(\frac{1}{\alpha}\left(\prod_{j=1}^n (1 + x_{j,i}\rho) - 1\right)\right)\right), \tag{29}$$

where all bits of Y are drawn independently given X. This definition is obtained from the original definition by substituting $\mu_i^n(x)/\mu_0^n(x)$ with $\prod_{j=1}^n (1 + x_{j,i}\rho)$. Indeed,

$$\mu_i^n(x)/\mu_0^n(x) = \prod_{j=1}^n \mu_i(x_j)/\mu_0(x_j) = \prod_{j=1}^n (1 + x_{j,i}\rho),$$

where the last inequality follows from the definition of a $\mathrm{BCD}(\rho)$ family, which requires that if $w = (w_1,\dots,w_k) \in \Omega$ then $\mu_i(w)/\mu_0(w) = 1 + \rho w_i$. We will use this definition of $P_{Y|X}$ in this lemma since it depends only on X and does not depend on $\mu_0,\dots,\mu_k$.

Fix some $y = (y_1,\dots,y_k) \in \{-1,1\}^k$. Let $\{\mu'_1,\dots,\mu'_k\}$ be the $\mathrm{BCD}(\rho)$ family of distributions such that μ'_0, its corresponding μ_0 distribution, is uniform over $\{-1,1\}^k$ and $\mu'_1,\dots,\mu'_k$ are derived from μ'_0 as in the definition of a $\mathrm{BCD}(\rho)$ family: for all $1 \leq i \leq k$ and for all $(w_1,\dots,w_k) \in \{-1,1\}^k$,

$$\mu_i((w_1,\dots,w_k)) = \mu'_0((w_1,\dots,w_k))(1 + w_i\rho) = 2^{-k}(1 + w_i\rho).$$

Define X' and Y' to be analogous to X and Y with respect to this family: $X' \sim (\mu_0')^n$ and $P_{Y'} = P_{Y|X} \circ P_{X'}$, using the new definition of $P_{Y|X}$ from Eq. (29). Since Y_i' is a function of the i'th column of X' and the columns of X' are independent, $Y_1', \dots, Y_k'$ are independent. Item 5 of Lemma 25 and Lemma 23 show that $|P_{Y_i'}(1) - 1/2| \le \frac{\alpha}{2k} \le \frac{1}{2k}$ for all $1 \le i \le k$ [8], therefore

$$P_{Y'}(y) = \prod_{i=1}^{k} P_{Y_i'}(y_i) \le \frac{1}{2^k}\left(1 + \frac{1}{k}\right)^k \le \frac{e}{2^k}. \tag{30}$$

We will bound $P_Y(y)/P_{Y'}(y)$ to complete the proof.

Recall that each row of X is a vector distributed according to μ_0 and each row of X' is a vector distributed according to μ_0'. Define intermediate random variables $X^{(0)}, X^{(1)}, \dots, X^{(n)}$ such that for all $0 \le j \le n$, rows 1 to j of $X^{(j)}$ are distributed according to μ_0 and rows $j+1$ to n are distributed according to μ_0', where all rows are independent. Define the random variables $Y^{(0)}, \dots, Y^{(n)} \in \{-1,1\}^k$ accordingly, namely $Y^{(j)} \sim P_{Y|X} \circ P_{X^{(j)}}$. It holds that $Y^{(0)}$ has the same distribution as Y' and $Y^{(n)}$ is distributed the same as Y.

Fix some $1 \le \ell \le n$ and we will bound $P_{Y^{(\ell)}}(y)/P_{Y^{(\ell-1)}}(y)$. Let $X_\ell^{(\ell)}$ be column ℓ of $X^{(\ell)}$, let $X_{-\ell}^{(\ell)}$ be $X^{(\ell)}$ without column ℓ and define $X_\ell^{(\ell-1)}$ and $X_{-\ell}^{(\ell-1)}$ similarly. Fix some $x_{-\ell} \in \{-1,1\}^{(n-1)\times k}$. For all $i \in [k]$, let

$$\begin{aligned}
p_i &= \Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell}\right] \\
&= \sum_{b \in \{-1,1\}} \Pr\left[X_{\ell,i}^{(\ell)} = b \middle| X_{-\ell}^{(\ell)} = x_{-\ell}\right] \Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell},\ X_{\ell,i}^{(\ell)} = b\right] \\
&= \sum_{b \in \{-1,1\}} \Pr\left[X_{\ell,i}^{(\ell)} = b\right] \Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell},\ X_{\ell,i}^{(\ell)} = b\right] && (31) \\
&= \frac{1}{2}\Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell},\ X_{\ell,i}^{(\ell)} = -1\right] + \frac{1}{2}\Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell},\ X_{\ell,i}^{(\ell)} = 1\right], && (32)
\end{aligned}$$

where Eq. (31) follows from the fact that the rows of $X^{(\ell)}$ are independent, and Eq. (32) follows from the fact that $X_\ell^{(\ell)}$ is distributed μ_0, and by definition of a $\mathrm{BCD}(\rho)$ family, each bit is uniform under μ_0. Recall that $Y^{(\ell)} = P_{Y|X} \circ X^{(\ell)}$. It holds that $p_i = \Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell}\right] > 0$ by definition of $P_{Y|X}$. Hence, one can define

$$\delta_i = \Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell},\ X_{\ell,i}^{(\ell)} = 1\right] / p_i - 1,$$

which implies that

$$\Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell},\ X_{\ell,i}^{(\ell)} = 1\right] = p_i\left(1 + \delta_i\right). \tag{33}$$

[8] Note that this item proves a corresponding statement on Y_i, however, we can also substitute it with Y_i': if we substitute $\mu_1, \dots, \mu_k$ with $\mu_1', \dots, \mu_k'$, all the requirements of Theorem 4 are satisfied: the only assumption on the family of distributions is Eq. (1), which the new family $\mu_1', \dots, \mu_k'$ satisfies, but since we haven't used this assumption yet, Lemma 25 applies to Y_i' even without requiring the new family to satisfy Eq. (1).

By Eq. (33) and by Eq. (32), for any value of $X^{(\ell)}_{\ell,i} \in \{-1,1\}$,

$$\Pr\left[Y^{(\ell)}_i = y_i \middle| X^{(\ell)}_{-\ell} = x_{-\ell},\ X^{(\ell)}_{\ell,i}\right] = p_i\left(1 + X^{(\ell)}_{\ell,i}\delta_i\right). \tag{34}$$

Furthermore, since $Y^{(\ell)}_i$ depends only on column i of $X^{(\ell)}$, for any value of $X^{(\ell)}_\ell \in \{-1,1\}^k$,

$$\Pr\left[Y^{(\ell)}_i = y_i \middle| X^{(\ell)}_{-\ell} = x_{-\ell},\ X^{(\ell)}_{\ell}\right] = p_i\left(1 + X^{(\ell)}_{\ell,i}\delta_i\right).$$

Since the bits of $Y^{(\ell)}$ are independent conditioned on $X^{(\ell)}$,

$$\Pr\left[Y^{(\ell)} = y \middle| X^{(\ell)}_{-\ell} = x_{-\ell},\ X^{(\ell)}_{\ell}\right] = \prod_{i=1}^{k} p_i\left(1 + X^{(\ell)}_{\ell,i}\delta_i\right). \tag{35}$$

Hence,

$$\begin{aligned}\Pr\left[Y^{(\ell)} = y \middle| X^{(\ell)}_{-\ell} = x_{-\ell}\right] &= \mathbb{E}_{X^{(\ell)}_\ell}\left[\Pr\left[Y^{(\ell)} = y \middle| X^{(\ell)}_{-\ell} = x_{-\ell},\ X^{(\ell)}_{\ell}\right]\right] \\ &= \mathbb{E}_{X^{(\ell)}_\ell}\prod_{i=1}^{k}(p_i(1 + X^{(\ell)}_{\ell,i}\delta_i)) \qquad (36) \\ &= \left(\prod_{i=1}^{k} p_i\right)\mathbb{E}_{A=(A_1,\ldots,A_k)\sim\mu_0}\prod_{i=1}^{k}(1 + A_i\delta_i) \qquad (37) \\ &= \left(\prod_{i=1}^{k} p_i\right)\sum_{S\subseteq[k]}\mathbb{E}_{A=(A_1,\ldots,A_k)\sim\mu_0}\left[\prod_{i\in S} A_i\delta_i\right] \\ &= \left(\prod_{i=1}^{k} p_i\right)\sum_{S\subseteq[k]}\mathbb{E}_{A=(A_1,\ldots,A_k)\sim\mu_0}\left[\prod_{i\in S}\left(\mu_i(A)/\mu_0(A) - 1\right)\delta_i/\rho\right]. \qquad (38)\end{aligned}$$

where Eq. (36) follows from Eq. (35), Eq. (37) follows from the fact that $X^{(\ell)}_\ell \sim \mu_0$ by definition of $X^{(\ell)}$, and Eq. (38) follows from the fact that by definition of a $\mathrm{BCD}(\rho)$ family, $\mu_i(A) = \mu_0(A)(1 + A_i\rho)$. It holds that

$$|\delta_i p_i| = \frac{1}{2}\left|\Pr\left[Y^{(\ell)}_i = y_i \middle| X^{(\ell)}_{-\ell} = x_{-\ell},\ X^{(\ell)}_{\ell,i} = 1\right] - \Pr\left[Y^{(\ell)}_i = y_i \middle| X^{(\ell)}_{-\ell} = x_{-\ell},\ X^{(\ell)}_{\ell,i} = -1\right]\right| \tag{39}$$

$$= \frac{1}{8}\left|\psi\left(\frac{1}{\alpha}\left((1+\rho)\prod_{1\le j\le n, j\ne\ell}(1 + (x_{-\ell})_{j,i}\rho) - 1\right)\right) - \psi\left(\frac{1}{\alpha}\left((1-\rho)\prod_{1\le j\le n, j\ne\ell}(1 + (x_{-\ell})_{j,i}\rho) - 1\right)\right)\right|. \tag{40}$$

where Eq. (39) follows from Eq. (34) and Eq. (40) follows from the fact that $Y^{(\ell)} = P_{Y|X} \circ X^{(\ell)}$ and from the definition of $P_{Y|X}$. If

$$\frac{1}{\alpha}\left((1-\rho)\prod_{1\le j\le n, j\ne\ell}(1 + (x_{-\ell})_{j,i}\rho) - 1\right) \ge 1$$

then by Eq. (40) and by definition of ψ in Eq. (18),

$$|\delta_i p_i| = \frac{1}{8}\,|1-1| = 0. \tag{41}$$

Otherwise,

$$\prod_{1\le j\le n, j\ne \ell} (1+(x_{-\ell})_{j,i}\rho) \le \frac{\alpha+1}{1-\rho} \le 3,$$

since $\rho \le \alpha \le 1/2$ by Lemma 23. Since ψ is 1-Lipschitz, Eq. (40) is at most

$$\frac{1}{8\alpha}((1+\rho)-(1-\rho)) \prod_{1\le j\le n, j\ne\ell} (1+(x_{-\ell})_{j,i}\rho) \le \frac{6\rho}{8\alpha}. \tag{42}$$

By Eq. (41) and Eq. (42), we conclude that $|\delta_i p_i| \le 3\rho/(4\alpha)$. It holds that $p_i = \Pr\left[Y_i^{(\ell)} = y_i \middle| X_{-\ell}^{(\ell)} = x_{-\ell}\right] \ge 1/4$ by definitions of p_i and $P_{Y|X}$, and $\alpha \ge 3\sqrt{n}\rho$ by Lemma 23, hence

$$|\delta_i| \le \frac{3\rho}{4\alpha p_i} \le \frac{3\rho}{\alpha} \le \frac{1}{\sqrt{n}}.$$

Hence, Eq. (38) is at most

$$\left(\prod_{i=1}^k p_i\right) \sum_{S\subseteq[k]} n^{-|S|/2}\rho^{-|S|} \left|\mathbb{E}_{A\sim\mu_0} \prod_{i\in S} (\mu_i(A)/\mu_0(A)-1)\right| \le \left(\prod_{i=1}^k p_i\right)\left(1+\frac{1}{n}\right),$$

where the last step follows from Eq. (1) and the fact that all terms corresponding to $|S|=1$ equal zero. This concludes that

$$\Pr\left[Y^{(\ell)} = y \middle| X_{-\ell}^{(\ell)} = x_{-\ell}\right] \le \left(\prod_{i=1}^k p_i\right)\left(1+\frac{1}{n}\right). \tag{43}$$

Since $Y^{(\ell-1)}$ is obtained from $X^{(\ell-1)}$ the same $Y^{(\ell)}$ is obtained from $X^{(\ell)}$ (using the conditioned probabilities of $P_{Y|X}$), it holds that

$$\Pr\left[Y^{(\ell-1)} = y \middle| X_{-\ell}^{(\ell-1)} = x_{-\ell},\ X_\ell^{(\ell-1)}\right] = \Pr\left[Y^{(\ell)} = y \middle| X_{-\ell}^{(\ell)} = x_{-\ell},\ X_\ell^{(\ell)}\right] = \prod_{i=1}^k p_i\left(1+X_{\ell,i}^{(\ell)}\delta_i\right) \tag{44}$$

where the last equation follows from Eq. (35). Since the entries of $X_\ell^{(\ell-1)}$ are distributed μ_0', they are independent, hence

$$\begin{aligned}
\Pr\left[Y^{(\ell-1)} = y \middle| X_{-\ell}^{(\ell-1)} = x_{-\ell}\right] &= \mathbb{E}_{X_\ell^{(\ell-1)}} \Pr\left[Y^{(\ell-1)} = y \middle| X_{-\ell}^{(\ell-1)} = x_{-\ell}, X_\ell^{(\ell-1)}\right] \\
&= \mathbb{E}_{X_\ell^{(\ell-1)}}\left[\prod_{i=1}^k p_i\left(1+X_{\ell,i}^{(\ell)}\delta_i\right)\right] && (45) \\
&= \prod_{i=1}^k \mathbb{E}_{X_\ell^{(\ell-1)}}\left[p_i\left(1+X_{\ell,i}^{(\ell)}\delta_i\right)\right] && (46) \\
&= \prod_{i=1}^k p_i && (47)
\end{aligned}$$

where Eq. (45) follows from Eq. (44), Eq. (46) follows from the fact that entries of $X_\ell^{(\ell-1)}$ are independent, and Eq. (47) follows from the fact that $X_\ell^{(\ell-1)}$ is distributed μ_0' and by definition of μ_0', each bit of $X_\ell^{(\ell-1)}$ is distributed uniformly. Eq. (47) and Eq. (43) imply that

$$\Pr\left[Y^{(\ell-1)}=y\middle|X_{-\ell}^{(\ell-1)}=x_{-\ell}\right]=\prod_{i=1}^{k}p_i\geq\left(1+\frac{1}{n}\right)^{-1}\Pr\left[Y^{(\ell)}=y\middle|X_{-\ell}^{(\ell)}=x_{-\ell}\right]. \tag{48}$$

Since $X_{-\ell}^{(\ell)}$ and $X_{-\ell}^{(\ell-1)}$ have the same distribution, we can take an expectation over $X_{-\ell}^{(\ell-1)}$ in the LHS of Eq. (48) and over $X_{-\ell}^{(\ell)}$ in the RHS, and obtain that

$$P_{Y^{(\ell-1)}}(y)\geq\left(1+\frac{1}{n}\right)^{-1}P_{Y^{(\ell)}}(y).$$

Therefore,

$$P_Y(y)=P_{Y^{(n)}}(y)\leq\left(1+\frac{1}{n}\right)^{n}P_{Y^{(0)}}(y)\leq eP_{Y^{(0)}}(y)=eP_{Y'}(y)\leq\frac{e^2}{2^k},$$

where the last inequality follows from Eq. (30). Eq. (28) implies that this concludes the proof. ■

Next, we show a chain of inequalities to conclude the proof. Fix some channel $P_{\Pi|X}$. Using this channel and $P_X=\mu_0^n$ we can define the joint distribution $P_{\Pi X}=P_{X,(\Pi|X)}$ and obtain the inverse channel $P_{X|\Pi}$ from this joint distribution. We can extend our Markov chain to $\Pi\to X\to Y\to Z$, where the conditional probability of X conditioned on Π is obtained from the channel $P_{X|\Pi}$. The data processing inequality (Proposition 18) implies that

$$I(\Pi;Y)\leq I(\Pi;X). \tag{49}$$

Lemma 27 enables us to bound $\sum_{i=1}^k I(\Pi;Y_i)$ in terms of $I(\Pi;Y)$. Formally, we obtain the following:

$$\begin{aligned}
\sum_{i=1}^{k}I(\Pi;Y_i) &= \sum_{i=1}^{k}\left(H(Y_i)-H(Y_i\mid\Pi)\right) && (50)\\
&\leq k-\sum_{i=1}^{k}\left(H(Y_i\mid\Pi)\right) && (51)\\
&\leq k-H(Y\mid\Pi) && (52)\\
&= k-H(Y)+I(\Pi;Y) && (53)\\
&\leq I(\Pi;Y)+C && (54)
\end{aligned}$$

where Eq. (50) is by definition of the mutual entropy[9], Eq. (51) follows from the fact that each Y_i is binary hence its entropy is at most 1, Eq. (52) follows from the inequality $H(AB\mid C)\leq H(A\mid C)+H(B\mid C)$ for all random variables A,B,C, Eq. (53) follows from the definition of the mutual entropy, Eq. (54) follows from Lemma 27 where C is the numeric constant from the lemma.

[9] The mutual entropy between two random variables A and B equals $I(A;B)=H(A)-H(A\mid B)=H(B)-H(B\mid A)$.

Next, we utilize the structure of the channel $P_{Z_i|Y_i}$ to strongly bound $I(\Pi; Z_i)$ in terms of $I(\Pi; Y_i)$. Recall that Z_i is a noisy version of Y_i. There exists a strong data processing for this channel (Ahlswede and Gács, 1976):

Proposition 28 *Let $A \to B \to C$ be a Markov chain such that B and C are binary random variables getting values in $\{-1, 1\}$. Let $0 \le q \le 1$ be a number and assume that the transition $B \to C$ is defined by $C = B$ with probability $(1 + q)/2$. Then $I(A; C) \le q^2 I(A; B)$.*

Since $\Pi \to Y_i \to Z_i$ is a Markov chain, applying Proposition 28 with $q = 2\alpha$ we get that

$$I(\Pi; Z_i) \le 4\alpha^2 I(\Pi; Y_i). \tag{55}$$

The following lemma by Bar-Yossef et al. (2004, Lemma 6.2), relates the Hellinger distance with the mutual information.

Lemma 29 *Let A and B be random variables such that A is uniform over $\{-1, 1\}$. Then*

$$\mathrm{h}^2(P_{B|A=-1}, P_{B|A=1}) \le I(A; B).$$

We would like to use this lemma in order to bound $\mathrm{h}^2(P_{\Pi|Z_i=-1}, P_{\Pi|Z_i=1})$ in terms of $I(\Pi; Z_i)$, however, Z_i is not necessarily uniform. On the other hand, Z_i is not very biased, hence one can reduce the case that Z_i is not very biased to the case that Z_i is uniform, as done in the following lemma.

Lemma 30 *Let A and B be random variables such that $A \in \{-1, 1\}$. Then*

$$\mathrm{h}^2(P_{B|A=-1}, P_{B|A=1}) \le \frac{I(A; B)}{2\min(\Pr[A = -1], \Pr[A = 1])}.$$

Proof Assume without loss of generality that $\Pr[A = 1] \ge \Pr[A = -1]$. Let $D \in \{0, 1\}$ be a random variable and we will define a Markov chain $D \to A \to B$, extending $A \to B$ [10]. Define the distribution of D and the conditional distribution of A (conditioned on D) as follows:

$$\begin{aligned}
&\Pr[D = 0] = 2\Pr[A = -1] = 2\min(\Pr[A = -1], \Pr[A = 1]) \\
&\Pr[A = 1 \mid D = 0] = 1/2 \\
&\Pr[A = 1 \mid D = 1] = 1.
\end{aligned}$$

Note that $P_{A|D} \circ P_D = P_A$ as required. Since A is deterministic when $D = 1$,

$$I(A; B \mid D = 1) = H(A \mid D = 1) - H(A \mid B, D = 1) = 0 - 0 = 0.$$

Since $D \to A \to B$ is a Markov chain, B is independent of D conditioned on A, hence

$$\begin{aligned}
\Pr[D = 0] I(A; B \mid D = 0) &= I(A; B \mid D) \;=\; H(B \mid D) - H(B \mid AD) \\
&\le H(B) - H(B \mid AD) \;=\; H(B) - H(B \mid A) \\
&= I(A; B),
\end{aligned} \tag{56}$$

[10] For any two random variables X and Y one can define a Markov chain $X \to Y$ by first drawing X and then drawing Y conditioned on X.

using the inequality $H(X \mid Y) \le H(X)$ for all random variables X, Y. Hence,

$$\begin{aligned} \mathrm{h}^2(P_{B|A=-1}, P_{B|A=1}) &= \mathrm{h}^2(P_{B|A=-1,D=0}, P_{B|A=1,D=0}) && (57)\\ &\le I(A;B \mid D=0) && (58)\\ &\le I(A;B)/\Pr[D=0]. && (59)\end{aligned}$$

where Eq. (57) follows from the fact that $B \to Z_i \to \Pi$ is a Markov chain hence Π is independent of B conditioned on Z_i, Eq. (58) follows from Lemma 29 and Eq. (59) follows from Eq. (56). ■

We apply Lemma 30 by setting $A = Z_i$ and $B = \Pi$. Lemma 25 and Lemma 23 imply that

$$1/2 - \min(\Pr[Z_i = 1], \Pr[Z_i = -1]) = |1/2 - \Pr[Z_i = 1]| \le \alpha^2/k \le 1/(4k) \le 1/4,$$

hence, Lemma 30 implies that

$$\mathrm{h}^2(P_{\Pi|Z_i=1}, P_{\Pi|Z_i=-1}) \le 2I(\Pi; Z_i). \tag{60}$$

Next, we claim that

$$\mathrm{h}^2(P_\Pi, P_{\Pi|Z_i=1}) \le \mathrm{h}^2(P_{\Pi|Z_i=-1}, P_{\Pi|Z_i=1}). \tag{61}$$

Indeed, P_Π is a convex combination of $P_{\Pi|Z_i=1}$ and $P_{\Pi|Z_i=-1}$, and one can verify that the following holds:

Proposition 31 *Let μ and ν be two probability distributions. Then, for any $0 \le \lambda \le 1$,*

$$\mathrm{h}((1-\lambda)\mu + \lambda\nu, \nu) \le \mathrm{h}(\mu, \nu).$$

In the following lemma we use the fact that conditioned on $Z_i = 1$, X is distributed similarly to μ_i^n, to bound $\mathrm{h}^2\left(P_\Pi, P_{\Pi|X\sim\mu_i^n}\right)$ in terms of $\mathrm{h}^2\left(P_\Pi, P_{\Pi|Z_i=1}\right)$.

Lemma 32 *It holds that*

$$\mathrm{h}^2\left(P_\Pi, P_{\Pi|X\sim\mu_i^n}\right) \le 2\mathrm{h}^2\left(P_\Pi, P_{\Pi|Z_i=1}\right) + 5\frac{\alpha^2}{k}.$$

Proof We inform the reader that any Markov chain can be reversed to get a new Markov chain: if $X_1 \to \cdots \to X_\ell$ is a Markov chain then the joint distribution of $X_1 \cdots X_\ell$ can be viewed as a Markov chain $X_\ell \to \cdots \to X_1$. Since $Z_i \to Y_i \to X \to \Pi$ is a Markov chain, $P_{\Pi|Z_i=1} = P_{\Pi|X} \circ P_{X|Z_i=1}$ and $P_{\Pi|X\sim\mu_i^n} = P_{\Pi|X} \circ \mu_i^n$. It holds that

$$\begin{aligned} \mathrm{h}^2\left(P_{\Pi|X\sim\mu_i^n}, P_{\Pi|Z_i=1}\right) &= \mathrm{h}^2\left(P_{\Pi|X} \circ \mu_i^n, P_{\Pi|X} \circ P_{X|Z_i=1}\right) \\ &\le \mathrm{h}^2\left(\mu_i^n, P_{X|Z_i=1}\right) && (62)\\ &\le d_{TV}\left(\mu_i^n, P_{X|Z_i=1}\right) && (63)\\ &= \frac{1}{2}\sum_x \left|\mu_i^n(x) - P_{X|Z_i=1}(x)\right| \\ &\le \frac{1}{2}\sum_x \left|\mu_i^n(x) - 2P_{Z_i}(1)P_{X|Z_i=1}(x)\right| + \frac{1}{2}\sum_x \left|2P_{Z_i}(1)P_{X|Z_i=1}(x) - P_{X|Z_i=1}(x)\right|, && (64)\end{aligned}$$

where Eq. (62) follows from the data processing inequality (Proposition 18) for the channel $X \to \Pi$ and Eq. (63) follows from Proposition 16. We will bound the two terms in Eq. (64) separately. First,

$$\begin{aligned}
\frac{1}{2}\sum_x \left|\mu_i^n(x) - 2P_{Z_i}(1)P_{X|Z_i=1}(x)\right| &= \frac{1}{2}\sum_x |\mu_i^n(x) - 2P_{XZ_i}(x,1)| \\
&= \frac{1}{2}\sum_{x\notin\mathcal{T}} |\mu_i^n(x) - 2P_{XZ_i}(x,1)| \qquad (65)\\
&\le \frac{1}{2}\sum_{x\notin\mathcal{T}} \mu_i^n(x) + \frac{1}{2}\sum_{x\notin\mathcal{T}} 2P_{XZ_i}(x,1) \\
&\le \frac{1}{2}\sum_{x\notin\mathcal{T}} \mu_i^n(x) + \sum_{x\notin\mathcal{T}} P_X(x) \\
&= \frac{1}{2}\sum_{x\notin\mathcal{T}} \mu_i^n(x) + \sum_{x\notin\mathcal{T}} \mu_0^n(x) \qquad (66)\\
&= \frac{1}{2}\mu_i^n(X\notin\mathcal{T}) + \mu_0^n(X\notin\mathcal{T}) \\
&\le \frac{3}{2}\alpha^2/k. \qquad (67)
\end{aligned}$$

where Eq. (65) follows from item 3 of Lemma 25, Eq. (66) follows from the definition of P_X and Eq. (67) follows from Lemma 24. Bounding the second term of Eq. (64), we get:

$$\frac{1}{2}\sum_x \left|2P_{Z_i}(1)P_{X|Z_i=1}(x) - P_{X|Z_i=1}(x)\right| = |P_{Z_i}(1) - 1/2|\sum_x P_{X|Z_i=1}(x) = |P_{Z_i}(1) - 1/2| \le \frac{\alpha^2}{k}, \qquad (68)$$

where the last inequality follows from the fourth item of Lemma 25. Eq. (68), Eq. (67) and Eq. (64) imply that

$$\mathrm{h}^2\left(P_{\Pi|X\sim\mu_i^n}, P_{\Pi|Z_i=1}\right) \le \frac{5\alpha^2}{2k}.$$

Since the (non-squared) Hellinger distance satisfies the triangle inequality, and $(a+b)^2 = a^2 + 2ab + b^2 \le 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{h}^2(P_\Pi, P_{\Pi|X\sim\mu_i^n}) &\le \left(\mathrm{h}(P_\Pi, P_{\Pi|Z_i=1}) + \mathrm{h}(P_{\Pi|Z_i=1}, P_{\Pi|X\sim\mu_i^n})\right)^2 \\
&\le 2\mathrm{h}^2(P_\Pi, P_{\Pi|Z_i=1}) + 2\mathrm{h}^2(P_{\Pi|Z_i=1}, P_{\Pi|X\sim\mu_i^n}) \\
&\le 2\mathrm{h}^2(P_\Pi, P_{\Pi|Z_i=1}) + 5\frac{\alpha^2}{k}.
\end{aligned}$$

■

To conclude the proof:

$$
\begin{aligned}
\sum_{i=1}^{k} \mathrm{h}^2\left(P_\Pi, P_{\Pi|X\sim\mu_i^n}\right) &\le 2\sum_{i=1}^{k} \mathrm{h}^2\left(P_\Pi, P_{\Pi|Z_i=1}\right) + 5\alpha^2 && (69)\\
&\le 2\sum_{i=1}^{k} \mathrm{h}^2(P_{\Pi|Z_i=-1}, P_{\Pi|Z_i=1}) + 5\alpha^2 && (70)\\
&\le 4\sum_{i=1}^{k} I(\Pi; Z_i) + 5\alpha^2 && (71)\\
&\le 16\alpha^2\sum_{i=1}^{k} I(\Pi; Y_i) + 5\alpha^2 && (72)\\
&\le 16\alpha^2 I(\Pi; Y) + (5+16C)\alpha^2 && (73)\\
&\le 16\alpha^2 I(\Pi; X) + (5+16C)\alpha^2 && (74)\\
&\le \rho\sqrt{n}\left(2\sqrt{2\ln(2k^2/(9n\rho^2))}+3\right)(16C+5)(I(\Pi;X)+1), && (75)
\end{aligned}
$$

where C is the constant from Lemma 27, Eq. (69) follows from Lemma 32, Eq. (70) follows from Eq. (61), Eq. (71) follows from Eq. (60), Eq. (72) follows from Eq. (55), Eq. (73) follows from Eq. (54), Eq. (74) follows from Eq. (49), and Eq. (75) follows from Lemma 23. Note that by definition $X \sim \mu_0^n$, hence $P_\Pi = P_{\Pi|X\sim\mu_0^n}$, which concludes the proof.

A.2.2. PROOF OF LEMMA 21

The core of the proof follows results of Braverman et al. (2016) and Jayram (2009). Let $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(m)}\right)$ be a random vector distributed $(\mu_0^n)^m$ where for all $j \in [m]$, $X^{(j)} \in \Omega^n$ is the input of player j. Let Π be the transcript of a $1/3$-error (m, n) protocol identifying $\mu \in \{\mu_1, \ldots, \mu_k\}$, distributed $P_{\Pi|\mathbf{X}}$ conditioned on the input of the players being $\mathbf{X}$. Given a vector $\mathbf{a} = (a_1, \ldots, a_m) \in \{0, 1, \ldots, k\}^m$, let $\Pi_{\mathbf{a}}$ be the random variable denoting the transcript Π when every player $j \in [m]$ receives an independent input distributed $\mu_{a_j}^n$. Formally, $\Pi_{\mathbf{a}} \sim P_{\Pi|\mathbf{X}\sim\left(\mu_{a_1}^n,\ldots,\mu_{a_m}^n\right)}$. For any $j \in [m]$ and $i \in [k]$, let $\mathbf{e}_{j,i}$ be the m-entry vector that equals i on coordinate j and all other coordinates are zero, and let $\mathbf{i}$ be the all-i vector.

Since $\mathbf{X} \sim \mu_0^{mn}$, for any $j \in [m]$, $P_{\Pi|X^{(j)}}$ is the distribution of Π conditioned on player j getting the input $X^{(j)} \in \Omega^n$ while all other players get an independent input distributed μ_0^n. Note that for all $j \in [m]$, $\Pi_{\mathbf{0}} \sim P_{\Pi|X^{(j)}\sim\mu_0^n}$, and for all $i \in [n]$, $\Pi_{\mathbf{e}_{j,i}} \sim P_{\Pi|X^{(j)}\sim\mu_i^n}$. Hence, the conditions of this lemma imply that

$$
\begin{aligned}
\sum_{j=1}^{m}\sum_{i=1}^{k} \mathrm{h}^2(\Pi_0, \Pi_{\mathbf{e}_{j,i}}) &= \sum_{j=1}^{m}\sum_{i=1}^{k} \mathrm{h}^2(P_{\Pi|X^{(j)}\sim\mu_0^n}, P_{\Pi|X^{(j)}\sim\mu_i^n})\\
&\le \sum_{j=1}^{m} \beta(I(\Pi; X^{(j)}) + 1). && (76)
\end{aligned}
$$

In order to bound the last term we present a known inequality in information theory.

Proposition 33 *If $X^{(1)}, \dots, X^{(m)}$ are independent random variables and Π is a random variable then*

$$\sum_{j=1}^{m} I\left(\Pi; X^{(j)}\right) \leq I\left(\Pi; X^{(1)} \cdots X^{(m)}\right).$$

Proof

$$\begin{aligned} I\left(\Pi; X^{(1)} \cdots X^{(m)}\right) &= \sum_{j=1}^{m} I\left(\Pi; X^{(j)} \mid X^{(1)} \cdots X^{(j-1)}\right) \\ &= \sum_{j=1}^{m} H\left(X^{(j)} \mid X^{(1)} \cdots X^{(j-1)}\right) - H\left(X^{(j)} \mid \Pi X^{(1)} \cdots X^{(j-1)}\right) \\ &\geq \sum_{j=1}^{m} H\left(X^{(j)}\right) - H\left(X^{(j)} \middle| \Pi\right) \\ &= \sum_{j=1}^{m} I\left(\Pi; X^{(j)}\right). \end{aligned}$$

where the first equation follows from the chain rule for mutual entropy and the first inequality follows from the independence of $X^{(1)} \cdots X^{(m)}$ and the fact that $H(A \mid BC) \leq H(A \mid B)$ for any random variables A, B, C. ∎

Hence, Eq. (76) implies that

$$\sum_{j=1}^{m} \sum_{i=1}^{k} \mathrm{h}^2(\Pi_0, \Pi_{\mathbf{e}_{j,i}}) \;\leq\; \beta(I(\Pi; \mathbf{X}) + m) \;\leq\; \beta(H(\Pi) + m) \;\leq\; \beta(|\Pi| + m), \tag{77}$$

where $|\Pi|$ is the communication complexity of the protocol. The following Lemma, Braverman et al. (2016, Lemma 2) lower bounds $\sum_{j=1}^{m} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_{j,i}})$.

Lemma 34 *For any $1 \leq i \leq m$,*

$$\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{i}}) \leq C \sum_{j=1}^{m} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_{j,i}})$$

for some numerical constant $C > 0$.

This and Eq. (77) implies that

$$\sum_{i=1}^{k} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{i}}) \leq C\beta(|\Pi| + m). \tag{78}$$

The next lemma states that for any protocol error $\varepsilon < 1/2$, the LHS of Eq. (78) is $\Omega(k)$.

Lemma 35 *Assume Π is a transcript of a protocol with a worst-case error (over μ_i) of at most $\varepsilon < 1/2$. Then there exists a subset $S \subseteq [k]$ of size $|S| = k-1$ such that for all $i \in S$,*

$$\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{i}}) \geq \frac{(1-2\varepsilon)^2}{8}.$$

In particular,

$$\sum_{i=1}^{k} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{i}}) \geq \frac{(k-1)(1-2\varepsilon)^2}{8}.$$

Proof First, note that for any $i \neq i' \in [k]$, $d_{TV}(\Pi_{\mathbf{i}}, \Pi_{\mathbf{i'}}) \geq 1 - 2\varepsilon$. Indeed, fix some $i \neq i'$ and let $\mathcal{A}$ be the set of all values of Π such that the protocol outputs i given these values. Since the protocol has ε-error, $\Pr[\Pi_{\mathbf{i}} \in \mathcal{A}] \geq 1 - \varepsilon$ and $\Pr[\Pi_{\mathbf{i'}} \in \mathcal{A}] \leq \varepsilon$. Hence, by definition of the total variation distance,

$$d_{TV}(\Pi_{\mathbf{i}}, \Pi_{\mathbf{i'}}) \geq \Pr[\Pi_{\mathbf{i}} \in \mathcal{A}] - \Pr[\Pi_{\mathbf{i'}} \in \mathcal{A}] \geq 1 - 2\varepsilon. \tag{79}$$

Assume for contradiction that there are $i \neq i' \in [k]$ such that

$$\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{i}}), \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{i'}}) < \frac{(1-2\varepsilon)^2}{8}.$$

Then, since the Hellinger distance $\mathrm{h}()$ obeys the triangle inequality and by Proposition 16,

$$d_{TV}(\Pi_{\mathbf{i}}, \Pi_{\mathbf{i'}}) \leq \sqrt{2}\mathrm{h}(\Pi_{\mathbf{i}}, \Pi_{\mathbf{i'}}) \leq \sqrt{2}\mathrm{h}(\Pi_{\mathbf{0}}, \Pi_{\mathbf{i}}) + \sqrt{2}\mathrm{h}(\Pi_{\mathbf{0}}, \Pi_{\mathbf{i'}}) < 1 - 2\varepsilon,$$

in contradiction to Eq. (79). ■

Lemma 35 and Eq. (78) conclude that any $1/3$-error protocol has a communication complexity of at least $Ck/\beta - m$, for some numerical constant $C > 0$. We conclude by showing that the communication complexity is at least $Ck/(2\beta)$. Assume for contradiction that the communication complexity is less than $Ck/(2\beta)$. Denote the parties in the protocol by $1, \ldots, m$ and assuming without loss of generality that party 1 is always the first to talk, party 2 is the first party to talk among parties $2, \ldots, m$, party 3 talks first among parties $3, \ldots, m$ etc.[11], then only a subset of parties $1, \ldots, \lfloor Ck/(2\beta) \rfloor$ participates in the protocol, hence we can assume that $m \leq Ck/(2\beta)$. The communication complexity is at least $Ck/\beta - m \geq Ck/(2\beta)$, which concludes the proof.

A.2.3. PROOF OF LEMMA 19

We start by defining an opposite channel $P_{\mu|\eta} \colon \Omega_\eta \to \{-1,1\}^k$: given some $y \in \Omega^\eta$, the channel sends it to $x = (x_1, \ldots, x_k) \in \{-1,1\}^k$, where each bit of x is set independently, such that:

$$(1+\rho)P_{\mu|\eta}(x_i = 1 \mid y) + (1-\rho)P_{\mu|\eta}(x_i = -1|y) = \eta_i(y)/\eta_0(y).^{12} \tag{80}$$

[11] The symmetries between the parties imply that if at some point in the protocol a new party is speaking, one can assume that this party has the lowest index among all parties that have not spoken yet.

[12] To avoid issues of devision by 0, assume that η_0 has full support. Indeed, one can remove from Ω_η all elements y for which $\eta_0(y) = 0$ to obtain Ω'_η and use Ω'_η as the joint sample space of $\eta_0, \ldots, \eta_k$. By definition of a CD(ρ) family, $\eta_i(y) = 0$ for any $y \in \Omega_\eta \setminus \Omega'_\eta$ and for all $i \in [k]$, hence $\eta_1, \ldots, \eta_k$ can be viewed as probability distributions over Ω'_η.

Such a definition is possible since, by the definition of a CD(ρ) family, $1-\rho \le \eta_i(y)/\eta_0(y) \le 1+\rho$. Define $\mu_0 = P_{\mu|\eta} \circ \eta_0$ and define Ω_μ as the support of μ_0. Note that taking an expectation over $y \sim \eta_0$ in Eq. (80), one obtains that

$$(1+\rho)\mu_0(x_i = 1) + (1-\rho)\mu_0(x_i = -1) = 1.$$

Hence, $\mu_0(x_i = 1) = \mu_0(x_i = -1) = 1/2$ for all $i \in [k]$, as required by the definition of a BCD(ρ) family. For $i = 1, \dots, k$, define the distribution μ_i as in the definition of a BCD(ρ) family: $\mu_i(x) = \mu_0(x)(1 + \rho x_i)$. It holds that $\{\mu_1, \dots, \mu_k\}$ is a BCD(ρ) family, as required.

Define the channel $P_{\eta|\mu} \colon \Omega_\mu \to \Omega_\eta$ as the channel sending μ_0 to η_0, namely,

$$P_{\eta|\mu}(y \mid x) = \frac{P_{\mu|\eta}(x \mid y)\eta_0(y)}{\mu_0(x)}$$

We will show that $\eta_i = P_{\eta|\mu} \circ \mu_i$ for any $1 \le i \le k$. Indeed,

$$\begin{aligned}
(P_{\eta|\mu} \circ \mu_i)(y) &= \sum_{x \in \Omega_\mu} \mu_i(x) P_{\eta|\mu}(y \mid x) \\
&= \sum_{x \in \Omega_\mu} \mu_i(x) \frac{P_{\mu|\eta}(x \mid y)\eta_0(y)}{\mu_0(x)} \\
&= \sum_{x \in \Omega_\mu} (1 + \rho x_i) P_{\mu|\eta}(x \mid y)\eta_0(y) && (81) \\
&= \eta_0(y) \sum_{b \in \{-1,1\}} (1 + b\rho) P_{\mu|\eta}(x_i = b \mid y) \\
&= \eta_0(y) \frac{\eta_i(y)}{\eta_0(y)}, && (82)
\end{aligned}$$

where Eq. (81) follows from the definition of μ_i and Eq. (82) follows from Eq. (80). In order to conclude the proof of this lemma, it remains to prove Eq. (6). Here is an auxiliary lemma:

Lemma 36 *Let $U = (U_1, \dots, U_k), V = (V_1, \dots, V_k) \in \mathbb{R}^k$ be random vectors. If for any possible value u of U, $\mathbb{E}[V \mid U = u] = u$ and $V_1, \dots, V_k$ are independent conditioned $U = u$, then for any subset $S \subseteq [k]$,*

$$\mathbb{E}\left[\prod_{i \in S} U_i\right] = \mathbb{E}\left[\prod_{i \in S} V_i\right].$$

Proof Fix some set $S \subseteq [k]$. It holds that:

$$\mathbb{E}\left[\prod_{i \in S} V_i\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{i \in S} V_i \,\middle|\, U\right]\right] = \mathbb{E}\left[\prod_{i \in S} \mathbb{E}\left[V_i \mid U\right]\right] = \mathbb{E}\left[\prod_{i \in S} U_i\right].$$

■

Let $X \sim \mu_0$ and $Y \sim \eta_0$. We conclude the proof of Eq. (6) by applying Lemma 36 with $U_i = \eta_i(Y)/\eta_0(Y) - 1$ and $V_i = \mu_i(X)/\mu_0(X) - 1$. Eq. (80) implies that the condition $\mathbb{E}[V \mid U = u] = u$ holds. By definition of $P_{\mu|\eta}$, the bits of X are independent conditioned on U. By definition of μ_i, $V_i = \rho X_i$, hence $V_1 \cdots V_k$ are independent conditioned on U, which implies that all conditions of Lemma 36 hold.

A.3. Proofs from Subsection 3.2

Fix some $0 < \rho < 1$ and define $\Omega = \{-1,1\}^d$ for some $d \geq 2$. Let $\mathcal{I}$ be the set of all nonempty subsets of $\{1,\dots,d\}$. For any $I \in \mathcal{I}$ and $0 < \rho < 1$, let $\mu_{I,\rho}$ be the distribution over Ω defined by

$$\mu_{I,\rho}((x_1,\dots,x_d)) = 2^{-d}(1 + \rho \prod_{i\in I} x_i).$$

We will write μ_I whenever ρ is implied from the context. Note that μ_I is almost uniform, with a small bias towards inputs that contain an even number of 1-values on I. For any subset $\mathcal{U} \subseteq \mathcal{I}$ and $0 < \rho < 1$, let $\mathcal{P}_{\mathcal{U},\rho} = \{\mu_{I,\rho} \colon I \in \mathcal{U}\}$. Note that $\mathcal{P}_{\mathcal{U},\rho}$ is a $\mathrm{CD}(\rho)$ family and the corresponding μ_0 distribution is the uniform distribution over Ω.

A.3.1. PROOF OF THEOREM 9

Let $A = (A_1,\dots,A_d) \sim \mu_0$ and for any $I \in \mathcal{I}$, define the random variable B_I as a function of A:

$$B_I = \prod_{i\in I} A_i. \tag{83}$$

Note that for all $0 < \rho < 1$,

$$B_I = (\mu_{I,\rho}(A)/\mu_0(A) - 1)/\rho, \tag{84}$$

a term which appears in Eq. (1) of Theorem 4. The next lemma states what are the correlations between these random variables B_I.

Lemma 37 *Let $\mathcal{J} \subseteq \mathcal{I}$. Then*

$$\mathbb{E}\left[\prod_{I\in\mathcal{J}} B_I\right] = \begin{cases} 1 & \triangle\mathcal{J} = \emptyset \\ 0 & \mathit{otherwise} \end{cases},$$

where $\triangle\mathcal{J}$ is the symmetric difference between all sets in $\mathcal{J}$ which contain all elements $i \in \{1,\dots,d\}$ which appear in an odd number of sets from $\mathcal{J}$. In particular, if $I_1, I_2 \in \mathcal{I}$ are distinct sets then $\mathbb{E}B_{I_1}B_{I_2} = 0$.

Proof Note that

$$\mathbb{E}\prod_{I\in\mathcal{J}} B_I = \mathbb{E}\prod_{I\in\mathcal{J}}\prod_{i\in I} A_i = \mathbb{E}\prod_{i\in\triangle\mathcal{J}} A_i = \prod_{i\in\triangle\mathcal{J}} \mathbb{E}A_i = \begin{cases} 1 & \triangle\mathcal{J} = \emptyset \\ 0 & \text{otherwise} \end{cases}$$

where the first equation follows from the definition of B_I, the third equation follows from the fact that the coordinates of A are independent and and the empty product is regarded as 1. ■

Lemma 37 states that the B_I are pairwise independent, hence, for any subset $\mathcal{U} \subseteq \mathcal{I}$ and suitable values of n and ρ one can apply Theorem 4 on the family of distributions $\mathcal{P}_{\mathcal{U},\rho}$: Lemma 37 imply that all the terms in Eq. (1) corresponding to $|S| = 2$ are zero, hence Thm. 4 can be applied for any $n \geq k^6$ (and a suitable ρ). This proves Theorem 9.

A.3.2. PROOF OF LEMMA 8

By definition of μ_I,

$$\begin{aligned}
\mathbb{E}_{x\sim\mu_I}\prod_{i\in I'} x_i &= \sum_{x\in\{-1,1\}^d} \mu_I(x)\prod_{i\in I'} x_i \\
&= \sum_{x\in\{-1,1\}^d} 2^{-d}(1+\rho\prod_{i\in I} x_i)\prod_{i\in I'} x_i \\
&= \mathbb{E}_{X\sim\mu_0}(1+\rho\prod_{i\in I} X_i)\prod_{i\in I'} X_i \qquad (85)\\
&= \prod_{i\in I'}\mathbb{E}_{X\sim\text{Uniform}(\{-1,1\})}[X_i]+\rho\prod_{i\in I\triangle I'}\mathbb{E}_{X\sim\text{Uniform}(\{-1,1\})}[X_i] \qquad (86)\\
&= \rho\prod_{i\in I\triangle I'}\mathbb{E}_{X\sim\text{Uniform}(\{-1,1\})}[X_i] \qquad (87)\\
&= \begin{cases}\rho & I=I'\\ 0 & I\neq I'\end{cases}
\end{aligned}$$

where Eq. (85) and Eq. (86) follow from the fact that μ_0 is the uniform measure over $\{-1,1\}^d$ and Eq. (87) follows from the fact that $I'\neq\emptyset$.

A.4. Proof of Theorem 12

First, we give an outline to the proof. Recall that $\eta_{I,\sigma}$ is defined as the Gaussian distribution over $\mathbb{R}^d$ with mean zero and its covariance matrix, $\Sigma_{I,\sigma}$, is almost the identity, except for two coordinates, i and j, with a covariance of σ. These coordinates satisfy $I=\{i,j\}$. Denote by η_0 the Gaussian distribution over $\mathbb{R}^d$ with zero mean and its covariance, Σ_0, is the identity matrix. For any $x\in\mathbb{R}^n$, let $\eta_{I,\sigma}(x)$ denote the density of $\eta_{I,\sigma}$ on x.

We start with some preliminaries in Subsection A.4.1. Then, we show that for any $I\neq I'$, $\eta_{I,\sigma}$ and $\eta_{I',\sigma}$ are pairwise uncorrelated with respect to η_0 in the following way:

$$\mathbb{E}_{X\sim\eta_0}\left[\left(\frac{\eta_{I,\sigma}(X)}{\eta_0(X)}-1\right)\left(\frac{\eta_{I',\sigma}(X)}{\eta_0(X)}-1\right)\right]=0,$$

which is equivalent to

$$\mathbb{E}_{X\sim\eta_0}\left[\frac{\eta_{I,\sigma}(X)}{\eta_0(X)}\frac{\eta_{I',\sigma}(X)}{\eta_0(X)}\right]=1.$$

This is proved by taking an integral and calculating a determinant. We denote $\Sigma_{I,I'}^{-1} = \Sigma_{I,\sigma}^{-1} + \Sigma_{I',\sigma}^{-1} - \Sigma_0^{-1}$. The following holds:

$$\begin{aligned}
\mathbb{E}_{X\sim\eta_0}\left[\frac{\eta_{I,\sigma}(X)}{\eta_0(X)}\frac{\eta_{I',\sigma}(X)}{\eta_0(X)}\right] &= \int_{x\in\mathbb{R}^d}\frac{\eta_I(x)\eta_{I'}(x)}{\eta_0(x)}dx \\
&= \int_{x\in\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_{I,\sigma})\det(\Sigma_{I',\sigma})}}\exp\left(-\frac{1}{2}x^t\Sigma_{I,I'}^{-1}x\right)dx \\
&= \frac{\sqrt{\det(\Sigma_{I,I'})}}{\sqrt{\det(\Sigma_{I,\sigma})\det(\Sigma_{I',\sigma})}}\int_{x\in\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_{I,I'})}}\exp\left(-\frac{1}{2}x^t\Sigma_{I,I'}^{-1}x\right)dx && (88) \\
&= \frac{\sqrt{\det(\Sigma_{I,I'})}}{\sqrt{\det(\Sigma_{I,\sigma})\det(\Sigma_{I',\sigma})}}, && (89)
\end{aligned}$$

where Eq. (89) follows from the fact that the integrand in Eq. (88) is a density function of a normal distribution with mean zero and covariance $\Sigma_{I,I'}$. The term in Eq. (89) equals 1 for any $I \neq I'$, as required. In the proof we also calculate higher order correlations, namely,

$$\mathbb{E}_{X\sim\eta_0}\left[\left(\prod_{i=1}^{r}\frac{\eta_{I_i,\sigma}(X)}{\eta_0(X)} - 1\right)\right], \tag{90}$$

for distinct $I_1,\ldots,I_r$. In order to calculate this expectation, we define the matrix $\Sigma_{I_1,\ldots,I_r}^{-1}$ in Eq. (93) similarly to $\Sigma_{I,I'}^{-1}$. In Lemma 43 we prove some properties of $\Sigma_{I_1,\ldots,I_r}^{-1}$ and in Lemma 44 we show that Eq. (90) equals zero for some collections $I_1,\ldots,I_r$. These two lemmas and other auxiliaries appear in Subsection A.4.2.

Note that we cannot apply Thm. 4 directly on the family of Gaussian distributions: the theorem requires that for any $x \in \mathbb{R}^d$, $|\eta_{I,\sigma}(x)/\eta_0(x) - 1| \le \rho$ for some $\rho > 0$, which is incorrect for the Gaussian distributions. Hence, we apply it on a family of truncated normal distributions, in Subsection A.4.3. The truncated Gaussian, $\eta_{I,\sigma,R}$, is defined as a truncation of $\eta_{I,\sigma}$ to $[-R,R]^d$, where R is logarithmic in the problem parameters. Indeed, it holds that $|\eta_{I,\sigma}(x)/\eta_0(x) - 1| \le \rho$ for some $\rho = \tilde{O}(\sigma)$. Additionally, we show that due to the fact that the truncated Gaussians are almost identical to the Gaussian distributions, their higher order correlations (Eq. (90)) are almost identical to those of Gaussian distributions. Hence, one can apply Thm. 4 as required.

Lastly, in Subsection A.4.4 we show that a communication lower bound on learning a truncated Gaussian implies a lower bound on learning a non-truncated Gaussian: due to the fact that high deviations in normal distributions are rare, with high probability all samples fall within $[-R,R]^d$. In that case, one cannot learn with little communication.

A.4.1. PRELIMINARIES

The next proposition states some basic properties of the determinant (denoted det).

Proposition 38 *The following hold for any matrix $M \in \mathbb{R}^{n\times n}$:*

1. $\det(cM) = c^n \det M$ *for any* $c \in \mathbb{R}$.

2. *If the matrix M can be written as $M = \begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$, where $A \in \mathbb{R}^{n_1 \times n_1}$, $B \in \mathbb{R}^{n_2 \times n_2}$, $C \in \mathbb{R}^{n_1 \times n_2}$ and the 0-block is of size $n_2 \times n_1$ for some integers n_1 and n_2 satisfying $n_1 + n_2 = n$, then* $\det M = \det A \det B$.

3. *Assume that M, M_1 and M_2 are $n \times n$ matrices which are identical except for column i (for some $1 \le i \le n$), such that column i of $M_1 + M_2$ equals column i of M. Then* $\det M = \det M_1 + \det M_2$.

4. *If A and B are squared matrices with the same dimension, then* $\det(AB) = \det A \det B$. *In particular, if A is invertible then* $\det A \det A^{-1} = \det(AA^{-1}) = \det \mathrm{I} = 1$.

5. $$\det M = \begin{cases} \sum_{i=1}^{n} (-1)^{i-1} M_{1i} \det M_{-1,-i} & n > 1 \\ M_{11} & n = 1 \end{cases}$$

   where $M_{-1,-i}$ is the $(n-1) \times (n-1)$ matrix obtained from M by removing its first row and column i.

Next, we define a *positive definite* matrix:

Definition 39 *Fix an integer $\ell \ge 1$. A squared matrix $M \in \mathbb{R}^{\ell \times \ell}$ is* positive definite *if one of the equivalent conditions hold:*

1. *For any nonzero vector $v \in \mathbb{R}^\ell$, $v^t M v > 0$.*
2. *All the eigenvalues of M are positive.*

Note that any positive definite matrix does not have the eigenvalue 0, hence it is invertible. Note that applying the same permutation on the rows and the columns of a matrix keeps many of its properties:

Proposition 40 *Fix an integer $\ell \ge 1$, a matrix $M \in \mathbb{R}^{\ell \times \ell}$ and a permutation $\pi \colon [\ell] \to [\ell]$. Let $\pi(M)$ be the matrix obtained after applying π on both the rows and columns of M: $(\pi(M))_{\pi(i),\pi(j)} = M_{i,j}$. The following hold:*

1. $\det \pi(M) = \det M$.
2. *$\pi(M)$ is positive definite if and only if M is positive definite.*
3. *If M is invertible then $\pi(M)$ is invertible and $\pi(M^{-1}) = (\pi(M))^{-1}$.*

Next, we define a multivariate normal distribution:

Definition 41 *For any integer $\ell \ge 1$ and a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{\ell \times \ell}$, the ℓ-variate normal distribution with mean 0 and covariance Σ is defined by the density function*

$$\frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-x^t \Sigma^{-1} x/2\right)$$

as a function of $x \in \mathbb{R}^\ell$.

For any $\ell \ge 1$, let I_ℓ be the identity matrix of dimension $\ell \times \ell$.

A.4.2. AUXILIARY TECHNICAL RESULTS

Let η_0 be the normal distribution in $\mathbb{R}^d$ with zero mean and its covariance matrix Σ_0 is the identity matrix. Recall the definition of $\Sigma_{I,\sigma}$ and $\eta_{I,\sigma}$ from Subsection 3.3. They will be written as Σ_I and η_I when σ is clear from context.

Lemma 42 *For any $I \in \mathcal{I}_2$ and $0 < \sigma < 1$, $\Sigma_{I,\sigma}$ is symmetric, positive definite,* $\det(\Sigma_{I,\sigma}) = 1-\sigma^2$ *and*

$$\Sigma_{I,\sigma}^{-1} = \frac{1}{1-\sigma^2}\begin{cases} 1 & i=j, i\in I \\ 1-\sigma^2 & i=j, i\notin I \\ -\sigma & i\neq j, I=\{i,j\} \\ 0 & i\neq j, I\neq\{i,j\} \end{cases}. \tag{91}$$

In particular, there exists a random vector with mean 0 *and covariance* $\Sigma_{I,\sigma}$.

Proof By proposition 40 it is sufficient to assume that $I = \{1,2\}$. Then, Σ_I is a block matrix

$$\Sigma_I = \begin{pmatrix} A & 0 \\ 0 & \mathrm{I}_{d-2} \end{pmatrix},$$

where

$$A = \begin{pmatrix} 1 & \sigma \\ \sigma & 1 \end{pmatrix}.$$

It holds that

$$\Sigma_I^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & \mathrm{I}_{d-2}^{-1} \end{pmatrix}$$

where

$$A^{-1} = \frac{1}{1-\sigma^2}\begin{pmatrix} 1 & -\sigma \\ -\sigma & 1 \end{pmatrix},$$

which concludes the proof for the formula of A^{-1}. To calculate the determinant, note that $\det \Sigma_I = \det A \det \mathrm{I}_{d-2} = 1-\sigma^2$. Lastly, $\Sigma_{I,\sigma}$ is positive definite because it is strictly diagonally dominant with positive diagonal entries and symmetric. ■

Fix $b = 5$, and assume that the constant C in Theorem 12 is sufficiently small to ensure that

$$\frac{4b^2\sigma}{1-\sigma^2} \le 1/2. \tag{92}$$

For any integer $2 \le r \le b$ and any distinct sets $I_1, \dots, I_r \in \mathcal{I}_2$, define

$$\Sigma_{I_1,\dots,I_r} := \left(\mathrm{I}_d + \sum_{i=1}^r \left(\Sigma_{I_i}^{-1} - \mathrm{I}_d\right)\right)^{-1}. \tag{93}$$

The following lemma shows that $\Sigma_{I_1,\dots,I_r}$ exists and estimates some of its properties. Given a matrix M let $\max|M|$ denote the maximal absolute value of an element of M.

Lemma 43 *Fix distinct pairs $I_1, \dots, I_r \in \mathcal{I}_2$ for some $2 \le r \le b$. The matrix $\Sigma_{I_1,\dots,I_r}$ defined in Eq. (93) exists and satisfies:*

1. $\Sigma_{I_1,\dots,I_r}$ is symmetric and positive definite.

2. $\det\left(\Sigma_{I_1,\dots,I_r}\right) \le 2$.

3. $\max|\Sigma_{I_1,\dots,I_r}| \le 2$.

In particular, there exists a multivariate normal distribution with mean 0 and covariance $\Sigma_{I_1,\dots,I_r}$.

Proof Since each I_i is a set of two elements, $|\cup_{i=1}^r I_i| \le 2r$. By Proposition 40, one can assume that $\bigcup_{i=1}^r I_i \subseteq [2r]$. Define

$$\Sigma_{I_1,\dots I_r}^{-1} = \mathrm{I}_d + \sum_{i=1}^r \left(\Sigma_{I_i}^{-1} - \mathrm{I}_d\right).$$

By Lemma 42, for any $I \in \mathcal{I}_2$,

$$(\Sigma_I^{-1} - \mathrm{I}_d)(i,j) = \frac{1}{1-\sigma^2}\begin{cases} \sigma^2 & i = j \in I \\ -\sigma & \{i,j\} = I \\ 0 & \text{otherwise} \end{cases}. \tag{94}$$

Hence,

$$\Sigma_{I_1,\dots,I_r}^{-1}(i,j) = \begin{cases} 1 + \frac{\sigma^2}{1-\sigma^2}\sum_{i=1}^r |\{i\} \cap I_i| & i = j \\ -\frac{\sigma}{1-\sigma^2} & \{i,j\} = I_i \text{ for some } 1 \le i \le r \\ 0 & \text{otherwise} \end{cases}. \tag{95}$$

By the assumption $\bigcup_{i=1}^r I_i \subseteq [2r]$,

$$\Sigma_{I_1,\dots,I_r}^{-1} = \begin{pmatrix} A & 0 \\ 0 & \mathrm{I}_{d-2r} \end{pmatrix} \tag{96}$$

where $A \in \mathbb{R}^{(2r)\times(2r)}$, I_{d-2r} is the identity matrix of size $(d-2r)\times(d-2r)$ and the two zero blocks are of sizes $(2r)\times(d-2r)$ and $(d-2r)\times(2r)$. We will start by showing that $\Sigma_{I_1,\dots,I_r}^{-1}$ is positive definite. Fix some nonzero $v \in \mathbb{R}^d$. Let v_A be the vector containing the first $2d$ coordinates of v and let v_{I} be the vector containing its remaining coordinates. It holds that

$$\begin{aligned} v^t\Sigma_{I_1,\dots,I_r}^{-1}v &= v_A^t A v_A + v_{\mathrm{I}}^t \mathrm{I}_{d-2r} v_{\mathrm{I}} \\ &= \sum_{i=1}^{2r}\sum_{j=1}^{2r} A_{i,j} v_i v_j + \|v_{\mathrm{I}}\|_2^2 \\ &= \sum_{i=1}^{2r} A_{i,i} v_i^2 + \sum_{\substack{i,j \in \{1,\dots,2r\} \\ i \ne j}} A_{i,j} v_i v_j + \|v_{\mathrm{I}}\|_2^2 \\ &\ge \sum_{i=1}^{2r} v_i^2 - \frac{\sigma}{1-\sigma^2}\sum_{i=1}^{2r}\sum_{j=1}^{2r} |v_i v_j| + \|v_{\mathrm{I}}\|_2^2 \\ &= \|v_A\|_2^2 - \frac{\sigma}{1-\sigma^2}\|v_A\|_1^2 + \|v_{\mathrm{I}}\|_2^2 \\ &\ge \left(1 - \frac{2r\sigma}{1-\sigma^2}\right)\|v_A\|_2^2 + \|v_{\mathrm{I}}\|_2^2 \end{aligned} \tag{97}$$

where Eq. (97) follows from the fact that for any vector v in $\mathbb{R}^\ell$, $\|v\|_1 \leq \sqrt{\ell}\|v\|_2$. By the assumption of this lemma, $1 - 2r\sigma/(1-\sigma^2) > 0$. Since v is nonzero, either $\|v_A\|_2 > 0$ or $\|v_{\mathrm{I}}\|_2 > 0$, which implies that the term in Eq. (97) is positive. By definition of positive definiteness (Definition 39), this implies that $\Sigma^{-1}_{I_1,\ldots,I_r}$ is positive definite. In particular, this implies that $\Sigma^{-1}_{I_1,\ldots,I_r}$ is invertible.

The positive definiteness of $\Sigma_{I_1,\ldots,I_r}$ follows from the positive definiteness of $\Sigma^{-1}_{I_1,\ldots,I_r}$: a matrix M is positive definite if and only if M^{-1} is positive definite.

Note that the calculation in Eq. (97) implies that lowest eigenvalue of A is at least $1-2r\sigma/\left(1-\sigma^2\right)$. Since the determinant is the multiplication of all eigenvalues, and using Eq. (96), it holds that

$$\det \Sigma^{-1}_{I_1,\ldots,I_r} = \det A \geq \left(1 - 2r\sigma/\left(1-\sigma^2\right)\right)^{2r} \geq 1 - 4r^2/\left(1-\sigma^2\right) \geq 1/2,$$

where the last inequality follows from Eq. (92) and the assumption of this lemma that $r \leq b$. This concludes the bound on $\det \Sigma_{I_1,\ldots,I_r} = \left(\det \Sigma^{-1}_{I_1,\ldots,I_r}\right)^{-1}$.

The matrix $\Sigma_{I_1,\ldots,I_r}$ is symmetric due to the fact that $\Sigma^{-1}_{I_1,\ldots,I_r}$ is symmetric (see Eq. (96) and Eq. (95)) and the fact that for any symmetric invertible matrix M, M^{-1} is symmetric.

Equation Eq. (95) implies that

$$\left|\left(A - \mathrm{I}_{2r}\right)(i,j)\right| \leq \begin{cases} \frac{\sigma^2}{1-\sigma^2} r \leq \frac{\sigma}{1-\sigma^2} & i = j \\ \frac{\sigma}{1-\sigma^2} & i \neq j \end{cases} \tag{98}$$

using the assumption using Eq. (92) and the assumption $r \leq b$ which imply $\sigma r \leq \sigma b \leq 1$. By induction on $\ell = 1, 2, \ldots$, one obtains that

$$\max\left|\left(A - \mathrm{I}_{2r}\right)^\ell\right| \leq \frac{\sigma}{1-\sigma^2}\left(\frac{2r\sigma}{1-\sigma^2}\right)^{\ell-1}. \tag{99}$$

For $\ell = 1$ it follows from Eq. (98) and for $\ell > 1$:

$$\max\left|\left(A - \mathrm{I}_{2r}\right)^\ell\right| = \max\left|\left(A - \mathrm{I}_{2r}\right)^{\ell-1}\left(A - \mathrm{I}_{2r}\right)\right| \leq$$
$$2r \max\left|\left(A - \mathrm{I}_{2r}\right)^{\ell-1}\right| \max\left|A - \mathrm{I}_{2r}\right| \leq \frac{\sigma}{1-\sigma^2}\left(\frac{2r\sigma}{1-\sigma^2}\right)^{\ell-1}$$

where the first inequality follows from the formula for matrix multiplication. Given a squared matrix M its Neumann series is defined as

$$\sum_{\ell=0}^{\infty} M^\ell.$$

If the Neumann series of M converges then $(\mathrm{I} - M)^{-1}$ exists and equals the Neumann series of M (I is the identity matrix). Substituting $M = \mathrm{I}_{2r} - A$, inequality Eq. (99) implies that

$$\sum_{\ell=0}^{\infty} \max\left|\left(I_{2r} - A\right)^\ell\right|_\infty \leq \sum_{\ell=0}^{\infty}\left(\frac{2r\sigma}{1-\sigma^2}\right)^\ell \leq \frac{1}{1 - 2r\sigma/(1-\sigma^2)} \leq 2, \tag{100}$$

hence, the series converges in absolute value, therefore it converges and equals $(\mathrm{I}_{2r} - (\mathrm{I}_{2r} - A))^{-1} = A^{-1}$. In particular, $\max|A^{-1}| \leq 2$. Using Eq. (96) one can verify that

$$\Sigma_{I_1,\ldots,I_r} = \begin{pmatrix} A^{-1} & 0 \\ 0 & \mathrm{I}_{d-2r} \end{pmatrix}$$

which concludes the proof. ■

Fix $I_1, \dots, I_r \in \mathcal{I}_2$ for some $r \le b$. It holds that

$$\eta_0(x) \prod_{i=1}^{r} \frac{\eta_{I_i}(x)}{\eta_0(x)} = \frac{1}{\sqrt{\det(2\pi\Sigma_0)}} \exp\left(-\frac{1}{2}x^t\Sigma_0^{-1}x\right) \prod_{i=1}^{r} \frac{1}{\sqrt{1-\sigma^2}} \exp\left(-\frac{1}{2}x^t\left(\Sigma_I^{-1} - \Sigma_0^{-1}\right)x\right) \tag{101}$$

$$= (2\pi)^{-d/2}(1-\sigma^2)^{-r/2} \exp\left(-\frac{1}{2}x^t\left(\sum_{i=1}^{r}\Sigma_{I_i}^{-1} - (r-1)\Sigma_0^{-1}\right)x\right)$$

$$= \sqrt{\frac{\det \Sigma_{I_1,\dots,I_r}}{(1-\sigma^2)^r}} \frac{1}{\sqrt{\det(2\pi\Sigma_{I_1,\dots,I_r})}} \exp\left(-\frac{1}{2}x^t\Sigma_{I_1,\dots,I_r}^{-1}x\right) \tag{102}$$

using $\det \Sigma_{I_i} = 1 - \sigma^2$ from Lemma 42. In particular, the RHS of Eq. (101) is the density of a d-variate normal distribution with mean 0 and covariance $\Sigma_{I_1,\dots,I_r}$, multiplied by a constant. Hence,

$$\mathbb{E}_{X\sim\eta_0} \prod_{i=1}^{r} \frac{\eta_{I_i}(X)}{\eta_0(X)} = \int_{x\in\mathbb{R}^d} \eta_0(x) \prod_{i=1}^{r} \frac{\eta_{I_i}(x)}{\eta_0(x)}$$

$$= \sqrt{\frac{\det \Sigma_{I_1,\dots,I_r}}{(1-\sigma^2)^r}} \int_{x\in\mathbb{R}^d} \frac{1}{\sqrt{\det(2\pi\Sigma_{I_1,\dots,I_r})}} \exp\left(-x^t\Sigma_{I_1,\dots,I_r}^{-1}x/2\right) \tag{103}$$

$$= \sqrt{\frac{\det \Sigma_{I_1,\dots,I_r}}{(1-\sigma^2)^r}}, \tag{104}$$

where the last equation holds since the integral in Eq. (103) is over the density function of a probability distribution.

Lemma 44 *Assume that $I_1, \dots, I_r \in \mathcal{I}_2$ for some $r \le b$ such that there exists $j \in \{1, \dots, d\}$ which satisfies $j \notin \bigcup_{i=1}^{r-1} I_i$ and $j \in I_r$. Then*

$$\mathbb{E}_{X\sim\eta_0} \prod_{i=1}^{r} \left(\frac{\eta_{I_i}(X)}{\eta_0(X)} - 1\right) = 0. \tag{105}$$

Proof We will start by showing that $\det \Sigma_{I_1,\dots,I_r}^{-1} = \frac{1}{1-\sigma^2} \det \Sigma_{I_1,\dots,I_{r-1}}^{-1}$. By Proposition 40 one can assume that the element j unique to I_r is 1 and that $I_r = \{1, 2\}$. Then, Eq. (95) implies that

$$\Sigma_{I_1,\dots,I_{r-1}}^{-1} = \frac{1}{1-\sigma^2} \begin{pmatrix} 1-\sigma^2 & 0 \\ 0 & A \end{pmatrix}$$

where $A \in \mathbb{R}^{d\times d}$ and the two 0-blocks contain $d-1$ zeros. From items 1 and 2 of Proposition 38,

$$\det \Sigma_{I_1,\dots,I_{r-1}}^{-1} = (1-\sigma^2)^{-d}(1-\sigma^2)\det A. \tag{106}$$

Additionally, from Eq. (93) and Eq. (94),

$$\Sigma_{I_1,\ldots,I_r}^{-1} = \Sigma_{I_1,\ldots,I_{r-1}}^{-1} + \Sigma_{I_r}^{-1} - \Sigma_0^{-1} = \frac{1}{1-\sigma^2}\begin{pmatrix} 1 & -\sigma & 0 \\ -\sigma & A_{00}+\sigma^2 & A_{01} \\ 0 & A_{10} & A_{11}\end{pmatrix}, \tag{107}$$

where

$$A = \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11}\end{pmatrix}$$

such that $A_{00} \in \mathbb{R}$, $A_{11} \in \mathbb{R}^{(d-2)\times(d-2)}$, $A_{01} \in \mathbb{R}^{1\times(d-2)}$ and $A_{10} \in \mathbb{R}^{(d-2)\times 1}$. Additionally, the two zero blocks in Eq. (107) contain $d-2$ zeros and the 1 and $-\sigma$ blocks contain one entry. Proposition 38 imply that

$$\begin{aligned}
(1-\sigma^2)^r \det \Sigma_{I_1,\ldots,I_r}^{-1} &= \det\begin{pmatrix} 1 & -\sigma & 0 \\ -\sigma & A_{00}+\sigma^2 & A_{01} \\ 0 & A_{10} & A_{11}\end{pmatrix} && (108)\\
&= \det\begin{pmatrix} A_{00}+\sigma^2 & A_{01} \\ A_{10} & A_{11}\end{pmatrix} - (-\sigma)\det\begin{pmatrix} -\sigma & A_{01} \\ 0 & A_{11}\end{pmatrix} && (109)\\
&= \det\begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11}\end{pmatrix} + \det\begin{pmatrix} \sigma^2 & A_{01} \\ 0 & A_{11}\end{pmatrix} - \sigma^2 \det A_{11} && (110)\\
&= \det A. && (111)
\end{aligned}$$

where Eq. (108) follows from item 1 of Proposition 38, Eq. (109) follows from item 5, Eq. (110) follows from items 3 and 2 and Eq. (111) follows from item 2. Equations Eq. (106) and Eq. (111) imply that

$$\det \Sigma_{I_1,\ldots,I_r}^{-1} = (1-\sigma^2)^{-1} \det \Sigma_{I_1,\ldots,I_{r-1}}^{-1},$$

hence

$$\det \Sigma_{I_1,\ldots,I_r} = (1-\sigma^2) \det \Sigma_{I_1,\ldots,I_{r-1}},$$

therefore Eq. (104) implies that

$$\mathbb{E}_{X\sim\eta_0} \prod_{i=1}^{r} \frac{\eta_{I_i}(X)}{\eta_0(X)} = \mathbb{E}_{X\sim\eta_0} \prod_{i=1}^{r-1} \frac{\eta_{I_i}(X)}{\eta_0(X)}. \tag{112}$$

Note that Eq. (112) can be applied when substituting $\{1,\ldots,r-1\}$ with any subset, namely, for any $S \subseteq \{1,\ldots,r-1\}$,

$$\mathbb{E}_{X\sim\eta_0} \prod_{i\in S\cup\{r\}} \frac{\eta_{I_i}(X)}{\eta_0(X)} = \mathbb{E}_{X\sim\eta_0} \prod_{i\in S} \frac{\eta_{I_i}(X)}{\eta_0(X)}. \tag{113}$$

Indeed, for any such S, $\{I_i\}_{i\in S\cup\{r\}}$ satisfy the conditions of Lemma 44. To conclude the proof, note that

$$\begin{aligned}
\mathbb{E}_{X\sim\eta_0} \prod_{i=1}^{r} \left(\frac{\eta_{I_i}(X)}{\eta_0(X)} - 1\right) &= \sum_{S\subseteq\{1,\ldots,r\}} (-1)^{r-|S|} \prod_{i\in S} \frac{\eta_{I_i}(X)}{\eta_0(X)} \\
&= \sum_{S\subseteq\{1,\ldots,r-1\}} (-1)^{r-|S|} \left(\prod_{i\in S} \frac{\eta_{I_i}(X)}{\eta_0(X)} - \prod_{i\in S\cup\{r\}} \frac{\eta_{I_i}(X)}{\eta_0(X)}\right) \\
&= 0,
\end{aligned}$$

where the last equation follows from Eq. (113). ■

A.4.3. APPLYING THEOREM 4 FOR TRUNCATED GAUSSIANS

We apply Theorem 4 on a family of truncated normal distributions, which is a $\mathrm{CD}(\rho)$ family for some small value of ρ. Recall that b is a constant integer defined above ($b = 5$). Define

$$p = \sigma^b \left(64bn\binom{n}{\le b}2^b\right)^{-1} \tag{114}$$

and

$$R = \max\left(\sqrt{2\ln(2dmn)}, \sqrt{4\ln\frac{2d}{p}}, \sqrt{2\ln(d/\sigma)}, 1\right). \tag{115}$$

For all $0 < \sigma < 1$ and $I \in \mathcal{I}_2 \cup \{0\}$, let $\eta_{I,\sigma,R}$ be the truncation of $\eta_{I,\sigma}$ to $[-R,R]^d$, namely,

$$\eta_{I,\sigma,R}(x) = \frac{1}{\eta_{I,\sigma}\left([-R,R]^d\right)}\begin{cases}\eta_{I,\sigma}(x) & x \in [-R,R]^d\\ 0 & x \notin [-R,R]^d.\end{cases}$$

We write shortly $\eta_{I,R}$ when σ is implied from context. Define $\mathcal{G}_{\sigma,R} = \{\eta_{I,\sigma,R} \colon I \in \mathcal{I}_2\}$.

Here is a well known tail bound for the normal distribution.

Proposition 45 *Let W be a random variable distributed normally with mean 0 and variance σ^2. Then, for any $w > 0$,*

$$\Pr[|W| \ge w] \le \frac{2e^{-w^2/(2\sigma^2)}}{(w/\sigma)\sqrt{2\pi}}. \tag{116}$$

Proof Start by assuming that $\sigma = 1$. Then,

$$\Pr[|W| \ge w] = 2\int_{t=w}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \le 2\int_{t=w}^{\infty}\frac{t}{w}\frac{1}{\sqrt{2\pi}}e^{-t^2/2} = \frac{2e^{-w^2/2}}{w\sqrt{2\pi}}.$$

Assuming that $\sigma \ne 1$, one can apply Eq. (116) on W/σ which is distributed normally with mean 0 and variance 1 and obtain

$$\Pr[|W| \ge w] = \Pr\left[\left|\frac{W}{\sigma}\right| \ge \frac{w}{\sigma}\right] \le \frac{2e^{-w^2/(2\sigma^2)}}{(w/\sigma)\sqrt{2\pi}}.$$

■

By substituting $w = R$, and recalling that $R \ge 1$ by definition, we get that for any normally distributed W with zero mean,

$$\Pr[|W| \ge R] \le \mathrm{Var}(W)^{1/2}e^{-\frac{1}{2}R^2/\mathrm{Var}(W)}. \tag{117}$$

In particular, this implies that if $X = (X_1, \dots, X_d) \sim \eta_I$ for some $I \in \mathcal{I}_2 \cup \{0\}$,

$$\Pr\left[X \notin [-R,R]^d\right] \le \sum_{i=1}^{d}\Pr\left[|X_i| \ge R\right] \le de^{-R^2/2} \le \min\left(\frac{1}{2mn}, p, \sigma\right). \tag{118}$$

Lemma 46 *Let $1 \le r \le b$ be an integer and let $I_1, \ldots, I_r \in \mathcal{I}_2$ be distinct sets. Then*

$$\left|\mathbb{E}_{X\sim\eta_0}\prod_{i=1}^{r}\left(\frac{\eta_{I_i}(X)}{\eta_0(X)}-1\right)-\mathbb{E}_{X\sim\eta_{0,R}}\prod_{i=1}^{r}\left(\frac{\eta_{I_i,R}(X)}{\eta_{0,R}(X)}-1\right)\right| \le \frac{\sigma^b}{4n\binom{n}{\le b}}.$$

Proof We start by showing that

$$\left|\mathbb{E}_{X\sim\eta_0}\prod_{i=1}^{r}\frac{\eta_{I_i}(X)}{\eta_0(X)}-\mathbb{E}_{X\sim\eta_{0,R}}\prod_{i=1}^{r}\frac{\eta_{I_i,R}(X)}{\eta_{0,R}(X)}\right| \le \frac{\sigma^b}{4n\binom{n}{\le b}2^b}. \tag{119}$$

Note that

$$\left|\mathbb{E}_{X\sim\eta_0}\prod_{i=1}^{r}\frac{\eta_{I_i}(X)}{\eta_0(X)}-\mathbb{E}_{X\sim\eta_{0,R}}\prod_{i=1}^{r}\frac{\eta_{I_i,R}(X)}{\eta_{0,R}(X)}\right| \tag{120}$$

$$\le\left|\mathbb{E}_{X\sim\eta_0}\left[\prod_{i=1}^{r}\frac{\eta_{I_i}(X)}{\eta_0(X)}\middle|X\in[-R,R]^d\right]\Pr_{X\sim\eta_0}\left[X\in[-R,R]^d\right]-\mathbb{E}_{X\sim\eta_{0,R}}\prod_{i=1}^{r}\frac{\eta_{I_i,R}(X)}{\eta_{0,R}(X)}\right| \tag{121}$$

$$+\left|\mathbb{E}_{X\sim\eta_0}\left[\prod_{i=1}^{r}\frac{\eta_{I_i}(X)}{\eta_0(X)}\middle|X\notin[-R,R]^d\right]\Pr_{X\sim\eta_0}\left[X\notin[-R,R]^d\right]\right|. \tag{122}$$

We will bound the terms Eq. (121) and Eq. (122) separately. Let $W = (W_1, \ldots, W_d)$ be a random variable distributed normally with mean 0 and covariance $\Sigma_{I_1,\ldots,I_r}$ and let P_W denote its density function.

$$Eq.~(121) = \left| \frac{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)}{\eta_0\left([-R,R]^d\right)^{r-1}} \mathbb{E}_{X\sim\eta_0}\left[\prod_{i=1}^r \frac{\eta_{I_i}(X)/\eta_{I_i}\left([-R,R]^d\right)}{\eta_0(X)/\eta_0\left([-R,R]^d\right)} \middle| X\in[-R,R]^d\right] - \mathbb{E}_{X\sim\eta_{0,R}} \prod_{i=1}^r \frac{\eta_{I_i,R}(X)}{\eta_{0,R}(X)}\right| \tag{123}$$

$$\begin{align}
&= \Bigg| \mathbb{E}_{X\sim\eta_0}\left[\prod_{i=1}^r \frac{\eta_{I_i}(X)/\eta_{I_i}\left([-R,R]^d\right)}{\eta_0(X)/\eta_0\left([-R,R]^d\right)} \middle| X\in[-R,R]^d\right] - \mathbb{E}_{X\sim\eta_{0,R}} \prod_{i=1}^r \frac{\eta_{I_i,R}(X)}{\eta_{0,R}(X)} \notag\\
&+ \left(\frac{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)}{\eta_0\left([-R,R]^d\right)^{r-1}} - 1\right) \frac{\eta_0\left([-R,R]^d\right)^r}{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)} \mathbb{E}_{X\sim\eta_0}\left[\prod_{i=1}^r \frac{\eta_{I_i}(X)}{\eta_0(X)} \middle| X\in[-R,R]^d\right]\Bigg| \notag\\
&= \left|\left(\frac{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)}{\eta_0\left([-R,R]^d\right)^{r-1}} - 1\right) \frac{\eta_0\left([-R,R]^d\right)^r}{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)} \mathbb{E}_{X\sim\eta_0}\left[\prod_{i=1}^r \frac{\eta_{I_i}(X)}{\eta_0(X)} \middle| X\in[-R,R]^d\right]\right| \notag\\
&= \left|\left(\frac{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)}{\eta_0\left([-R,R]^d\right)^{r-1}} - 1\right) \frac{\eta_0\left([-R,R]^d\right)^r}{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)} \int_{x\in[-R,R]^d} \frac{\eta_0(x)}{\eta_0\left([-R,R]^d\right)} \prod_{i=1}^r \frac{\eta_{I_i}(x)}{\eta_0(x)}\right| \notag\\
&= \left|\left(\frac{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)}{\eta_0\left([-R,R]^d\right)^{r-1}} - 1\right) \frac{\eta_0\left([-R,R]^d\right)^{r-1}}{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)} \sqrt{\frac{\det \Sigma_{I_1,\ldots,I_r}}{(1-\sigma^2)^r}} \int_{x\in[-R,R]^d} P_W(x)\right| \tag{124}\\
&\le \left|\left(1 - \frac{\eta_0\left([-R,R]^d\right)^{r-1}}{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)}\right) \sqrt{\frac{\det \Sigma_{I_1,\ldots,I_r}}{(1-\sigma^2)^r}}\right| \notag\\
&\le \left|\left(1 - \frac{\eta_0\left([-R,R]^d\right)^{r-1}}{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)}\right) \sqrt{\frac{2}{1-\sigma^2 r}}\right| \tag{125}\\
&\le \frac{2}{\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right)} \left|\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right) - \eta_0\left([-R,R]^d\right)^{r-1}\right| \tag{126}\\
&\le \frac{2}{(1-p)^r} \left|\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right) - \eta_0\left([-R,R]^d\right)^{r-1}\right| \tag{127}\\
&\le 4\left|\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right) - \eta_0\left([-R,R]^d\right)^{r-1}\right|. \tag{128}
\end{align}$$

where Eq. (124) follows from Eq. (102), Eq. (125) follows from Lemma 43, Eq. (126) follows Eq. (92) which implies that $\sigma^2 r \le \sigma r \le \sigma b \le 1/2$,Eq. (127) follows from Eq. (118) and Eq. (128) follows from $(1-p)^r \ge 1 - rp \ge 1 - bp \ge 1/2$. We will estimate the term in Eq. (128). From Eq. (118),

$$\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right) - \eta_0\left([-R,R]^d\right)^{r-1} \le 1 - (1-p)^{r-1} \le 1 - (1-p)^b \le 1 - (1-bp) = bp.$$

Additionally,

$$\prod_{i=1}^r \eta_{I_i}\left([-R,R]^d\right) - \eta_0\left([-R,R]^d\right)^{r-1} \ge (1-p)^r - 1 \ge (1-p)^b - 1 \ge (1-bp) - 1 = -bp.$$

Hence, by definition of p in Eq. (114),

$$Eq.\ (121) \le Eq.\ (128) \le 4bp \le \frac{\sigma^b}{8n\binom{n}{\le b}2^b}. \tag{129}$$

Next, we will bound Eq. (122).

It follows from Lemma 43 that $\max|\Sigma_{I_1,\dots,I_r}| \le 2$ hence the variance of each coordinate of W (the d-variate normally distributed random variable with mean 0 and covariance $\Sigma_{I_1,\dots,I_r}$) is at most 2. Hence,

$$\Pr\left[W \notin [-R,R]^d\right] \le \sum_{i=1}^{d} \Pr\left[|W_i| > d\right] \le p, \tag{130}$$

using Eq. (117) and $R \ge \sqrt{4\ln\frac{2}{dp}}$. Therefore,

$$\begin{aligned} Eq.\ (122) &= \left(1-\eta_0\left([-R,R]^d\right)\right)\int_{x\in\mathbb{R}^d\setminus[-R,R]^d} \frac{\eta_0(x)}{1-\eta_0\left([-R,R]^d\right)} \prod_{i=1}^{r}\frac{\eta_{I_i}(X)}{\eta_0(X)} \\ &= \sqrt{\frac{\det\Sigma_{I_1,\dots,I_r}}{(1-\sigma^2)^r}}\int_{x\in\mathbb{R}^d\setminus[-R,R]^d} P_W(x) && (131) \\ &\le 4\Pr\left[W\notin[-R,R]^d\right] && (132) \\ &\le 4p && (133) \\ &\le \frac{\sigma^b}{8n\binom{n}{\le b}2^b} && (134) \end{aligned}$$

where Eq. (131) follow from Eq. (102), Eq. (132) follows from the same calculation as in Eq. (125) and Eq. (126), Eq. (133) follows from Eq. (130) and Eq. (134) follows from the definition of p in Eq. (114).

Note that Eq. (120), Eq. (129) and Eq. (134) imply Eq. (119). To conclude the proof,

$$\begin{aligned} &\left|\mathbb{E}_{X\sim\eta_0}\prod_{i=1}^{r}\left(\frac{\eta_{I_i}(X)}{\eta_0(X)}-1\right) - \mathbb{E}_{X\sim\eta_{0,R}}\prod_{i=1}^{r}\left(\frac{\eta_{I_i,R}(X)}{\eta_{0,R}(X)}-1\right)\right| \\ &\le \sum_{S\subseteq\{1,\dots,r\}}\left|\mathbb{E}_{X\sim\eta_0}\prod_{i\in S}\frac{\eta_{I_i}(X)}{\eta_0(X)} - \mathbb{E}_{X\sim\eta_{0,R}}\prod_{i=1}^{r}\frac{\eta_{I_i,R}(X)}{\eta_{0,R}(X)}\right| \le \frac{\sigma^b}{4n\binom{n}{\le b}}, \end{aligned}$$

where the last inequality follows from Eq. (119). ■

Define

$$\rho = \frac{2\sigma^2}{1-\sigma^2} + \frac{4\sigma R^2}{(1-\sigma^2)^2} + 2\sigma. \tag{135}$$

We will show that $\mathcal{G}_{\sigma,R}$, the family of truncated distributions, is $\mathrm{CD}(\rho)$ for the value of ρ presented in the following lemma.

Lemma 47 *For any $x \in [-R, R]^d$ and any $I \in \mathcal{I}_2$,*

$$\left| \frac{\eta_{I,R}(x)}{\eta_{0,R}(x)} - 1 \right| \leq \rho.$$

Proof Fix some $I = \{i, j\}$ and $x \in [-R, R]^d$.

$$\begin{aligned}
\frac{\eta_{I,\sigma}(x)}{\eta_{0,\sigma}(x)} &= \sqrt{\frac{\det \Sigma_0}{\det \Sigma_I}} \exp\left(-\frac{1}{2} x^t \left(\Sigma_I^{-1} - \mathrm{I}_d\right) x\right) \\
&= \frac{1}{1-\sigma^2} \exp\left(-\frac{x_i^2\sigma^2 - 2\sigma x_i x_j + x_j^2 \sigma^2}{2\left(1-\sigma^2\right)}\right) && (136) \\
&\leq \frac{1}{1-\sigma^2} \exp\left(\frac{\sigma R^2}{1-\sigma^2}\right) \\
&\leq \frac{1}{1-\sigma^2} \left(1 + \frac{2\sigma R^2}{1-\sigma^2}\right) && (137) \\
&= 1 + \frac{\sigma^2}{1-\sigma^2} + \frac{2\sigma R^2}{\left(1-\sigma^2\right)^2} && (138)
\end{aligned}$$

where Eq. (136) follows from Eq. (94) and from $\det \Sigma_I = 1 - \sigma^2$ which is proved in Lemma 42, Eq. (137) follows from $\sigma R^2 / \left(1 - \sigma^2\right) \leq 1$ which holds if the constant C in Theorem 12 is sufficiently large and the fact that $e^x \leq 1 + 2x$ for all $0 \leq x \leq 1$. Additionally,

$$\begin{aligned}
\frac{\eta_{I,\sigma}}{\eta_{0,\sigma}} &= \frac{1}{1-\sigma^2} \exp\left(-\frac{x_i^2\sigma^2 - 2\sigma x_i x_j + x_j^2 \sigma^2}{2\left(1-\sigma^2\right)}\right) && (139) \\
&\geq \frac{1}{1-\sigma^2} \left(1 - \frac{x_i^2\sigma^2 - 2\sigma x_i x_j + x_j^2 \sigma^2}{2\left(1-\sigma^2\right)}\right) && (140) \\
&\geq 1 - \frac{R^2\sigma^2}{\left(1-\sigma^2\right)^2} && (141)
\end{aligned}$$

where Eq. (139) is Eq. (136) and Eq. (140) follows from $e^{-x} \geq 1 - x$ for all $x \in \mathbb{R}$. Together, Eq. (138) and Eq. (141) imply that

$$\left| \frac{\eta_{I,\sigma}}{\eta_{0,\sigma}} - 1 \right| \leq \rho' := \frac{\sigma^2}{1-\sigma^2} + \frac{2\sigma R^2}{\left(1-\sigma^2\right)^2}. \tag{142}$$

Hence,

$$\begin{aligned}\left|\frac{\eta_{I,R}(x)}{\eta_{0,R}(x)}-1\right| &\leq \left|\frac{\eta_{I,R}(x)}{\eta_{0,R}(x)}-\frac{\eta_I(x)}{\eta_0(x)}\right|+\left|\frac{\eta_I(x)}{\eta_0(x)}-1\right| && (143)\\ &= \frac{\eta_I(x)}{\eta_0(x)}\left|\frac{\eta_0\left([-R,R]^d\right)}{\eta_I\left([-R,R]^d\right)}-1\right|+\left|\frac{\eta_I(x)}{\eta_0(x)}-1\right| \\ &\leq (1+\rho')\left|\frac{\eta_0\left([-R,R]^d\right)}{\eta_I\left([-R,R]^d\right)}-1\right|+\rho' && (144)\\ &= (1+\rho')\frac{\left|\eta_0\left([-R,R]^d\right)-\eta_I\left([-R,R]^d\right)\right|}{\eta_I\left([-R,R]^d\right)}+\rho' \\ &\leq (1+\rho')\frac{\sigma}{1-\sigma}+\rho' && (145)\\ &\leq (1+\rho')2\sigma+\rho' && (146)\\ &\leq 2\sigma+2\rho' && (147)\\ &= \rho && (148)\end{aligned}$$

where Eq. (144) follows from Eq. (142), Eq. (145) follows from Eq. (118) and Eq. (146) and Eq. (147) hold if the constant C from Theorem 12 is sufficiently large such that $\sigma \leq 1/2$. ■

Next, we show that Eq. (1) holds. We start with an auxiliary lemma.

Lemma 48 *For any $3 \leq \ell \leq d$, $2 \leq r \leq d$, the number of collections of sets $\mathcal{J} \subseteq \mathcal{I}_r$, $|\mathcal{J}| = \ell$, for which no element of $\{1,\ldots,d\}$ appears in exactly one set is at most $d^{\ell r/2}C(\ell,r)$, where*

$$C(\ell,r)=\begin{cases}\frac{1}{(\ell r/2)!}\binom{\binom{\ell r/2}{r}}{\ell} & \ell r \text{ is even}\\ 0 & \ell r \text{ is odd}\end{cases}.$$

Proof For any $\mathcal{J}$ satisfying the condition of the lemma, every index $i \in \bigcup\mathcal{J}$ is a member of at least 2 sets $I \in \mathcal{J}$. Therefore,

$$\left|\bigcup\mathcal{J}\right| \leq \frac{1}{2}\sum_{I\in\mathcal{J}}|I| = \frac{1}{2}\ell r,$$

since $\mathcal{J}$ contains ℓ sets, each of size r. Hence, each such $\mathcal{J}$ satisfies that $\bigcup\mathcal{J}$ is contained in some set J of size $r\ell/2$. There are $\binom{d}{r\ell/2}$ sets J of size $r\ell/2$. Each such J is a super-set of $\binom{|J|}{r}$ sets of size r, hence there are

$$\binom{\binom{|J|}{r}}{\ell}=\binom{\binom{\ell r/2}{r}}{\ell}$$

collections of ℓ subsets of J of size r. In total, there can be no more than

$$\binom{d}{\ell r/2}\binom{\binom{\ell r/2}{r}}{\ell} \leq \frac{d^{\ell r/2}}{(\ell r/2)!}\binom{\binom{\ell r/2}{r}}{\ell}$$

collections $\mathcal{J} \subseteq \mathcal{I}_r$ of size ℓ for which $\triangle\mathcal{J} = \emptyset$. This completes the proof for the case that ℓr is even. If $r\ell$ is odd, there is no collection $\mathcal{J} \subseteq \mathcal{I}_r$ of ℓ sets which satisfies $\triangle\mathcal{J} = \emptyset$: at least one of the elements in $\bigcup\mathcal{J}$ has to appear in an odd number of sets. ■

Lemma 49 *The following holds:*

$$\sum_{S\subseteq\mathcal{I}_2\colon |S|\ge 2} n^{-|S|/2}\rho^{-|S|}\left|\mathbb{E}_{A\sim\eta_0}\prod_{i\in S}(\eta_{I,\sigma,R}(A)/\eta_0(A)-1)\right| \le \frac{1}{n}. \tag{149}$$

Proof First, Lemma 22 implies that the sum of terms corresponding to $|S|>5$ is at most $1/(2n)$, assuming that the constant C of Theorem 12 is sufficiently large. Recall that $b=5$ by definition and we will bound the sum of terms corresponding to $2\le|S|\le b$. For any $2\le\ell\le b$, let $\mathcal{U}_\ell$ be the set of all collections of pairs $S\subseteq\mathcal{I}_2$ of size $|S|=\ell$ for which no element $i\in[d]$ appears in exactly one pair $I\in S$. We will bound the sum of terms corresponding to $S\notin\bigcup_{\ell=2}^{b}\mathcal{U}_\ell$, $2\le|S|\le b$: Lemma 44 and Lemma 46 imply that each such S contributes to the LHS of Eq. (149) at most

$$n^{-|S|/2}\rho^{-|S|}\frac{\rho^b}{4n\binom{n}{\le b}} \le \frac{1}{4n\binom{n}{\le b}},$$

where the inequality follows from $|S|\le b$ and $\rho\le 1$, assuming that the constant C from Theorem 12 is sufficiently large. The number of such sets is at most $\binom{n}{\le b}$, hence the total contribution of these sets is at most $1/(4n)$. Lastly, we bound the contribution of sets $S\in\bigcup_{\ell=2}^{b}\mathcal{U}_\ell$. It follows from Lemma 48 that there is a numerical constant C' such that $|\mathcal{U}_\ell|\le C'd^\ell$ for all $2\le\ell\le b$. Furthermore, it trivially holds that $|\mathcal{U}_2|=0$. Each set S contributes to the sum at most $n^{-|S|/2}$, from Lemma 47. Hence, the total contribution of sets $S\in\bigcup_{\ell=2}^{b}\mathcal{U}_\ell$ is at most

$$\sum_{\ell=3}^{b}|\mathcal{U}_\ell|\,n^{-\ell/2}\le\frac{1}{n}\sum_{\ell=3}^{b}C'd^\ell n^{-\ell/2-1}\le\frac{1}{n}\sum_{\ell=3}^{b}C'd^\ell n^{-\ell/6}\le\frac{1}{n}\sum_{\ell=3}^{b}\frac{C'}{C}\le\frac{1}{4n},$$

where C is the constant from Theorem 4 and the last inequality holds if C is sufficiently large. ■

We apply Thm. 4 on $\mathcal{G}_{R,\sigma}$. Lemma 47 implies that $\mathcal{G}_{\sigma,R}$ is a CD(ρ) family. Lemma 49 implies that Eq. (1) holds. Additionally, the definition of ρ in Eq. (135) and the definition of R in Eq. (115) imply that if the constant C from Thm. 12 is sufficiently large then the requirement that ρ is sufficiently small with respect to n and k holds.

A.4.4. FROM TRUNCATED TO STANDARD GAUSSIANS

To conclude the proof, we reduce the hardness of identifying a truncated normal distribution to the hardness of identifying a normal distribution, using the fact that with high probability, if we draw mn samples from a normal distribution $\eta\in\mathcal{G}_\sigma$, they are all in $[-R,R]^d$.

Lemma 50 *For any $0<\sigma<1$ and $0<\varepsilon<1$, if a protocol identifies $\eta\in\mathcal{G}_\sigma$ with a worst case error of ε then it identifies $\eta\in\mathcal{G}_{\sigma,R}$ with a worst case error of at most 2ε.*

Proof Let π be a protocol for identifying $\eta\in\mathcal{G}_\sigma$ with a worst case error of ε. Given an input containing the samples $x^{(1)},\dots,x^{(mn)}$ distributed by the m parties, let $\Pi(x^{(1)},\dots,x^{(mn)})\in\mathcal{I}_2$ be the random variable denoting the output of π when the input is $x^{(1)},\dots,x^{(mn)}$. Then, for any

$I \in \mathcal{I}_2$,

$$\begin{aligned}
\varepsilon \geq & \Pr_{(X^{(1)},\ldots,X^{(mn)})\sim\eta_{I,\sigma}^{mn}}\left[\Pi(X^{(1)},\ldots,X^{(mn)}) \neq I\right] \\
\geq & \Pr_{(X^{(1)},\ldots,X^{(mn)})\sim\eta_{I,\sigma}^{mn}}\left[\Pi(X^{(1)},\ldots,X^{(mn)}) \neq I, (X^{(1)},\ldots,X^{(mn)}) \in \left([-R,R]^d\right)^{mn}\right] \\
= & \Pr_{(X^{(1)},\ldots,X^{(mn)})\sim\eta_{I,\sigma}^{mn}}\left[\left(X^{(1)},\ldots,X^{(mn)}\right) \in \left([-R,R]^d\right)^{mn}\right] \cdot \\
& \Pr_{(X^{(1)},\ldots,X^{(mn)})\sim\eta_{I,\sigma}^{mn}}\left[\Pi(X^{(1)},\ldots,X^{(mn)}) \neq I \mid (X^{(1)},\ldots,X^{(mn)}) \in \left([-R,R]^d\right)^{mn}\right] \\
\geq & \frac{1}{2}\Pr_{(X^{(1)},\ldots,X^{(mn)})\sim\eta_{I,\sigma}^{mn}}\left[\Pi(X^{(1)},\ldots,X^{(mn)}) \neq I \mid (X^{(1)},\ldots,X^{(mn)}) \in \left([-R,R]^d\right)^{mn}\right] \qquad (150) \\
= & \frac{1}{2}\Pr_{(X^{(1)},\ldots,X^{(mn)})\sim\eta_{I,\sigma,R}^{mn}}\left[\Pi(X^{(1)},\ldots,X^{(mn)}) \neq I\right]. \qquad (151)
\end{aligned}$$

where Eq. (150) follows from Eq. (118). ■

Appendix B. Improved Results for Identifying Order-r Correlations

As discussed in Subsection 3.2, Theorems 9 and 10 assume that the correlation ρ is sufficiently small compared to the other problem parameters (and in the communication-constrained case, that n is sufficiently large). The following two theorems show that the assumptions can be somewhat relaxed, if we consider specifically the case of detecting order-r correlations (that is, the family of coordinate subsets we consider are $\mathcal{U} = \{I \in \mathcal{I} \colon |I| = r\}$, for some $r \geq 2$). In that case, we only require $n = \Omega(d^{3r})$ for r even and $n = \Omega(d^{2r+\varepsilon})$ for r odd in the communication-constrained case (whereas Thm. 9 requires $n = \Omega(d^{6r})$), and in the memory-constrained case, only $\rho = O(d^{-3r/2})$ for r even or even $\rho = O(d^{-(1+\epsilon)r})$ for r odd (whereas Thm. 10 requires $\rho = O(d^{-3r})$).

Theorem 51 *There exist numerical constants C', C'' and a positive function $C(r) \colon \mathbb{N} \to \mathbb{R}_+$ such that the following holds. Fix $2 \leq r \leq d-1$, and let $k = \binom{d}{r}$. Let n be an integer such that $n \geq d^{3r}C(r)$. Fix a number $0 < \rho \leq (n \ln k)^{-1/2}/C'$. Let $m \geq 1$ be an integer. Then, any (m,n) protocol identifying $\mu \in \mathcal{P}_{\mathcal{I}_r,\rho}$ has a communication complexity of at least*

$$\frac{k}{C''\rho^2 n \log(k^2/(n\rho^2))}. \qquad (152)$$

Furthermore, if r is odd then for any $0 < \varepsilon < 1$ there exists a number $C(r,\varepsilon)$ which depends only on r and ε such that Eq. (152) holds whenever $n \geq d^{(2+\varepsilon)r}C(r,\varepsilon)$.

Theorem 52 *There exist a numerical constant C' and a positive function $C(r) \colon \mathbb{N} \to \mathbb{R}_+$ such that the following holds. Fix $2 \leq r \leq d-1$ and fix a number $0 < \rho \leq d^{-3r/2}\ln^{-1/2} d/C(r)$. For any integers $t, s \geq 1$, any (t,s)-algorithm identifying $\mu \in \mathcal{P}_{\mathcal{I}_r,\rho}$ satisfies*

$$ts \geq \frac{\binom{d}{r}}{C'\rho^2 \ln\binom{d}{r}}.$$

Furthermore, if r is odd then for any $0 < \varepsilon < 1$ there exists a number $C(r, \varepsilon) \geq 1$ which depends only on r and ε such that Eq. (152) holds whenever $\rho \leq d^{-(1+\varepsilon)r}/C(r, \varepsilon)$.

Thm. 51 is derived from our general result (Thm. 4), by more delicately bounding the expression in Eq. (1), allowing us to use larger values of ρ and smaller values of n. A full proof is presented below. Thm. 52 is derived as a direct corollary of Thm. 51, using the same communication-to-memory reduction that we used for proving Thm. 5 based on Thm. 4.

B.1. Proof of Thm. 51

Lemma 48 implies that:

$$\sum_{\mathcal{J} \subseteq \mathcal{I}_\nabla : \, |\mathcal{J}| = \ell} \left| \mathbb{E} \prod_{I \in \mathcal{J}} B_I \right| \leq k^{\ell r/2} C(\ell, r), \tag{153}$$

for the value $C(\ell, r)$ appearing in this lemma. Indeed, from Lemma 37, the LHS of Eq. (153) equals the number of collections of sets $\mathcal{J} \subseteq \mathcal{I}_r$ of size $|\mathcal{J}| = \ell$ for which $\triangle \mathcal{J} = \emptyset$. For any such $\mathcal{J}$, every index $i \in \bigcup \mathcal{J}$ is a member of an even number of sets $I \in \mathcal{J}$, hence there is no element in $\{1, \ldots, d\}$ appearing in exactly one set from $\mathcal{J}$. This implies that the LHS of Eq. (153) is at most the term bounded in Lemma 48, hence Eq. (153) holds.

To prove Theorem 51, it is sufficient to show that Eq. (1) holds. Under the conditions of Theorem 51, the requirement that $n \geq d^{3r} \geq \binom{d}{r}^3 = k^3 = k^{2(5+1)/(5-1)}$ and Lemma 22 imply that the sum of all terms in Eq. (1) corresponding to $|S| > 5$ is at most $1/(2n)$. From Eq. (84) and Lemma 37 it holds that the sum of terms in Eq. (1) corresponding to $|S| = 2$ is zero, and by Eq. (84) and Eq. (153) the sum of terms corresponding to $|S| = \ell$ is at most $n^{-\ell/2} C(\ell, r) d^{\ell r/2}$, where $C(\ell, r)$ is the number from Lemma 48 . Hence, the sum of terms corresponding to $3 \leq |S| \leq 5$ is at most

$$\begin{aligned}
\sum_{\ell=3}^{5} n^{-\ell/2} d^{\ell r/2} C(\ell, r) &= \frac{1}{n} \sum_{\ell=3}^{5} n^{-\ell/2+1} d^{\ell r/2} C(\ell, r) \\
&\leq \frac{1}{n} \sum_{\ell=3}^{5} d^{-3\ell r/2+3r} C(r)^{-\ell/2+1} d^{\ell r/2} C(\ell, r) \\
&= \frac{1}{n} \sum_{\ell=3}^{5} d^{-\ell r+3r} C(r)^{-\ell/2+1} C(\ell, r) \\
&\leq \frac{1}{n} \sum_{\ell=3}^{5} C(r)^{-\ell/2+1} C(\ell, r) \\
&\leq \frac{1}{2n},
\end{aligned}$$

where the last inequality holds whenever $C(r)$ is sufficiently large as a function of $C(\ell, r)$, for $3 \leq \ell \leq 5$. Whenever this holds, Eq. (1) holds.

Next, assume that r is odd, fix $0 < \varepsilon < 1$, and let $\ell(\varepsilon)$ be the smallest integer which satisfies $2(\ell(\varepsilon)+1)/(\ell(\varepsilon)-1) \leq 2 + \varepsilon$. Since $n \geq d^{(2+\varepsilon)r} \geq \binom{d}{r}^{2+\varepsilon} \geq \binom{d}{r}^{2(\ell(\varepsilon)+1)/(\ell(\varepsilon)-1)}$, Lemma 22 implies that the sum of all terms in Eq. (1) corresponding to $|S| > \ell(\varepsilon)$ is at most $1/(2n)$. As in the case of a general r, all terms corresponding to $|S| = 2$ are zero. Inequalities Eq. (153) and Eq. (84)

imply that the sum of terms corresponding to $|S| = 3$ is zero. Similarly to the calculation in the case of a general r, the sum of terms corresponding to $4 \leq |S| \leq \ell(\varepsilon)$ is at most

$$\begin{aligned}\frac{1}{n}\sum_{\ell=4}^{\ell(\varepsilon)} n^{-\ell/2+1} d^{\ell r/2} C(\ell, r) &\leq \frac{1}{n}\sum_{\ell=4}^{\ell(\varepsilon)} d^{-\ell r+2r} C(r,\varepsilon)^{-\ell/2+1} d^{\ell r/2} C(\ell, r) \\ &= \frac{1}{n}\sum_{\ell=4}^{\ell(\varepsilon)} d^{-\ell r/2+2r} C(r,\varepsilon)^{-\ell/2+1} C(\ell, r) \\ &\leq \frac{1}{n}\sum_{\ell=4}^{\ell(\varepsilon)} C(r,\varepsilon)^{-\ell/2+1} C(\ell, r) \\ &\leq \frac{1}{2n},\end{aligned}$$

where the last inequality holds whenever $C(r, \varepsilon)$ is sufficiently large, which concludes the proof.

Appendix C. Comparison to Raz (2016)

Raz (2016) studied the problem of learning a linear function over $\mathbb{Z}_2^d$ (namely, d dimensional vectors of integers modulo 2): given samples (x, y), where x is picked uniformly at random from $\mathbb{Z}_2^d$ and $y = \langle w, x\rangle (\mod 2)$ for some unknown $w \in \mathbb{Z}_2^d$, the goal is learn w. He showed that with less than $d^2/10$ bits of memory, exponentially many samples are required. Intuitively, this memory requirement follows from the fact that one has to store $\Omega(d)$ samples in memory in order to learn w.

One can view this problem as a problem of learning a distribution over $(x_1, \ldots, x_{d+1}) \in \mathbb{Z}_2^{d+1}$, where $x = (x_1, \ldots, x_d)$ and $y = x_{d+1}$. There are 2^d possible distributions, each distribution corresponding to some $w \in \mathbb{Z}_2^d$. Furthermore, each distribution is $\mu_{I,\rho}$ for some $I \subseteq \{1, 2, \ldots, d+1\}$ and [13] $\rho = 1$. Moreover, the memory requirement is $\Theta(d^2)$ and $d + O(1)$ samples are required. In contrast, we use different techniques to study a very different regime: There are k distributions for some $k \in \{2, 3, \ldots, 2^d\}$, ρ is polynomially small in k, the memory requirement is $\tilde{\Theta}(k)$ and $\tilde{\Theta}(1/\varepsilon^2)$ samples are required. Additionally, our threshold is soft: one can learn with less memory and more samples, as opposed to requiring exponentially many samples already for $d^2/10$.

[13] Given $w = (w_1, \ldots, w_d)$, let $I = \{i \colon w_i = 1\} \cup \{d+1\}$. The distribution corresponding to w is uniform over all $(x_1, \ldots, x_{d+1}) \in \mathbb{Z}_2^{d+1}$ for which $\sum_{i \in I} x_i = 0$, which is equivalent to $\mu_{I,\rho}$ with $\rho = 1$ (the only difference is that in our setting, the elements are from $\{-1, 1\}$ instead of $\{0, 1\}$ and the operation is multiplication instead of addition).