

# Communication Lower Bounds for Statistical Estimation Problems via a Distributed Data Processing Inequality\*

Mark Braverman  
Princeton University  
35 Olden Street  
Princeton, USA  
mbraverm@princeton.edu

Ankit Garg  
Princeton University  
35 Olden Street  
Princeton, USA  
garg@cs.princeton.edu

Tengyu Ma  
Princeton University  
35 Olden Street  
Princeton, USA  
tengyu@cs.princeton.edu

Huy L. Nguyễn  
Toyota Technological Institute  
at Chicago  
6045 S Kenwood Ave  
Chicago, USA  
hlnghuyen@cs.princeton.edu

David P. Woodruff  
IBM Research Almaden  
650 Harry Rd  
San Jose  
CA 95120  
dpwoodru@us.ibm.com

## ABSTRACT

We study the tradeoff between the statistical error and communication cost of distributed statistical estimation problems in high dimensions. In the distributed sparse Gaussian mean estimation problem, each of the  $m$  machines receives  $n$  data points from a  $d$ -dimensional Gaussian distribution with unknown mean  $\theta$  which is promised to be  $k$ -sparse. The machines communicate by message passing and aim to estimate the mean  $\theta$ . We provide a tight (up to logarithmic factors) tradeoff between the estimation error and the number of bits communicated between the machines. This directly leads to a lower bound for the distributed *sparse linear regression* problem: to achieve the statistical minimax error, the total communication is at least  $\Omega(\min\{n, d\}m)$ , where  $n$  is the number of observations that each machine receives and  $d$  is the ambient dimension. These lower bound results improve upon [Sha14, SD15] by allowing a multi-round interactive communication model. We also give the first optimal simultaneous protocol in the dense case for mean estimation.

As our main technique, we prove a *distributed data processing inequality*, as a generalization of usual data processing inequalities, which might be of independent interest and useful for other problems.

## Categories and Subject Descriptors

H.4 [Information Systems Applications]: communication complexity, secure computation; F.1.3 [Software Engineering]: information and communication complexity;

\*This is an extended abstract. Full version appears at <http://arxiv.org/abs/1506.07216>

G.3 [Probability and Statistics]: Multivariate statistics

## General Terms

Theory

## Keywords

Communication complexity, Information complexity, statistical estimation.

## 1. INTRODUCTION

Rapid growth in the size of modern data sets has fueled a lot of interest in solving statistical and machine learning tasks in a distributed environment using multiple machines. Communication between the machines has emerged as an important resource and sometimes the main bottleneck. A lot of recent work has been devoted to design communication-efficient learning algorithms [DAW12, ZDW13, ZX15, KVW14, LBKW14, SSZ14, LSLT15, BWZ15, WZ16].

In this paper we consider statistical estimation problems in the distributed setting, which can be formalized as follows. There is a family of distributions  $\mathcal{P} = \{\mu_\theta : \theta \in \Omega \subset \mathbb{R}^d\}$  that is parameterized by  $\theta \in \mathbb{R}^d$ . Each of the  $m$  machines is given  $n$  i.i.d samples drawn from an unknown distribution  $\mu_\theta \in \mathcal{P}$ . The machines communicate with each other by message passing, and do computation on their local samples and the messages that they receives from others. Finally one of the machines needs to output an estimator  $\hat{\theta}$  and the statistical error is usually measured by the mean-squared loss  $\mathbb{E}[\|\hat{\theta} - \theta\|^2]$ . We count the communication between the machines in bits.

This paper focuses on understanding the fundamental tradeoff between communication and the statistical error for high-dimensional statistical estimation problems. Modern large datasets are often equipped with a high-dimensional statistical model, while communication of high dimensional vectors could potentially be expensive. It has been shown by Duchi et al. [DJWZ14] and Garg et al. [GMN14] that for the linear regression problem, the communication cost must

scale with the dimensionality for achieving optimal statistical minimax error – not surprisingly, the machines have to communicate high-dimensional vectors in order to estimate high-dimensional parameters.

These negative results naturally lead to the interest in high-dimensional estimation problems with additional sparse structure on the parameter  $\theta$ . It has been well understood that the statistical minimax error typically depends on the intrinsic dimension, that is, the sparsity of the parameters, instead of the ambient dimension<sup>1</sup>. Thus it is natural to expect that the same phenomenon also happens for communication.

However, this paper disproves this possibility in the interactive communication model by proving that for the *sparse Gaussian mean estimation* problem (where one estimates the mean of a Gaussian distribution which is promised to be sparse, see Section 2 for the formal definition), in order to achieve the statistical minimax error, the communication must scale with the ambient dimension. On the other end of the spectrum, if alternatively the communication only scales with the sparsity, then the statistical error must scale with the ambient dimension (see Theorem 4.5). Shamir [Sha14] establishes the same result for the 1-sparse case under a non-iterative communication model.

Our lower bounds for the Gaussian mean estimation problem imply lower bounds for the *sparse linear regression* problem (Corollary 4.8) via the reduction of [ZDJW13]: for a Gaussian design matrix, to achieve the statistical minimax error, the communication per machine needs to be  $\Omega(\min\{n, d\})$  where  $d$  is the ambient dimension and  $n$  is the dimension of the observations that each machine receives. This lower bound matches the upper bound in [LSLT15] when  $n$  is larger than  $d$ . When  $n$  is less than  $d$ , we note that it is not clear whether  $O(n)$  or  $O(d)$  should be the minimum communication cost per machine needed. In any case, our contribution here is in proving a lower bound that does not depend on the sparsity. Compared to previous work of Steinhardt and Duchi [SD15], which proves the same lower bounds for a memory-bounded model, our results work for a stronger communication model where multi-round iterative communication is allowed. Moreover, our techniques are possibly simpler and potentially easier to adapt to related problems. For example, we show that the result of Woodruff and Zhang [WZ12] on the information complexity of distributed gap majority can be reproduced by our technique with a cleaner proof (see the full paper for the proof).

We complement our lower bounds for this problem in the dense case by providing a new simultaneous protocol, improving the number of rounds of the previous communication-optimal protocol from  $O(\log m)$  to 1 (see Theorem 4.6). Our protocol is based on a certain combination of many bits from a few Gaussian samples, together with roundings (to a single bit) of the fractional parts of many Gaussian samples.

Our proof techniques are potentially useful for other questions along these lines. We first use a modification of the direct-sum result of [GMN14], which is tailored towards sparse problems, to reduce the estimation problem to a detection problem. Then we prove what we call a *distributed data processing inequality* for bounding from below the cost

<sup>1</sup>the dependency on the ambient dimension is typically logarithmic.

of the detection problem. The latter is the crux of our proofs. We elaborate more on it in the next subsection.

## 1.1 Distributed Data Processing Inequality

We consider the following distributed detection problem. As we will show in Section 4 (by a direct-sum theorem), it suffices to prove a tight lower bound in this setting, in order to prove a lower bound on the communication cost for the sparse linear regression problem.

**Distributed detection problem:** We have a family of distributions  $\mathcal{P}$  that consist of only two distributions  $\{\mu_0, \mu_1\}$ , and the parameter space  $\Omega = \{0, 1\}$ . To facilitate the use of tools from information theory, sometimes it is useful to introduce a prior over the parameter space. Let  $V \sim B_q$  be a Bernoulli random variable with probability  $q$  of being 1. Given  $V = v \in \{0, 1\}$ , we draw i.i.d. samples  $X_1, \dots, X_m$  from  $\mu_v$  and the  $j$ -th machine receives one sample  $X_j$ , for  $j = 1, \dots, m$ . We use  $\Pi \in \{0, 1\}^*$  to denote the sequences of messages that are communicated by the machines. We will refer to  $\Pi$  as a “transcript”, and the distributed algorithm that the machines execute as a “protocol”.

The final goal of the machines is to output an estimator for the hidden parameter  $v$  which is as accurate as possible. We formalize the estimator as a (random) function  $\hat{v} : \{0, 1\}^* \rightarrow \{0, 1\}$  that takes the transcript  $\Pi$  as input. We require that given  $V = v$ , the estimator is correct with probability at least  $3/4$ , that is,  $\min_{v \in \{0, 1\}} \Pr[\hat{v}(\Pi) = v \mid V = v] \geq 3/4$ . When  $q = 1/2$ , this is essentially equivalent to the statement that the transcript  $\Pi$  carries  $\Omega(1)$  information about the random variable  $V$ . Therefore, the mutual information  $I(V; \Pi)$  is also used as a convenient measure for the quality of the protocol when  $q = 1/2$ .

**Strong data processing inequality:** The mutual information viewpoint of the accuracy naturally leads us to the following approach for studying the simple case when  $m = 1$  and  $q = 1/2$ . When  $m = 1$ , we note that the parameter  $V$ , data  $X$ , and transcript  $\Pi$  form a simple Markov chain  $V \rightarrow X \rightarrow \Pi$ . The channel  $V \rightarrow X$  is defined as  $X \sim \mu_v$ , conditioned on  $V = v$ . The strong data processing inequality (SDPI) captures the relative ratio between  $I(V; \Pi)$  and  $I(X; \Pi)$ .

**DEFINITION 1 (SPECIAL CASE OF SDPI).** *Let  $V \sim B_{1/2}$  and the channel  $V \rightarrow X$  be defined as above. Then there exists a constant  $\beta \leq 1$  that depends on  $\mu_0$  and  $\mu_1$ , such that for any  $\Pi$  that depends only on  $X$  (that is,  $V \rightarrow X \rightarrow \Pi$  forms a Markov Chain), we have*

$$I(V; \Pi) \leq \beta \cdot I(X; \Pi). \quad (1)$$

*An inequality of this type is typically referred to as a strong data processing inequality for mutual information when  $\beta < 1$ <sup>2</sup>. Let  $\beta(\mu_0, \mu_1)$  be the infimum over all possible  $\beta$  such that (1) is true, which we refer to as the **SDPI constant**.*

Observe that the LHS of (1) measures how much information  $\Pi$  carries about  $V$ , which is closely related to the accuracy of the protocol. The RHS of (1) is a lower bound on the expected length of  $\Pi$ , that is, the expected communication cost. Therefore the inequality relates two quantities that we are interested in – the statistical quality of the

<sup>2</sup>Inequality (1) is always true for a Markov chain  $V \rightarrow X \rightarrow \Pi$  with  $\beta = 1$  and this is called the data processing inequality.

protocol and the communication cost of the protocol. Concretely, when  $q = 1/2$ , in order to recover  $V$  from  $\Pi$ , we need that  $I(V; \Pi) \geq \Omega(1)$ , and therefore inequality (1) gives that  $I(X; \Pi) \geq \Omega(\beta^{-1})$ . Then it follows from Shannon's source coding theory that the expected length of  $\Pi$  (denoted by  $|\Pi|$ ) is bounded from below by  $\mathbb{E}[|\Pi|] \geq \Omega(\beta^{-1})$ . We refer to [Rag14] for a thorough survey of SDPI.<sup>3</sup>

In the multiple machine setting, Duchi et al. [DJWZ14] links the distributed detection problem with SDPI by showing from scratch that for any  $m$ , when  $q = 1/2$ , if  $\beta$  is such that  $(1 - \sqrt{\beta})\mu_1 \leq \mu_0 \leq (1 + \sqrt{\beta})\mu_1$ , then

$$I(V; \Pi) \leq \beta \cdot I(X_1 \dots X_m; \Pi).$$

This results in the bounds for the Gaussian mean estimation problem and the linear regression problem. The main limitation of this inequality is that it requires the prior  $B_q$  to be unbiased (or close to unbiased). For our target application of high-dimensional problems with sparsity structures, like sparse linear regression, in order to apply this inequality we need to put a very biased prior  $B_q$  on  $V$ . The proof technique of [DJWZ14] seems also hard to extend to this case with a tight bound<sup>4</sup>. Moreover, the relation between  $\beta$ ,  $\mu_0$  and  $\mu_1$  may not be necessary (or optimal), and indeed for the Gaussian mean estimation problem, the inequality is only tight up to a logarithmic factor, while potentially in other situations the gap is even larger.

Our approach is essentially a prior-free multi-machine SDPI, which has the same SDPI constant  $\beta$  as is required for the single machine one. We prove that, as long as the SDPI (1) for a single machine is true with parameter  $\beta$ , and  $\mu_0 \leq O(1)\mu_1$ , then the following prior-free multi-machine SDPI is true with the same constant  $\beta$  (up to a constant factor).

**THEOREM 1.1 (DISTRIBUTED SDPI).** *Suppose  $\frac{1}{c} \cdot \mu_0 \leq \mu_1 \leq c\mu_0$  for some constant  $c \geq 1$ , and let  $\beta(\mu_0, \mu_1)$  be the SDPI constant defined in Definition 1. Then in the distributed detection problem, we have the following distributed strong data processing inequality,*

$$\begin{aligned} & h^2(\Pi|_{V=0}, \Pi|_{V=1}) \\ & \leq Kc\beta(\mu_0, \mu_1) \cdot \min_{v \in \{0,1\}} \{I(X_1 \dots X_m; \Pi | V = v)\} \end{aligned} \quad (2)$$

where  $K$  is a universal constant, and  $h(\cdot, \cdot)$  is the Hellinger distance between two distributions and  $\Pi|_{V=v}$  denotes the distribution of  $\Pi$  conditioned on  $V = v$ .

Moreover, for any  $\mu_0$  and  $\mu_1$  which satisfy the condition of the theorem, there exists a protocol that produces transcript  $\Pi$  such that (2) is tight up to a constant factor.

As an immediate consequence, we obtain a lower bound on the communication cost for the distributed detection problem.

<sup>3</sup>Also note that in information theory, SDPI is typically interpreted as characterizing how information decays when passed through the reverse channel  $X \rightarrow V$ . That is, when the channel  $X \rightarrow V$  is lossy, then information about  $\Pi$  will decay by a factor of  $\beta$  after passing  $X$  through the channel. However, in this paper we take a different interpretation that is more convenient for our applications.

<sup>4</sup>We note, though, that it seems possible to extend the proof to the situation where there is only one-round of communication.

**COROLLARY 1.2.** *Suppose the protocol and estimator  $(\Pi, \hat{v})$  are such that for any  $v \in \{0, 1\}$ , given  $V = v$ , the estimator  $\hat{v}$  (that takes  $\Pi$  as input) can recover  $v$  with probability  $3/4$ . Then*

$$\max_{v \in \{0,1\}} \mathbb{E}[|\Pi| | V = v] \geq \Omega(\beta^{-1}).$$

Our theorem suggests that to bound the communication cost of the multi-machine setting from below, one could simply work in the single machine setting and obtain the right SDPI constant  $\beta$ . Then, a lower bound of  $\Omega(\beta^{-1})$  for the multi-machine setting immediately follows. In other words, multi-machines need to communicate a lot to fully exploit the  $m$  data points they receive (1 on each single machine) regardless of however complicated their multi-round protocol is.

**REMARK 1.** *Note that our inequality differs from the typical data processing inequality on both the left and right hand sides. First of all, the RHS of (2) is always less than or equal to  $I(X_1 \dots X_m; \Pi | V)$  for any prior  $B_q$  on  $V$ . This allows us to have a tight bound on the expected communication  $\mathbb{E}[|\Pi|]$  for the case when  $q$  is very small.*

*Second, the squared Hellinger distance (see Definition 4) on the LHS of (2) is not very far away from  $I(\Pi; V)$ , especially for the situation that we consider. It can be viewed as an alternative (if not more convenient) measure of the quality of the protocol than mutual information – the further  $\Pi|_{V=0}$  from  $\Pi|_{V=1}$ , the easier it is to infer  $V$  from  $\Pi$ . When a good estimator is possible (which is the case that we are going to apply the bound in), Hellinger distance, total variation distance between  $\Pi|_{V=0}$  and  $\Pi|_{V=1}$ , and  $I(V; \Pi)$  are all  $\Omega(1)$ . Therefore in this case, the Hellinger distance does not make the bound weaker.*

*Finally, suppose we impose a uniform prior for  $V$ . Then the squared Hellinger distance is within a constant factor of  $I(V; \Pi)$  (see Lemma 4, and the lower bound side was proved by [BYJKS04]),*

$$2h^2(\Pi|_{V=0}, \Pi|_{V=1}) \geq I(V; \Pi) \geq h^2(\Pi|_{V=0}, \Pi|_{V=1}).$$

*Therefore, in the unbiased case, (2) implies the typical form of the data processing inequality.*

**REMARK 2.** *The tightness of our inequality does not imply that there is a protocol that solves the distributed detection problem with communication cost (or information cost)  $O(\beta^{-1})$ . We only show that inequality (2) is tight for some protocol but solving the problem requires having a protocol such that (2) is tight and that  $h^2(\Pi|_{V=0}, \Pi|_{V=1}) = \Omega(1)$ . In fact, a protocol for which inequality (2) is tight is one in which only a single machine sends a message  $\Pi$  which maximizes  $I(\Pi; V)/I(\Pi; X)$ .*

**Organization of the paper:** Section 2 formally sets up our model and problems and introduces some preliminaries. Then we prove our main theorem in Section 3. In Section 4 we state the main applications of our theory to the sparse Gaussian mean estimation problem and to the sparse linear regression problem. The next three sections are devoted to the proofs of results in Section 4. In Section 5, we prove Theorem 4.4. The other missing proofs appear in the full paper.

## 2. PROBLEM SETUP, NOTATIONS AND PRELIMINARIES

### 2.1 Distributed Protocols and Parameter Estimation Problems

Let  $\mathcal{P} = \{\mu_\theta : \theta \in \Omega\}$  be a family of distributions over some space  $\mathcal{X}$ , and  $\Omega \subset \mathbb{R}^d$  be the space of all possible parameters. There is an unknown distribution  $\mu_\theta \in \mathcal{P}$ , and our goal is to estimate a parameter  $\theta$  using  $m$  machines. Machine  $j$  receives  $n$  i.i.d samples  $X_j^{(1)}, \dots, X_j^{(n)}$  from distribution  $\mu_\theta$ . For simplicity we will use  $X_j$  as a shorthand for all the samples machine  $j$  receives, that is,  $X_j = (X_j^{(1)}, \dots, X_j^{(n)})$ . Therefore  $X_j \sim \mu_\theta^n$ , where  $\mu_\theta^n$  denotes the product of  $n$  copies of  $\mu$ . When it is clear from context, we will use  $X$  as a shorthand for  $(X_1, \dots, X_m)$ . We define the problem of estimating parameter  $\theta$  in this distributed setting formally as task  $T(n, m, \mathcal{P})$ . When  $\Omega = \{0, 1\}$ , we call this a detection problem and refer it to as  $T_{det}(n, m, \mathcal{P})$ .

The machines communicate via a publicly shown blackboard. That is, when a machine writes a message on the blackboard, all other machines can see the content. The messages that are written on the blackboard are counted as communication between the machines. Note that this model captures both point-to-point communication as well as broadcast communication. Therefore, our lower bounds in this model apply to both the message passing setting and the broadcast setting.

We denote the collection of all the messages written on the blackboard by  $\Pi$ . We will refer to  $\Pi$  as the transcript and note that  $\Pi \in \{0, 1\}^*$  is written in bits and the communication cost is defined as the length of  $\Pi$ , denoted by  $|\Pi|$ . We will call the algorithm that the machines follow to produce  $\Pi$  a protocol. With a slight abuse of notation, we use  $\Pi$  to denote both the protocol and the transcript produced by the protocol.

One of the machines needs to estimate the value of  $\theta$  using an estimator  $\hat{\theta} : \{0, 1\}^* \rightarrow \mathbb{R}^d$  which takes  $\Pi$  as input. The accuracy of the estimator on  $\theta$  is measured by the mean-squared loss:

$$R(\Pi, \hat{\theta}, \theta) = \mathbb{E} \left[ \|\hat{\theta}(\Pi) - \theta\|_2^2 \right],$$

where the expectation is taken over the randomness of the data  $X$ , and the estimator  $\hat{\theta}$ . The error of the estimator is the supremum of the loss over all  $\theta$ ,

$$R(\Pi, \hat{\theta}) = \sup_{\theta \in \Omega} \mathbb{E} \left[ \|\hat{\theta}(\Pi) - \theta\|_2^2 \right]. \quad (3)$$

The communication cost of a protocol is measured by the expected length of the transcript  $\Pi$ , that is,  $CC(\Pi) = \sup_{\theta \in \Omega} \mathbb{E}[|\Pi|]$ . The information cost  $IC$  of a protocol is defined as the mutual information between transcript  $\Pi$  and the data  $X$ ,

$$IC(\Pi) = \sup_{\theta \in \Omega} I_\theta(\Pi; X | R_{pub}) \quad (4)$$

where  $R_{pub}$  denotes the public coin used by the algorithm and  $I_\theta(\Pi; X | R_{pub})$  denotes the mutual information between random variable  $X$  and  $\Pi$  when the data  $X$  is drawn from distribution  $\mu_\theta$ . We will drop the subscript  $\theta$  when it is clear from context.

For the detection problem, we need to define minimum information cost, a stronger version of information cost

$$\min\text{-IC}(\Pi) = \min_{v \in \{0, 1\}} I_v(\Pi; X | R_{pub}) \quad (5)$$

**DEFINITION 2.** We say that a protocol and estimator pair  $(\Pi, \hat{\theta})$  solves the distributed estimation problem  $T(m, n, d, \Omega, \mathcal{P})$  with information cost  $I$ , communication cost  $C$ , and mean-squared loss  $R$  if  $IC(\Pi) \leq I$ ,  $CC(\Pi) \leq C$  and  $R(\Pi, \hat{\theta}) \leq R$ .

When  $\Omega = \{0, 1\}$ , we have a detection problem, and we typically use  $v$  to denote the parameter and  $\hat{v}$  as the (discrete) estimator for it. We define the communication and information cost the same as (2.1) and (4), while defining the error in a more meaningful and convenient way,

$$R_{det}(\Pi, \hat{v}) = \max_{v \in \{0, 1\}} \Pr[\hat{v}(\Pi) \neq v | V = v]$$

**DEFINITION 3.** We say that a protocol and estimator pair  $(\Pi, \hat{v})$  solves the distributed detection problem  $T_{det}(m, n, d, \Omega, \mathcal{P})$  with information cost  $I$ , if  $IC(\Pi) \leq I$ ,  $R_{det}(\Pi, \hat{v}) \leq 1/4$ .

Now we formally define the concrete questions that we are concerned with.

**Distributed Gaussian detection problem:** We call the problem with  $\Omega = \{0, 1\}$  and  $\mathcal{P} = \{\mathcal{N}(0, \sigma^2)^n, \mathcal{N}(\delta, \sigma^2)^n\}$  the Gaussian mean detection problem, denoted by  $GD(n, m, \delta, \sigma^2)$ .

**Distributed (sparse) Gaussian mean estimation problem:** The distributed statistical estimation problem defined by  $\Omega = \mathbb{R}^d$  and  $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2 I_{d \times d}) : \theta \in \Omega\}$  is called the distributed Gaussian mean estimation problem, abbreviated  $GME(n, m, d, \sigma^2)$ . When  $\Omega = \{\theta \in \mathbb{R}^d : |\theta|_0 \leq k\}$ , the corresponding problem is referred to as distributed sparse Gaussian mean estimation, abbreviated  $SGME(n, m, d, k, \sigma^2)$ .

**Distributed sparse linear regression:** For simplicity and the purpose of lower bounds, we only consider sparse linear regression with a random design matrix. To fit into our framework, we can also regard the design matrix as part of the data. We have a parameter space  $\Omega = \{\theta \in \mathbb{R}^d : |\theta|_0 \leq k\}$ . The  $j$ -th data point consists of a row of design matrix  $A_j$  and the observation  $y_j = \langle A_j, \theta \rangle + w_j$  where  $w_j \sim \mathcal{N}(0, \sigma^2)$  for  $j = 1, \dots, mn$ , and each machine receives  $n$  data points among them<sup>5</sup>. Formally, let  $\mu_\theta$  denote the joint distribution of  $(A_j, y_j)$  here, and let  $\mathcal{P} = \{\mu_\theta : \theta \in \Omega\}$ . We use  $SLR(n, m, d, k, \sigma^2)$  as shorthand for this problem.

### 2.2 Hellinger distance and cut-paste property

In this subsection, we introduce Hellinger distance, and the key property of protocols that we exploit here, the so-called ‘‘cut-paste’’ property developed by [BYJKS04] for proving lower bounds for set-disjointness and other problems. We also introduce some notation that will be used later in the proofs.

**DEFINITION 4 (HELLINGER DISTANCE).** Consider two distributions with probability density functions  $f, g : \Omega \rightarrow \mathbb{R}$ . The square of the Hellinger distance between  $f$  and  $g$  is defined as  $h^2(f, g) := \frac{1}{2} \cdot \int_\Omega \left( \sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx$

<sup>5</sup>We note that here for convenience, we use subscripts for samples, which is different from the notation convention used for previous problems.

A key observations regarding the property of a protocol by [BYJKS04, Lemma 16] is the following: fixing  $X_1 = x_1, \dots, X_m = x_m$ , the distribution of  $\Pi|_{X=x}$  can be factored in the following form,

$$\Pr[\Pi = \pi \mid X = x] = p_{1,\pi}(x_1) \dots p_{m,\pi}(x_m) \quad (6)$$

where  $p_{i,\pi}(\cdot)$  is a function that only depends on  $i$  and the entire transcript  $\pi$ . To see this, one could simply write the density of  $\pi$  as a product of densities of each message of the machines and group the terms properly according to machines (and note that  $p_{i,\pi}(\cdot)$  is allowed to depend on the entire transcript  $\pi$ ).

We extend equation (6) to the situation where the inputs are from product distributions. For any vector  $\mathbf{b} \in \{0, 1\}^m$ , let  $\mu_{\mathbf{b}} := \mu_{b_1} \times \dots \times \mu_{b_m}$  be a distribution over  $\mathcal{X}^m$ . We denote by  $\Pi_{\mathbf{b}}$  the distribution of  $\Pi(X_1, \dots, X_m)$  when  $(X_1, \dots, X_m) \sim \mu_{\mathbf{b}}$ .

Therefore if  $X \sim \mu_{\mathbf{b}}$ , using the fact that  $\mu_{\mathbf{b}}$  is a product measure, we can marginalize over  $X$  and obtain the marginal distribution of  $\Pi$  when  $X \sim \mu_{\mathbf{b}}$ ,

$$\Pr_{X \sim \mu_{\mathbf{b}}} [\Pi = \pi] = q_{1,\pi}(b_1) \dots q_{m,\pi}(b_m), \quad (7)$$

where  $q_{j,\pi}(b_j)$  is the marginalization of  $p_{j,\pi}(x)$  over  $x \sim \mu_{b_j}$ , that is,  $q_{j,\pi}(b_j) = \int_x p_{j,\pi}(x) d\mu_{b_j}$ .

Let  $\Pi_{\mathbf{b}}$  denote the distribution of  $\Pi$  when  $X \sim \mu_{\mathbf{b}}$ . Then by the decomposition (7) of  $\Pi_{\mathbf{b}}(\pi)$  above, we have the following cut-paste property for  $\Pi_{\mathbf{b}}$  which will be the key property of a protocol that we exploit.

**PROPOSITION 2.1 (CUT-PASTE PROPERTY OF A PROTOCOL).** *For any  $\mathbf{a}, \mathbf{b}$  and  $\mathbf{c}, \mathbf{d}$  with  $\{a_i, b_i\} = \{c_i, d_i\}$  (in a multi-set sense) for every  $i \in [m]$ ,*

$$\Pi_{\mathbf{a}}(\pi) \cdot \Pi_{\mathbf{b}}(\pi) = \Pi_{\mathbf{c}}(\pi) \cdot \Pi_{\mathbf{d}}(\pi) \quad (8)$$

and therefore,

$$h^2(\Pi_{\mathbf{a}}, \Pi_{\mathbf{b}}) = h^2(\Pi_{\mathbf{c}}, \Pi_{\mathbf{d}}) \quad (9)$$

### 3. DISTRIBUTED STRONG DATA PROCESSING INEQUALITIES

In this section we prove our main Theorem 1.1. We state a slightly weaker looking version here but in fact it implies Theorem 1.1 by symmetry. The same proof also goes through for the case when the RHS is conditioned on  $V = 1$ .

**THEOREM 3.1.** *Suppose  $\mu_1 \leq c \cdot \mu_0$ , and  $\beta(\mu_0, \mu_1) = \beta$ . We have*

$$h^2(\Pi|_{V=0}, \Pi|_{V=1}) \leq K(c+1)\beta \cdot I(X; \Pi \mid V = 0). \quad (10)$$

where  $K$  is an absolute constant.

Note that the RHS of (10) naturally tensorizes (by Lemma 1 that appears below) in the sense that

$$\sum_{i=1}^m I(X_i; \Pi \mid V = 0) \leq I(X; \Pi \mid V = 0), \quad (11)$$

since conditioned on  $V = 0$ , the  $X_i$ 's are independent. Our main idea consists of the following two steps a) We tensorize the LHS of (10) so that the target inequality (10) can be written as a sum of  $m$  inequalities. b) We prove each of these  $m$  inequalities using the single machine SDPI. To this

end, we do the following thought experiment: Suppose  $W$  is a random variable that takes value from  $\{0, 1\}$  uniformly. Suppose data  $X'$  is generated as follows:  $X'_i \sim \mu_W$ , and for any  $j \neq i$ ,  $X'_j \sim \mu_0$ . We apply the protocol on the input  $X'$ , and view the resulting transcript  $\Pi'$  as communication between the  $i$ -th machine and the remaining machines. Then we are in the situation of a single machine case, that is,  $W \rightarrow X'_i \rightarrow \Pi'$  forms a Markov Chain. Applying the data processing inequality (1), we obtain that

$$I(W; \Pi') \leq \beta I(X'_i; \Pi'). \quad (12)$$

Using Lemma 4, we can lower bound the LHS of (12) by the Hellinger distance and obtain

$$h^2(\Pi'|_{W=0}, \Pi'|_{W=1}) \leq \beta \cdot I(X'_i; \Pi')$$

Let  $\mathbf{e}_i = (0, 0, \dots, 1, \dots, 0)$  be the unit vector that only takes 1 in the  $i$ th entry, and  $\mathbf{0}$  the all zero vector. Using the notation defined in Section 2.2, we observe that  $\Pi'|_{W=0}$  has distribution  $\Pi_{\mathbf{0}}$  while  $\Pi'|_{W=1}$  has distribution  $\Pi_{\mathbf{e}_i}$ . Then we can rewrite the equation above as

$$h^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}) \leq \beta \cdot I(X'_i; \Pi') \quad (13)$$

Observe that the RHS of (13) is close to the first entry of the LHS of (11) since the joint distribution of  $(X'_i, \Pi')$  is not very far from  $X, \Pi \mid V = 0$ . (The only difference is that  $X'_i$  is drawn from a mixture of  $\mu_0$  and  $\mu_1$ , and note that  $\mu_0$  is not too far from  $\mu_1$ ). On the other hand, the sum of LHS of (13) over  $i \in [m]$  is lower-bounded by the LHS of (10). Therefore, we can tensorize equation (10) into inequality (13) which can be proved by the single machine SDPI. We formalize the intuition above by the following two lemmas,

**LEMMA 1.** *Suppose  $\mu_1 \leq c \cdot \mu_0$ , and  $\beta(\mu_0, \mu_1) = \beta$ , then*

$$h^2(\Pi_{\mathbf{e}_i}, \Pi_{\mathbf{0}}) \leq \frac{(c+1)\beta}{2} \cdot I(X_i; \Pi \mid V = 0) \quad (14)$$

**LEMMA 2.** *Let  $\mathbf{0}$  be the  $m$ -dimensional all 0's vector, and  $\mathbf{1}$  the all 1's vector, we have that*

$$h^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}) \leq O(1) \cdot \sum_{i=1}^m h^2(\Pi_{\mathbf{e}_i}, \Pi_{\mathbf{0}}) \quad (15)$$

Using Lemma 1 and Lemma 2, we obtain Theorem 3.1 straightforwardly by combining inequalities (11), (14) and (15)<sup>6</sup>.

Finally we provide the proof of Lemma 1. Lemma 2 is a direct corollary of Theorem A.1 (which is in turn a direct corollary of Theorem 7 of [Jay09]) and Proposition 2.1.

**PROOF OF LEMMA 1.** Let  $W$  be a uniform Bernoulli random variable and define  $X'$  and  $\Pi'$  as follows: Conditioned on  $W = 0$ ,  $X' \sim \mu_{\mathbf{0}}$  and conditioned on  $W = 1$ ,  $X' \sim \mu_{\mathbf{e}_i}$ . We run protocol on  $X'$  and get transcript  $\Pi'$ .

Note that  $V \rightarrow X' \rightarrow \Pi'$  is a Markov chain and so is  $V \rightarrow X'_i \rightarrow \Pi'$ . Also by definition, the conditional random variable  $X'_i|V$  has the same distribution as the random variable  $X|V$  in Definition 1. Therefore by Definition 1, we have that

$$\beta \cdot I(X'_i; \Pi') \geq I(V; \Pi'). \quad (16)$$

<sup>6</sup>Note that  $\Pi_{\mathbf{0}}$  is the same distribution as  $\Pi|_{V=0}$  under the notation introduced in Section 2.2.

It is known that mutual information can be expressed as the expectation of KL divergence, which in turn is lower-bounded by Hellinger distance. We invoke a technical variant of this argument, Lemma 6.2 of [BJKS04], restated as Lemma 4, to lower bound the right hand side. Note that  $Z$  in Lemma 4 corresponds to  $V$  here and  $\phi_{z_1}, \phi_{z_2}$  correspond to  $\Pi_{e_i}$  and  $\Pi_0$ . Therefore,

$$\mathbb{I}(V; \Pi') \geq h^2(\Pi_{e_i}, \Pi_0). \quad (17)$$

It remains to relate  $\mathbb{I}(X'_i; \Pi')$  to  $\mathbb{I}(X_i; \Pi \mid V = 0)$ . Note that the difference between joint distributions of  $(X'_i, \Pi')$  and  $(X_i, \Pi)|_{V=0}$  is that  $X'_i \sim \frac{1}{2}(\mu_0 + \mu_1)$  and  $X_i|_{V=0} \sim \mu_0$ . We claim (by Lemma 5) that since  $\mu_0 \geq \frac{2}{c+1}(\frac{\mu_0 + \mu_1}{2})$ , we have

$$\mathbb{I}(X_i; \Pi \mid V = 0) \geq \frac{2}{c+1} \cdot \mathbb{I}(X'_i; \Pi'). \quad (18)$$

Combining equations (16), (17) and (18), we obtain the desired inequality.

□

## 4. APPLICATIONS TO PARAMETER ESTIMATION PROBLEMS

### 4.1 Warm-up: Distributed Gaussian mean detection

In this section we apply our main technical Theorem 3.1 to the situation when  $\mu_0 = \mathcal{N}(0, \sigma^2)$  and  $\mu_1 = \mathcal{N}(\delta, \sigma^2)$ . We are also interested in the case when each machine receives  $n$  samples from either  $\mu_0$  or  $\mu_1$ . We will denote the product of  $n$  i.i.d copies of  $\mu_v$  by  $\mu_v^n$ , for  $v \in \{0, 1\}$ .

Theorem 3.1 requires that a)  $\beta = \beta(\mu_0, \mu_1)$  can be calculated/estimated b) the densities of distributions  $\mu_0$  and  $\mu_1$  are within a constant factor with each other at every point.

Certainly b) is not true for any two Gaussian distributions. To this end, we consider  $\mu'_0, \mu'_1$ , the truncation of  $\mu_0$  and  $\mu_1$  on some support  $[-\tau, \tau]$ , and argue that the probability mass outside  $[-\tau, \tau]$  is too small to make a difference.

For a), we use tools provided by Raginsky [Rag14] to estimate the SDPI constant  $\beta$ . [Rag14] proves that Gaussian distributions  $\mu_0$  and  $\mu_1$  have SDPI constant  $\beta(\mu_0, \mu_1) \leq O(\delta^2/\sigma^2)$ , and more generally it connects the SDPI constants to transportation inequalities. We use the framework established by [Rag14] and apply it to the truncated Gaussian distributions  $\mu'_0$  and  $\mu'_1$ . Our proof essentially uses the fact that  $(\mu'_0 + \mu'_1)/2$  is a log-concave distribution and therefore it satisfies the log-Sobolev inequality, and equivalently it also satisfies the transportation inequality. The details and connections to concentration of measures are provided in the full version.

**THEOREM 4.1.** *Let  $\mu'_0$  and  $\mu'_1$  be the distributions obtained by truncating  $\mu_0$  and  $\mu_1$  on support  $[-\tau, \tau]$  for some  $\tau > 0$ . If  $\delta \leq \sigma$ , we have  $\beta(\mu'_0, \mu'_1) \leq \delta^2/\sigma^2$ .*

As a corollary, the SDPI constant between  $n$  copies of  $\mu'_0$  and  $\mu'_1$  is bounded by  $n\delta^2/\sigma^2$ .

**COROLLARY 4.2.** *Let  $\tilde{\mu}_0$  and  $\tilde{\mu}_1$  be the distributions over  $\mathbb{R}^n$  that are obtained by truncating  $\mu_0^n$  and  $\mu_1^n$  outside the ball  $\mathcal{B} = \{x \in \mathbb{R}^n : |x_1 + \dots + x_n| \leq \tau\}$ . Then when  $\sqrt{n}\delta \leq \sigma$ , we have  $\beta(\tilde{\mu}_0, \tilde{\mu}_1) \leq n\delta^2/\sigma^2$ .*

Applying our distributed data processing inequality (Theorem 3.1) on  $\tilde{\mu}_0$  and  $\tilde{\mu}_1$ , we obtain directly that to distinguish  $\tilde{\mu}_0$  and  $\tilde{\mu}_1$  in the distributed setting,  $\Omega(\frac{\sigma^2}{n\delta^2})$  communication is required. By properly handling the truncation of the support, we can prove that it is also true with the true Gaussian distribution.

**THEOREM 4.3.** *Any protocol estimator pair  $(\Pi, \hat{v})$  that solves the distributed Gaussian mean detection problem  $\text{GD}(n, m, \delta, \sigma^2)$  with  $\delta \leq \sigma/\sqrt{n}$  requires communication cost and minimum information cost at least,*

$$\mathbb{E}[|\Pi|] \geq \text{min-IC}(\Pi) \geq \Omega\left(\frac{\sigma^2}{n\delta^2}\right).$$

**REMARK 3.** *The condition  $\delta \leq \sigma/\sqrt{n}$  captures the interesting regime. When  $\delta \gg \sigma/\sqrt{n}$ , a single machine can even distinguish  $\mu_0$  and  $\mu_1$  by its local  $n$  samples.*

**PROOF OF THEOREM 4.3.** Let  $\Pi_0$  and  $\Pi_1$  be the distributions of  $\Pi|V=0$  and  $\Pi|V=1$ , respectively, as defined in Section 2.2. Since  $\hat{v}$  solves the detection problem, we have that  $\|\Pi_0 - \Pi_1\|_{\text{TV}} \geq 1/4$ . It follows from Lemma 3 that  $h(\Pi_0, \Pi_1) \geq \Omega(1)$ .

We pick a threshold  $\tau = 20\sigma$ , and let  $\mathcal{B} = \{z \in \mathbb{R}^n : |z_1 + \dots + z_n| \leq \sqrt{n}\tau\}$ . Let  $F = 1$  denote the event that  $X = (X_1, \dots, X_n) \in \mathcal{B}$ , and otherwise  $F = 0$ . Note that  $\Pr[F = 1] \geq 0.95$  and therefore even if we conditioned on the event that  $F = 1$ , the protocol estimator pair should still be able to recover  $v$  with good probability in the sense that

$$\Pr[\hat{v}(\Pi(X)) = v \mid V = v, F = 1] \geq 0.6 \quad (19)$$

We run our whole argument conditioned on the event  $F = 1$ . First note that for any Markov chain  $V \rightarrow X \rightarrow \Pi$ , and any random variable  $F$  that only depends on  $X$ , the chain  $V|_{F=1} \rightarrow X|_{F=1} \rightarrow \Pi|_{F=1}$  is also a Markov Chain. Second, the channel from  $V$  to  $X|_{F=1}$  satisfies that random variable  $X|_{V=v, F=1}$  has the distribution  $\tilde{\mu}_v$  as defined in the statement of Corollary 4.2. Note that by Corollary 4.2, we have that  $\beta(\tilde{\mu}_0, \tilde{\mu}_1) \leq n\delta^2/\sigma^2$ . Also note that by the choice of  $\tau$  and the fact that  $\delta \leq O(\sigma/\sqrt{n})$ , we have that for any  $z \in \mathcal{B}$ ,  $\tilde{\mu}_0(z) \leq O(1) \cdot \tilde{\mu}_1(z)$ .

Therefore we are ready to apply Theorem 3.1 and conclude that

$$\mathbb{I}(X; \Pi \mid V = 0, F = 1) \geq \Omega(\beta(\tilde{\mu}_0, \tilde{\mu}_1)^{-1}) = \Omega\left(\frac{\sigma^2}{n\delta^2}\right)$$

Note that  $\Pi$  is independent of  $F$  conditioned on  $X$  and  $V = 0$ . Therefore we have that

$$\begin{aligned} \mathbb{I}(X; \Pi \mid V = 0) &\geq \mathbb{I}(X; \Pi \mid F, V = 0) \\ &\geq \mathbb{I}(X; \Pi|F = 1, V = 0) \Pr[F = 1 \mid V = 0] \\ &= \Omega\left(\frac{\sigma^2}{n\delta^2}\right). \end{aligned}$$

Note that by construction, it is also true that  $\tilde{\mu}_0 \leq O(1)\tilde{\mu}_1$ , and therefore if we switch the positions of  $\tilde{\mu}_0, \tilde{\mu}_1$  and run the argument above we will have

$$\mathbb{I}(X; \Pi \mid V = 1) = \Omega\left(\frac{\sigma^2}{n\delta^2}\right)$$

Hence the proof is complete. □

## 4.2 Sparse Gaussian mean estimation

In this subsection, we prove our lower bound for the sparse Gaussian mean estimation problem via a variant of the direct-sum theorem of [GMN14] tailored towards sparse mean estimation.

Our general idea is to make the following reduction argument: Given a protocol  $\Pi'$  for  $d$ -dimensional  $k$ -sparse estimation problem with information cost  $I$  and loss  $R$ , we can construct a protocol  $\Pi$  for the detection problem with information cost roughly  $I/d$  and loss  $R/k$ . The protocol  $\Pi$  embeds the detection problem into one random coordinate of the  $d$ -dimensional problem, prepares fake data on the remaining coordinates, and then runs the protocol  $\Pi$  on the high dimensional problem. It then extracts information about the true data from the corresponding coordinate of the high-dimensional estimator.

The key distinction from the construction of [GMN14] is that here we are not able to show that  $\Pi'$  has small information cost, but only able to show that  $\Pi'$  has a small minimum information cost<sup>7</sup>. This is the reason why in Theorem 4.3 we needed to bound the minimum information cost instead of the information cost.

To formalize the intuition, let  $\mathcal{P} = \{\mu_0, \mu_1\}$  define the detection problem. Let  $\Omega_{d,k,\delta} = \{\theta : \theta \in \{0, \delta\}^d, |\theta|_0 \leq k\}$  and  $\mathcal{Q}_{d,k,\delta} = \{\mu_\theta = \mu_{\theta_1/\delta} \times \cdots \times \mu_{\theta_d/\delta} : \theta \in \Omega_{d,k,\delta}\}$ . Therefore  $\mathcal{Q}$  is a special case of the general  $k$ -sparse high-dimensional problem. We have that

**THEOREM 4.4 (DIRECT-SUM FOR SPARSE PARAMETERS).** *Let  $d \geq 2k$ , and  $\mathcal{P}$  and  $\mathcal{Q}$  defined as above. If there exists a protocol estimator pair  $(\Pi, \hat{\theta})$  that solves the detection task  $T(n, m, \mathcal{Q})$  with information cost  $I$  and mean-squared loss  $R \leq \frac{1}{16}k\delta^2$ , then there exists a protocol estimator pair  $(\Pi', \hat{v}')$  (shown in Protocol 1 in Section 5) that solves the task  $T_{det}(n, m, \mathcal{P})$  with minimum information cost  $\frac{I}{d-k+1}$ .*

The proof of the theorem is deferred to Section 5. Combining Theorem 4.3 and Theorem 4.4, we get the following theorem:

**THEOREM 4.5.** *Suppose  $d \geq 2k$ . Any protocol estimator pair  $(\Pi, \hat{v})$  that solves the  $k$ -sparse Gaussian mean problem  $\text{SGME}(n, m, d, k, \sigma^2)$  with mean-squared loss  $R$  and information cost  $I$  and communication cost  $C$  satisfies that*

$$\begin{aligned} R &\geq \Omega \left( \min \left\{ \frac{\sigma^2 k}{n}, \max \left\{ \frac{\sigma^2 dk}{nI}, \frac{\sigma^2 k}{nm} \right\} \right\} \right) \\ &\geq \Omega \left( \min \left\{ \frac{\sigma^2 k}{n}, \max \left\{ \frac{\sigma^2 dk}{nC}, \frac{\sigma^2 k}{nm} \right\} \right\} \right). \end{aligned} \quad (20)$$

Intuitively, to parse equation (20), we remark that the term  $\frac{\sigma^2 k}{n}$  comes from the fact that any local machine can achieve this error  $O(\frac{\sigma^2 k}{n})$  using only its local samples, and the term  $\frac{\sigma^2 k}{nm}$  is the minimax error that the machines can achieve with infinite amount of communication. When the target error is between these two quantities, equation (20) predicts that the minimum communication  $C$  should scale inverse linearly in the error  $R$ .

<sup>7</sup>This might be inevitable because protocol  $\Pi$  might reveal a lot information for the nonzero coordinate of  $\theta$  but since there are very few non-zeros, the total information revealed is still not too much.

Our theorem gives a tight tradeoff between  $C$  and  $R$  up to a logarithmic factor, since it is known [GMN14]

that for any communication budget  $C$ , there exists a protocol which uses  $C$  bits and has error  $R \leq O \left( \min \left\{ \frac{\sigma^2 k}{n}, \max \left\{ \frac{\sigma^2 dk}{nC}, \frac{\sigma^2 k}{nm} \right\} \right\} \cdot \log d \right)$ .

As a side product, in the case when  $k = d/2$ , our lower bound improves previous works [DJWZ14] and [GMN14] by a logarithmic factor, and turns out to match the upper bound in [GMN14] up to a constant factor.

**PROOF OF THEOREM 4.5.** If  $R \leq \frac{1}{16} \frac{k\sigma^2}{n}$  then we are done. Otherwise, let  $\delta := \sqrt{16R/k} \leq \sigma/\sqrt{n}$ . Let  $\mu_0 = \mathcal{N}(0, \sigma^2)$  and  $\mu_1 = \mathcal{N}(\delta, \sigma^2)$  and  $\mathcal{P} = \{\mu_0, \mu_1\}$ . Let  $\mathcal{Q}_{d,k,\delta} = \{\mu_\theta = \mu_{\theta_1/\delta} \times \cdots \times \mu_{\theta_d/\delta} : \theta \in \Omega_{d,k,\delta}\}$ . Then  $T(n, m, \mathcal{Q})$  is just a special case of the sparse Gaussian mean estimation problem  $\text{SGME}(n, m, d, k, \sigma^2)$ , and  $T(n, m, \mathcal{P})$  is the distributed Gaussian mean detection problem  $\text{GD}(n, m, \delta, \sigma^2)$ . Therefore, by Theorem 4.4, there exists  $(\Pi', \hat{v}')$  that solves  $\text{GD}(n, m, \delta, \sigma^2)$  with minimum information cost  $I' = O(I/d)$ . Since  $\delta \leq O(\sigma/\sqrt{n})$ , by Theorem 4.3 we have that  $I' \geq \Omega(\sigma^2/(n\delta^2))$ . It follows that  $I \geq \Omega(d\sigma^2/(n\delta^2)) = \Omega(kd\sigma^2/(nR))$ . To derive (20), we observe that  $\Omega(\sigma^2 k/nm)$  is the minimax lower bound for  $R$ , which completes the proof.  $\square$

To complement our lower bounds, we also give a new protocol for the Gaussian mean estimation problem achieving communication optimal up to a constant factor in any number of dimensions in the dense case. Our protocol is a *simultaneous protocol*, whereas the only previous protocol achieving optimal communication requires  $\Omega(\log m)$  rounds [GMN14]. This resolves an open question in Remark 2 of [GMN14], improving the trivial protocol in which each player sends its truncated Gaussian to the coordinator by an  $O(\log m)$  factor.

**THEOREM 4.6.** *For any  $0 \leq \alpha \leq 1$ , there exists a protocol that uses one round of communication for the Gaussian mean estimation problem  $\text{GME}(n, m, d, \sigma^2)$  with communication cost  $C = \alpha dm$  and mean-squared loss  $R = O \left( \frac{\sigma^2 d}{\alpha mn} \right)$ .*

The protocol and proof of this theorem are deferred to Section 6, though we mention a few aspects here. We first give a protocol under the assumption that  $|\theta|_\infty \leq \frac{\sigma}{\sqrt{n}}$ . The general protocol is in the full version of the paper. The protocol trivially generalizes to  $d$  dimensions so we focus on 1 dimension. The protocol coincides with the first round of the multi-round protocol in [GMN14], yet we can extract all necessary information in only one round, by having each machine send a single bit indicating if its input Gaussian is positive or negative. Since the mean is on the same order as the standard deviation, one can bound the variance and give an estimator based on the Gaussian density function. If the mean of the Gaussian is allowed to be much larger than the variance, and this no longer works. Instead, a few machines send their truncated inputs so the coordinator learns a crude approximation. To refine this approximation, in parallel the remaining machines each send a bit which is 1 with probability  $x - \lfloor x \rfloor$ , where  $x$  is the machine's input Gaussian. This can be viewed as rounding a sample of the "sawtooth wave function"  $h$  applied to a Gaussian. For technical reasons each machine needs to send two bits, another which is 1 with probability  $(x + 1/5) - \lfloor (x + 1/5) \rfloor$ . We give an estimator based on an analysis using the Fourier series of  $h$ .

### Sparse Gaussian estimation with signal strength lower bound.

Our techniques can also be used to study the optimal rate-communication tradeoffs in the presence of a strong signal in the non-zero coordinates, which is sometimes assumed for sparse signals. That is, suppose the machines are promised that the mean  $\theta \in \mathcal{R}^d$  is  $k$ -sparse and also if  $\theta_i \neq 0$ , then  $|\theta_i| \geq \eta$ , where  $\eta$  is a parameter called the signal strength. We get tight lower bounds for this case as well.

**THEOREM 4.7.** *For  $d \geq 2k$  and  $\eta^2 \geq 16R/k$ , any protocol estimator pair  $(\Pi, \hat{v})$  that solves the  $k$ -sparse Gaussian mean problem  $\text{SGME}(n, m, d, k, \sigma^2)$  with signal strength  $\eta$  and mean-squared loss  $R$  requires information cost (and hence expected communication cost) at least  $\Omega\left(\frac{\sigma^2 d}{n\eta^2}\right)$ .*

Note that there is a protocol for  $\text{SGME}(n, m, d, k, \sigma^2)$  with signal strength  $\eta$  and mean-squared loss  $R$  that has communication cost  $\tilde{O}\left(\min\left\{\frac{\sigma^2 d}{n\eta^2} + \frac{\sigma^2 k^2}{nR}, \frac{\sigma^2 dk}{nR}\right\}\right)$ . In the regime where  $\eta^2 \geq 16R/k$ , the first term dominates and by Theorem 4.7, and the fact that  $\frac{\sigma^2 k^2}{nR}$  is a lower bound even when the machines know the support [GMN14], we also get a matching lower bound. In the regime where  $\eta^2 \leq 16R/k$ , second term dominates and it is a lower bound by Theorem 4.5.

**PROOF OF THEOREM 4.7.** The proof is very similar to the proof of Theorem 4.4. Given a protocol estimator pair  $(\Pi, \hat{v})$  that solves  $\text{SGME}(n, m, d, k, \sigma^2)$  with signal strength  $\eta$ , mean-squared loss  $R$  and information cost  $I$  (where  $\eta^2 \geq 16R/k$ ), we can find a protocol  $\Pi'$  that solves the Gaussian mean detection problem  $\text{GD}(n, m, \eta, \sigma^2)$  with information cost  $\leq O(I/d)$  (as usual the information cost is measured when the mean is 0).  $\Pi'$  would be exactly the same as Protocol 1 but with  $\mu_0$  replaced by  $\mathcal{N}(0, \sigma^2)$ ,  $\mu_1$  replaced by  $\mathcal{N}(\eta, \sigma^2)$  and  $\delta$  replaced by  $\eta$ . We leave the details to the reader.  $\square$

### 4.3 Lower bound for Sparse Linear Regression

In this section we consider the sparse linear regression problem  $\text{SLR}(n, m, d, k, \sigma^2)$  in the distributed setting as defined in Section 2. Suppose the  $i$ -th machine receives a subset  $S_i$  of the  $mn$  data points, and we use  $A_{S_i} \in \mathbb{R}^{n \times d}$  to denote the design matrix that the  $i$ -th machine receives and  $y_{S_i}$  to denote the observed vector. That is,  $y_{S_i} = A_{S_i}\theta + w_{S_i}$ , where  $w_{S_i} \sim \mathcal{N}(0, \sigma^2 I_{n \times n})$  is Gaussian noise.

This problem can be reduced from the sparse Gaussian mean problem, and thus its communication can be lower-bounded. It follows straightforwardly from our Theorem 4.5 and the reduction in Corollary 2 of [DJWZ14]. To state our result, we assume that the design matrices  $A_{S_i}$  have uniformly bounded spectral norm  $\lambda\sqrt{n}$ . That is,

$$\lambda = \max_{1 \leq i \leq m} \|A_{S_i}\|/\sqrt{n}.$$

**COROLLARY 4.8.** *Suppose machines receive data from the sparse linear regression model. Let  $\lambda$  be as defined above. If there exists a protocol under which the machines can output an estimator  $\hat{\theta}$  with mean squared loss  $R = \mathbb{E}[\|\hat{\theta} - \theta\|^2]$  with communication  $C$ , then  $R \cdot C \geq \Omega\left(\frac{\sigma^2 kd}{\lambda^2 n}\right)$ .*

When  $A_{S_i}$  is a Gaussian design matrix, that is, the rows of  $A_{S_i}$  are i.i.d drawn from distribution  $\mathcal{N}(0, I_{d \times d})$ , we have  $\lambda = O\left(\max\{\sqrt{d/n}, 1\}\right)$  and Corollary 4.8 implies that to

achieve the statistical minimax rate  $R = O\left(\frac{k\sigma^2}{nm}\right)$ , the algorithm has to communicate  $\Omega(m \cdot \min\{n, d\})$  bits. The point is that we get a lower bound that doesn't depend on  $k$ —that is, with sparsity assumptions, it is impossible to improve both the loss and communication so that they depend on the intrinsic dimension  $k$  instead of the ambient dimension  $d$ . Moreover, in the regime when  $d/n \rightarrow c$  for a constant  $c$ , our lower bound matches the upper bound of [LSLT15] up to a logarithmic factor. The proof follows Theorem 4.5 and the reduction from Gaussian mean estimation to sparse linear regression of [ZDJW13] straightforwardly and is deferred to the full version of the paper.

## 5. DIRECT-SUM THEOREM FOR SPARSE PARAMETERS

Unknown parameter:  $v \in \{0, 1\}$

Inputs: Machine  $j$  gets  $n$  samples  $X_j = (X_j^{(1)}, \dots, X_j^{(n)})$ , where  $X_j$  is distributed according to  $\mu_v^n$ .

1. All machines publicly sample  $k$  independent coordinates  $I_1, \dots, I_k \subset [d]$  (without replacement).
2. Each machine  $j$  locally prepares data  $\tilde{X}_j = (\tilde{X}_{j,1}, \dots, \tilde{X}_{j,d})$  as follows: The  $I_1$ -th coordinate is embedded with the true data,  $\tilde{X}_{j,I_1} = X_j$ . For  $r = 2, \dots, k$ ,  $j$ -th the machine draws  $\tilde{X}_{j,I_r}$  privately from distribution  $\mu_1^n$ . For any coordinate  $i \in [d] \setminus \{I_1, \dots, I_k\}$ , the  $j$ -th machine draws privately  $\tilde{X}_{j,i}$  from the distribution  $\mu_0^n$ .
3. The machines run protocol  $\Pi$  with input data  $\tilde{X}$ .
4. If  $|\hat{\theta}(\Pi)_{I_1}| \geq \delta/2$ , then the machines output 1, otherwise they output 0.

**Protocol 1:** direct-sum reduction for sparse parameter

We prove Theorem 4.4 in this section. Let  $\Pi'$  be the protocol described in Protocol 1. Let  $\theta \in \mathbb{R}^d$  be such that  $\theta_{I_1} = v\delta$  and  $\theta_{I_r} = \delta$  for  $r = 2, \dots, k$ , and  $\theta_i = 0$  for  $i \in [d] \setminus \{I_1, \dots, I_k\}$ . We can see that by our construction, the distribution of  $\tilde{X}_j$  is the same as  $\mu_\theta^n$ , and all  $X_j$ 's are independent. Also note that  $\theta$  is  $k$ -sparse. Therefore when  $\Pi'$  invokes  $\Pi$  on data  $\tilde{X}$ ,  $\Pi$  will have loss  $R$  and information cost  $I$  with respect to  $\tilde{X}$ .

We first verify that the protocol  $\Pi$  does distinguish between  $v = 0$  and  $v = 1$ .

**PROPOSITION 5.1.** *Under the assumption of Theorem 4.4, when  $v = 1$ , we have that*

$$\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1} - \delta|^2\right] \leq \frac{R}{k} \quad (21)$$

and when  $v = 0$ , we have

$$\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1}|^2\right] \leq \frac{R}{d - k + 1} \quad (22)$$

Moreover, with probability at least  $3/4$ ,  $\Pi'$  outputs the correct answer  $v$ .

The proof appears in the full version of the paper.

## 6. TIGHT UPPER BOUND WITH ONE-WAY COMMUNICATION

In this section, we describe a one-way communication protocol achieving the tight minimal communication for Gaussian mean estimation problem  $\text{GME}(n, m, d, \sigma^2)$  with the assumption that  $|\theta|_\infty \leq \frac{\sigma}{\sqrt{n}}$ . We defer the protocol without this assumption to the full version of the paper.

Note that for the design of protocol, it suffices to consider a one-dimensional problem. Protocol 2 solves the one-dimensional Gaussian mean estimation problem, with each machine sending exactly 1 bit, and therefore the total communication is  $m$  bits. To get a  $d$ -dimensional protocol, we just need to apply Protocol 2 to each dimension. In order to obtain the tradeoff as stated in Theorem 4.6, one needs to run Protocol 2 on the first  $\alpha m$  machines, and let the other machines be idle.

Unknown parameter  $\theta \in [-\sigma/\sqrt{n}, \sigma/\sqrt{n}]$   
 Inputs: Machine  $i$  gets  $n$  samples  $(X_i^{(1)}, \dots, X_i^{(n)})$  where  $X_i^{(j)} \sim \mathcal{N}(\theta, \sigma)$ .

- Simultaneously, each machine  $i$ 
  1. Computes  $X_i = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^n X_i^{(j)}$
  2. Sends  $B_i$
$$B_i = \begin{cases} 1 & \text{if } X_i \geq 0 \\ -1 & \text{otherwise} \end{cases}$$
- Machine 1 computes
 
$$T = \sqrt{2} \cdot \text{erf}^{-1} \left( \frac{1}{m} \sum_{i=1}^m B_i \right)$$
 where  $\text{erf}^{-1}$  is the inverse of the Gauss error function.
- It returns the estimate  $\hat{\theta} = \frac{\sigma}{\sqrt{n}} \hat{\theta}'$  where  $\hat{\theta}' = \max(\min(T, 1), -1)$  is obtained by truncating  $T$  to the interval  $[-1, 1]$ .

**Protocol 2:** A simultaneous algorithm for estimating the mean of a normal distribution in the distributed setting.

The correctness of the protocol follows from the following theorem.

**THEOREM 6.1.** *The algorithm described in Protocol 2 uses  $m$  bits of communication and achieves the following mean squared loss.*

$$\mathbb{E} \left[ (\hat{\theta} - \theta)^2 \right] = O \left( \frac{\sigma^2}{mn} \right)$$

where the expectation is over the random samples and the random coin tosses of the machines.

**PROOF.** Let  $\bar{\theta} = \theta\sqrt{n}/\sigma$ .

Notice that  $X_i$  is distributed according to  $\mathcal{N}(\bar{\theta}, 1)$ . Our goal is to estimate  $\bar{\theta}$  from the  $X_i$ 's. By our assumption on  $\theta$ , we have  $\bar{\theta} \in [-1, 1]$ .

The random variables  $B_i$  are independent with each other. We consider the mean and variance of  $B_i$ 's. For the mean we have that,

$$\mathbb{E}[B_i] = \mathbb{E}[2 \cdot \Pr[0 \leq X_i] - 1]$$

For any  $i \in [m]$ ,  $\Pr[0 \leq X_i] = \Pr[-X_i \leq 0] = \Phi_{-\bar{\theta}, 1}(0)$ , where  $\Phi_{\mu, \sigma^2}$  is the CDF of normal distribution  $\mathcal{N}(\mu, \sigma^2)$ . Note the following relation between the error function and the CDF of a normal random variable

$$\Phi_{\mu, \sigma^2}(x) = \frac{1}{2} + \frac{1}{2} \text{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right)$$

Hence,

$$\mathbb{E}[B_i] = \text{erf}(\bar{\theta}/\sqrt{2}).$$

Let  $B = \frac{1}{m} \sum_{i=1}^m B_i$ , then we have that  $\mathbb{E}[B] = \text{erf}(\bar{\theta}/\sqrt{2}) \leq \text{erf}(1/\sqrt{2})$  and therefore by a Chernoff bound, the probability that  $B > \text{erf}(1)$  or  $B \leq \text{erf}(-1)$  is  $\exp(-\Omega(m))$ . Thus, with probability at least  $1 - \exp(-\Omega(m))$ , we have  $\text{erf}(-1) \leq B \leq \text{erf}(1)$  and therefore  $|T| \leq \sqrt{2}$ .

Let  $\mathcal{E}$  be the event that  $|T| \leq \sqrt{2}$ , then we have that the error of  $\bar{\theta}$  is bounded by

$$\begin{aligned} \mathbb{E}[|\hat{\theta}' - \bar{\theta}|^2] &= \mathbb{E}[|\hat{\theta}' - \bar{\theta}|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + \mathbb{E}[|\hat{\theta}' - \bar{\theta}|^2 \mid \bar{\mathcal{E}}] \Pr[\bar{\mathcal{E}}] \\ &\leq \mathbb{E}[|\sqrt{2} \text{erf}^{-1}(B) - \sqrt{2} \text{erf}^{-1}(\mathbb{E}[B])|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + 2 \Pr[\bar{\mathcal{E}}] \\ &= \mathbb{E}[|\sqrt{2} \text{erf}^{-1}(B) - \sqrt{2} \text{erf}^{-1}(\mathbb{E}[B])|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + 2 \exp(-\Omega(m)) \end{aligned}$$

Let  $M = \max_{\text{erf}^{-1}(x) \in [-1, 1]} \frac{d \text{erf}^{-1}(x)}{dx} < 3$ . Then we have that  $|\text{erf}^{-1}(x) - \text{erf}^{-1}(y)| \leq M|x - y| \leq O(1) \cdot |x - y|$  for any  $x, y \in [-1, 1]$ . Therefore it follows that

$$\begin{aligned} \mathbb{E}[|\hat{\theta}' - \bar{\theta}|^2] &\leq \mathbb{E}[|\sqrt{2} \text{erf}^{-1}(B) - \sqrt{2} \text{erf}^{-1}(\mathbb{E}[B])|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + 2 \exp(-\Omega(m)) \\ &\leq \mathbb{E}[2M^2|B - \mathbb{E}[B]|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + 2 \exp(-\Omega(m)) \\ &\leq \mathbb{E}[2M^2|B - \mathbb{E}[B]|^2] + 2 \exp(-\Omega(m)) \\ &\leq O \left( \frac{1}{m} \right) + 2 \exp(-\Omega(m)) \\ &\leq O \left( \frac{1}{m} \right) \end{aligned}$$

Hence we have that

$$\mathbb{E} \left[ |\hat{\theta} - \theta|^2 \right] = \frac{\sigma^2}{n} \mathbb{E} \left[ |\hat{\theta}' - \bar{\theta}|^2 \right] = O \left( \frac{\sigma^2}{mn} \right)$$

□

## ACKNOWLEDGMENTS.

We thank Yuchen Zhang for suggesting to us the version of sparse Gaussian mean estimation with signal strength assumption. We are indebted to Ramon van Handel for helping us for proving transportation inequality for truncated Gaussian distribution. Mark Braverman would like to thank the support in part by an NSF CAREER award (CCF-1149888), NSF CCF-1525342, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry. Ankit Garg would like to thank the support by a Simons Award in Theoretical Computer Science and a Siebel Scholarship. Tengy Ma would like to thank the support by a Simons Award in Theoretical Computer Science and IBM PhD Fellowship. D. Woodruff would like to thank the support from XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory FA8750-12-C-0323.

## 7. REFERENCES

- [BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.
- [BWZ15] Christos Boutsidis, David P. Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. *CoRR*, abs/1504.06729, 2015.
- [BYJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4), 2004.
- [DAW12] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *Automatic Control, IEEE Transactions on*, 57(3):592–606, 2012.
- [DJWZ14] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Yuchen Zhang. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *CoRR*, abs/1405.0782, 2014.
- [GMN14] Ankit Garg, Tengyu Ma, and Huy L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2726–2734, 2014.
- [Jay09] T.S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In Irit Dinur, Klaus Jansen, Joseph Naor, and José Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 5687 of *Lecture Notes in Computer Science*, pages 562–573. Springer Berlin Heidelberg, 2009.
- [KVW14] Ravi Kannan, Santosh Vempala, and David P. Woodruff. Principal component analysis and higher correlations for distributed data. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1040–1057, 2014.
- [LBKW14] Yingyu Liang, Maria-Florina Balcan, Vandana Kanchanapally, and David P. Woodruff. Improved distributed principal component analysis. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3113–3121, 2014.
- [LSLT15] Jason D Lee, Yuekai Sun, Qiang Liu, and Jonathan E Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015.
- [Rag14] Maxim Raginsky. Strong data processing inequalities and  $\Phi^2$ -sobolev inequalities for discrete channels. *CoRR*, abs/1411.3575, 2014.
- [SD15] Jacob Steinhardt and John C. Duchi. Minimax rates for memory-bounded sparse linear regression. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 1564–1587, 2015.
- [Sha14] Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 163–171. Curran Associates, Inc., 2014.
- [SSZ14] Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1000–1008, 2014.
- [WZ12] David P. Woodruff and Qin Zhang. Tight bounds for distributed functional monitoring. *STOC*, 2012.
- [WZ16] David P. Woodruff and Peilin Zhong. Distributed low rank approximation of implicit functions of a matrix. *CoRR*, abs/1601.07721, 2016.
- [ZDJW13] Yuchen Zhang, John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *NIPS*, pages 2328–2336, 2013.
- [ZDW13] Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2013.
- [ZX15] Yuchen Zhang and Lin Xiao. Communication-efficient distributed optimization of self-concordant empirical loss. *CoRR*, abs/1501.00263, 2015.

## APPENDIX

### A. TOOLBOX

LEMMA 3 (FOLKLORE). *For any two distributions  $P, Q$ , we have*

$$h^2(P, Q) \leq \|P - Q\|_{TV} \leq \sqrt{2}h(P, Q)$$

LEMMA 4. *Let  $\phi(z_1)$  and  $\phi(z_2)$  be two random variables. Let  $Z$  denote a random variable with uniform distribution in  $\{z_1, z_2\}$ : Suppose  $\phi(z)$  is independent of  $Z$  for each  $z \in \{z_1, z_2\}$ : Then,*

$$2h^2(\phi_{z_1}, \phi_{z_2}) \geq I(Z; \phi(Z)) \geq h^2(\phi_{z_1}, \phi_{z_2})$$

THEOREM A.1 (COROLLARY OF THEOREM 7 OF [JAY09]). Then note that  
 Suppose a family of distribution  $\{P_{\mathbf{b}} : \mathbf{b} \in \{0, 1\}^m\}$  satisfies the cut-paste property: for any  $\mathbf{a}, \mathbf{b}$  and  $\mathbf{c}, \mathbf{d}$  with  $\{a_i, b_i\} = \{c_i, d_i\}$  (in a multi-set sense) for every  $i \in [m]$ ,  $h^2(\Pi_{\mathbf{a}}, \Pi_{\mathbf{b}}) = h^2(\Pi_{\mathbf{c}}, \Pi_{\mathbf{d}})$ . Then we have

$$\sum_{i=1}^m h^2(P_{\mathbf{0}}, P_{\mathbf{e}_i}) \geq \Omega(1) \cdot h^2(P_{\mathbf{0}}, P_{\mathbf{1}}) \quad (23)$$

where  $\mathbf{0}$  and  $\mathbf{1}$  are all 0's and all 1's vectors respectively, and  $\mathbf{e}_i$  is the unit vector that only takes 1 in the  $i$ th entry.

LEMMA 5. Suppose two distributions  $\mu, \mu'$  satisfies  $\mu \geq c \cdot \mu'$ . Let  $\Pi(X)$  be a random function that only depends on  $X$ . If  $X \sim \mu$  and  $X' \sim \mu'$ , then we have that

$$I(X; \Pi(X)) \geq c \cdot I(X'; \Pi(X')) \quad (24)$$

PROOF. Since  $\mu \geq c \cdot \mu'$ , we have that

$$I(X; \Pi(X)) = \mathbb{E}_{X \sim \mu} [D_{\text{kl}}(\Pi_X \| \Pi)] \geq c \cdot \mathbb{E}_{X' \sim \mu'} [D_{\text{kl}}(\Pi_{X'} \| \Pi)]$$

$$\mathbb{E}_{X' \sim \mu'} [D_{\text{kl}}(\Pi_{X'} \| \Pi)] = \mathbb{E}_{X' \sim \mu'} [D_{\text{kl}}(\Pi_{X'} \| \Pi')] + D_{\text{kl}}(\Pi' \| \Pi)$$

It follows that

$$I(X; \Pi(X)) \geq c \cdot \mathbb{E}_{X' \sim \mu'} [D_{\text{kl}}(\Pi_{X'} \| \Pi')] = c \cdot I(X'; \Pi(X'))$$

□

LEMMA 6 (FOLKLORE). When  $X$  is drawn from a product distribution, then

$$\sum_{i=1}^m I(X_i; \Pi) \leq I(X; \Pi).$$